Mathematische Verfahren zur Aufklärung der Struktur, Dynamik und Biologischen Aktivität von Molekülen unter Verwendung von NMR Spektroskopischen und Empirischen Parametern

Dissertation zur Erlangung des
Doktorgrades der Naturwissenschaften

vorgelegt beim Fachbereich Chemische und Pharmazeutische Wissenschaften
der Johann Wolfgang Goethe – Universität
in Frankfurt am Main

von
Jens Meiler
aus Leipzig

Frankfurt 2001
DF1

vom Fachbereich Chemische und Pharmazeutische Wissenschaften
der Johann Wolfgang Goethe – Universität als Dissertation angenommen

Dekan:	Prof. Dr. J. W. Engels
Gutachter:	Prof. Dr. C. Griesinger P. D. Dr. R. Meusinger
Datum der Disputation:	07.09.2001

Meinen Eltern und Großeltern

Ein Sumpf zieht am Gebirge hin,
Verpestet alles schon Errungene;
Den faulen Pfuhl auch abzuziehn,
Das Letzte wär das Höchsterrungene.
Eröffn ich Räume vielen Millionen,
Nicht sicher zwar, doch tätig-frei zu wohnen.
Grün das Gefilde, fruchtbar; Mensch und Herde
Sogleich behaglich auf der neusten Erde,
Gleich angesiedelt an des Hügels Kraft,
Den aufgewälzt kühn-emsige Völkerschaft.
Im Innern hier ein paradiesisch Land,
Da rase draußen Flut bis auf zum Rand,
Und wie sie nascht, gewaltsam einzuschließen,
Gemeindrang eilt, die Lücke zu verschließen.
Ja! Diesem Sinne bin ich ganz ergeben,
Das ist der Weisheit letzter Schluß:
Nur der verdient sich Freiheit wie das Leben,
Der täglich sie erobern muß.
Und so verbringt, umrungen von Gefahr,
Hier Kindheit, Mann und Greis sein tüchtig Jahr.
Solch ein Gewimmel möcht ich sehn,
Auf freiem Grund mit freiem Volk zu stehn.
Zum Augenblicke dürft ich sagen:
Verweile doch, du bist so schön!
Es kann die Spur von meinen Erdetagen
Nicht in Äonen untergehn. –
Im Vorgefühl von solch hohem Glück
Genieß ich jetzt den höchsten Augenblick.
...
„Wer immer strebend sich bemüht,
Den können wir erlösen.“

Johann Wolfgang Goethe

Die vorliegende Arbeit wurde am Institut für organische Chemie der Johann Wolfgang Goethe – Universität in Frankfurt am Main unter Anleitung von Herrn Prof. Dr. C. Griesinger angefertigt.

Herrn Prof. Dr. C. Griesinger danke ich für die Aufnahme in den Arbeitskreis und die Überlassung der interessanten Fragestellungen. Sein stetes Interesse am Fortgang der Projekte führte zu vielen konstruktiven und gewinnbringenden Diskussionen. Im besonderen bin ich ihm für die Freiheit und seine Unterstützung, während meiner Promotion auch eigenen Ideen, Fragestellungen und Projekten nachzugehen, zu Dank verpflichtet.

Herr P. D. Dr. Reinhard Meusinger unterstützte mich während meiner Promotion, durch Anregung und Diskussion bei der Erarbeitung und Erprobung von Verfahren, die neuronale Netze und genetische Algorithmen auf chemische und biologische Zusammenhänge anwenden.

Ich danke Herrn Dr. Alexander Diener, Herrn Wolfgang Peti, Herrn Dr. Martin Will, Herrn Dr. Jochen Junker, Herrn P. D. Dr. Matthias Köck, Herrn Dr. Michael Nilges, Herrn Dr. Andreas Kamlowski, Frau Dr. Tanja Parac, Frau Dr. Teresa Carlomagno, Herrn Dr. Marcel J. J. Blommers und Herrn Prof. Dr. Rafael Brüschweiler für die sehr gute und konstruktive Zusammenarbeit bei den gemeinsamen Projekten. Herr Prof. Dr. Rafael Brüschweiler nahm mich für einen Monat in seine Arbeitsgruppe an der Clarc Universität in Worcester auf. Michael Müller, Annalen Bleckmann, Felix Schmäschke und Anita Zeidler danke ich für die erfrischende Zusammenarbeit während ihrer Schnupperpraktika an der Universität Frankfurt. Ich möchte allen Mitarbeitern des Arbeitskreises Griesinger für die gute Zusammenarbeit während der Promotion danken.

Ich danke meiner Familie, Freunden und Bekannten für Unterstützung, Zuspruch und Gespräch. Insbesondere bin ich Dorit, Birgit, Heinz und Julius für freundliche Aufnahme und aufopferungsvolle Pflege zu Dank verpflichtet.

Dem Fond der chemischen Industrie danke ich für ein Doktorandenstipendium. Weiterhin bin ich dem Organisationskomitee der „experimental NMR conference“ und dem Graduiertenkolleg „Chemische und biologische Synthese von Wirkstoffen“ für finanzielle Unterstützung verpflichtet.

Inhaltsverzeichnis

1 Einleitung

Im Informationszeitalter kommt auch in den Naturwissenschaften Chemie und Biochemie der Anwendung moderner mathematischer Methoden und der elektronischen Datenverarbeitung eine ständig wachsende Bedeutung zu. Datenmengen, wie sie zum Beispiel durch das „human genome project“ anfallen, könnten ohne elektronische Datenbanken überhaupt nicht erfasst werden und ihre Auswertung in Folgeprojekten, wie „structural genomics“, wäre undenkbar ohne ausgefeilte mathematische Methoden.

In dieser Arbeit werden verschiedene Projekte vorgestellt, die mathematische Methoden implementieren, nutzen und weiterentwickeln, um Fragestellungen der Chemie und Biologie zu lösen. Die Kapitel 3 bis 6 leiten dabei jeweils in eine Fragestellung ein und fassen anschließend Ergebnisse zusammen, die in Publikationen detailliert beschrieben sind. Die Manuskripte dieser Publikationen sind im Anhang H (eigene Publikationen) aufgenommen und ergänzen das jeweilige Kapitel.

Das Protein Tubulin hat neben vielen anderen Aufgaben in der Zelle entscheidende Bedeutung bei der Zellteilung, da es durch Polymerisation zu Mikrotubuli den Spindelapparat ausbildet. Da sich Tumore durch intensive und beschleunigte Zellteilung auszeichnen, stellt das Polymerisationsgleichgewicht des Tubulin konsequenterweise einen Angriffspunkt für potentielle *anti* – Tumor Wirkstoffe dar. Bekanntester Vertreter dieser Stoffe ist das Taxol[1], welches bereits in der Chemotherapie eingesetzt wird. Zwei weitere Substanzklassen, die das Polymerisationsgleichgewicht gegenteilig beeinflussen, sind die Dolastatine[2] und die Epothilone[3]. Die dreidimensionale Struktur von freiem Dolastatin 10 sowie von freiem und dem an Tubulin gebundenen Epothilon A werden in Kapitel 2 aus NMR spektroskopisch bestimmten Parametern mit Hilfe von beschränkter Molekulardynamik berechnet. Während für die freien Formen der Verbindungen jeweils bereits Strukturmodelle existieren bzw. parallel bestimmt wurden[4-6], wird erstmals die Struktur eines an Tubulin gebundenen Wirkstoffes bestimmt und mit der freien Form verglichen.

In partiell orientierenden Systemen werden dipolare Kopplungen nicht mehr durch eine isotrope Bewegung der Moleküle ausgemittelt, sondern auch in Lösung mit NMR Spektroskopie anteilig messbar[7;8]. Die Größe der beobachteten dipolaren Kopplung hängt dabei sowohl von der Orientierung als auch von der zeitlichen Bewegung des Vektors zwischen den beiden koppelnden Kernen relativ zum Magnetfeld ab. Somit beinhalten residuale dipolare Kopplungen Informationen über Struktur und Dynamik. Die Strukturinformation kann zur Implementierung einer 3D Homologiesuche ähnlicher Proteine in der PDB genutzt werden, wie in Kapitel 3.2 beschrieben ist. Dazu werden die experimentellen dipolaren Kopplungen eines Proteins – basierend auf einem von Losonczi et. al.[9] in dieses Gebiet eingeführten Algorithmus – mit den 3D Strukturen verglichen. Ein Verfahren zur Beschleunigung der Strukturbestimmung von Proteinen, welches parallel auch von Annila et. al.[10] vorgeschlagen wurde. Basierend auf diesen Publikationen ergeben sich verschiedene weitere Ansätze, die Strukturbestimmung von Proteinen essentiell zu beschleunigen[11,12]. Tjandra et. al. beschreibt zuerst eine Methode, die in den dipolaren Kopplungen enthaltene Strukturinformation für die Verfeinerung von Proteinstrukturmodellen zu nutzen[13]. Die Komplexität der Energiehyperfläche, auf der die Strukturoptimierung erfolgt, nimmt nach der Einführung der dipolaren Kopplungen entscheidend zu. Dies verhindert aber eine effektive Nutzung der dipolaren Kopplungen während des „simulated annealing“ Verfahrens[14], welches zum Absuchen des zugänglichen Strukturraumes genutzt wird. In Kapitel 3.3 wird daher ein mathematisch neuer Ansatz verfolgt, der die indirekte Strukturinformation der dipolaren Kopplungen in intramolekulare Projektionswinkelbeschränkungen übersetzt, und die Energiehyperfläche als Konsequenz stark vereinfacht. Dipolare Kopplungen sind durch Bewegungen des die koppelnden Kerne verbindenden Vektors beeinflusst, die bis in den µs – oder gar ms – Bereich reichen. Damit enthalten sie Informationen über die relative Bewegung von Sekundärstrukturelementen und Domänen von Proteinen, die bisher kaum zugänglich waren. Tolman et. al.[15] stellen die Ableitung eines generalisierten Ordnungsparameters vor, mit dem alle dipolaren Kopplungen skaliert werden, unabhängig von ihrer relativen Orientierung. Parallel dazu wird in Kapitel 3.4 ein modellfreier Ansatz entwickelt, der eine detailliertere Analyse dipolarer Kopplungen im Hinblick auf Dynamik zulässt.

Die ^{13}C NMR chemische Verschiebung ist aufgrund ihrer guten Zugänglichkeit kombiniert mit dem komplexen Informationsgehalt einer der wichtigsten Parameter für die Strukturaufklärung organischer Moleküle[16]. Die ^{13}C chemische Verschiebung muss entsprechend häufig zur Validierung von Strukturvorschlägen berechnet werden. Dies kann mit *ab initio* quantenchemischen Rechnungen geschehen[17], die aber sehr rechenzeitintensiv sind und oft durch systematische Fehler der Methode qualitativ zurückstehen. Ihre Vorteile liegen bei ungewöhnlichen Strukturen, die in den existierenden Datenbanken unterrepräsentiert sind. Die Suche in den bestehenden Datenbanken ^{13}C NMR chemischer Verschiebungen[18,19] mittels sphärischer Codierungsformate[20] stellt eine schnellere und exaktere Methode für die meisten in der organischen Chemie relevanten Verbindungen dar. Diese Methode bleibt aber von der Präsenz einer großen Datenbank abhängig. Durch die notwendige Suche in der Datenbank ist sie immer noch zu langsam für das Screaning großer Zahlen von Strukturen in kurzer Zeit. Als dritte Methode gelingt es mittels semiempirisch abgeleiteter Parameter die ^{13}C NMR chemische Verschiebung durch Inkrementmethoden[21] und neuerdings neuronale Netze[22-24] schnell und Datenbank unabhängig zu berechnen. In Kapitel 4.2 wird ein Ansatz vorgestellt, der die Berechnung der ^{13}C NMR chemischen Verschiebung aller organisch relevanten Moleküle mit neuronalen Netzen ermöglicht. Dabei werden mit dem Erreichen hoher Präzision und hoher Geschwindigkeit die Vorteile der existierenden Datenbank- und Inkrementmethoden kombiniert. Dieses Verfahren wird in Kapitel 4.3 genutzt, um Ergebnisse von Strukturgeneratoren nachträglich zu überprüfen. Darüber hinaus wird in Kapitel 4.4 ein genetischer Algorithmus entwickelt und implementiert, der Konstitutionsformeln von Molekülen als Individuen nutzt und in der Lage ist, diese auf die Übereinstimmung ihres berechneten mit dem experimentellen ^{13}C NMR Spektrum zu optimieren. Praktikabel werden beide Verfahren erst durch die schnelle und exakte Berechnung der ^{13}C NMR chemischen Verschiebungen der vorgeschlagenen Strukturen mit den künstlichen neuronalen Netzen.

Im Zuge der elektronischen Verarbeitung tausender Primärsequenzen von Proteinen spielen numerische Darstellungen von Aminosäuren eine wichtige Rolle[25], da ihre Aneinanderreihung eine rechnerlesbare Codierung eines Proteins ergibt. Solche Codierungen wurden erfolgreich genutzt, um mit neuronalen Netzen die Sekundärstruktur von Proteinen zu berechnen[26]. Die Hauptkomponentenanalyse ist ein weit verbreitetes Verfahren, um vieldimensionale Datensätze in niederdimensionale Räume zu projizieren[27]. Dabei werden aber nur lineare Abhängigkeiten zwischen Parameter und Hauptkomponente berücksichtigt. Symmetrische neuronale Netze wurden von Livingstone et. al. als Methode zum Mapping von Datensätzen in niederdimensionale Räume vorgeschlagen[28], um komplexere Abhängigkeiten berücksichtigen zu können. In Kapitel 5 werden solche Netze genutzt, um die große Vielfalt der existierenden Parameter für die verschiedenen Aminosäuren zu analysieren und in niederdimensionale Räume zu projizieren. Die so erhaltenen reduzierten

Parameterrepräsentationen für Aminosäuren werden zur Vorhersage von Sekundärstrukturelementen in Proteinen, ausgehend von deren primärer Sequenz, genutzt.

Künstliche neuronale Netze werden in steigendem Maße genutzt, um quantitative Struktur – Aktivitäts – Beziehungen in Chemie und Biochemie zu untersuchen[29]. Die hohe Flexibilität dieses mathematischen Modells erlaubt es, die oft hochgradig nichtlinearen und komplexen Zusammenhänge zwischen verschiedenen Einflussgrößen auf eine Aktivität zu beschreiben. Derartige quantitative Struktur – Aktivitäts – Beziehungen zwischen Strukturmerkmalen eines *anti* – Tumor Wirkstoffes und seiner biologischen Aktivität werden in Kapitel 6 mittels künstlicher neuronaler Netze erstellt. Die gefundenen mathematischen Modelle werden in einem neuen Ansatz genutzt, um die Struktur des Wirkstoffes zu optimieren, indem die Aktivitäten bisher nicht synthetisierter Derivate des Wirkstoffes berechnet werden.

2 Strukturbestimmung von Wirkstoffen durch NMR Spektroskopie

2.1 Durch NMR Parameter beschränkte Molekulardynamik mit X-Plor

Die Aufklärung der dreidimensionalen (3D) Struktur von Biomolekülen spielt eine entscheidende Rolle für das Verständnis ihrer Wirkmechanismen und damit ihrer Funktion im biologischen Gleichgewicht. Dies gilt sowohl für die drei wichtigsten Stoffklassen der Biochemie: Proteine, RNA/DNA und Zucker, als auch für die große Zahl von Molekülen, die wir als Medikamente und Wirkstoffe einsetzen.

Sind Konstitution und Stereochemie aufgeklärt, können 3D Strukturmodelle einer Verbindung erstellt werden. Diese sollen zunächst die bekannten Gleichgewichtsbindungslängen, Gleichgewichtsbindungswinkel, Dihedralwinkel sowie abgeschätzte VAN DER WAALS und elektrostatische Wechselwirkungen berücksichtigen. Die genannten Parameter schränken aber schon bei kleinen Molekülen den möglichen Strukturraum nicht ausreichend ein, um eine eindeutige 3D Struktur zu erhalten. Somit wird die Verwendung weiterer, experimentell zu bestimmender Strukturinformationen nötig. Die NMR Spektroskopie ist eine Methode, solche Informationen zu gewinnen und ermöglicht damit die Bestimmung eines besser definierten 3D Strukturmodells.

Dazu liefert die NMR Spektroskopie eine Vielzahl von Parametern, deren wichtigster nach wie vor der NOE ist. Mit Hilfe von NOESY-[30] und ROESY- Experimenten[31-33] ist es möglich, aus Signalintensitäten Atomabstände zu bestimmen und Aussagen zum Relaxationsverhalten zu treffen. Mit Austausch- und Relaxationsexperimenten können Wasserstoffbrückenbindungen detektiert werden, die wiederum indirekt zusätzliche Abstandsinformation liefern. Dihedralwinkel sind durch die Auswertung skalarer ^3J Kopplungen zugänglich, die von der Größe dieses Winkels abhängig sind[34]. Die Messung kreuzkorrelierter Relaxationsraten[35,36] und dipolarer Kopplungen[8] ermöglicht den Zugang zu weiterer wertvoller Strukturinformation. Die beiden letztgenannten neueren Parameter gilt es nun effektiv in die Bestimmung der Struktur einzubeziehen. Diener diskutiert in seiner Dissertation[37] die Implementierung kreuzkorrelierter Relaxationsraten. Die Nutzung dipolarer Kopplungen zur Bestimmung und Verbesserung der 3D Struktur wird in Kapitel 3.3 vorgestellt.

Ziel der Strukturbestimmung ist es, ein 3D Modell der Struktur zu gewinnen, das sowohl mit den genannten empirischen Parametern über den Bau von Molekülen, als auch mit den experimentell gewonnenen Informationen im Einklang steht. Zwei prinzipielle Rechenverfahren, die dieses leisten können, sind Distanzgeometrie[38-40] und Molekulardynamik[41].

Die Distanzgeometrie ist ein mathematisches Verfahren, das Koordinaten von Punkten berechnet, die bestimmte vorgegebene Abstände erfüllen. Diese werden aus empirischen und experimentellen Parametern berechnet. Das Verfahren sucht einen großen Strukturraum ab, ist aber nicht in der Lage, VAN DER WAALS und elektrostatische Wechselwirkungen zu berücksichtigen, was eine nachträgliche Verfeinerung der Struktur mit anderen Methoden nötig macht.

Molekulardynamik benötigt deutlich längere Rechenzeiten, und eine spezielle Implementierung ist notwendig, um einen großen Strukturraum absuchen zu können. Dafür können aber alle relevanten Wechselwirkungen während der gesamten Rechnung genutzt werden. Dadurch wird der zugängliche Konformationsraum besser repräsentiert. Zudem bietet die Molekulardynamik die Möglichkeit, Dynamik von Molekülen zu studieren bzw. sogar in die Rechnungen einzubeziehen.

Um in Molekulardynamikrechnungen die Wahrscheinlichkeit zu erhöhen, das globale Energieminimum zu finden, beginnt die Simulation bei einer hohen Temperatur, die langsam gesenkt wird. Dieses Verfahren wird „simulated annealing“ genannt (Abbildung 1). Durch die hohe Temperatur reicht die kinetische Energie in der Anfangsphase der Rechnung aus, Energiebarrieren zu überwinden, die möglicherweise auf dem Weg zum Energieminimum liegen. Die Hochtemperaturphase (6500 Schritte á 5fs bei 2000K) kann als das Absuchen des Lösungsraumes betrachtet werden, während die anschließende Abkühlphase (7000 Schritte á 5fs auf 100K) und die Energieminimierung als Verfeinerung angesehen werden können. Es gibt keine Garantie, mit diesem Verfahren das globale Minimum zu finden. Ein guter Hinweis ist, beim Wiederholen der Rechnung ähnliche Strukturen zu erhalten. Unterscheiden sich die gefundenen Strukturen aber deutlich, konvergieren sie in mehrere verschiedene Minima, von denen konsequenterweise nur eines das globale sein kann.

Da quantenmechanische Rechnungen ab einer bestimmten Molekülgröße rechnerisch nicht mehr zu bewältigen sind, ist es nötig, ein molekulares System mit den Gesetzen der klassischen Mechanik zu beschreiben. Zwar werden dann nur die Atomkerne und nicht die Elektronen in die Rechnung einbezogen, dafür lässt sich aber die Struktur großer Moleküle mit hinreichender Genauigkeit in sinnvoller Zeit berechnen. Mathematisch gesehen ist die NEWTONsche Bewegungsgleichung

$$m_i \frac{d^2\vec{x}_i}{dt^2}(t) = -\vec{\nabla}_{x_i} E_{total} \tag{2.1}$$

für alle Atome *i* zu lösen[42]. Um die Temperatur kontrollieren zu können, wird in einer sogenannten LANGEVIN Dynamik die NEWTONsche Bewegungsgleichung um einen Reibungsterm erweitert. Dessen Größe ist abhängig von der Differenz aus Soll- und Istwert der Temperatur.

Die Energieminimierung ist eine statische Methode, die das nächstliegende Minimum auf der Energiehyperfläche erreicht. Sie wird verwandt, um innere Spannungen abzubauen, die zum Beispiel nach einer Molekulardynamik vorhanden sein können (Abbildung 1). Für die eigentliche Strukturfindung ist sie ungeeignet, da sie nicht in der Lage ist, einen großen Konformationsraum abzusuchen. Bekannte Verfahren erster Ordnung sind zum Beispiel das Verfahren des steilsten Abstiegs („steepest decent“) und das konjugierte Gradienten Verfahren („conjugated gradient“).

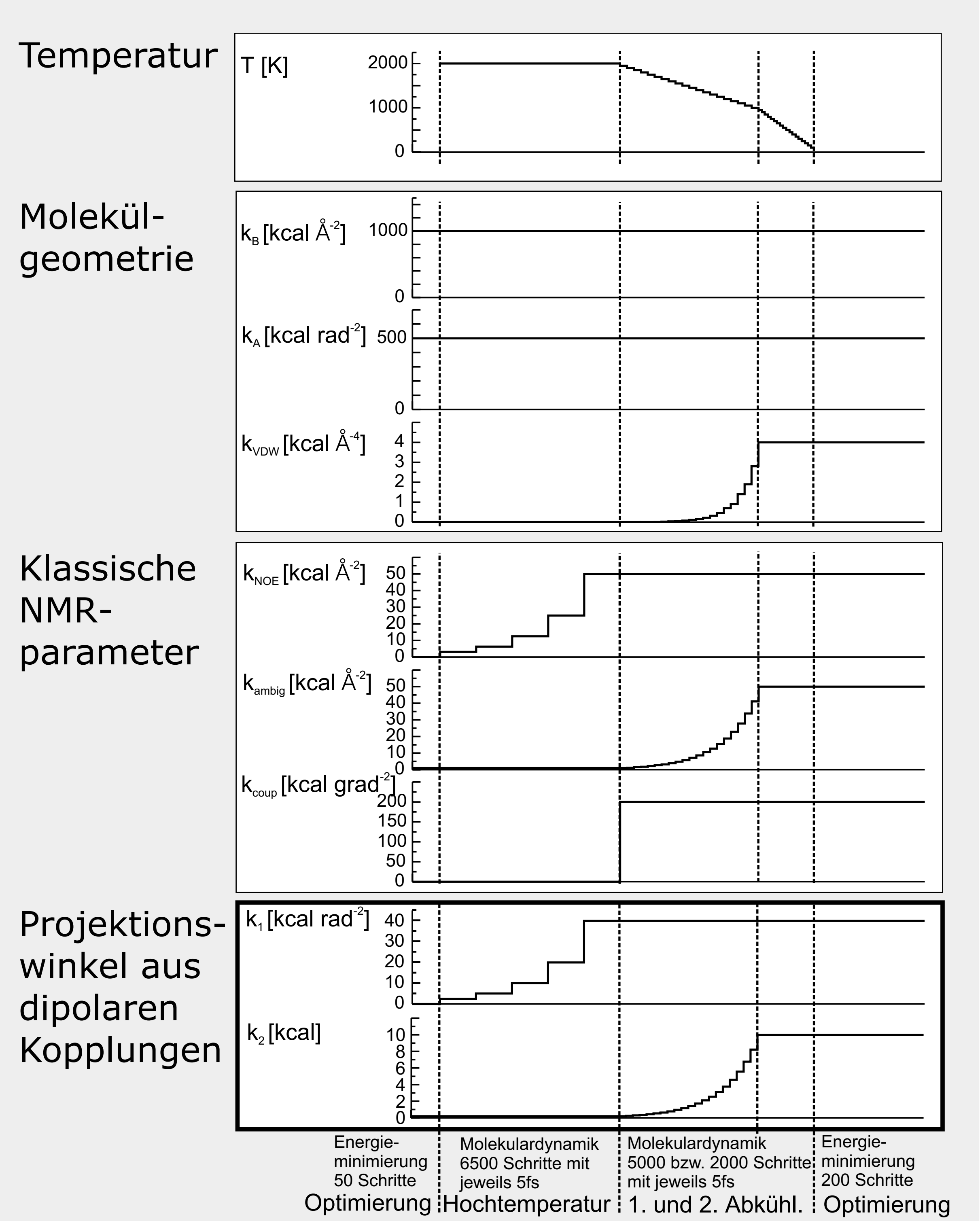


Abbildung 1: **Kraftkonstanten während des X-Plor *"simulated annealing"* Protokolls zur Bestimmung von 3D Strukturmodellen von Molekülen unter Berücksichtigung NMR spektroskopisch bestimmter Parameter (k_B - Bindungen / k_A - Winkel / k_{VDW} - VAN DER WAALS Wechselwirkungen / k_{NOE} - Distanzen aus NOE Intensitäten / $k_{a\ mbig}$ - mehrdeutige Distanzen aus NOE Intensitäten / $k_{co\ p}$ - Dihedralwinkel aus skalaren Kopplungen / k_1, k_2 vgl. Kapitel 3 und Abbildung 8).**

Ausgehend von einer zufälligen Struktur, welche aber die empirisch vorgegebenen Parameter der Molekülgeometrie weitgehend erfüllt, werden in einer Hochtemperaturphase die Kraftkonstanten für NOE - Distanzen und dipolare Kopplungen stufenweise exponentiell angehoben. Hier wird die Faltung der Struktur erreicht.

In einer ersten Abkühlphase werden nun auch die weiteren experimentellen Parameter (mehrdeutige NOE - Distanzen und die Barriere zwischen den beiden erlaubten Projektionswinkelbereichen, die sich aus den dipolaren Kopplungen ergeben) sowie die VAN DER WAALS Wechselwirkungen aktiviert. Anschließend wird die Struktur in einer zweiten Abkühlphase in das energetische Minimum getrieben und in einer letzten Energieminimierung optimiert.

Die Nutzung der Projektionswinkel aus dipolaren Kopplungen wird später ausführlich diskutiert.

Das für die Rechnung verwandte Kraftfeld sollte alle für die Struktur des Moleküls relevanten Wechselwirkungen sowie alle experimentell bestimmten Parameter berücksichtigen. Demzufolge ergibt sich:

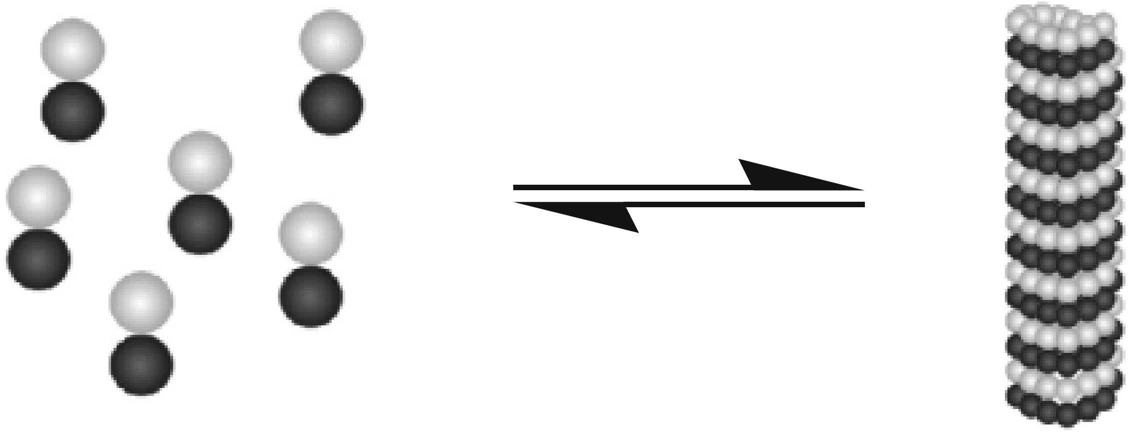
$$E_{total} = E_{empirisch} + E_{experimentell}\,. \tag{2.2}$$

Die empirischen Energieterme werden vorzugsweise durch einfache, oft quadratische Funktionen beschrieben und Kreuzterme werden vernachlässigt. Dies geschieht, um die Rechnung zu beschleunigen. Solange das Hauptaugenmerk das Finden von Strukturen ist, die mit den experimentellen Daten im Einklang stehen, sind diese Vereinfachungen zulässig. Mit der gleichen Begründung wird bei diesen Rechnungen auf die explizite Betrachtung von Lösungsmittelmolekülen verzichtet und lediglich eine allgemeine Dielektrizitätskonstante bei der Berechnung der elektrostatischen Wechselwirkung durch einen Faktor $(\varepsilon_0 r)^{-1}$ berücksichtigt, wobei ε_0 die Dielektrizitätskonstante des Lösungsmittels ist. Entsprechend ist die Güte der Lösungsstruktur entscheidend von der Dichte und der Genauigkeit der experimentellen Daten abhängig.

In allen vorgestellten Rechnungen wird das Programm X-Plor[14] verwendet bzw. erweitert. Das genutzte Protokoll für das „simulated annealing“ wurde gegenüber[37] bzw.[14] nicht geändert. Alle geänderten Topologien und Kraftfelder werden unten im Detail diskutiert und sind im Anhang gegeben. Abbildung 1 stellt die Temperatur, die Zeiträume sowie die Größe der genutzten Energiekonstanten während des „simulated annealing“ dar.

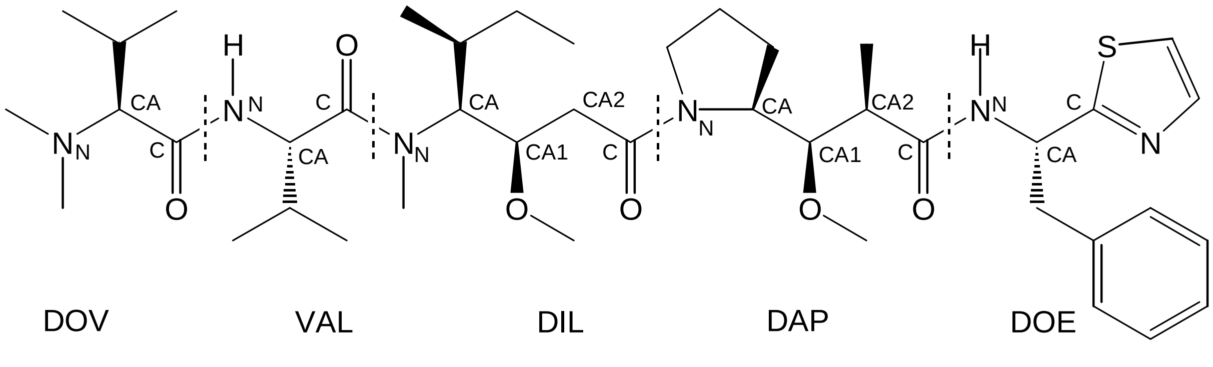
2.2 Strukturbestimmung des Dolastatin 10

Marine Lebewesen gewinnen als Quelle für neue, bisher unbekannte Leitstrukturen immer mehr an Bedeutung. Dolastatin 10 ist gemeinsam mit einer Reihe weiterer Pseudopeptide aus der Meeresschnecke *dolabella auricularia* isoliert worden[2,43]. Es ist in dieser Reihe das stärkste Zellgift mit einem subnanomolaren IC_{50} Wert und dient den ansonsten wehrlosen Tieren zur Verteidigung. Seine Wirkung beruht auf der Unterbindung der Mitose durch reversible Bindung an Tubulin, was dessen Polymerisation verhindert (Schema 1). Dadurch ist die Ausbildung des zur Mitose notwendigen Spindelapparates unmöglich und die Zellteilung wird somit gestoppt. Damit ist Dolastatin 10 ebenso wie die viel diskutierten Taxoide (Kapitel 2.3 und 6) ein potentielles Krebstherapeutikum.



Schema 1: Tubulinmonomere bestehend aus einer α- (grau) und einer β- (schwarz) Untereinheit polymerisieren zu Mikrotubuli

Um den Mechanismus der Bindung an Tubulin zu verstehen und die Struktur des Wirkstoffes weiter zu optimieren, ist die Bestimmung der 3D Struktur von Dolastatin 10 interessant. Die Primärsequenz sowie die Konfiguration von Dolastatin 10 sind im Schema 2 gegeben.



Schema 2: Die Primärsequenz des Dolastatin 10 sowie die Konstitution und die Konfiguration der ungewöhnlichen Aminosäuren Dolavalin (DOV), Dolaisoleuin (DIL), Dolaproin (DAP) und Dolaphenin (DOE)

Dolastatin 10 ist aus den Aminosäuren Dolavalin (DOV), Valin, Dolaisoleuin (DIL), Dolaproin (DAP) und Dolaphenin (DOE) aufgebaut. Bis auf Valin sind die enthaltenen Aminosäuren ungewöhnlich: DOV ist N,N-Dimethylvalin, DAP und DIL sind an Stickstoff und Sauerstoff voll methylierte γ–Aminosäuren und DOE ist ein Thiazolderivat des Phenylalanin.

Durch Röntgenkristallstrukturanalyse kann die Struktur von Dolastatin 10 nicht bestimmt werden, da es nicht kristallisiert. Pettit et. al.[44] publizierten eine Röntgenkristallstrukturanalyse des DOE(α-R) Enantiomeren. Auf experimentellen NMR Daten basierende Strukturuntersuchungen an Dolastatin 10 wurden bereits von Benedetti et. al.[4], Alattia et. al.[5] und Quant[45] durchgeführt. Die im folgenden dargestellte Lösung der Struktur beruht auf den in letztgenannter Publikation beschriebenen experimentellen Daten.

Dolastatin 10 liegt in Methanol in zwei Konformationen vor, die sich durch die *cis* – bzw. *trans* – Stellung der Peptidbindung DIL – DAP unterscheiden. Beide Konformationen stehen im chemischen Gleichgewicht. Der Austausch ist aber so langsam, dass für beide Konformationen getrennte Sätze von NOE Signalen bestimmt werden können. Für die Strukturbestimmung werden nur die Kreuzsignale genutzt, die eindeutig einem der beiden Konformeren zugeordnet werden können. Die Rechnungen werden mit dem Programm X-Plor[14] unter Nutzung der Standardprotokolle für Proteine durchgeführt. Um die ungewöhnlichen Aminosäuren handhaben zu können, muss deren Topologie in Anhang A (Definition der Topologie und des Kraftfeldes für Dolastatin 10) definiert werden. Die Definition neuer Kraftfeldparameter war nicht notwendig, da alle notwendigen Atomtypen bereits im in X-Plor genutzten Kraftfeld für Proteine genutzt waren.

Für die *cis* – Konformation gingen 54 NOE – Intensitäten sowie vier Dihedralwinkel aus skalaren J – Kopplungen in die Rechnung ein. Für die *trans* – Konformation werden 48 NOE – Intensitäten und ebenfalls vier Dihedralwinkel genutzt. Die Abstandsinformationen ergeben sich entsprechend den relativen NOE – Intensitäten bei einer Eichung auf den H,H – Abstand zweier Wasserstoffatome im aromatischen Ring des DOE bzw. zweier geminaler Wasserstoffatome in Methylengruppen gemäß Gleichung (2.3). Die Daten wurden aus der Literatur[45] übernommen und sind in Tabelle 1 und Tabelle 2 zusammengefasst. Die skalaren Kopplungen werden X-Plor direkt vorgegeben und an die Gleichung

$J_{HCCH} = \left[11.0\cos^2\left(\varphi_{HCCH}\right) - 0.5\cos\left(\varphi_{HCCH}\right) + 0.5\right]Hz$ angepasst, wobei φ_{HCCH} der zugehörige Dihedralwinkel ist.

$$NOE^{ij} \sim \frac{1}{r_{ij}^6} \tag{2.3}$$

Nicht zyklische Pentapeptide sind sehr flexibel in ihrer 3D Struktur. Im Vergleich zu diesen, schränkt der Fünfring des DAP den zugänglichen Konformationsraum etwas ein. Die ungewöhnlichen Aminosäuren des Dolastatin 10 mit den zusätzlichen Bindungen im Rückgrat des Moleküls, werden die absolute Flexibilität von Dolastatin 10 gegenüber klassischen Pentapeptiden erhöhen. Relativ zur Länge des Rückgrats kann die Flexibilität sinken, da die neu eingeführten Bindungen im Rückgrat ausschließlich sp^3-hybridisierte Kohlenstoffatome verbinden. Die möglichen Dihedralwinkeleinstellungen sind bei solchen Bindungen besser definiert und durch tiefere Energieminima gekennzeichnet als bei Bindungen zwischen sp^3- und sp^2-hybridisierten Atomen, wie sie sonst im Rückgrat des Proteins vorherrschen.

Die *cis* – Konformation ist aufgrund des abgewinkelten Rückgrats etwas globulärer, was zu einer höheren Zahl von NOE – Intensitäten führt (54 bei der *cis* - Konformation statt nur 48 in der *trans* – Konformation). Die zusätzlichen NOE Intensitäten definieren zusätzlich auch Distanzen zwischen Atomen, die im Bindungsnetzwerk weit voneinander entfernt liegen, zwischen den Aminosäuren DIL und DOE. Daher ist die *cis* – Konformation mit einem RMSD – Wert von 1.042 Å auch besser definiert als die *trans* – Konformation mit 1.220 Å. Tabelle 3 fasst die erhaltenen RMSD – Werte zusammen. Die Dihedralwinkel des Rückgrats für die *cis* – und die *trans* – Konformation sind in Tabelle 4 gegeben. Aus den Standardabweichungen dieser Dihedralwinkel geht hervor, das die Struktur des Rückgrats der *cis* – Konformation in der ersten Aminosäuren DOV (ψ) und VAL (φ und ψ), im zentralen DIL (Dihedralwinkel um CA1 – CA2) sowie auch in den letzten Aminosäuren DAP (ψ) und DOE (φ) sehr schlecht definiert ist. Die *trans* – Konformation wird ebenfalls in DOV (ψ) und VAL (φ), DIL (Dihedralwinkel um CA1 – CA2) sowie DAP (ψ) und DOE (φ) mit einer sehr großen Standardabweichung in den Dihedralwinkeln gefunden. Besser beschrieben ist diese Konformation lediglich in VAL (ψ), während im zentralen DIL auch ψ schlechter definiert ist. Die sehr große Standardabweichung einiger Winkel ergibt sich als Mittelung verschiedener Konformere. Dies trifft für DOV (ψ): ~ 60 ° (Konformer 1) und ~ 180 ° (Konformer 2) sowie VAL (φ): ~ –60 ° (Konformer 1) und ~ –120 ° (Konformer 2) jeweils sowohl in der *cis* – als auch in der *trans* – Konformation zu. Der Dihedralwinkel DOV (ψ) ist sicher durch hohe Flexibilität gekennzeichnet. Entsprechend sind in der Tabelle 4 für diesen Winkel in den mit ⇒ gekennzeichneten Zeilen zusätzlich die Mittelwerte beider Konformationen mit den jeweiligen Standardabweichungen gegeben. Der Dihedralwinkel VAL (φ) sollte durch die HN – HA skalare Kopplung bestimmbar sein, ist aber in der genutzten Quelle nicht angegeben[45]. Das dort abgebildete eindimensionale ^{1}H NMR Spektrum erlaubt es, diese Kopplung abzuschätzen. Mit dem erhaltenen kleinen Wert stimmt die Konformation um –60 ° deutlich besser überein als die Konformation um –120 °. In der mit ⇒ gekennzeichneten Zeile in Tabelle 4 wurden zur Berechnung des Dihedralwinkels nur die Strukturen genutzt, die diese Einstellung haben. Ansonsten müssen Dihedralwinkel mit großer Standardabweichung als flexiblere Bereiche oder weniger gut definierte Bereiche interpretiert werden. Insbesondere fällt hier die größere Standardabweichung der Rückgratwinkel des DIL in der *trans* – Konformation auf, die aber hauptsächlich auf die zusätzlichen NOE Signale zwischen DIL und DOE in der *cis* – Konformation zurückzuführen sind. Die energieärmsten Strukturen beider Konformationen sind in Abbildung 2 dargestellt. Eine Überlagerung der 10 energieärmsten aus 100 gerechneten Strukturen ist in Abbildung 3 gegeben.

Alle vorgegebenen Distanzen aus den NOE Intensitäten werden in den erstellten Strukturmodellen mit der vorgegebenen oberen Toleranz von 0.5 Å. Kürzere Distanzen, als aus den NOE Intensitäten errechnet, werden nicht als Verletzung gewertet. Die skalaren Kopplungen werden aufgrund der recht großen Unschärfe der Karpluskurve mit einer Toleranz von 3.0 Hz berücksichtigt. Die Winkel, die durch skalare Kopplungen zwischen HA(DIL) und HA1(DIL) eingeschränkt werden, ergeben in beiden Konformationen, mit Abweichungen von 4.3 Hz bzw. 3.5 Hz etwas höhere Werte, als die vorgegebene Toleranz. Die mittlere *trans* – Konformation ergibt zusätzlich eine Abweichung von 3.5 Hz für die skalare Kopplung zwischen HN(DOE) und HA(DOE). Acht der 10 energieärmsten Strukturen erfüllen jedoch die Kopplung innerhalb der vorgegebenen Toleranz mit –150 ° bzw. +150 °. Die Unsicherheit von ±23.0 ° zeigt den weiten Bereich für diesen Dihedralwinkel an. Die zusätzlichen geringen Abweichungen der skalaren Kopplungen, die über die Toleranz hinausgehen, rufen in dem verwandten Kraftfeld nur geringe Kräfte hervor, die von aus NOE Distanzen hervorgerufenen Kräften überlagert werden.

Das *trans* – Konformer ist der von Allatia et. al.[5] vorgestellten Struktur in DMSO sehr ähnlich. Alle von Allatia et. al. bestimmten NOE Distanzen werden von der hier berechneten Struktur mit Abweichungen kleiner 1 Å erfüllt. Das *cis* – Konformer ist in der in DMSO berechneten Struktur noch deutlich mehr gefaltet, als es mit den in Methanol gemessenen Daten erhalten wird. Die von Allatia et. al. beschriebenen NOE Signale zwischen den Aminosäuren DOE – VAL, DOE – DIL, DOV – DAP werden sämtlich nicht beobachtet und demzufolge von den berechneten Strukturen nicht erfüllt. Alle weiteren NOE Intensitäten werden ebenfalls mit Abweichungen kleiner 1 Å erfüllt. Sie beschreiben aber Distanzen zwischen Atompaaren, die in der Konstitution sehr nahe liegen, so dass sich die Änderungen in der Struktur hier nur geringfügig auf die Distanzen auswirken. Die Struktur in DMSO ist offensichtlich deutlich rigider als die in Methanol. Da Methanol dem Lösungsmittel Wasser ähnlicher ist als DMSO, ist die hier vorgestellte Struktur das realistischere Modell für Dolastatin 10 in wässriger Lösung. Die Anzahl der gemessenen NOE Intensitäten und $^3J_{HH}$ skalaren Kopplungen nimmt in Methanol deutlich ab, was einer höheren Flexibilität des Moleküls in Methanol gleich kommt. Dies spiegeln auch die erhöhten RMSD – Werte der Atompositionen wieder und ist begründet durch die deutlich schlechtere Solvatation des Dolastatin 10 in Methanol[45].

Um dieses Strukturmodell weiter zu verbessern, müssten Studien zur Dynamik des Moleküls angefertigt bzw. diese durch Relaxationsexperimente untersucht werden. Weitere experimentelle Daten, wie zum Beispiel skalare und dipolare Kopplungen, können den zugänglichen Strukturraum weiter einschränken. Im Vergleich zur Bestimmung der Struktur von Proteinen, spielen im besonderen skalare Kopplungen, die Dihedralwinkel definieren, eine große Rolle bei der Bestimmung der Struktur kleinerer Moleküle. Hier fehlen häufig die NOE Signale zwischen im Bindungsnetzwerk weit entfernten Atomen, die die 3D Grundstruktur definieren.

Die sehr flexible Struktur des Dolastatin 10 lässt ebenfalls eine deutliche Änderung der Struktur bei der Bindung des Moleküls an Tubulin („induced fit“) erwarten.

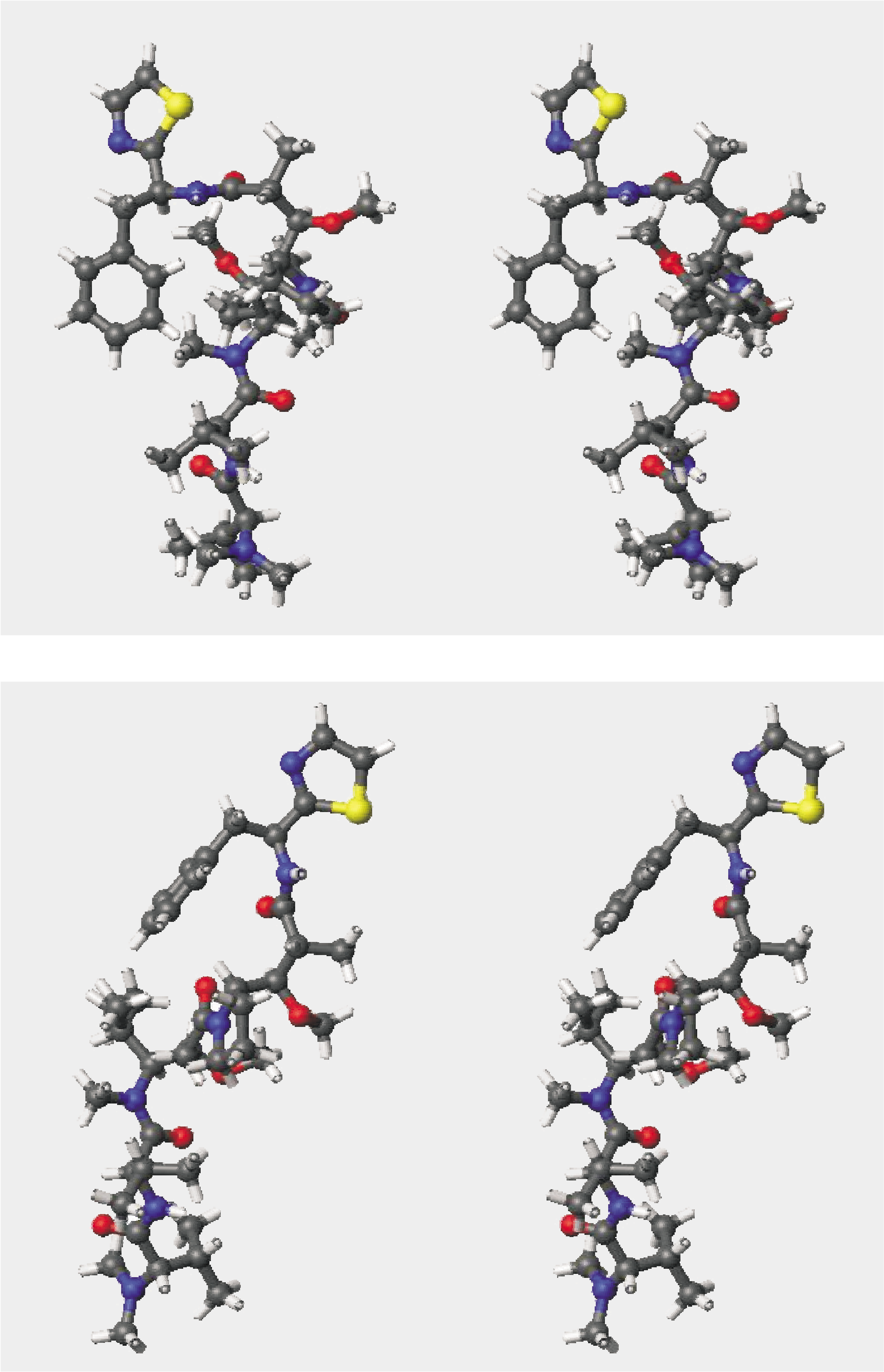


Abbildung 2: **Energieärmste 3D Struktur in Stereodarstellung des *cis* - Dolastatin 10 (oben) und des *trans*- Dolastatin 10 (unten), die aus der Molekulardynamik unter Berücksichtigung der experimentellen NMR Parameter erhalten wurden.**

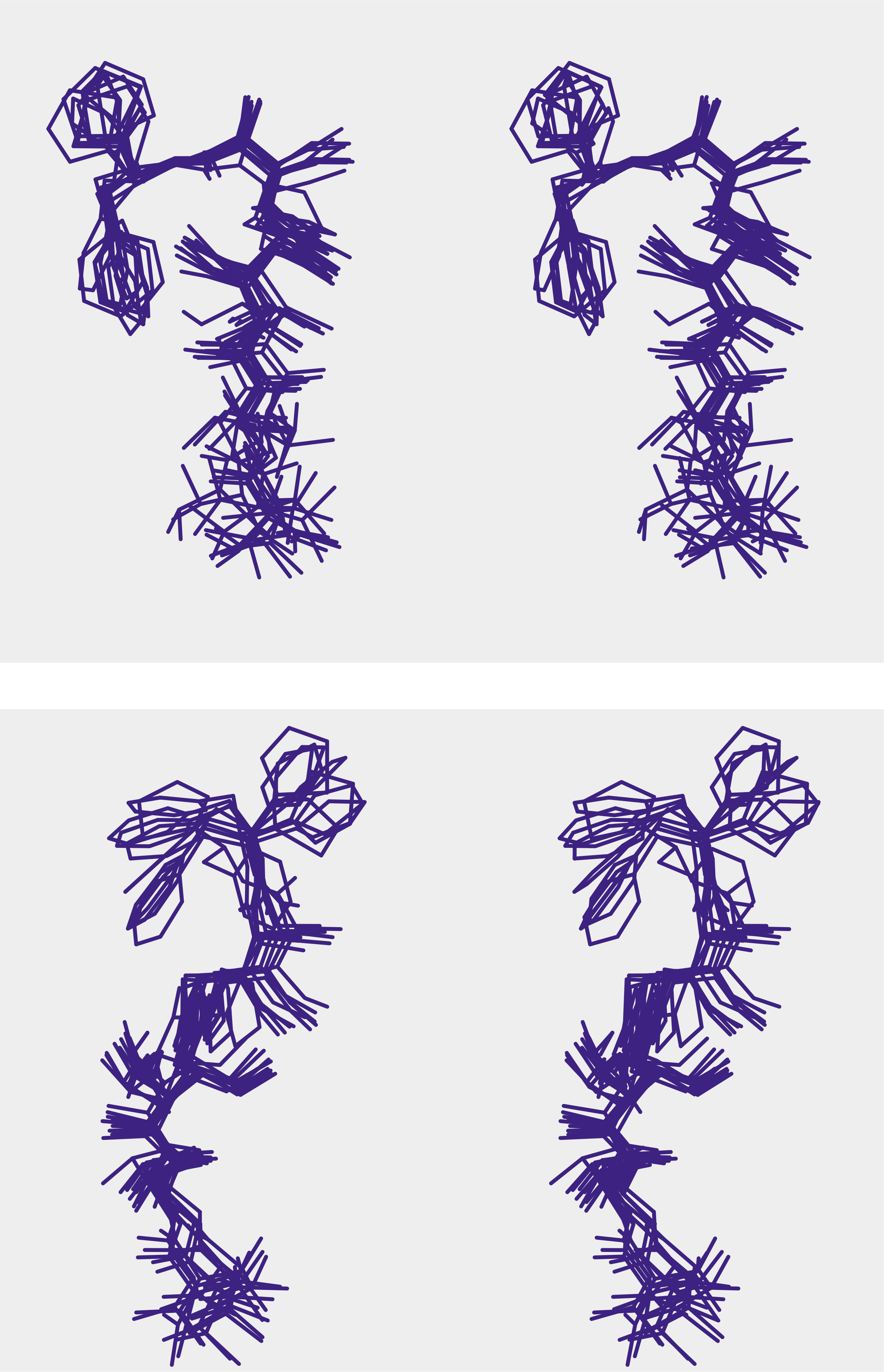
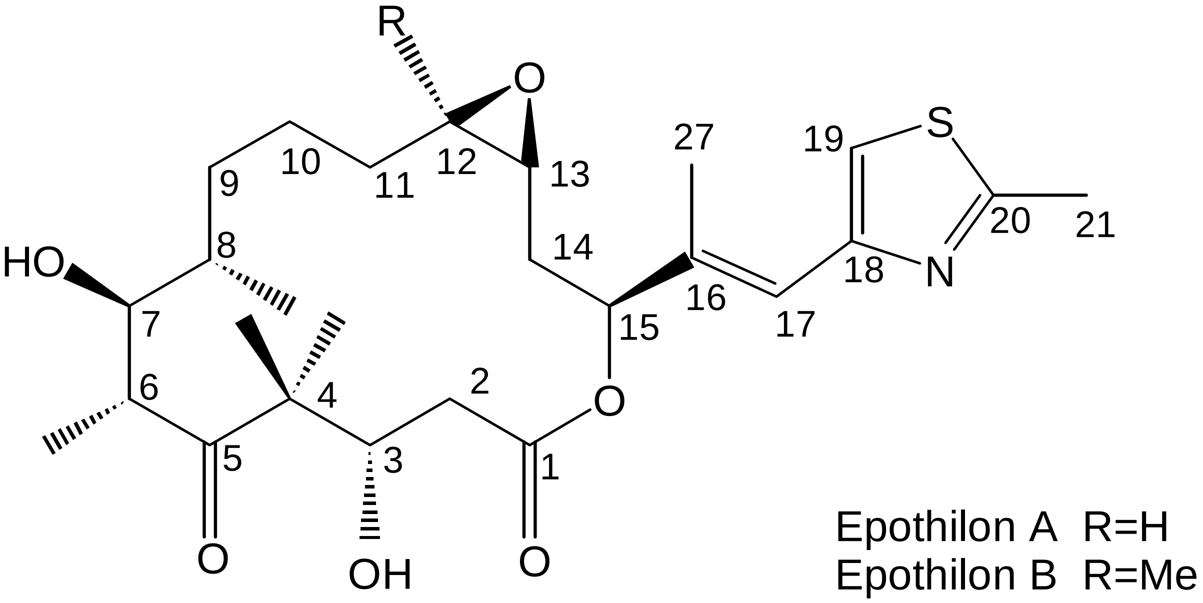


Abbildung 3: Überlagerung der 10 energieärmsten 3D Struktur des *cis* - Dolastatin 10 (oben) und des *trans*- Dolastatin 10 (unten), die aus der Molekulardynamik unter Berücksichtigung der experimentellen NMR Parameter erhalten werden (Stereodarstellung).

2.3 Strukturbestimmung des Epothilon A

Die Epothilone A and B wurden vom Myxobacterium *sorangium cellulosum* strain 90 von Höfle et. al.[46,47] isoliert. Die Entdeckung ihrer zelltoxischen Wirkung gegen Tumorzellen führte zur intensiven Erforschung ihrer Chemie und Biologie. Bollag et al.[3] entdeckten die induzierende Wirkung auf die Tubulin[48,49] Polymerisation dieser Substanzklasse, ähnlich dem Taxol[1]. Der Stabilisierungseffekt auf Mikrotubuli in Taxol resistenten Krebszellen[50] erhöhte ihr Potential für die Chemotherapie weiter[3,51,52].

Die komplette Struktur mit gelöster Stereochemie (Schema 3) wurde von Höfle et. al.[53] publiziert. Die 3D Struktur des Epothilons ist durch Röntgenkristallstrukturanalyse[54] und durch NMR Spektroskopie untersucht[6].



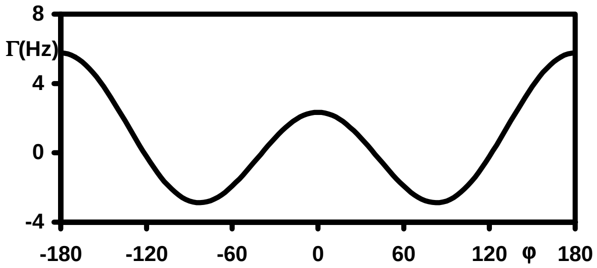
Schema 3: Konstitution und Konfiguration des Epothilons

Um den Wirkungsmechanismus des Epothilons untersuchen zu können, ist die Aufklärung der 3D Struktur unabdingbar. Im Gegensatz zum Dolastatin 10, lässt die globuläre Struktur des Moleküls eine Vielzahl von NOE – Signalen und somit eine deutlich besser definierte Struktur erwarten. Im besonderen ist ein Vergleich der an Tubulin gebundenen Struktur mit der freien Form des Epothilon A wünschenswert, da Änderungen zwischen diesen beiden Formen auf eine Wechselwirkung mit Tubulin zurückzuführen sind und somit direkt mit der biochemischen Wirkung in Verbindung stehen.

Epothilon A wird im Verhältnis 100:1 mit Tubulin gemischt und in wässriger Lösung untersucht. Die Bindung von Epothilon A ist schwach genug (im Gegensatz zu Epothilon B), dass in der Lösung ein ständiger Austausch zwischen der gebundenen und der freien Form existiert. Dieser Austausch macht die Beobachtung eines NOESY Spektrums möglich, welches NOE – Intensitäten zeigt, wie sie für die gebundene Form signifikant sind. Dieser Effekt tritt auf, da in die Mittelung die der NOE – Intensitäten die Korrelationszeit τ_c der Gesamtbewegung der Moleküle eingeht. Sie ist für die gebundene Form viel größer als für die freie Form, da sich die Masse von ca. 100kDa im Komplex auf 0.4 kDa für die freie Form reduziert. Analog können kreuzkorrelierte Relaxationsraten beobachtet werden, allerdings können zum Beispiel keine skalaren Kopplungen transferiert werden, da ihr zeitliche Mittelung nicht mit τ_c skaliert ist. 27 NOE – Intensitäten sowie 3 Dihedralwinkel, die aus kreuzkorrelierten Relaxationsraten abgeleitet sind, werden in der Rechnung verwendet[55]. Alle NOE – Intensitäten wurden drei Gruppen zugeordnet und mit Distanzen identifiziert: stark < 2.5 Å, mittel < 3.5 Å und schwach < 5.0 Å. Die angegebenen Distanzen wurden um 0.5 Å, 0.7 Å und 1.0 Å erhöht, wenn eine Methylgruppe beteiligt ist, um die zu hohe Intensität dieser Signale zu korrigieren. Durch kreuzkorrelierte Relaxation wird der mittlere Winkel zwischen zwei wechselwirkenden Dipolen beschrieben:

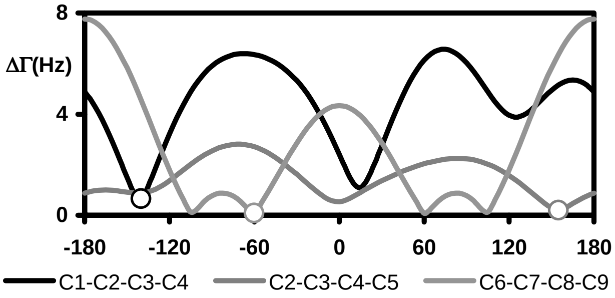
$$\Gamma^{C}_{C_iH_i-C_jH_j} = \frac{2}{5}\left(\frac{\gamma_C\gamma_H\mu_0}{4\pi\left\langle r^3_{CH}\right\rangle}\right)^2 \tau_C \left\langle\frac{3\cos^2\theta_{ij}-1}{2}\right\rangle \tag{2.4}$$

Handelt es sich dabei um vicinale Protonen *i* und *j*, lässt sich θ_{ij} mit dem Dihedralwinkel $\varphi_{H_iC_iC_jH_j}$ als $\theta_{ij} = \sin^2(109.5°)\bullet\cos\varphi_{H_iC_iC_jH_j} - \cos^2(109.5°)$ beschreiben. Schema 4 zeigt den Graph dieser Funktion für eine effektive Korrelationszeit τ_C von 0.74 ns.



Schema 4: Kreuzkorrelierte Rate Γ(Hz) in Abhängigkeit vom Dihedralwinkel φ

Fünf kreuzkorrelierten Relaxationsraten wurden von Marcel J. J. Blommers und Teresa Carlomagno bestimmt[55]. Sie sind in Tabelle 6 zusammengefasst. Zwei der experimentellen Raten wurden zwischen dem Dipol C03 – H03 und den Methylgruppen C22 – H22* bzw. C23 – H23* gemessen. Hier muss die oben eingeführte Abhängigkeit (2.4) zusätzlich mit $\frac{1}{2}\left(3\cos^2\left(109.5^{\mathrm{o}}\right)-1\right)=-0.3296^{\mathrm{o}}$ skaliert werden, wird eine isotrope Rotation der Methylgruppen zugrundegelegt. Bezieht man die gemessenen Raten nun auf die Winkel im Rückgrat des Epothilon A, ergeben sich je nach Einstellung dieser Dihedralwinkel die im Schema 5 dargestellten RMSD – Werte für die kreuzkorrelierten Raten.



Schema 5: RMSD der Raten in Abhängigkeit von den entsprechenden Dihedralwinkeln

Daraus werden die gekennzeichneten Minima als Dihedralwinkelbeschränkungen für die Strukturrechnung genutzt. Die Einstellung des Dihedralwinkels C07 – C08 ist nach der Berücksichtigung weiterer NOE – Signale eindeutig und stimmt auch mit der Röntgenstruktur überein (siehe hinten).

Zur Bestimmung der Struktur der freien Form wird Epothilon A in DMSO gelöst und NOESY – Spektren mit 300 ms, 500 ms und 700 ms Mischzeit aufgenommen. Die sich ergebenden Intensitäten der NOE – Intensitäten werden von Andreas Kamlowski übernommen[56] und gemäß Gleichung (2.3) auf den Mittelwert der vier beobachteten geminalen H,H – NOE – Intensitäten skaliert. Genutzt werden dabei die Werte bei 300ms Mischzeit. Bei den beiden längeren Mischzeiten wird für einige Signale eine deutliche Abweichung vom linearen Anstieg der Intensität mit der Mischzeit „initial rate approximation" beobachtet, die auf Spindiffusion zurückzuführen ist. 61 Distanzen werden aus NOE – Intensitäten gewonnen und wie oben beschrieben prozessiert. Zusätzlich werden in der Rechnung 11 skalare H,H – Kopplungen und 15 skalare H,C – Kopplungen verwendet. Dabei wird wiederum direkt die beobachtete Kopplung als zu optimierender Parameter eingesetzt. Analog der bereits im Dolastatin genutzten Beziehung $J_{HCCH} = \left[11\cos^2(\varphi_{HCCH}) - 0.5\cos(\varphi_{HCCH}) + 0.5\right] Hz$ wird zusätzlich für die H,C – Kopplungen $J_{HCCC} = \left[7\cos^2(\varphi_{HCCC}) - 0.9\cos(\varphi_{HCCC})\right] Hz$ eingeführt.

Um die Strukturen mit X-Plor behandeln zu können, müssen die Topologie definiert und die benötigten Atome in das vereinfachte Kraftfeld für X-Plor eingefügt werden. Die entsprechende Topologie und das Kraftfeld sind im Anhang C (Definition der Topologie und des Kraftfeldes für Epothilone A) gegeben. Die Vorzugsbindungslängen und Vorzugsbindungswinkel wurden aus der durch Röntgenkristallstrukturanalyse bestimmten Struktur entnommen.

Alle experimentellen Parameter für die gebundene und auch die freie Form des Epothilon A sowie die entsprechenden erhaltenen Größen in den Rechnungen und aus der Röntgenkristallstrukturanalyse[54] sind in Tabelle 5, Tabelle 6 und Tabelle 7 zusammengefasst. Abbildung 4 stellt jeweils die energieärmste Struktur von 100 gerechneten Strukturen für die gebundene und die freie Form gemeinsam mit der aus der Röntgenkristallstrukturanalyse erhaltenen Struktur dar. Abbildung 5 zeigt eine Überlagerung der 10 energieärmsten Strukturen für beide Formen. Der RMSD – Wert der gebundenen Form liegt mit 0,103 Å für die Nichtwasserstoffatome und 0,537 Å für alle Atome höher, als für die freie Form mit 0.497 Å und 0.016 Å (Tabelle 8). Die freie Form ist durch die deutlich größere Zahl der verwendeten NOE – Intensitäten (61 für die freie statt 27 für die gebundene Form des Epothilon A) und skalaren Kopplungen (26 für die freie statt 3 kreuzkorrelierte Relaxationsraten für die gebundene Form des Epothilon A) deutlich besser definiert. Die

Dihedralwinkel des Rückgrats sind in Tabelle 9 gegeben und mit der Röntgenstruktur verglichen.

Die freie Form des Epothilon A ist im wesentlichen mit der durch Röntgenstrukturanalyse bestimmten Struktur identisch, wie aus dem Vergleich der Dihedralwinkel in Tabelle 9 ersichtlich ist. Die Struktur ist ebenfalls identisch mit der von Taylor et. al.[6] publizierten Lösungsstruktur in der Konformation A (ca. 85%). Alle publizierten Distanzen aus NOE – Intensitäten stimmen mit den hier bestimmten Werten mit einer Abweichung kleiner 0.5 Å überein. Ebenfalls kann die von Taylor et. al.[6] postulierte Konformation B (ca. 15%) bestätigt werden. Entsprechend verkürzte NOE – Distanzen wurden für die in Analogie mit den publizierten Werten für die Distanzen H03 – H07, H03 – H06 und H06 – H25* beobachtet (vgl. Tabelle 5). Diese Distanzen werden für die Bestimmung der Struktur der am stärksten populierten Konformation A mit um 1.5 Å größeren Toleranzen verwandt. Danach werden alle experimentell bestimmten Distanzen für die gebundene und für die freie Form erfüllt. Die drei Dihedralwinkel, die durch kreuzkorrelierte Relaxation definiert sind, werden innerhalb der vorgegebenen Toleranz von 10 ° erfüllt. Die vorgegebenen Winkel betragen –140 °, 170 ° und –60 °. Die sich aus den resultierenden Einstellungen ergebenden kreuzkorrelierten Raten sind in Tabelle 6 gegeben und mit der Röntgenstruktur verglichen. Die für die freie Form experimentell erhaltenen skalaren Kopplungen werden ebenfalls innerhalb der vorgegebenen Toleranz von 3 Hz erfüllt, bis auf die Kopplungen zwischen H15 und C27, H03 und H23 sowie H03 und C05. Die erste Kopplung beschreibt den Dihedralwinkel um die Bindungen C15 – C16. Für diese Bindung ist weiterhin die skalare Kopplung zwischen H15 und C17 gemessen worden, die durch das Strukturmodell erfüllt ist. Die daraus resultierende Einstellung des Dihedralwinkels lässt aber eine Erfüllung der Kopplung zwischen H15 und C27 nicht zu. Diese Einstellung des Dihedralwinkels wird weiterhin durch das intensive NOE Signal zwischen H15 und H17 bestätigt, welches die Distanz auf 2.5 Å festlegt. Die skalaren Kopplungen zwischen H03 und H23 sowie H03 und C05 widersprechen sich ebenfalls direkt in geringfügiger Weise. Der in dem Strukturmodell gefundene Kompromiss verletzt die experimentellen Kopplungen um 4.8 Hz bzw. 3.2 Hz, erfüllt aber die weiterhin bestimmte skalare Kopplung zwischen H03 und H22.

Während sich die Struktur der freien Form gegenüber der Röntgenkristallstrukturanalyse kaum verändert, zeigt die gebundene Form deutliche

Abweichungen. Die freie Form des Epothilon A zeigt gegenüber der durch Röntgenkristallstrukturanalyse bestimmten Struktur Dihedralwinkeländerungen von über 30 ° für die Dihedralwinkel um die Bindungen um C01 – C02, C03 – C04 und C15 – C16. Während der Dihedralwinkel um C01 – C02 NMR spektroskopisch weder durch NOE Intensitäten noch durch skalare Kopplungen eingeschränkt ist und deshalb einer größeren experimentellen Unsicherheit unterliegt, sind die Dihedralwinkel und die Bindungen C03 – C04 sowie C15 – C16 durch skalare Kopplungen bestimmt, die diese Abweichungen hervorrufen. Auffällig ist, dass beide Abweichungen für Dihedralwinkel gefunden werden, für die mehr als eine skalare Kopplung bestimmt ist, und diese sich teilweise widersprechen. Der Röntgenkristallstrukturanalyse widersprechen die Kopplungen zwischen H15 und C27 sowie H03 und H05, was in der NMR spektroskopisch bestimmten Struktur korrigiert ist (vgl. Tabelle 7).

Die gebundene Form des Epothilon A erfährt erwartungsgemäß stärkere Veränderungen. Der Makrozyklus ändert seine Konformation in den Dihedralwinkeln um die Bindungen C01 – C02, C02 – C03, C03 – C04, C04 – C05, C11 – C12, O011 – C01 um mehr als 30 °. Während die Konformation um die Bindungen C04 – C05 und O011 – C01 noch im Bereich von ±40 ° liegt und durch die dramatischen Änderungen um die benachbarten Bindungen hervorgerufen wird, sind die Änderungen um die Bindungen C02 – C03 und C03 – C04 durch die experimentell bestimmten kreuzkorrelierten Relaxationsraten direkt begründet. Die Änderung um die Bindung C01 – C02 ist mit über 80 ° deutlich, aber durch die benachbarten, definierten Winkel hervorgerufen und nicht durch weitere NMR spektroskopische Parameter eingeschränkt. Auch die Bindung um C11 – C12 ist in der gebundenen Form weder durch Distanzen aus NOE Signalen noch durch kreuzkorrelierte Relaxation bestimmt, so dass die hier beschriebene Änderung ausschließlich durch das Kraftfeld hervorgerufen ist. Trotzdem verfügt der Makrozyklus des Epothilon A in der gebundenen Form eine sichtbare Ähnlichkeit zur freien Form, wie in Abbildung 4 dargestellt ist.

Eine entscheidende und gut belegte Änderung der Konformation ist aber in der Seitenkette zu beobachten. Der Dihedralwinkel um C15 – C16 erfährt eine 40 ° Drehung, und die konjugierte Bindung C17 – C18 erfährt eine 180 ° Drehung. Die erste Änderung ist durch eine aus NOE Intensitäten bestimmte Distanz zwischen H141 und H17 hervorgerufen. Die Drehung um die Bindung C17 – C18 ist durch die Änderung der Intensitäten der NOE Signale zwischen H17 und H19 sowie H19 und H27* gegenüber der freien Form bewiesen. Während die Intensität des ersten Signals sich stark verringert, steigt die des zweiten deutlich an, wechselt man vom freien in den gebundenen Zustand (vgl. Tabelle 5). Diese gebundene Konformation liegt energetisch höher als die Konformation der Seitenkette in der freien Form, da hier H19 und H27* sterisch wechselwirken. Die Änderung muss daher auf die Wechselwirkung des Moleküls mit Tubulin zurückgeführt werden.

Durch die Drehung wird das Stickstoffatom des Thiazolrings zugänglich, welches vorher durch die beiden benachbarten Methylgruppen gut abgeschirmt war. Zusätzlich bewegt es sich auf die Seite des Moleküls, auf der sich ebenfalls potentielle Substituenten an den Kohlenstoffatomen C12 und C13 befinden. Alle drei Strukturmerkmale spielen eine entscheidende Rolle bei der Wechselwirkung des Epothilons mit Tubulin, wie in Kapitel 6 diskutiert ist und augenblicklich Gegenstand weiterer Untersuchungen ist.

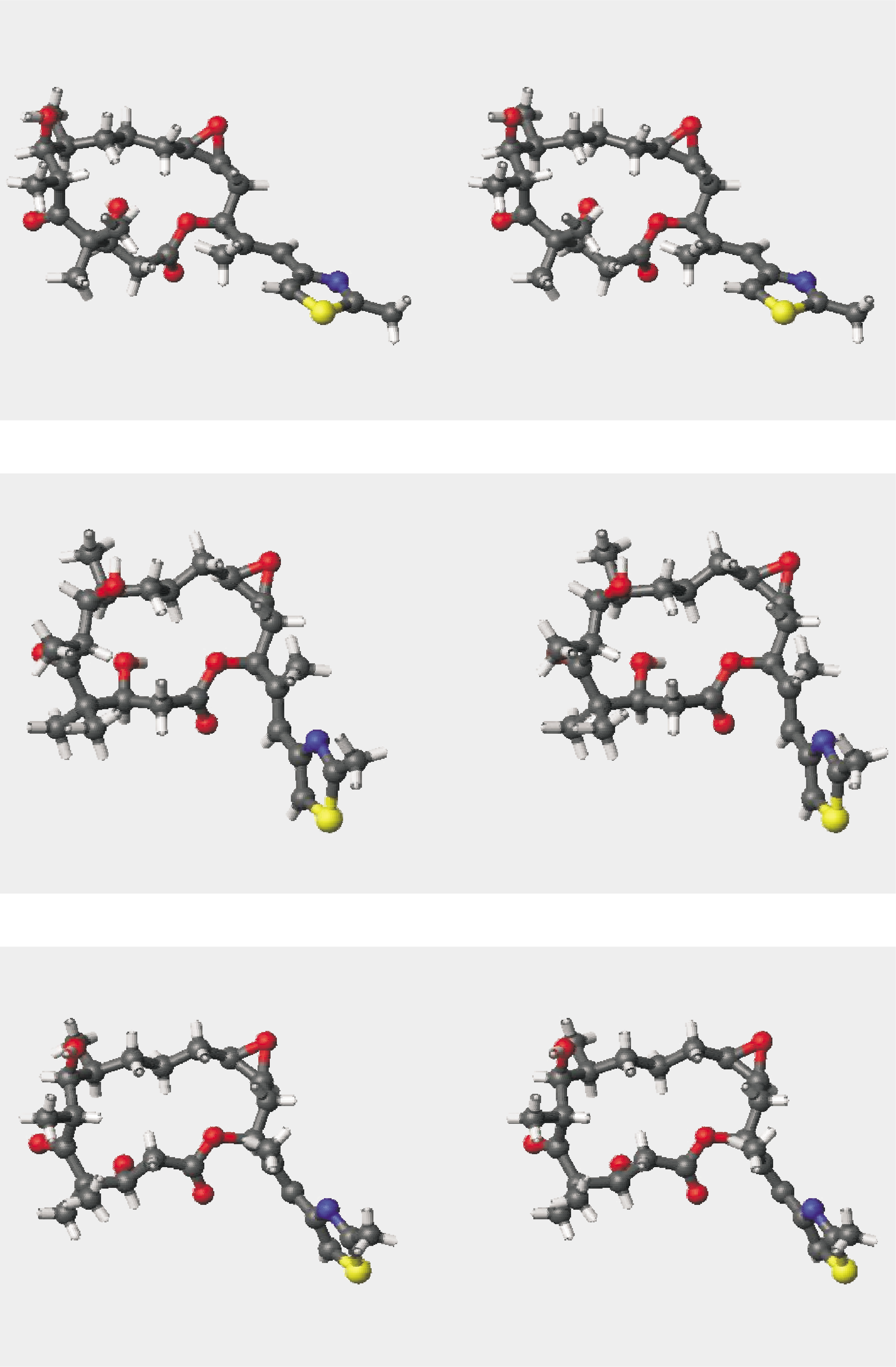


Abbildung 4: **Energieärmste 3D Struktur des gebundenen Epothilon A (oben) und des freien Epothilon A (mitte) die aus der Molekulardynamik unter Berücksichtigung der experimentellen NMR Parameter erhalten wurden. Unten ist die 3D Struktur aus der Röntgenkristallstrukturanalyse zum Vergleich gegeben. Alle Abbildungen sind Stereodarstellungen.**

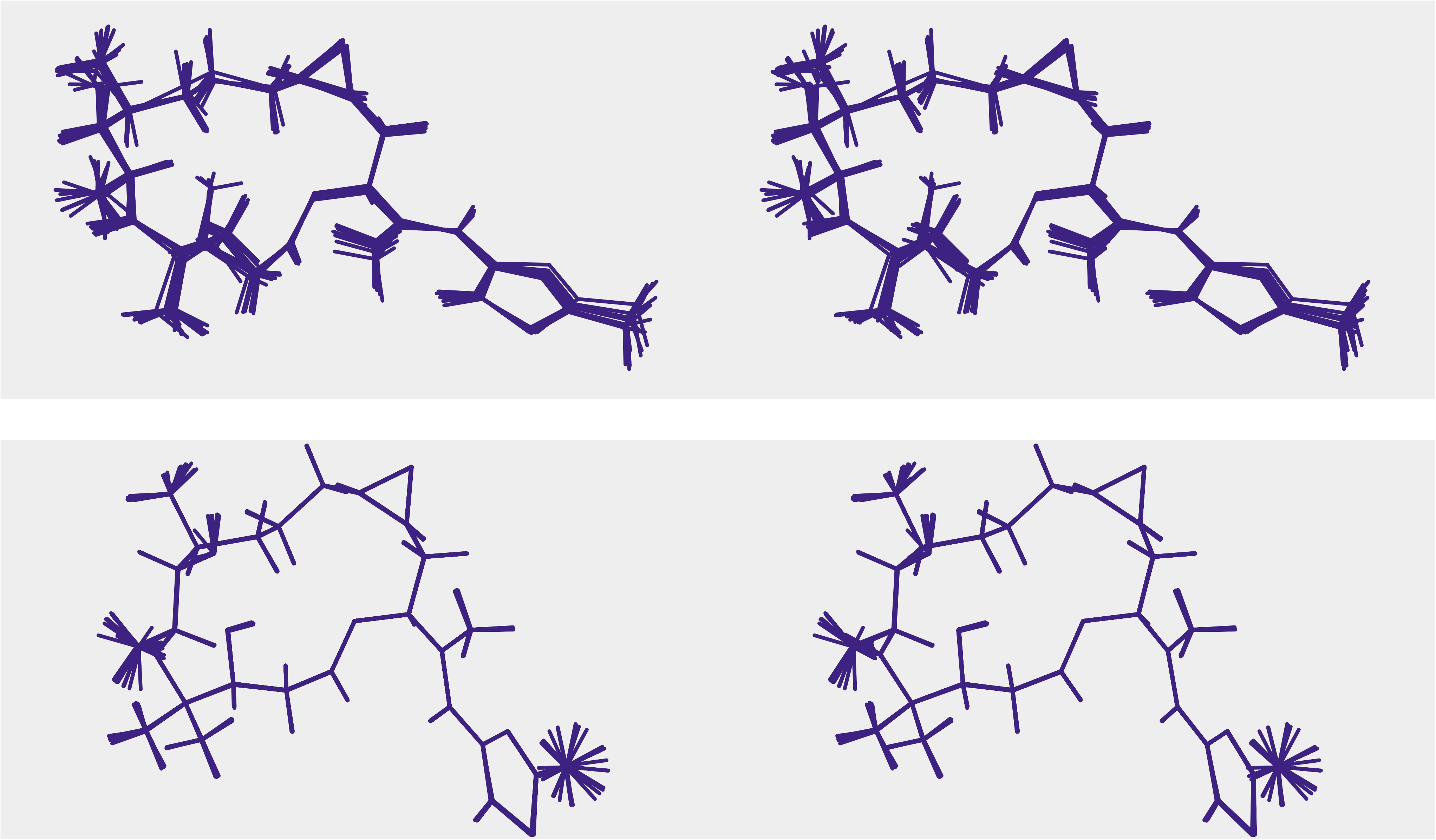


Abbildung 5: Überlagerung der 10 energieärmsten 3D Strukturen des gebundenen Epothilon A (oben) und des freien Epothilon A (unten), die aus der Molekulardynamik unter Berücksichtigung der experimentellen NMR Parameter erhalten werden (Stereodarstellung).

3 Dipolare Kopplungen in partiell orientierenden Systemen

3.1 Theoretische Grundlagen

Dipolare Kopplungen ergeben sich aus Wechselwirkungen zwischen zwei Dipolen in einem äußeren Magnetfeld gemäß der Gleichung:

$$\left\langle D_{ij}\right\rangle = -\frac{\mu_0}{4\pi^2}\frac{h}{2\pi}\frac{\gamma_i\gamma_j}{\left\langle r_{ij}^3\right\rangle}\left\langle\frac{3\cos^2\chi-1}{2}\right\rangle \tag{3.1}$$

wobei χ der Winkel zwischen dem Vektor und dem Magnetfeld B_0 ist, $\gamma_{i,j}$ die gyromagnetischen Verhältnisse der Kerne *i* und *j* und r_{ij} der Abstand der Kerne *i* und *j* ist. Wird eine isotrope Bewegung eines Moleküls in Lösung angenommen, mitteln sich die dipolaren Kopplungen zu Null. Gelingt aber eine partielle Orientierung, werden die dipolaren Kopplungen nicht vollständig ausgemittelt, sondern sind anteilig messbar.

Eine sehr kleine Vorzugsorientierung eines Moleküls ergibt sich, falls das Molekül selbst ein Dipolmoment hat. Jedoch sind die daraus resultierenden dipolaren Kopplungen sehr klein und liegen häufig im Bereich der Messungenauigkeit. Daher wird versucht, dem Molekül eine zusätzliche Orientierung zu geben. Dies kann durch den Zusatz von Bizellen[57-66], durch Wechselwirkung an Purpurmembranen[67,68] bzw. Zellulose Kristalliten[69], der Orientierung durch Gele[70] bzw. Phagen[71,72] oder der Bindung des Moleküls an ein paramagnetisches Zentrum[73-78] erreicht werden. In allen Fällen lässt sich die erreichte Orientierung durch einen Tensor zweiten Ranges beschreiben, womit sich für die Größe der dipolaren Kopplung für einen Vektor im Koordinatensystem des Tensors:

$$\left\langle D_{ij}\right\rangle = D_{zz}\sqrt{\tfrac{4\pi}{5}}\left(\left\langle Y_{20}(\theta,\varphi)\right\rangle+\sqrt{\tfrac{3}{8}}R\left(\left\langle Y_{22}(\theta,\varphi)\right\rangle+\left\langle Y_{22}{}^*(\theta,\varphi)\right\rangle\right)\right) \tag{3.2}$$

ergibt. Während (θ,φ) die Orientierung des Vektors im Koordinatensystem des Tensors darstellt, gibt D_{zz} die Größe der erreichten anteiligen Orientierung an. Im Falle der Bizellen gilt:

$$D_{zz} = A_{zz}\frac{\gamma_i\gamma_j}{r_{ij}^3}, \tag{3.3}$$

wobei A_{zz} den Anteil der Orientierung darstellt ($A_{zz}=1 \rightarrow$ Festkörper).

Wird näherungsweise von einem intern starren Molekül ausgegangen, ergibt sich mit Gleichung (3.2):

$$\begin{aligned}
\left\langle D_{ij}\right\rangle &= D_{zz}\sqrt{\tfrac{4\pi}{5}}\left(Y_{20}(\theta,\varphi)+\sqrt{\tfrac{3}{8}}R\left(Y_{22}(\theta,\varphi)+Y_{22}{}^{*}(\theta,\varphi)\right)\right)\\
&= D_{zz}\left(\frac{3\cos^2\theta-1}{2}+\frac{3}{2}R\sin^2\theta\cos 2\varphi\right)\\
&= \begin{pmatrix}\sin\theta\cos\varphi\\ \sin\theta\sin\varphi\\ \cos\theta\end{pmatrix}^T \begin{pmatrix}D_{xx}&0&0\\0&D_{yy}&0\\0&0&D_{zz}\end{pmatrix}\begin{pmatrix}\sin\theta\cos\varphi\\ \sin\theta\sin\varphi\\ \cos\theta\end{pmatrix}\quad,\\
&= \begin{pmatrix}\cos\theta_x\\ \cos\theta_y\\ \cos\theta_z\end{pmatrix}^T \begin{pmatrix}D_{xx}&0&0\\0&D_{yy}&0\\0&0&D_{zz}\end{pmatrix}\begin{pmatrix}\cos\theta_x\\ \cos\theta_y\\ \cos\theta_z\end{pmatrix}
\end{aligned}\tag{3.4}$$

mit $\left(\theta_x,\theta_y,\theta_z=\theta\right)$ als den drei Projektionswinkeln des Vektors auf die Achsen des Tensors, $D_{xx}=-D_{zz}\left(\tfrac{1}{2}-\tfrac{3}{4}R\right)$, $D_{yy}=-D_{zz}\left(\tfrac{1}{2}+\tfrac{3}{4}R\right)$ und folglich $D_{xx}+D_{yy}+D_{zz}=0$.

Häufig ist es notwendig, den Vektor nicht im Koordinatensystem des Tensors, sondern in einem anderen Molekülkoordinatensystem zu betrachten. Dies kann durch die Berücksichtigung der entsprechenden EULER-Rotation $R(\alpha,\beta,\gamma)$ in den Gleichungen (3.2) und (3.4) erfolgen. Die sich ergebende Darstellung auf Basis der drei Projektionswinkel $\left(\theta_x,\theta_y,\theta_z\right)$ mit einer rotierten Darstellung der Tensormatrix, der sogenannten SAUPE-Matrix, wird in Kapitel 3.2 verwandt:

$$\begin{aligned}
\left\langle D_{ij}\right\rangle &= \begin{pmatrix}\cos\theta_x\\ \cos\theta_y\\ \cos\theta_z\end{pmatrix}^T R^T(\alpha,\beta,\gamma)\begin{pmatrix}D_{xx}&0&0\\0&D_{yy}&0\\0&0&D_{zz}\end{pmatrix}R(\alpha,\beta,\gamma)\begin{pmatrix}\cos\theta_x\\ \cos\theta_y\\ \cos\theta_z\end{pmatrix}\\
&= \begin{pmatrix}\cos\theta_x\\ \cos\theta_y\\ \cos\theta_z\end{pmatrix}^T \begin{pmatrix}S_{xx}&S_{xy}&S_{xz}\\S_{xy}&S_{yy}&S_{yz}\\S_{xz}&S_{yz}&S_{zz}\end{pmatrix}\begin{pmatrix}\cos\theta_x\\ \cos\theta_y\\ \cos\theta_z\end{pmatrix}
\end{aligned}.\tag{3.5}$$

Hierbei wird $R(\alpha,\beta,\gamma)$ als entsprechende 3x3 Rotationsmatrix betrachtet.

Auch denkbar ist eine entsprechende EULER-Rotation $R(\alpha,\beta,\gamma)$ der Kugelflächenfunktionen in Gleichung (3.2):

$$Y_{2M} = \sum_{M'=-2}^{2} e^{-i\alpha M'} d^{2}_{M'M}\left(\beta\right) e^{-i\gamma M} Y_{2M'}\,, \tag{3.6}$$

wie in Kapitel 3.4 beschrieben wird. Die $d^{2}_{M'M}\left(\beta\right)$ sind hierbei die Elemente der reduzierten WIGNER-Rotationsmatrix[79]. In beiden Fällen ergeben sich lineare Gleichungen für die dipolare Kopplung. Im ersten Fall sind die Koeffizienten reell, im zweiten Fall komplex. Beide Gleichungen werden in den folgenden Kapiteln verwandt werden.

3.2 3D Homologiesuche

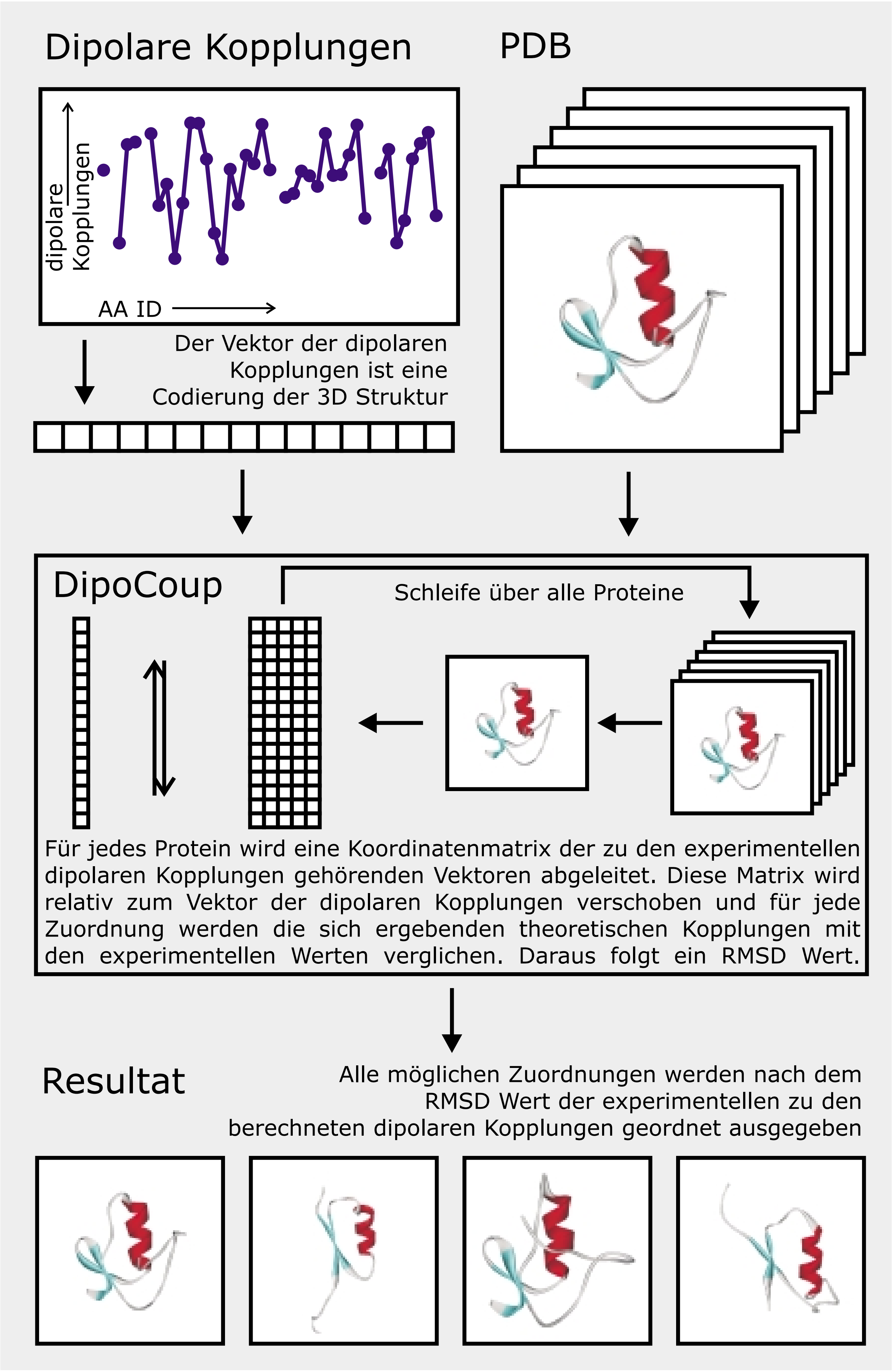
Die Zahl der alljährlich untersuchten Proteine nimmt stetig zu. Primärstrukturen von Tausenden Proteinen werden durch die Sequenzanalyse von Genen zugänglich. Zum Beispiel ergeben sich allein aus dem „human genome projekt“ die Primärstruktur einiger 100 000 Proteine des menschlichen Körpers[80,81]. Damit steigt zwangsläufig auch die Zahl der Proteine, die zum Ziel intensiver Forschung werden, da das Verständnis ihrer Funktion notwendig für eine Erweiterung unseres Bildes komplexer biochemischer Prozesse ist. Häufig ist in solchen Fällen die Aufklärung der 3D Struktur des Proteins, also der relativen Position aller Atome, unabdingbar. Heutige Methoden und Kapazitäten zur Grundlage genommen, würde die Bestimmung aller relevanten Strukturen allerdings einige Jahrhunderte dauern. Eine essentielle Beschleunigung dieser Verfahren ist daher nötig.

Häufig ist für das Verständnis der Funktion eines Proteins eine „low resolution structure“, also eine Struktur mit schlechterer Auflösung der Atompositionen, ausreichend. Eine solche Struktur lässt das Erkennen von Sekundär-, Tertiär- sowie Quartärstruktur zu und kann oftmals in erheblich kürzerer Zeit als eine hochaufgelöste Struktur bestimmt werden. Dies wird besonders deutlich, wird berücksichtig, dass sich z. Zt. ca. 50% der neu bestimmten Proteinstrukturen auf eine bekannte Faltung zurückführen lassen. Es wird daher von einer oberen Grenze in der Zahl der möglichen Proteinfaltungen ausgegangen. Die Bestimmung einer „low resolution structure“ lässt sich dann im einfachsten Fall durch die Zuordnung eines Proteins unbekannter Struktur zu einer bekannten Familie von Faltungen erzielen.

Die mit Abstand meisten Strukturen von Proteinen (~84%) werden durch Röntgenkristallstrukturanalyse bestimmt. Nachteile der Methode sind, dass die Substanzen nicht in ihrer natürlichen Umgebung, also einer Lösung, untersucht werden und dass sich deshalb ihre Struktur im Festkörper gegenüber der biologischen Realität verändert darstellen kann. Jegliche Dynamik, die in der Lösung eine erhebliche Rolle für die Funktion eines Enzyms spielt, ist nicht sichtbar. Des weiteren stellt die Kristallisation eines Proteins eine erhebliche Herausforderung dar und ist auch nicht in allen Fällen möglich. Ein Vorteil der Methode ist, dass die zu bestimmenden Proteine in ihrer Größe weit weniger limitiert sind, als bei Benutzung der zweiten bedeutenden Methode: der NMR Spektroskopie (~14%).

Die NMR Spektroskopie erlaubt das Studium eines Proteins in Lösung (ebenso wie auch im Festkörper) und ist somit in der Lage, ein realistischeres Bild von Struktur und Dynamik unter biologischen Randbedingungen zu liefern. Die Limitierung in der Größe des untersuchten Proteins wird durch Verbesserungen der Methodik[82] verringert. Die Nutzung partiell orientierender Medien hat zu vielversprechenden Ansätzen für die Beschleunigung der Strukturaufklärung mittels NMR Spektroskopie geführt. Sie sind nach der Zuordnung der Signale einfach und daher schnell auswertbare Quellen der Information. „long range" Strukturinformation wird zur Verfügung gestellt, die bisher für die NMR Spektroskopie nur indirekt zugänglich war. Damit sind relative Orientierungen von Sekundärstrukturelementen (z. B. Helix oder Faltblatt) und Tertiärstrukturelementen (Domänen) in einer einfach codierten Form erhältlich und können neben der direkten Verwendung für die Bestimmung der Struktur (Kapitel 3.3) auch zur Identifizierung der Faltung verwandt werden.

Dabei wird der experimentell bestimmte Satz dipolarer Kopplungen als dreidimensionale Repräsentation der Struktur verstanden und mit bekannten dreidimensionalen Strukturen (z. B. aus der PDB) verglichen (Abbildung 6). Ein Satz experimenteller $N - H^N$ für ein Protein enthält für jede Aminosäure (außer Prolin) eine reelle Zahl und kann somit als Vektor geschrieben werden. Sukzessive wird jedes Protein der Datenbank einzeln analysiert. Die Aminosäuresequenz eines Proteins stellt ebenfalls einen Vektor dar und kann somit mit dem Vektor der dipolaren Kopplungen überlagert werden. Um alle möglichen Zuordnungen der beiden Vektoren zueinander auf ihre 3D Strukturähnlichkeit hin überprüfen zu können, müssen beide Vektoren gegeneinander verschoben werden.



Der Vektor der dipolaren Kopplungen ist eine Codierung der 3D Struktur

Für jedes Protein wird eine Koordinatenmatrix der zu den experimentellen dipolaren Kopplungen gehörenden Vektoren abgeleitet. Diese Matrix wird relativ zum Vektor der dipolaren Kopplungen verschoben und für jede Zuordnung werden die sich ergebenden theoretischen Kopplungen mit den experimentellen Werten verglichen. Daraus folgt ein RMSD Wert.

Alle möglichen Zuordnungen werden nach dem RMSD Wert der experimentellen zu den berechneten dipolaren Kopplungen geordnet ausgegeben

Abbildung 6: Prinzip der Bestimmung 3D homologer Proteinfaltungen auf Basis dipolarer Kopplungen und der PDB. Die experimentellen Daten werden mit theoretisch berechneten Werten bekannter Proteinstrukturen verglichen und gemäß ihrer quadratischen Abweichung geordnet.

Für jede einzelne der sich ergebenden Zuordnungen führt die lineare Darstellung dipolarer Kopplungen im Molekülkoordinatensystem (Gleichung (3.5)) zu einem überbestimmten System linearer Gleichungen, welches durch Pseudoinversion der zugehörigen rechteckigen Matrix (Moore-Penrose-Inversion) näherungsweise in $\begin{pmatrix} S_{xx} & S_{yy} & S_{xy} & S_{xz} & S_{yz} \end{pmatrix}$ gelöst werden kann[9,10] (Gleichung (3.7)).

$$\begin{pmatrix} S_{xx} \\ S_{yy} \\ S_{xy} \\ S_{xz} \\ S_{yz} \end{pmatrix} = \begin{pmatrix} \cos^2\theta_x^1 - \cos^2\theta_z^1 & \cos^2\theta_y^1 - \cos^2\theta_z^1 & \cos\theta_x^1\cos\theta_y^1 & \cos\theta_x^1\cos\theta_z^1 & \cos\theta_y^1\cos\theta_z^1 \\ \cos^2\theta_x^2 - \cos^2\theta_z^2 & \cos^2\theta_y^2 - \cos^2\theta_z^2 & \cos\theta_x^2\cos\theta_y^2 & \cos\theta_x^2\cos\theta_z^2 & \cos\theta_y^2\cos\theta_z^2 \\ \cos^2\theta_x^3 - \cos^2\theta_z^3 & \cos^2\theta_y^3 - \cos^2\theta_z^3 & \cos\theta_x^3\cos\theta_y^3 & \cos\theta_x^3\cos\theta_z^3 & \cos\theta_y^3\cos\theta_z^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cos^2\theta_x^n - \cos^2\theta_z^n & \cos^2\theta_y^n - \cos^2\theta_z^n & \cos\theta_x^n\cos\theta_y^n & \cos\theta_x^n\cos\theta_z^n & \cos\theta_y^n\cos\theta_z^n \end{pmatrix}^{-1} \bullet \begin{pmatrix} D_{exp}^1 \\ D_{exp}^2 \\ D_{exp}^3 \\ \vdots \\ D_{exp}^n \end{pmatrix} \quad (3.7)$$

Daraus ergibt sich ein Satz theoretischer dipolarer Kopplungen D_{calc}^i, die die experimentell bestimmten Werte mit minimalen RMSD – Werten erfüllen:

$$\begin{pmatrix} D_{calc}^1 \\ D_{calc}^2 \\ D_{calc}^3 \\ \vdots \\ D_{calc}^n \end{pmatrix} = \begin{pmatrix} \cos^2\theta_x^1 - \cos^2\theta_z^1 & \cos^2\theta_y^1 - \cos^2\theta_z^1 & \cos\theta_x^1\cos\theta_y^1 & \cos\theta_x^1\cos\theta_z^1 & \cos\theta_y^1\cos\theta_z^1 \\ \cos^2\theta_x^2 - \cos^2\theta_z^2 & \cos^2\theta_y^2 - \cos^2\theta_z^2 & \cos\theta_x^2\cos\theta_y^2 & \cos\theta_x^2\cos\theta_z^2 & \cos\theta_y^2\cos\theta_z^2 \\ \cos^2\theta_x^3 - \cos^2\theta_z^3 & \cos^2\theta_y^3 - \cos^2\theta_z^3 & \cos\theta_x^3\cos\theta_y^3 & \cos\theta_x^3\cos\theta_z^3 & \cos\theta_y^3\cos\theta_z^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \cos^2\theta_x^n - \cos^2\theta_z^n & \cos^2\theta_y^n - \cos^2\theta_z^n & \cos\theta_x^n\cos\theta_y^n & \cos\theta_x^n\cos\theta_z^n & \cos\theta_y^n\cos\theta_z^n \end{pmatrix} \bullet \begin{pmatrix} S_{xx} \\ S_{yy} \\ S_{xy} \\ S_{xz} \\ S_{yz} \end{pmatrix} \quad (3.8)$$

Diese Rechnung wird für alle Proteine einer repräsentativen Auswahl von Faltungen aus der PDB (z. B.[83,84]) wiederholt. Die Resultate jeder einzelnen Zuordnung werden gemäß dem Q – Wert geordnet. Dieser ist ein normierter RMSD – Wert (Gleichung (3.9)), für den $0 \le Q \le 1$ der sinnvolle Wertebereich ist. Je geringer der Q – Wert ist, umso besser stimmen die experimentellen Werte mit einer Proteinstruktur in der jeweiligen Zuordnung überein. Abbildung 6 stellt das Prinzip der Homologiesuche dar.

$$Q = \sqrt{\frac{\sum_{i=1}^{n}\left(D_{exp}^i - D_{calc}^i\right)^2}{\sum_{i=1}^{n}\left(D_{exp}^i\right)^2}} \quad (3.9)$$

Diese Prozedur wurde im Programm „DipoCoup“[85] implementiert und im WWW zur wissenschaftlichen Nutzung zur Verfügung gestellt[86]. „DipoCoup“ ist in C++ programmiert und läuft unter Windows95/98/2000/NT®.

Die Suche nach 3D homologen Strukturen spielt vor allen Dingen in der frühen Phase der Strukturaufklärung von Proteinen eine Rolle. Gesetzt den Fall, dass keine Proteine mit einer hohen Primärsequenzhomologie gefunden werden können, handelt es sich mit großer Wahrscheinlichkeit um eine unbekannte Proteinfaltung. Der Strukturaufklärung solcher neuen Faltungen kommt eine besondere Bedeutung zu. Nach der Sequenzierung des Proteins können $N - H^N$ residuale dipolare Kopplungen relativ schnell bestimmt werden[87] und im besonderen wesentlich schneller als zum Beispiel NOE Signale zugeordnet und ausgewertet sind. Die beschriebene Prozedur erlaubt es nun, diese experimentellen Daten unabhängig von Primärsequenzhomologie über bekannte Proteine oder Proteinteile zu legen und damit die Ähnlichkeit in der Faltung des Proteinrückgrats zu prüfen. Die Methode ist somit konträr zum Vergleich der Primärsequenz. Sie ist in der Lage, Proteine oder zumindest Fragmente mit ähnlicher 3D Struktur zu finden, auch wenn die Primärsequenzhomologie gering ist (vgl. Liteatur[85]). Ein weiterer entscheidender Vorteil zum Vergleich der Primärsequenz ist, dass 3D Strukturinformation direkt verglichen wird. Bei Verfahren, die auf Basis einer Primärsequenzhomologie eine Struktur bestimmen, wird die Ähnlichkeit der 3D Struktur nur angenommen, kann aber nicht bewiesen werden.

Logische Weiterentwicklungen dieser Methode sind Verfahren, die nur auf Basis dipolarer Kopplungsinformation die Struktur eines Proteins ableiten. Zwei Ansätze wurden hier vorgestellt: Delaglio et. al.[12] generieren aus der PDB eine Vielzahl von Bruchstücken mit jeweils sieben Aminosäuren. Oben vorgestelltes mathematisches Verfahren wird nun genutzt, um jeweils sieben aufeinanderfolgende dipolare $N - H^N$ Kopplungen über diese Bruchstücke zu legen. Aus den Strukturen, die am besten übereinstimmen, werden die Rückgratwinkel für einen Strukturvorschlag des Proteins gewonnen. Zumindest im Falle von Ubiquitin stimmt die so gewonnene Struktur gut mit der Röntgenstruktur überein. Hus et. al.[88] nutzen ebenfalls dieses mathematische Verfahren, um aus einem Satz von mindestens fünf verschiedenen dipolaren Kopplungen (z.B. $N - H^N$, $N - C'$, $C' - H^N$, $C^\alpha - C'$ in zwei verschiedenen Orientierungen gemessen), die für eine Aminosäure bestimmt wurden, die 3D Lage der Peptidebene im Koordinatensystem des Tensors abzuleiten. Aus diesen Orientierungen lässt sich ebenfalls ein Strukturvorschlag für das Protein ableiten, indem die einzelnen Peptidebenen in der bestimmten Orientierung aneinandergereiht werden. Auch dieses Verfahren wurde bis jetzt nur an Ubiquitin erfolgreich getestet.

Alle drei Methoden zielen auf eine Beschleunigung der Bestimmung von Proteinstrukturen ab und sind als wichtige Schritte in diese Richtung einzuordnen. So wie hier beschrieben, bleiben sie jedoch auf der Stufe einer sehr groben Struktur stehen. Seitenketten werden zum Beispiel überhaupt nicht in die Rechnung einbezogen. Des weiteren werden alle drei Verfahren in der genannten Reihenfolge auch anfälliger für experimentelle Fehler. Man benötigt mindesten fünf experimentelle Werte, um das Gleichungssystem mit fünf unbekannten Größen lösen zu können. Dieser unteren Grenze nähert man sich in den beiden letztgenannten Beispielen. Liegen einige der genutzten Vektoren nahezu parallel, kann die Matrix auch für mehr als fünf Vektoren nahezu singulär werden, und kleine experimentelle Fehler auf den dipolaren Kopplungen führen dann zu großen Abweichungen in den bestimmten Strukturen. Außerdem vernachlässigen alle Verfahren Dynamik, die vor allen Dingen in flexiblen Regionen des Proteins die dipolare Kopplung entscheidend beeinflussen kann[89,90]. Eine Einbeziehung dynamischer Modelle in diese Ansätze sowie die Nutzung mehrerer dipolarer Kopplungen, die in verschiedenen Medien an verschiedenen Vektoren gemessen sind, kann diese Verfahren sicherlich verbessern und robuster gestalten.

3.3 Nutzung dipolarer Kopplungen in der Strukturaufklärung

Die Aufklärung der 3D Struktur eines Proteins erfolgt ganz analog dem in Kapitel 2.1 vorgestellten Verfahren. Es wird eine Molekulardynamik in einem stark vereinfachten Kraftfeld als „simulated annealing“ durchgeführt, die alle experimentell bestimmten Parameter in das Kraftfeld integriert und darauf ausgerichtet ist, diese zu erfüllen. Um dipolare Kopplungen zur Bestimmung bzw. Verfeinerung der 3D Struktur nutzen zu können, müssen die entsprechenden Programme so modifiziert werden, dass sie die Information der dipolaren Kopplungen sinnvoll in Energien und resultierende Kräfte umrechnen. Das verbreitetste dieser Programme ist das in Kapitel 2.1 beschriebene „X-Plor“[14].

Die naheliegendste Implementierung würde, gemäß Gleichung (3.4), direkt die relative Orientierung des dipolaren Vektors zu dem Orientierungstensor so optimieren, dass die beobachtete dipolare Kopplung erfüllt ist. Dieser direkte Ansatz benötigt die Definition eines Tensors, relativ zu dem die Orientierung optimiert werden kann. Tjandra, et. al.[13] schlagen ein solches Protokoll für X-Plor erstmalig vor, indem sie vier zusätzliche Atome einführen, die das Koordinatensystem des Tensors definieren und weit vom Protein entfernt starr zueinander gehalten werden. Junker[91] implementierte ein ähnliches Verfahren, welches das Molekülkoordinatensystem selbst als Koordinatensystem des Tensors nutzt. Beide Methoden haben jedoch zwei entscheidende Nachteile: Der zusätzliche Tensor muss definiert werden, und die Beschränkungen sind demnach nicht intramolekular, sondern wirken vielmehr relativ zu diesem Tensor und nur indirekt auf relative Orientierungen der Vektoren innerhalb des Moleküls. Hinzu kommt die in Abbildung 7 sichtbare Komplexität der dipolaren Kopplungsinformation. In den allermeisten Fällen gibt es zwei generelle Einstellungen des Vektors relativ zur z- bzw. y-Achse des Tensors und dann wiederum unendlich viele Möglichkeiten auf einem Ellipsoiden der Sphäre. Daraus ergibt sich das entscheidende Problem dieser Implementierung: Wird, wie im Kapitel 2.1 beschrieben, von einer Zufallsstruktur des Proteins gestartet, werden die Vektoren statistisch im Raum verteilt sein. Sie werden daher jeweils zu ca. 50% auf der Nord- bzw. der Südhalbkugel liegen. Obwohl der Vektor in der realen Struktur eine definierte Orientierung haben wird, gibt es wegen der Symmetrie des Tensors auf beiden Halbkugeln Positionen, die die dipolare Kopplung erfüllen. Der Tensor selbst ist in seiner Orientierung durch dieses Zufallsmolekül gänzlich unbeschrieben. Somit sind alle Kräfte unsinnig, die aus den dipolaren Kopplungen abgeleitet werden. Zwangsläufig versucht ein angelegtes Kraftfeld die Vektoren auf der jeweiligen Halbkugel zu halten, die aber eben in 50% der Fälle gerade die falsche ist. Dies führt zu einer deutlichen Verschlechterung der Konvergenzrate (ca. 95 % → ca. 30 % der berechneten Strukturen). Somit werden dipolare Kopplungen entweder nur zur Strukturverfeinerung[91] oder mit so kleinen Kraftkonstanten in sehr langen Protokollen genutzt, dass ihr Einfluss zu Beginn der Rechnung vernachlässigbar klein ist[13].

Die Lösung der beschriebenen Probleme liefert das Übersetzen der dipolaren Kopplungen in intramolekulare Strukturinformation. In Abbildung 8 wird gezeigt, dass die eigentlich relevante Strukturinformation intramolekular ist, im einfachsten Fall der Winkel δ^{ij} zwischen zwei Vektoren, der als das Skalarprodukt berechnet werden kann:

$$\cos\delta^{ij} = \begin{pmatrix} \cos\varphi^{i}\sin\theta^{i} \\ \cos\varphi^{i}\sin\theta^{i} \\ \cos\theta^{i} \end{pmatrix}^{T} \begin{pmatrix} \cos\varphi^{j}\sin\theta^{j} \\ \cos\varphi^{j}\sin\theta^{j} \\ \cos\theta^{j} \end{pmatrix} \tag{3.10}$$

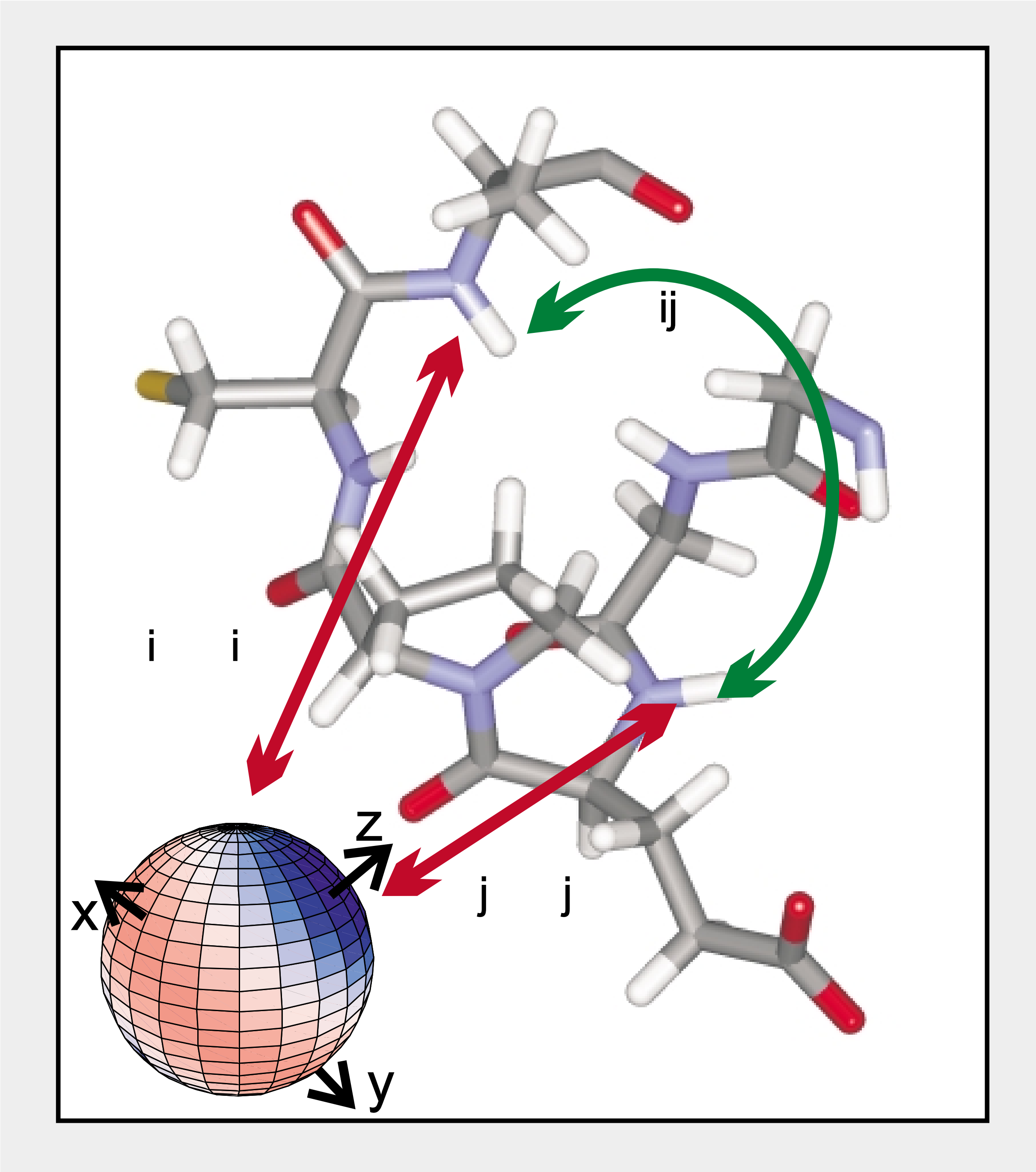


Abbildung 7: Charakter der Strukturinformation aus dipolaren Kopplungen: Die Orientierung der Vektoren zwischen den koppelnden Kernen, relativ zu dem die partielle Orientierung beschreibenden Tensor, ist durch die dipolare Kopplung bekannt. Es ergeben sich zwei generell mögliche Einstellungen für den Vektor gegenüber der z- (bzw. der y-) Achse des Tensors ( ) und dort wiederum unendlich viele relative Orientierungen der Projektion des Vektors in die xy-Ebene gegenüber der x-Achse ( ).

Für die Bestimmung der 3D Struktur ist die Übersetzung dieser indirekten Strukturinformation in eine direkte Information über den Winkel zwischen zwei dieser Vektoren vorteilhaft. Dadurch kann die Energiehyperfläche stark vereinfacht werden und eine deutlich bessere Konvergenz der Molekulardynamik wird erreicht.

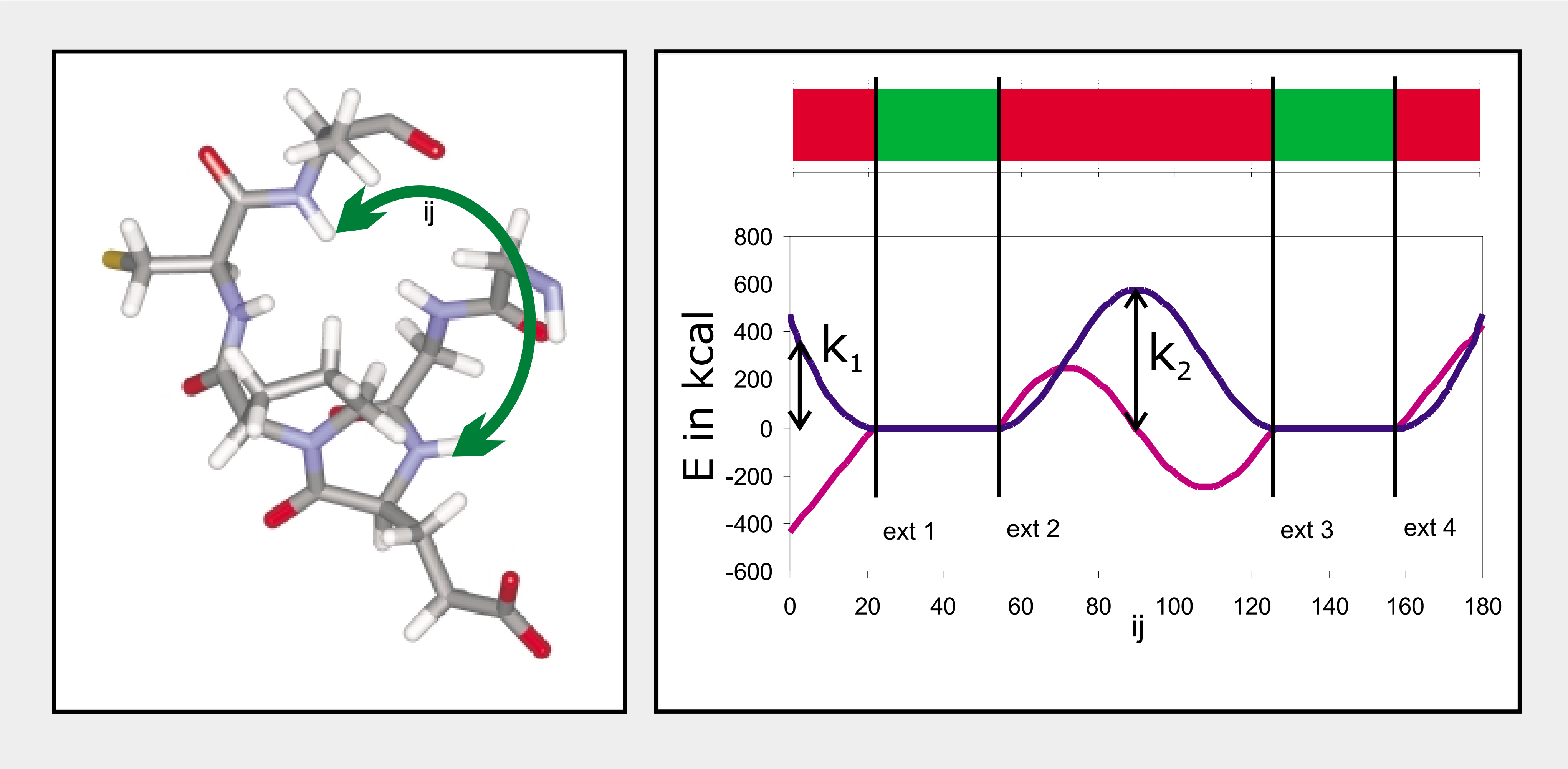


Abbildung 8: Übersetzung zweier experimentell bestimmter dipolarer Kopplungen in eine Einschränkung des theoretisch möglichen Bereiches für den Winkel zwischen beiden Vektoren. Gemäß der Symmetrie des Orientierungstensors sind zwei zueinander symmetrische Bereiche für den Winkel erlaubt und der Rest des Bereiches ist verboten. Entsprechend diesem Verhaltens wurde eine Energie (blau) und ein resultierende Kraft (rot) abgeleitet, die in X-Plor eingeführt sind. Damit können dipolare Kopplungen auf die beschriebene Art und Weise in der Strukturoptimierung genutzt werden. k_1 und k_2 sind zwei unabhängig steuerbare Kraftkonstanten. k_1 kontrolliert ein quadratisch ansteigendes Potential im Bereich zwischen den Rändern des erlaubten Bereiches und 0° bzw. 180°. k_2 steuert die Höhe der Barriere zwischen den beiden erlaubten Winkelbereichen, die als Ausschnitt der Sinusfunktion gestaltet ist. Die unabhängige Kontrolle beider Kräfte ist vorteilhaft, da die Barriere die Energiehyperfläche kompliziert und dadurch die Faltung des Proteins in der Hochtemperaturphase behindern kann. Sie wird entsprechend erst in der ertsen Abkühlphase aktiviert, während k_1 schon in der Hochtemperaturphase zum Finden der korrekten Faltung beiträgt.

Dieser Winkel δ^{ij} hängt bei bekannter Länge des Vektors von vier Freiheitsgraden $\varphi^i,\varphi^j,\theta^i,\theta^j$ ab. Er kann im Molekül zwischen 0 ° und 180 ° liegen, wird aber durch die beiden gemessenen Werte der dipolaren Kopplungen auf einen kleineren, „erlaubten" Bereich beschränkt, indem zwei der vier Freiheitsgrade eliminiert werden:

$$
\begin{aligned}
\cos\delta^{ij} = & \left(\sqrt{\frac{2(3D^i - 2D_{zz} + D_{xx} + D_{yy})}{3(\tfrac{3}{2}(D_{xx} - D_{yy})\cos\varphi^i - 2D_{zz} + D_{xx} + D_{yy})}} \sqrt{\frac{2(3D^j - 2D_{zz} + D_{xx} + D_{yy})}{3(\tfrac{3}{2}(D_{xx} - D_{yy})\cos\varphi^j - 2D_{zz} + D_{xx} + D_{yy})}}\right)\cos(\varphi^i \pm \varphi^j) \\
& \pm \left(\sqrt{1 - \frac{2(3D^i - 2D_{zz} + D_{xx} + D_{yy})}{3(\tfrac{3}{2}(D_{xx} - D_{yy})\cos\varphi^i - 2D_{zz} + D_{xx} + D_{yy})}} \sqrt{1 - \frac{2(3D^j - 2D_{zz} + D_{xx} + D_{yy})}{3(\tfrac{3}{2}(D_{xx} - D_{yy})\cos\varphi^j - 2D_{zz} + D_{xx} + D_{yy})}}\right)
\end{aligned}
\tag{3.11}
$$

Diese erlaubten Bereiche werden durch vier Winkel $\delta_{ext1},\delta_{ext2},\delta_{ext3},\delta_{ext4}$ begrenzt und werden für alle Vektoren paarweise bestimmt (Abbildung 8). Sie können anschließend als Projektionswinkelbeschränkungen direkt in der Molekulardynamikrechnung genutzt werden. Sie sind rein intramolekulare Beschränkungen und können ohne Konvergenzprobleme (95 % → 90 % der berechneten Strukturen) ab Beginn der Rechnung parallel mit den NOE – Intensitäten genutzt werden. Zu beachten ist, dass eine eventuell auftretende Barriere zwischen den zwei symmetrischen erlaubten Bereichen erst in der ersten Abkühlphase aktiviert wird, um die Energiehyperfläche in der Hochtemperaturphase einfach zu halten. Dazu wird das in Abbildung 1 vorgestellte X-Plor Protokoll aus Kapitel 2.1 mit zwei zusätzlichen Kräften ergänzt (fett gedruckter Rahmen in Abbildung 1, Gleichung (3.12)).

$$
\begin{aligned}
E^{ij}_{0\to\delta_{ext1}} &= k_1\left(\delta^{ij} - \delta^{ij}_{ext1}\right)^2 \\
E^{ij}_{\delta_{ext1}\to\delta_{ext2}} &= 0 \\
E^{ij}_{\delta_{ext2}\to\delta_{ext3}} &= k_2\cos^2\left(\pi\left(\frac{\delta^{ij} - \delta^{ij}_{ext2}}{\delta^{ij}_{ext3} - \delta^{ij}_{ext2}} - \frac{1}{2}\right)\right) \\
E^{ij}_{\delta_{ext3}\to\delta_{ext4}} &= 0 \\
E^{ij}_{\delta_{ext4}\to 180^\circ} &= k_1\left(\delta^{ij} - \delta^{ij}_{ext4}\right)^2
\end{aligned}
\tag{3.12}
$$

Der vollständige Programmcode dieser Erweiterung ist im Anhang E (Implementierung der Projektionswinkelbeschränkungen in X-Plor) gegeben. Die Implementierung ist ausführlich beschrieben und an Beispielen erläutert[92]. Die benötigten Winkelbereiche können wahlweise direkt in X-Plor oder auch vorab mit dem Programm „DipoCoup“ (siehe 3.2) berechnet werden.

Die Konvergenz der Rechnung wird durch die so eingeführte dipolare Kopplungsinformation kaum beeinflusst. Die dipolaren Kopplunge werden nach der Rechnung mit nur kleinen Abweichungen erfüllt, die innerhalb der vorgegebenen Toleranz liegen. Eine interessante Frage ist, welcher Anteil der verhältnismäßig aufwendig zu bestimmenden Abstandsinformation aus NOE – Intensitäten sich durch dipolare Kopplungen ersetzen lässt. Neben den unter Kapitel 3.2 beschriebenen Ansätzen auf NOE – Intensitäten ganz zu verzichten, kann auch ohne die Nutzung bereits bekannter Protein- oder Fragmentstrukturen ein erheblicher Teil der NOE – Information substituiert werden, wie in[92] beschrieben ist. Es wird in allen Fällen eine erhebliche Orthogonalität zwischen dipolarer und NOE – Information beobachtet. Viele Strukturen erfüllen alle NOE – Intensitäten und haben trotzdem erhebliche Fehler in den dipolaren Kopplungen. Dies resultiert aus dem sich ergänzendem Charakter der Daten. Während NOE – Intensitäten Abstandsinformation liefern, lassen sich dipolare Kopplungen wie beschrieben in intramolekulare Winkelinformation übersetzen. Dipolare Kopplungen liefern relevante zusätzliche und neue Strukturinformation für Proteine und andere biologisch relevante Verbindungen, wie zum Beispiel Zucker[93]. Darüber hinaus lässt sich aus dipolaren Kopplungen auch Information über die Dynamik von Molekülen erhalten.

3.4 Dynamische Betrachtung dipolarer Kopplungen

Das Verständnis der Moleküldynamik ist von wesentlicher Bedeutung für die Erklärung biochemischer Prozesse, da diese meist mit Änderungen in der 3D Struktur einhergehen, die von der Beweglichkeit einzelner Molekülteile wesentlich mitbestimmt werden. Dynamische Information ist aus NMR spektroskopischen Daten über Relaxations- und Austauschexperimente zugänglich. Während Relaxationsexperimente Informationen auf *ns* – Zeitskalen für Proteine liefern, ergeben Austauschexperimente Informationen im *ms* – Bereich. Der bisher kaum zugängliche μs – Bereich ist nun durch dipolare Kopplungen abgedeckt. Die experimentell bestimmten dipolaren Kopplungen ergeben sich als Mittel der tatsächlichen dipolaren Kopplung während der gesamten Messperiode. Eine Trennung der Gesamtbewegung des Moleküls von der lokalen Bewegung einzelner Vektoren oder Molekülteile führt zur gesuchten Informationen über lokale Dynamik.

Es kann gezeigt werden, dass eine rein axialsymmetrische Bewegung des Vektors eine Skalierung der experimentell beobachteten dipolaren Kopplung mit dem Ordnungsparameter S der Bewegung gleichkommt. Kompliziertere Bewegungsmodelle ergeben aber deutlich komplexere Änderungen der experimentell beobachtbaren dipolaren Kopplung.

Abbildung 9 zeigt die Größe der N – H^{N} dipolaren Kopplung während einer 10 *ns* Molekulardynamikrechnung des Proteins Ubiquitin für Asp 52. Deutliche Änderungen in der Größe der Kopplung sind ersichtlich. Der Kopplungswert ist abhängig von der jeweiligen Projektion des Vektor auf die Hauptachsen. Da jedoch nur der Mittelwert der entsprechenden Verteilung experimentell zugänglich ist, wird Information über Moleküldynamik erst durch die Kombination mehrerer dipolarer Kopplungen für einen Vektor möglich, die in verschiedenen Orientierungen gemessen wurden.

$$\frac{\langle D\rangle}{D_{zz}}=\sqrt{\tfrac{4\pi}{5}}\begin{pmatrix} \sum_{M'=-2}^{2} e^{-iM'\alpha} d_{M'0}^{2(\beta)} e^{-i0\gamma}\left\langle Y_{2M'}(\theta,\varphi)\right\rangle \\ +\sqrt{\tfrac{3}{8}}R\sum_{M'=-2}^{2} e^{-iM'\alpha} d_{M'2}^{2(\beta)} e^{-i2\gamma}\left\langle Y_{2M'}(\theta,\varphi)\right\rangle \\ +\sqrt{\tfrac{3}{8}}R\sum_{M'=-2}^{2} e^{-iM'\alpha} d_{M'-2}^{2(\beta)} e^{i2\gamma}\left\langle Y_{2M'}(\theta,\varphi)\right\rangle \end{pmatrix} \tag{3.13}$$

Gleichung (3.13) enthält die Abhängigkeit der dipolaren Kopplung eines Vektors im Molekülkoordinatensystem zu einem Orientierungstensor, der sich in einem um $R(\alpha,\beta,\gamma)$ rotierten Koordinatensystem befindet (vgl. Kapitel 3.1). Die entscheidenden Größen sind die Mittelwerte der Kugelflächenfunktionen $\left\langle Y_{2M}(\theta,\varphi)\right\rangle$ mit $m=-2,...,+2$. Sind experimentelle dipolare Kopplungen $\langle D_i\rangle$ zusammen mit den Rotationen $R(\alpha_i,\beta_i,\gamma_i)$ von mindestens fünf ($i=1,..,n\geq 5$) verschiedenen Orientierungstensoren sowie deren Größe $D_{zz,i}$ und Rhombizität R_i bekannt, ergibt sich aus Gleichung (3.13) ein komplexes lineares

Gleichungssystem. Dieses kann durch (Pseudo-)-Inversion der entsprechenden Matrix gelöst werden und liefert die zum Vektor gehörenden $\left\langle Y_{2,m}\left(\theta,\varphi\right)\right\rangle$ -werte.

Aus diesen Mittelwerten lässt sich zunächst in vollständiger Analogie zu dem von Liparo und Szabo[94,95] eingeführten Ordnungsparameter für die Relaxation

$$S^2 = \tfrac{4\pi}{5}\sum_{m=-2}^{2}\left\langle Y_{2,m}\left(\theta,\varphi\right)\right\rangle\left\langle Y_{2,m}^{*}\left(\theta,\varphi\right)\right\rangle, \tag{3.14}$$

ein Ordnungsparameter für dipolare Kopplungen ableiten. Darüber hinaus ist aus den $\left\langle Y_{2,m}\left(\theta,\varphi\right)\right\rangle$ weit mehr Information über die Bewegung des zugehörigen Vektors ableitbar. So ist zum Beispiel ein Vergleich dieser Werte mit den verschiedensten Bewegungsmodellen denkbar. Im einfachsten Fall kann zum Beispiel eine Asymmetrie η der Bewegung definiert werden:

$$\eta = \sum_{m=-2,-1,+1,+2}\left\langle Y_{2,m}\left(\theta,\varphi\right)\right\rangle\left\langle Y_{2,m}^{*}\left(\theta,\varphi\right)\right\rangle \Big/ \sum_{m=-2,-1,0,+1,+2}\left\langle Y_{2,m}\left(\theta,\varphi\right)\right\rangle\left\langle Y_{2,m}^{*}\left(\theta,\varphi\right)\right\rangle. \tag{3.15}$$

Weiterhin ist durch die Minimierung des Terms

$$\sum_{m=-2}^{2}\left(\left\langle Y_{2,m}\left(\theta,\varphi\right)\right\rangle - Y_{2,m}\left(\theta_{eff},\varphi_{eff}\right)\right)^2 \tag{3.16}$$

eine effektive Orientierung des Vektors $\left(\theta_{eff},\varphi_{eff}\right)$ zugänglich, die der mittleren Orientierung $\left(\theta_{av},\varphi_{av}\right)$ sehr nahe kommt. Im besonderen liegt sie näher als jede Orientierung, die aus nur einer dipolaren Kopplung abgeleitet wird. Somit ist also auch eine Verbesserung des starren Modells einer mittleren Proteinstruktur durch eine Berücksichtigung der Moleküldynamik kein Widerspruch.

Diese Methodik der dynamischen Diskussion dipolarer Kopplungen wird anhand einer freien 10 ns Dynamik des Proteins Ubiquitin in einem CHARMM 24 Kraftfeld entwickelt und erprobt[89]. In[90] werden die entwickelten Verfahren auf experimentelle Daten des Proteins Ubiquitin angewandt.

Die Dynamik von Proteinen spielt eine essentielle Rolle bei Faltung, Transport und Funktion. Ihr Verständnis ist daher Grundlage weitergehender Untersuchungen. Dipolare Kopplungen sind geeignet, Dynamik von Proteinen zu untersuchen. Im besonderen lassen sie die Ableitung eines modellfreien Ordnungsparameters S^2 zu, der in Analogie zum von Liparo und Szabo[94,95] eingeführten Ordnungsparameter für die Relaxation definiert werden kann. Im Unterschied zu diesem wird er aber nicht auf einer ns Zeitskala, sondern im µs und ms Bereich bestimmt, eine Zeitskala, die für die Untersuchung der Dynamik bisher nur schwer zugänglich war. Über die Ordnungsparameter hinaus lässt sich ebenso modellfrei eine mittlere Orientierung des Vektors, axialsymmetrische Anteile und nichtaxialsymmetrische Anteile der Dynamik ableiten und auswerten. Diese Daten lassen eine detaillierte Interpretation der Bewegung eines Vektors sowie eine Bestimmung seiner mittleren Position zu. Die Anwendung der so entwickelten Protokolle auf experimentelle Daten[90] lässt Proteine deutlich dynamischer erscheinen, als auf der Zeitskala der Relaxationsexperimente zu erkennen ist. Der mittlere Ordnungsparameter sinkt von 0.8 auf 0.6. Dies entspricht einer Vergrößerung des mittleren Öffnungswinkels der Bewegung von 22 ° auf 33 °. Die Bewegungen weichen teilweise bis zu 40% und im Mittel 15% von der Axialsymmetrie ab. Weitere Einsichten in dynamische Effekte auf dieser Zeitskala würde die Wiederholung der Experimente und Berechnungen mit einem weniger globulären und starren Protein als Ubiquitin liefern.

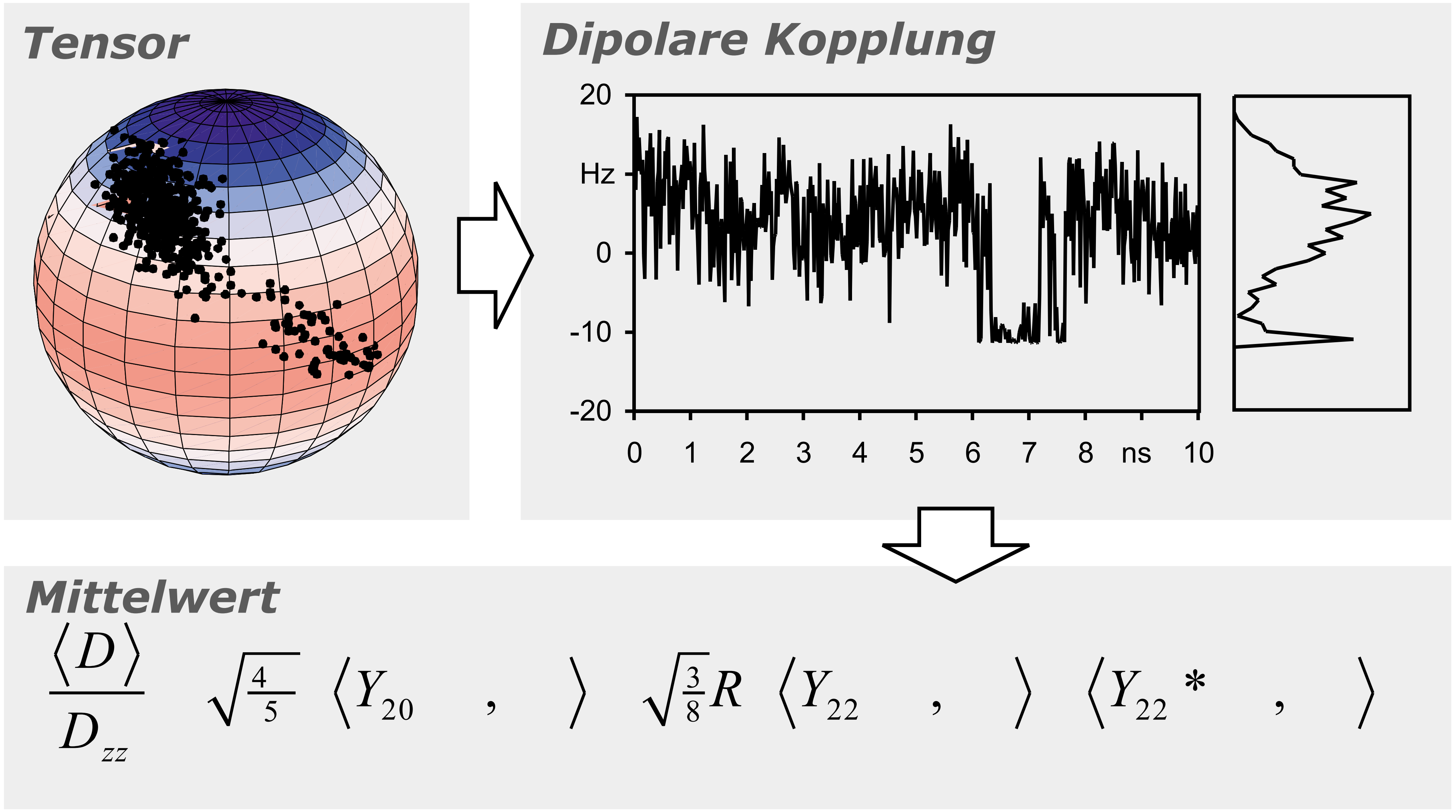

Tensor
Dipolare Kopplung
20
Hz
0
-10
-20
0
1
2
3
4
5
6
7
8
ns
10
Mittelwert
$\frac{\langle D \rangle}{D_{zz}} \quad \sqrt{\frac{4}{5}} \left\langle Y_{20} \quad , \quad \right\rangle \quad \sqrt{\frac{3}{8}}R \left\langle Y_{22} \quad , \quad \right\rangle \left\langle Y_{22}{}^{*} \quad , \quad \right\rangle$

Abbildung 9: Dynamische Information aus dipolaren Kopplungen: Experimentell zugänglich ist der Mittelwert der dipolaren Kopplung über den Zeitraum der Messung. Diese Information spiegelt neben der Geometrie, auch die Dynamik des Vektors zwischen den koppelnden Kernen wieder. Misst man sie in verschiedenen partiellen Orientierungen, kann man die die Dynamik beschreibenden Mittelwerte $<Y_{2\ M}>$ und einen Ordnungsparameter S^2 aus dipolaren Kopplungen ableiten.

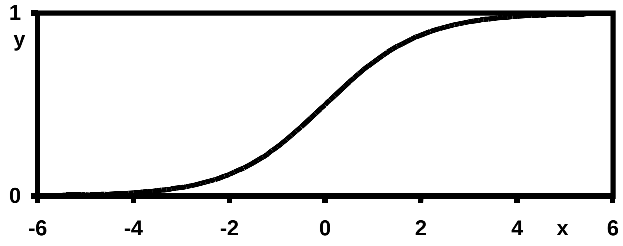
4 Kohlenstoff ^{13}C NMR chemische Verschiebungen in der Strukturaufklärung

4.1 Neuronale Netze und genetische Algorithmen in der Chemie

Neuronale Netze und genetische Algorithmen spielen eine ständig wachsende Rolle bei der Auswertung von numerischen Daten in der Chemie und Biochemie[29,96,97].

Künstliche Neuronen werden in Computerprogrammen simuliert, sind aber in ihrem Aufbau und ihrer Funktionsweise dem natürlichen Vorbild entlehnt. Abbildung 10 zeigt ein künstliches und ein natürliches Neuron im Vergleich. Die drei essentiellen Schritte der Datenverarbeitung: die Wichtung der Information verschiedener Quellen, ihre Kombination zu einem Signal sowie der Sprungcharakter der Transferfunktion sind dargestellt. Als Transferfunktion wird häufig die Sigmoidfunktion (Gleichung (4.1), Schema 6) verwandt. Neuronen dieses grundsätzlich immer äquivalenten Aufbaus lassen sich zu künstlichen neuronalen Netzen zusammenschalten und analog zu natürlichen neuronalen Netzen trainieren, bestimmte Zusammenhänge zu „verstehen“. Obwohl sie in ihrer Komplexität (noch) weit hinter jedem natürlichen neuronalen Netz zurückstehen, können sie, auf verhältnismäßig kleine Probleme spezialisiert, bessere Ergebnisse als diese erzielen.

$$y = \frac{1}{1+e^{-x}} \tag{4.1}$$



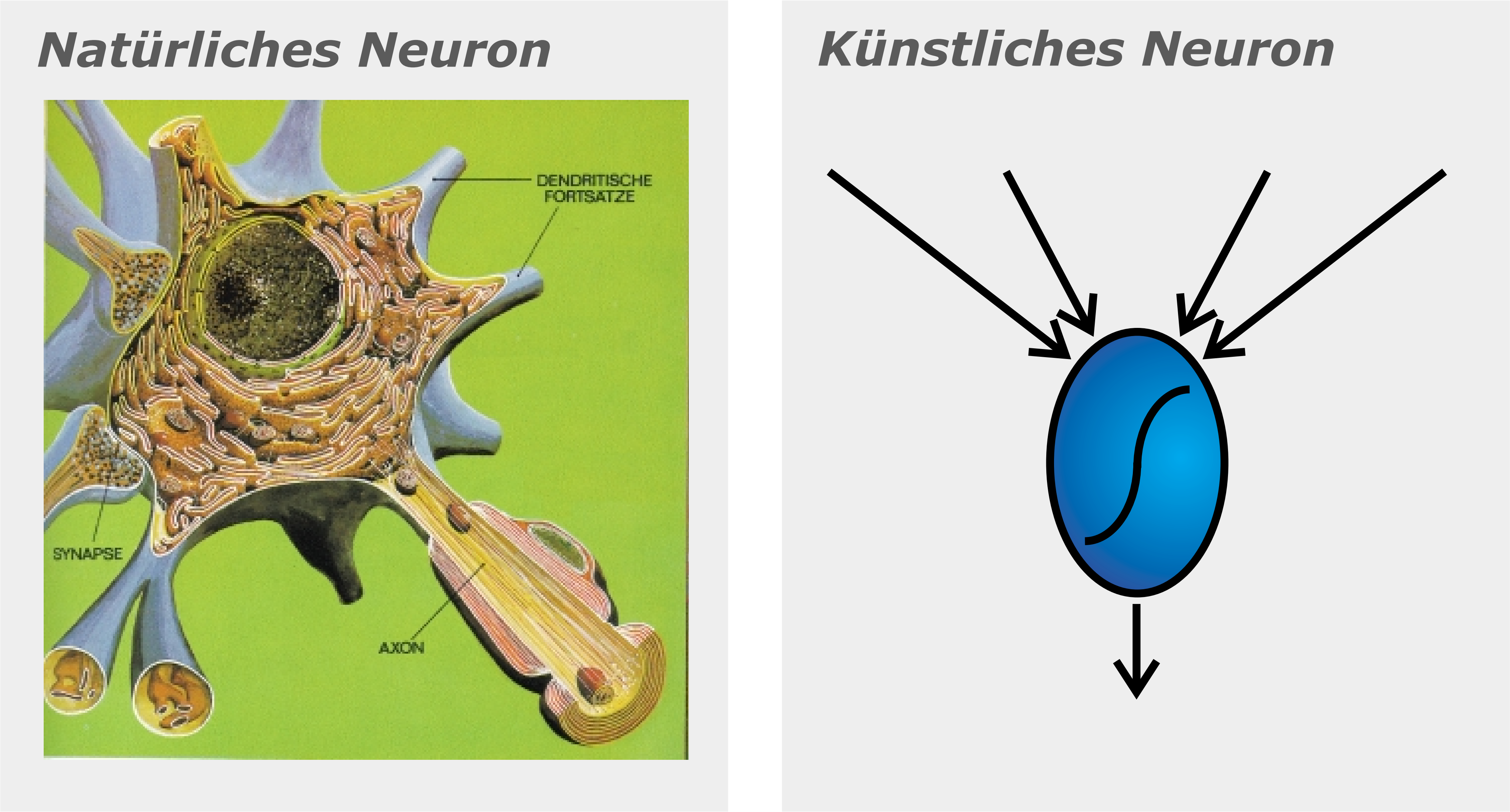
Schema 6: Graph der Sigmoidfunktion

Ein prädestiniertes Anwendungsgebiet neuronaler Netze sind quantitative Struktur – Eigenschafts – Beziehungen (QSPR) und quantitative Struktur – Aktivitäts – Beziehungen (QSAR). Insbesondere wenn kein Model für den mathematischen Zusammenhang besteht, also eine direkte Anpassung der Parameter des Modells an die experimentellen Daten unmöglich ist, sind neuronale Netze die Methode der Wahl. Sie sind aufgrund ihrer komplexen Struktur flexibel genug, auch komplizierte mathematische Zusammenhänge zu beschreiben. Die mathematische Form der Beschreibung ergibt sich während des Trainings und wird daher oft als „sich selbst organisierend (self organizing)" bezeichnet. Diese Anpassung eines sehr flexiblen Modells an das Problem stellt den eigentlichen Quantensprung bei der Verwendung eines künstlichen neuronalen Netzes im Vergleich zu herkömmlichen Verfahren mit einem vorgegebenen mathematischen Modell dar. Diese Eigenschaften erlauben neuronalen Netzen mit „unscharfer" bzw. „verwaschener" Information zu arbeiten (also Datensätzen, wo der direkte Zusammenhang zwischen einem einzelnem Parameter und der Wirkung unklar ist und nur durch Kombination einer Vielzahl verschiedener Parameter erklärt werden kann). Im besonderen erzielen sie hervorragende Ergebnisse bei der Interpolation und auch bei der Extrapolation unbekannter Datensätze.

Der verbreitetste Typ neuronaler Netze wird nach seiner Trainingsmethode als „backpropagation" – Netzwerk bezeichnet. Die Neuronen werden in Schichten angeordnet, und alle Neuronen zweier benachbarter Schichten sind paarweise miteinander verbunden (Abbildung 11). Die Information fließt von den Eingängen ohne Rückkopplung zu den Ausgängen des Netzwerks. Das Training erfolgt durch die Anpassung der Gewichte an einen Datensatz mit vorgegebenen Eigenschaften (Trainingsdatensatz). Aus der Abweichung des berechneten Wertes vom Vorgabewert dieses Trainingssatzes wird eine Veränderung der Gewichte berechnet, die diese Differenz verkleinert. Das Training eines neuronalen Netzwerkes ist somit ein iterativer Prozess, der durch zwei wesentliche Trainingsparameter bestimmt wird. Gleichung (4.2) gibt die Änderung der Gewichte in einem Trainingsschritt an:

$$\Delta w_n = \eta \cdot \delta y + \alpha \cdot \Delta w_{n-1}. \tag{4.2}$$

η ist die sogenannte Lernrate („learning rate") und bestimmt den Anteil der berechneten Veränderung, welcher tatsächlich im n – ten Schritt der Optimierung auf die Gewichte angewandt wird. α bezeichnet den Anteil der Änderung des (*n-1*) – ten Schritts, der im n – ten Schritt übernommen wird („momentum") und ist eine Art Trägheit der Bewegung, die es dem Optimierungsverfahren erlaubt, lokale Minima zu überwinden.



Natürliches Neuron

Künstliches Neuron

Abbildung 10: Vergleich eines natürlichen und eines künstlichen Neurons. Im natürlichen Neuron wird eingehende Information zunächst durch die Gestaltung des synaptischen Spaltes gewichtet, anschließend auf der Membranoberfläche der Zelle summiert und anschließend weitergeleitet, wenn das Potential einen Grenzwert übersteigt. Im künstlichen Neuron finden sich die gleichen drei Schritte der Datenverarbeitung: Wichtung (durch Multiplikation mit einer Zahl), Summenbildung und anschließende Anwendung einer Schwellwertfunktion. Somit lehnen sich künstliche Neuronen an ihre natürlichen Vorbilder an. Im Gegensatz zum natürlichen Neuron ist im künstlichen Neuron die Information aber numerisch.

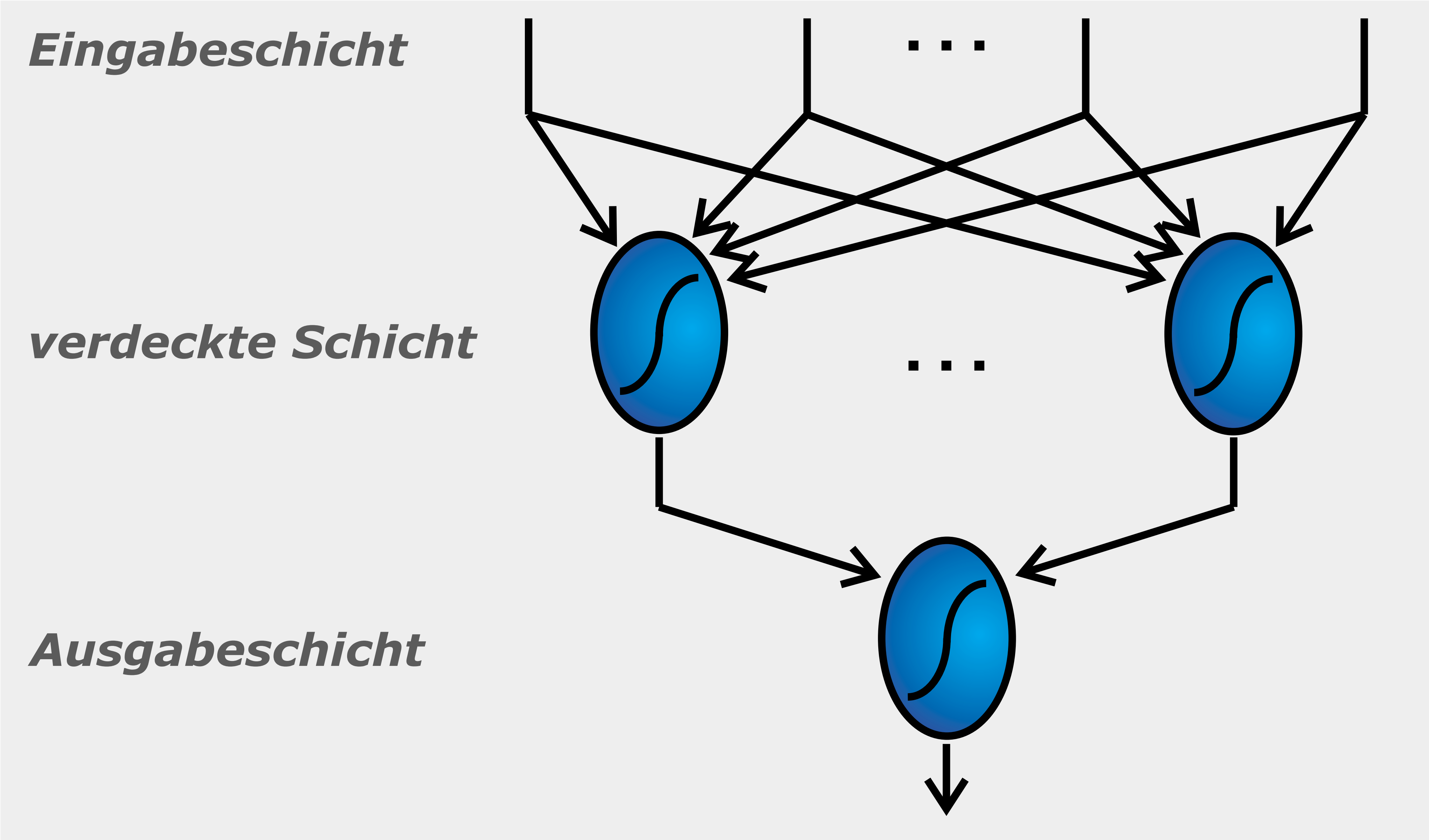


Abbildung 11: Aufbau eines dreischichtigen künstlichen neuronalen Netzwerks. An den Eingängen wird die Information in Form von Zahlen angeboten, dann im Netzwerk prozessiert und am Ausgang wird die Eigenschaft bzw. die Aktivität abgerufen. Bei diesem Netztyp sind alle Neuronen in zwei benachbarten Schichten paarweise miteinander verbunden.

Mit dem Programm „Smart“ wird ein Werkzeug zum Training und zur Nutzung von neuronalen Netzen diesen Typs im WWW für die wissenschaftliche Nutzung verfügbar gemacht[86]. Das Programm „Smart“ ist in C++ programmiert und läuft unter Windows95/98/2000/NT®.

Genetische Algorithmen sind Optimierungsverfahren. Sie beruhen auf der Anpassung eines genetischen Codes an eine Umgebung (Fitnessfunktion), ähnlich wie er in der Natur angewandt wird. Drei wesentliche Schritte: Selektion, Rekombination und Mutation müssen dazu implementiert werden. Lässt sich die Information zum Beispiel in einem Vektor mit den Elementen „Null“ und „Eins“ speichern, wird noch eine „Fitnessfunktion“ benötigt, welche aus diesem Vektor den zu optimierenden Parameter errechnet. Ein Anfangsensemble von n Individuen wird durch zufälliges Besetzen der Vektoren mit „Null“ und „Eins“ generiert. Anschließend wird für jedes dieser Individuen der zu optimierende Parameter mit der definierten „Fitnessfunktion“ berechnet. Die nächste Generation wird durch n – fache Rekombination jeweils zweier ausgewählter Individuen erhalten. Dabei werden Individuen mit einer kleineren Abweichung vom Vorgabewert (hohe Fitness) mit einer größeren Wahrscheinlichkeit ausgewählt. Die Rekombination selbst erfolgt durch das Besetzen des neuen Vektors mit jeweils einer der entsprechenden Zahlen aus den Vektoren der Eltern. Diese wird zufällig gewählt. Eine Mutation wäre dann das willkürliche Ändern einer der Zahlen. Anschließend wird wieder für jedes der neuen Individuen eine Fitness berechnet und der Prozess beginnt von vorne.

Ein Vorteil genetischer Algorithmen ist, dass sie den Raum vollständig absuchen können, ohne von lokalen Minima aufgehalten zu werden. Im Kontrast zu einem „simulated annealing“, wird hier durch die indirekte Bestimmung der Optimierungsrichtung (Selektion über Fitness) ein direkter Zug zu lokalen Minima vermieden. Dies ist vorteilhaft bei sehr komplexen Hyperflächen, aber auch zeitaufwendiger. Nachteilig wirkt sich aus, dass sie zwar im Regelfall recht nahe am absoluten Minimum, aber oft nicht genau im absoluten Minimum enden und eine anschließende Optimierung mit anderen Methoden erforderlich wird.

4.2 Berechnung ^{13}C chemischer Verschiebungen mittels neuronaler Netze

Die effektive computergestützte Nutzung der Kohlenstoff ^{13}C chemischen Verschiebung zur Strukturaufklärung setzt eine schnelle und exakte Möglichkeit zur Berechnung selbiger voraus, wie in den beiden folgenden Kapiteln sichtbar werden wird.

Die Kohlenstoff ^{13}C chemische Verschiebung ist seit der Einführung der Computertechnik eine der bestarchivierten Größen in der Chemie überhaupt. Dies liegt an der Kombination der bestechenden Einfachheit einer einzelnen Zahl mit der komplexen Information über die chemische Umgebung eines Kohlenstoffatoms, die in ihr enthalten ist. Entsprechend werden Kohlenstoff ^{13}C chemische Verschiebungen in großen Datenbanksystemen, wie zum Beispiel SPECINFO[19], gespeichert und so der weiteren Nutzung zugänglich gemacht.

Mit der Etablierung der Datenbanken wurden auch Verfahren zur Berechnung bzw. zur Abschätzung der chemischen Verschiebung entwickelt[16,17,21,98-100]. Diese Verfahren setzen bereits eine Codierung der chemischen Umgebung eines Kohlenstoffatoms voraus. Dies erfolgt meist in sphärischer Form (Abbildung 12), da der Einfluss von Substituenten auf die chemische Verschiebung mit wachsender Entfernung im Bindungsnetzwerk normalerweise abnimmt. Ausnahmen sind hier Effekte auf die chemische Verschiebung, die durch den Raum wirken, wie Wasserstoffbrückenbindungen oder sterische Wechselwirkungen. Sie sind gleichzeitig die Schwachpunkte der allermeisten Verfahren.

Alternativ werden auch ^{13}C chemische Verschiebungen aus der Molekülgeometrie durch *ab initio* Verfahren berechnet[101-103]. Beide Verfahren ergänzen sich, da die Berechnung über semiempirische Verfahren deutlich schneller, aber von der dem Verfahren zugrunde gelegten Datenbasis abhängig ist.

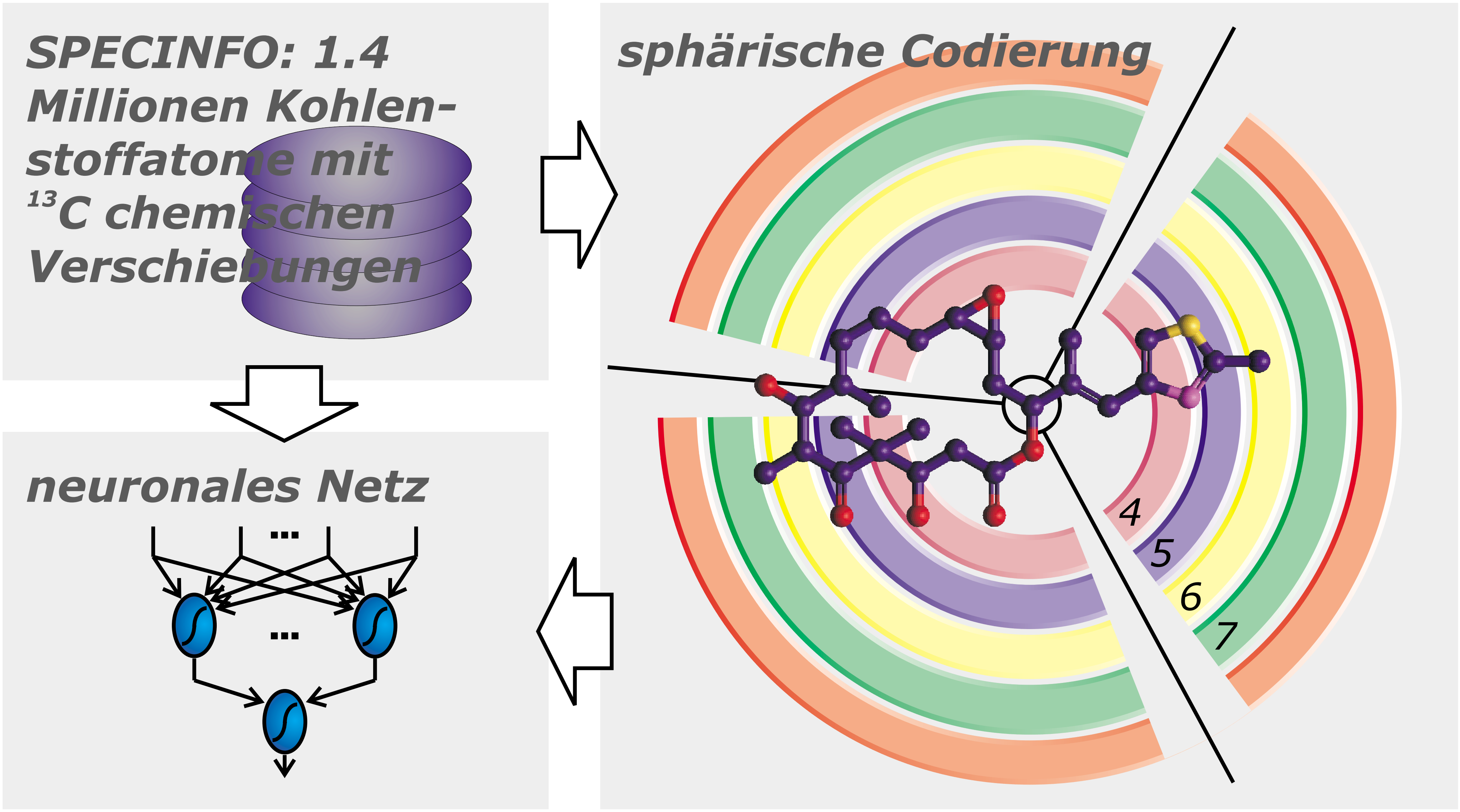
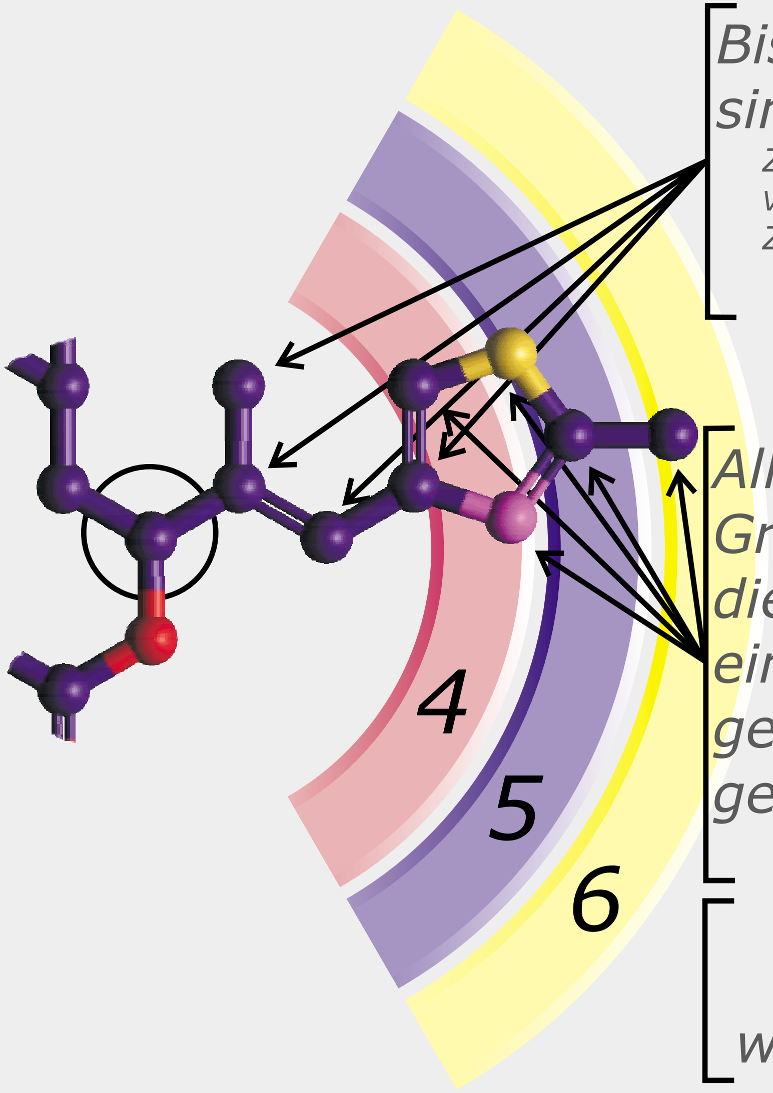

SPECINFO: 1.4 Millionen Kohlenstoffatome mit ^{13}C chemischen Verschiebungen
sphärische Codierung
4
5
6
7
neuronales Netz
...
...

Abbildung 12: Prinzip des Trainings der neuronalen Netze zur Berechnung der ^{13}C NMR chemischen Verschiebung: Aus den in der SPECINFO Datenbank abgelegten Strukturen werden numerische, sphärische Codierungen der Umgebung einzelner Kohlenstoffatome errechnet. Diese werden an den Eingängen eines neuronalen Netzes angelegt. Unter Nutzung der bekannten zugehörigen ^{13}C NMR chemischen Verschiebungen, kann das Netz nun trainiert werden, diese zu berechnen.



Bis zu 13 Atome in den ersten drei Sphären sind individuell codiert mit jeweils 8 Zahlen:

Zahl der Außenelektronen, Periode, Elektronegativität,
VAN DER WAALS Radius, Hybridisierung, Bindungstyp zum vorherigen Atom,
Zahl der gebundenen Wasserstoffatome, Zahl der Ringschlüsse

Alle Atome in weiteren Sphären sind in 32 Gruppen unterteilt. Die Zahl der Atome in diesen Gruppen in vier weiteren Sphären und einer "Summensphäre" wird als Eingabe genutzt. Die Zahlen werden für - und - gebundene Atome getrennt betrachtet.

424 Zahlen pro Substituent

werden für bis zu vier Substituenten genutzt.

Abbildung 13: **Sphärische Codierung der Umgebung eines Kohlenstoffatoms. Während in den ersten drei Sphären die Information jedes einzelnen Atoms codiert ist, werden für die folgenden Sphären nur Summenparameter errechnet. Es eregeben sich so für das Kohlenstoffatom einer Methylgruppe 424 Descriptoren und für ein quartäres Kohlenstoffatom entsprechend 1696 Descriptoren, da vier Substituenten codiert werden müssen.**

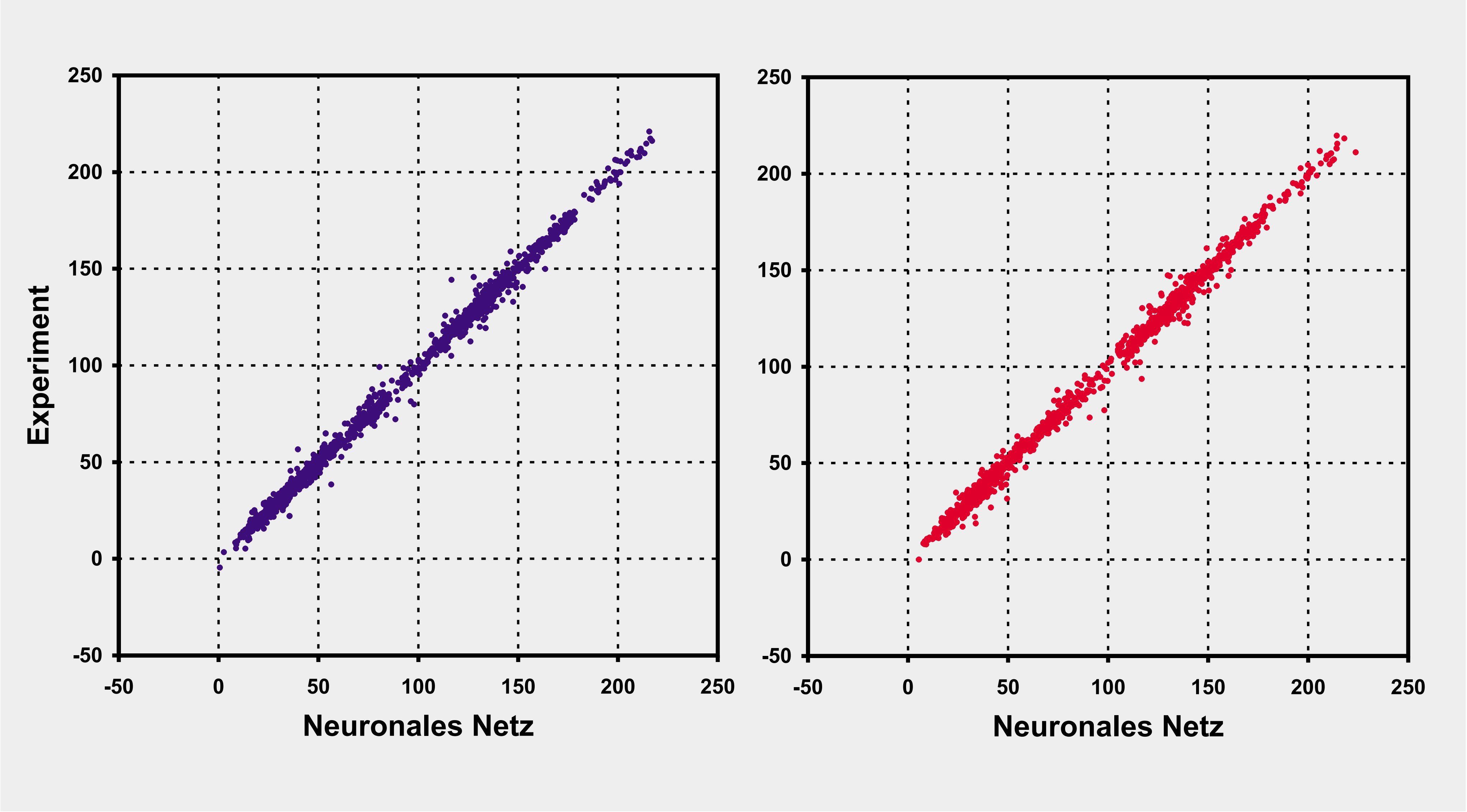


Abbildung 14: Vergleich der berechneten mit den experimentellen ^{13}C NMR chemischen Verschiebungen für 2 000 zufällig gewählte Kohlenstoffatome des Trainingsdatensatzes (links) und des unabhängigen Testdatensatzes (rechts). Die Korrelationskoeffizienten betragen in beiden Fällen R = 0.998.

Die semiempirische Berechnung der Kohlenstoff ^{13}C chemischen Verschiebung aus sphärischen Codierungen war auch eines der ersten Anwendungsgebiete für neuronale Netze in der Chemie[22,24,104-115]. Abbildung 13 zeigt, wie eine numerische, sphärische Codierung der Umgebung eines Kohlenstoffatoms abgeleitet und als Eingabevektor für ein künstliches neuronales Netz verwandt wird. Dieses wurde mit Beispielen aus über zwei Millionen chemischen Verschiebungen und den zugehörigen Atomumgebungen trainiert, die chemische Verschiebung zu berechnen. Die mittlere Abweichung der Methode beträgt 1.6 ppm. Bei einer durchschnittlichen Berechnung von 5 000 chemischen Verschiebungen pro Sekunde, ist dies mit Abstand das schnellste und exakteste bekannte Verfahren. Abbildung 14 zeigt die Korrelation der berechneten mit den experimentellen chemischen Verschiebungen. Mit der umfangreichen Datenbasis wird die bekannte organische Chemie vollständig abdeckt. Die Durchführung und die Ergebnisse sind detailliert in der Literatur[115] vorgestellt. Mit dem Programm „C_shift" wurde ein Werkzeug entwickelt, welches die Eingabe einer Struktur und die Berechnung des Kohlenstoff ^{13}C NMR Spektrums erlaubt. Das Programm „C_shift" ist in C++ programmiert und läuft unter Windows95/98/2000/NT®.

Die chemische Verschiebung eines Kohlenstoffatoms wird hauptsächlich durch seine konstitutionelle chemische Umgebung bestimmt. Sie bedingt Schwankungen im Bereich von bis zu 250 ppm. Räumliche Wechselwirkungen, wie sterische Hinderungen, Lösungsmittel und stereochemische Effekte, spielen nur bedingt in einer Minderheit der organischen Moleküle eine entscheidende Rolle. In den meisten Fällen übersteigen diese Einflüsse nicht 2 ppm (ca. 1 % der Gesamtskala), in Ausnahmefällen sind aber bis zu 25 ppm (cis – trans Stereoisomerie, ca. 10 % der Gesamtskala) möglich. Neuronale Netze können mittels einer numerischen Codierung der konstitutionellen Umgebung und ausreichend vielen Beispielen trainiert werden, die chemische Verschiebung der Kohlenstoffatome organischer Verbindungen zu berechnen. Dabei kombinieren sie die Vorteile bisher bekannter Datenbankabschätzungen (hohe Genauigkeit) und Inkrementverfahren (hohe Geschwindigkeit). Darüber hinaus sind sie auch in der Lage, gute Vorhersagen für organische Moleküle zu treffen, die in der Datenbank schlecht repräsentiert bzw. durch Inkremente nicht beschrieben sind. Ein genereller Nachteil dieser Implementierung ist, dass man im Gegensatz zu Datenbankverfahren nicht auf die Rohdaten zurückgreifen kann. Außerdem werden die meist kleinen Effekte durch räumliche Wechselwirkungen vernachlässigt. Die hohe

Geschwindigkeit prädestiniert das Verfahren zur Kombination mit Strukturgeneratoren wie im folgenden deutlich werden wird.

4.3 Auswertung der Resultate von Strukturgeneratoren

Strukturgeneratoren erstellen große Ensemble möglicher Konstitutionen unter Berücksichtigung von Randbedingungen. Diese Randbedingungen können im einfachsten Fall nur die Summenformel umfassen (Programm MOLGEN[116,117]), aber auch komplexere Informationen wie Bindungsmuster aus zweidimensionalen NMR Spektren nutzen (Programm COCON[118-121]).

Schon für eine relativ kleine Anzahl von z. B. 20 Nichtwasserstoffatomen werden Milliarden von Konstitutionen möglich. Selbst unter Einbeziehung so komplexer Randbedingungen wie zweidimensionale NMR Spektren, lässt sich diese Zahl oft nicht auf eine einzige Lösung reduzieren. Diese kombinatorische Explosion beschränkt alle diese Verfahren in ihrer Nutzung, da eine nachträgliche Analyse von einigen 1 000 generierten Strukturvorschlägen per Hand unmöglich ist.

Die Zahl der generierten Strukturen kann letztendlich nur durch die effiziente Nutzung weiterer experimenteller Informationen verkleinert werden. Zwei Möglichkeiten sind die Einführung weiterer Randbedingungen (z. B. Nutzung weiterer 2D Spektren in COCON) oder die nachträgliche Validierung der generierten Strukturen mit weiteren experimentellen Daten. Diese Validierung ist umso effektiver, je orthogonaler die verwandte Information zu den bereits genutzten Randbedingungen ist.

So wird zum Beispiel im Programm COCON die Kohlenstoff ^{13}C chemische Verschiebung nicht genutzt, stellt aber eine deutlich orthogonale Information zu den Verknüpfungsmustern dar, die aus 2D NMR Spektren gewonnen werden. Eine naheliegende Anwendung der im Kapitel 4.2 vorgestellten schnellen Berechnung chemischer Verschiebungen ist der nachträgliche Vergleich des experimentellen ^{13}C NMR Spektrums und der durch das neuronale Netz berechneten Spektren für alle generierten Strukturen. Mit dieser Methode lassen sich alle Strukturvorschläge gemäß ihrer Übereinstimmung mit dem Experiment ordnen. Für kleine Ensemble (je nach Beispiel bis einige 1 000 Strukturen) lässt sich direkt die richtige Konstitution ableiten. Im Falle größerer Ensemble (einige 10 000 Strukturen) kann die Zahl der Strukturvorschläge auf diejenigen reduziert werden, die mit dem experimentellen Spektrum in Einklang zu bringen sind.

Die Methode lässt sich weiterhin durch die Kombination mit einer Substruktursuche ergänzen. Diese erlaubt die Suche nach Basisstrukturen und deren individuellen Vergleich mit dem ^{13}C NMR Spektrum. Damit können im Falle sehr großer Strukturensemble zumindest die wahrscheinlichsten Basisstrukturelemente (z. B. bestimmte Ringssysteme) extrahiert werden. Die Durchführung und die Ergebnisse sind detailliert in der Literatur[122] vorgestellt. Das Programm „Analyze“ führt die Berechnung aus, ist in C++ programmiert und läuft unter Windows95/98/2000/NT®.

Das ^{13}C NMR Spektrum einer organischen Verbindung stellt eine detaillierte Beschreibung seiner Struktur dar. Demzufolge kann es genutzt werden, um Vorschläge von Strukturgeneratoren zu überprüfen. Ist das experimentelle ^{13}C NMR Spektrum bekannt, kann für alle Moleküle eines Ensembles, welches zuvor mit dem Strukturgenerator CoCon erstellt wurde, das ^{13}C NMR Spektrum berechnet werden. Die Zahl der wahrscheinlichen Strukturen kann so in wenigen Sekunden oder Minuten auf ca. 1 ‰ der Strukturen eingeschränkt werden, die eine geringe Abweichung zum experimentellen NMR Spektrum haben. Die Kombination mit einer Substrukturanalyse erlaubt weiterhin die Erkennung wahrscheinlicher, geschlossener Ringsysteme und gibt einen Überblick über die Struktur des generierten Konstitutionssubraumes.

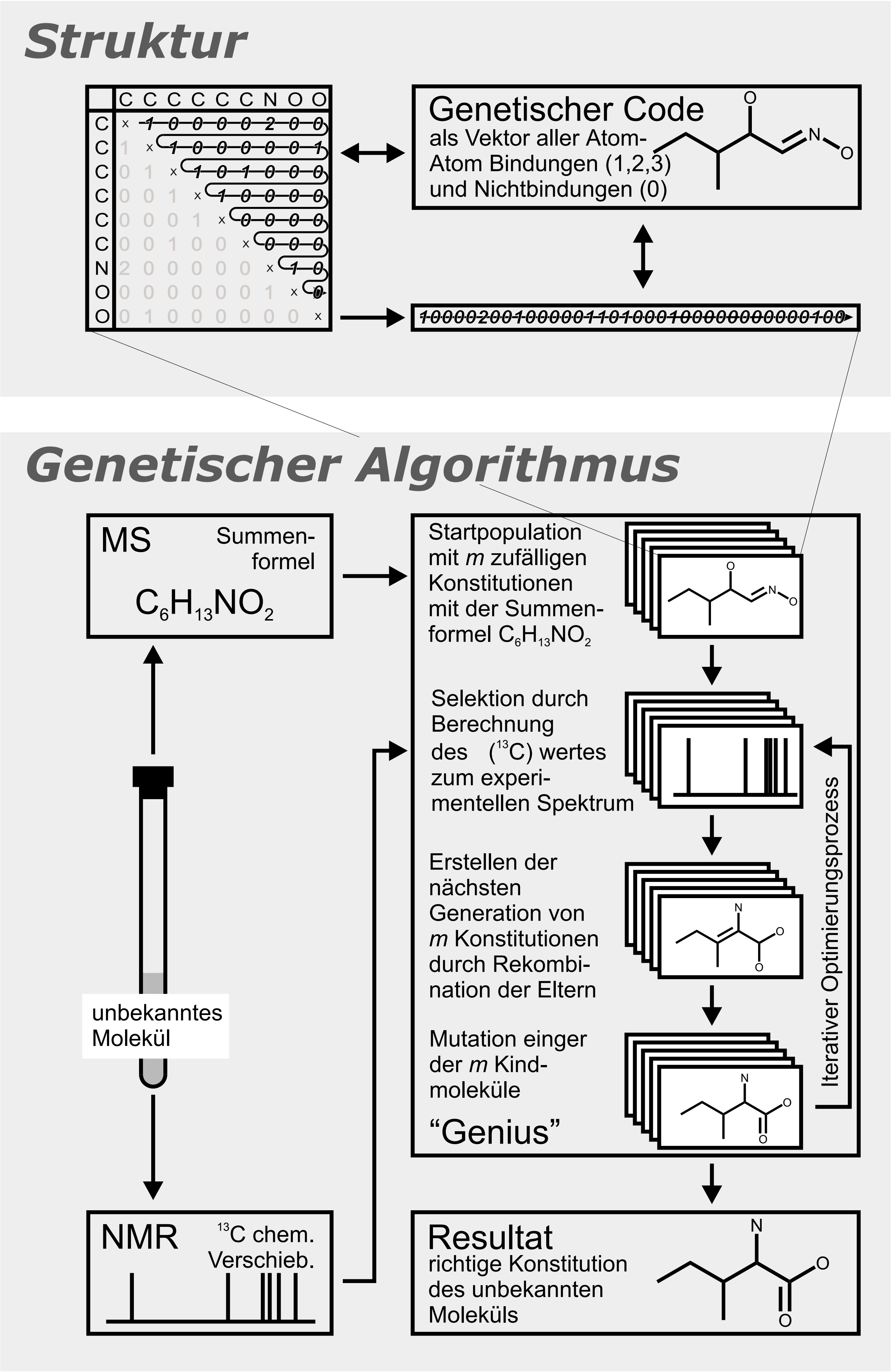
4.4 Ab initio Strukturbestimmung mit Hilfe eines genetischen Algorithmus

Genetische Algorithmen sind in der Chemie noch weit weniger verbreitet als künstliche neuronale Netze (vgl. Kapitel 4.1). Eine der wichtigsten Fragestellungen bei der Nutzung genetischer Algorithmen ist die sinnvolle Implementierung der Individuen: Die beschriebenen Schritte eines genetischen Algorithmus müssen auf die speziellen Individuen abgestimmt werden. Moleküle lassen sich auch (mit etwas Umstand) als Vektor von Nullen und Einsen darstellen, wie im Kapitel 4.1 beschrieben wurde. Die dort erklärten Mechanismen für Mutation und Rekombination sind aber dann nicht sinnvoll. Sie führen in den allermeisten

Fällen nicht zu Molekülen und konsequenterweise nur in den allerwenigsten Fällen zu sinnvollen und besseren Strukturvorschlägen im Sinne der Fitnessfunktion.

Offenbar ist eine neue Implementierung genetischer Algorithmen notwendig, sollen Konstitutionsformeln von Molekülen als Individuen betrachtet werden. Mit einer solchen Implementierung ist es dann möglich, die Struktur von Molekülen direkt auf experimentell bestimmte Eigenschaften hin zu optimieren. Abbildung 15 zeigt, wie sich eine Konstitutionsformel in einen Vektor von Zahlen codieren lässt. Dazu wird ein Teil der Bindungsmatrix des Moleküls als Vektor geschrieben. Dieser Vektor enthält für jedes Atom – Atom Paare genau eine Zahl, also gerade $\frac{1}{2}N\left(N-1\right)$ Werte, wenn N die Zahl der Atome ist. Für jedes Atom – Atom Paar ist der Bindungszustand gegeben mit 0 (ungebunden), 1 (Einfachbindung), 2 (Doppelbindung) oder 3 (Dreifachbindung). Dieser Vektor wird als genetischer Code des Moleküls betrachtet. Er berücksichtigt in dieser Form nur die Konstitution, kann aber leicht erweitert werden, um auch Stereochemie zu codieren.

Um eine Kindgeneration erzeugen zu können, müssen zwei Moleküle zu einem neuen Molekül rekombiniert werden. Diese Funktion betrachtet nacheinander alle Paare von jeweils zwei Atomen, und somit alle Positionen in den entsprechenden Vektoren der beiden Elternmoleküle. Für die Bindung zwischen diesen beiden Atomen im neuen Molekül wählt sie zufällig eine der beiden Möglichkeiten, die in den Elternmolekülen vorgegeben ist. Zu testende Randbedingungen sind, dass die Struktur chemisch sinnvoll ist, am Ende ein einziges zusammenhängendes Molekül entstanden ist und dass die Zahl der nicht explizit behandelten Wasserstoffatome unverändert bleibt. Mutation wird durch das Ändern einer Bindung erreicht. So kann zum Beispiel eine Bindung entfernt und dafür eine neue eingefügt oder auch der Typ einer Bindung geändert werden.



	C	C	C	C	C	C	N	O	O
C	x	1	0	0	0	0	2	0	0
C	1	x	1	0	0	0	0	0	1
C	0	1	x	1	0	1	0	0	0
C	0	0	1	x	1	0	0	0	0
C	0	0	0	1	x	0	0	0	0
C	0	0	1	0	0	x	0	0	0
N	2	0	0	0	0	0	x	1	0
O	0	0	0	0	0	0	1	x	0
O	0	1	0	0	0	0	0	0	x

Struktur
Genetischer Code
als Vektor aller Atom-Atom Bindungen (1,2,3)
und Nichtbindungen (0)
O
N
O
100002001000001101000100000000000100
Genetischer Algorithmus
MS
Summenformel
$C_6H_{13}NO_2$
unbekanntes Molekül
NMR
^{13}C chem. Verschieb.
Startpopulation mit *m* zufälligen Konstitutionen mit der Summenformel $C_6H_{13}NO_2$
Selektion durch Berechnung des (^{13}C) wertes zum experimentellen Spektrum
Erstellen der nächsten Generation von *m* Konstitutionen durch Rekombination der Eltern
Mutation einger der *m* Kindmoleküle
“Genius”
Iterativer Optimierungsprozess
Resultat
richtige Konstitution des unbekannten Moleküls
N
O
O

Abbildung 15: Im oberen Teil der Abbildung ist der genetische Code eines Moleküls gegeben, welcher die Konstitution des Moleküls vollständig beschreibt. Ein Teil der symmetrischen Konnektivitätsmatrix wird hierfür in einen Vektor transformiert.

Im unteren Teil der Abbildung ist ein genetischer Algorithmus dargestellt, der unter Verwendung dieser Codiereung die Konstitution eines Moleküls auf das ^{13}C NMR Spektrum optimiert. Damit wird eine weitgehend automatische Aufklärung der Struktur möglich.

Entscheidendes Moment eines genetischen Algorithmus ist aber die Auswahl der Moleküle, die für die Rekombination verwandt werden. Diese Selektion bestimmt, auf welche Eigenschaften die Strukturen optimiert werden sollen. In Kombination mit den Ergebnissen aus den Kapiteln 4.2 und 4.3 liegt die Verwendung des ^{13}C NMR Spektrums nahe. Dieses kann, wie im Kapitel 4.2 beschrieben, schnell berechnet werden und stellt, wie in Kapitel 4.3 gezeigt, eine detaillierte Beschreibung der Struktur dar. Es eignet sich daher in besonderem Maße als Fitnessfunktion. Durch diese Kombination kann eine Konstitution gefunden werden, die das ^{13}C NMR Spektrum mit möglichst geringer Abweichung erfüllt, womit ein Ansatz zur automatisierten Strukturaufklärung gegeben ist. In Kombination mit weiteren intelligenten Interventionsmöglichkeiten, wie z. B. Listen verbotener und notwendiger Fragmente, kann ein solcher Algorithmus auch ein unterstützendes Werkzeug zur Strukturaufklärung werden. Da die Berechnung des ^{13}C NMR Spektrums nur die Konstitution einbezieht, ist aber keine Optimierung der Stereochemie oder gar der räumlichen Struktur möglich. Daher enthält der genetische Code auch diese Information nicht.

Die Implementierung und erste Ergebnisse einer Strukturaufklärung mit dem beschriebenen Ansatz sind detailliert vorgestellt[123]. Das Programm „Genius" führt die Berechnung aus, ist in C++ programmiert und läuft unter Windows95/98/2000/NT®.

Die Konstitution von Molekülen kann durch einen Vektor der Bindungszustände zwischen allen Atom – Atom Paaren beschrieben werden. Selbige Vektoren sind geeignet, in einem genetischen Algorithmus als genetischer Code von Konstitutionen betrachtet zu werden. Rekombinationsoperatoren und Mutationsoperatoren lassen sich unter Verwendung dieser Vektoren definieren. Existiert eine Funktion, die es erlaubt aus diesem Vektor eine Eigenschaft des Moleküls zu berechnen und ist weiterhin der experimentelle Wert zugänglich, kann die Konstitutionsformel optimiert werden, diese Eigenschaft zu erfüllen. Das ^{13}C NMR Spektrum stellt eine detaillierte Beschreibung der Struktur dar und ist daher im besonderen als Selektionskriterium geeignet. Mit dieser Kombination des genetischen Algorithmus mit der Spektrenberechnung durch neuronale Netze ist die Aufklärung der Konstitution von Molekülen mit bis zu 20 Nichtwasserstoffatomen möglich. Die Größe des zu durchsuchenden Strukturraumes ist die Limitierung des Verfahrens. Zum einen kann die notwendige Zeitspanne zu groß werden, zum anderen steigt die Wahrscheinlichkeit, Konstitutionen zu finden, die eine kleinere Abweichung zum experimentellen Spektrum haben als die richtige Konstitution, da auch Berechnung und Messung der ^{13}C NMR chemischen Verschiebung fehlerbehaftet sind. Beide Punkte hängen direkt mit der Berechnung der ^{13}C NMR chemischen Verschiebung zusammen. Sie ist zum einen der langsamste Schritt im Algorithmus und zum anderen auch verantwortlich für den Eintrag des Berechnungsfehlers. Die Geschwindigkeit und Genauigkeit neuronaler Netze lassen eine solchen Implementierung erstmals zu. Die Einführung weiterer Randbedingungen, wie zum Beispiel Listen verbotener und notwendiger Fragmente, kann den zugänglichen Konstitutionsraum verkleinern und dadurch diese Limitierungen zu größeren Summenformeln hin verschieben. Der Vorteil genetischer Algorithmen besteht darin, das sie große, komplexe Hyperflächen absuchen können, ohne Information über den Gradienten an einem bestimmten Punkt zu benötigen. Dies ist im besonderen wichtig, wenn die Hyperfläche wie in diesem Fall unstetig ist, und somit der Gradient nicht bestimmt werden kann. Dies unterscheidet genetische Algorithmen auch von „simulated annealing“ (vgl. Kapitel 2.1) oder „Monte Carlo“ Verfahren, die ähnlich große Hyperflächen absuchen können.

5 Sekundärstrukturbestimmung von Proteinen mittels neuronaler Netze

In der statistischen Auswertung experimenteller Daten spielen Verfahren eine große Rolle, welche es erlauben, vieldimensionale Datensätze in niederdimensionale Räume zu projizieren. Zu diesen Verfahren zählen Faktor-, Hauptkomponenten- und Clusteranalysen, die auch in der Chemometrie genutzt werden[27]. Alle diese Verfahren legen jedoch der Ableitung der niederdimensionalen Parameterrepräsentationen ein Modell (meist eine lineare Abhängigkeit) zugrunde. Eine von Livingstone et. al.[28,124] und Kocjancic et. al.[125] eingeführte Methode nutzt „symmetrische“ neuronale Netze, um Parameterrepräsentationen in ihrer Dimension zu reduzieren. Dabei wird, wie in Abbildung 16 gezeigt, ein insofern symmetrisches neuronales Netz erstellt, als dass die Zahl seiner Ausgangsneuronen *m* mit der Zahl der Eingangsneuronen übereinstimmt. In der verdeckten Schicht des dreischichtigen Netzwerkes befindet sich eine kleinere Anzahl Neuronen *n* als in der Eingangs- und Ausgangsschicht $(n < m)$. Trainiert wird dieses Netz mit der *m*-dimensionalen Parameterrepräsentationen von *l* Individuen darauf, die *m* Parameter wieder vorherzusagen. So werden alle *m* Parameter in der verdeckten Schicht durch *n* Zahlen repräsentiert. Diese *n* Zahlen stellen eine in der Dimension reduzierte Repräsentation der *m* vorgegebenen Parameter. Die Verwendung neuronaler Netze erlaubt die Einführung nichtlinearer speziell auf den Datensatz abgestimmter Transformationen in diese Verfahren. Abgesehen von der nichtlinearen Übertragung der Information ist aber die Analogie zwischen einem symmetrischen dreischichtigen Netzwerk und einer Hauptkomponentenanalyse erheblich. Komplexere Modelle werden durch die Einführung zweier weiterer Schichten von Neuronen zugänglich. Sie werden als sogenannte Codierungs- und Decodierungsschicht vor bzw. hinter der verdeckten Schicht eingefügt[124].

Eine Projektion in $n \leq 3$ Dimensionen erlaubt die Visualisierung der Daten und eröffnet damit die Möglichkeit eines besseren Verständnisses der Beziehungen zwischen den Individuen. Darüber hinaus sind die resultierenden reduzierten Parametersätze geeignet, Rechenzeit zu ersparen, ohne Information zu verlieren. Durch ihre Nutzung kann der wesentliche Informationsinhalt jedes Parameters extrahiert und für eine weitere Verarbeitung genutzt werden. Die Zahl der Parameter bleibt aber begrenzt und Überschneidungen werden vermieden.

Die ständig steigende Zahl der Primärstrukturen von Proteinen (Kapitel 3.2) lässt Parameterrepräsentationen von Aminosäuren eine besondere Bedeutung zukommen. Ein Aneinanderreihen dieser Parameter, entsprechend der Aminosäuresequenz im Protein, führt zu einer numerischen Codierung des Proteins. Diese kann nun mit dessen Eigenschaften (z. B. Bindungsverhalten gegenüber Wirkstoffen) korreliert werden. Für Aminosäuren bzw. deren Seitenketten wurden entsprechend eine Vielzahl von Parametern eingeführt und teilweise auch ähnliche Größen recht unterschiedlich definiert[25]. Hinzu kommen statistische Parameter, die sich aus den bekannten Proteinstrukturen ableiten lassen. Die Wahrscheinlichkeit, mit der eine Aminosäure in einer α-Helix bzw. in einem β-Faltblatt auftritt, ist ein Beispiel.

Eine der ersten und am intensivsten diskutierten Fragestellungen, bei der Parameter für Aminosäuren eine entscheidende Rolle spielen, ist die Vorhersage von Sekundärstrukturelementen auf Basis der Primärstruktur mittels künstlicher neuronaler Netze[26,126-133]. Verfahren, die immerhin teilweise eine Vorhersagegenauigkeit von bis zu 80% im sogenannten Q_3-wert erreichen. Dieser ist als Anteil der richtig berechneten Sekundärstruktur für die einzelnen Aminosäuren definiert, wobei nur zwischen α-Helix, β-Faltblatt und anderen Bereichen unterschieden wird.

Die Eingabeschicht dieser neuronalen Netze codiert ein Fenster in der Primärsequenz um eine bestimmte Aminosäure herum, für welche die Sekundärstruktur berechnet werden soll. Durch Verschieben dieses Fensters über die Sequenz wird die Sekundärstruktur des gesamten Proteins berechnet. Die Zahl der Eingänge des neuronalen Netzes ergibt sich konsequenterweise als Produkt aus der Größe des Fensters und der Zahl der genutzten Parameter. Die Vielfalt der existierenden experimentellen, empirischen und statistisch abgeleiteten Aminosäureparameter lassen aber eine parallele Nutzung aller dieser Parameter nicht zu, da die rechentechnisch verarbeitbare Anzahl von Netzwerkeingängen begrenzt ist. Der im folgenden vorgestellte Ansatz umgeht dieses Problem, indem zunächst aus den existierenden vieldimensionalen Parameterrepräsentationen niederdimensionale reduzierte Parameterrepräsentationen mittels der symmetrischen neuronalen Netze extrahiert werden. Diese können dann für die Berechnung der Sekundärstruktur genutzt werden und zusätzlich lassen sie eine ein-, zwei- oder dreidimensionale Visualisierung der Daten zu.

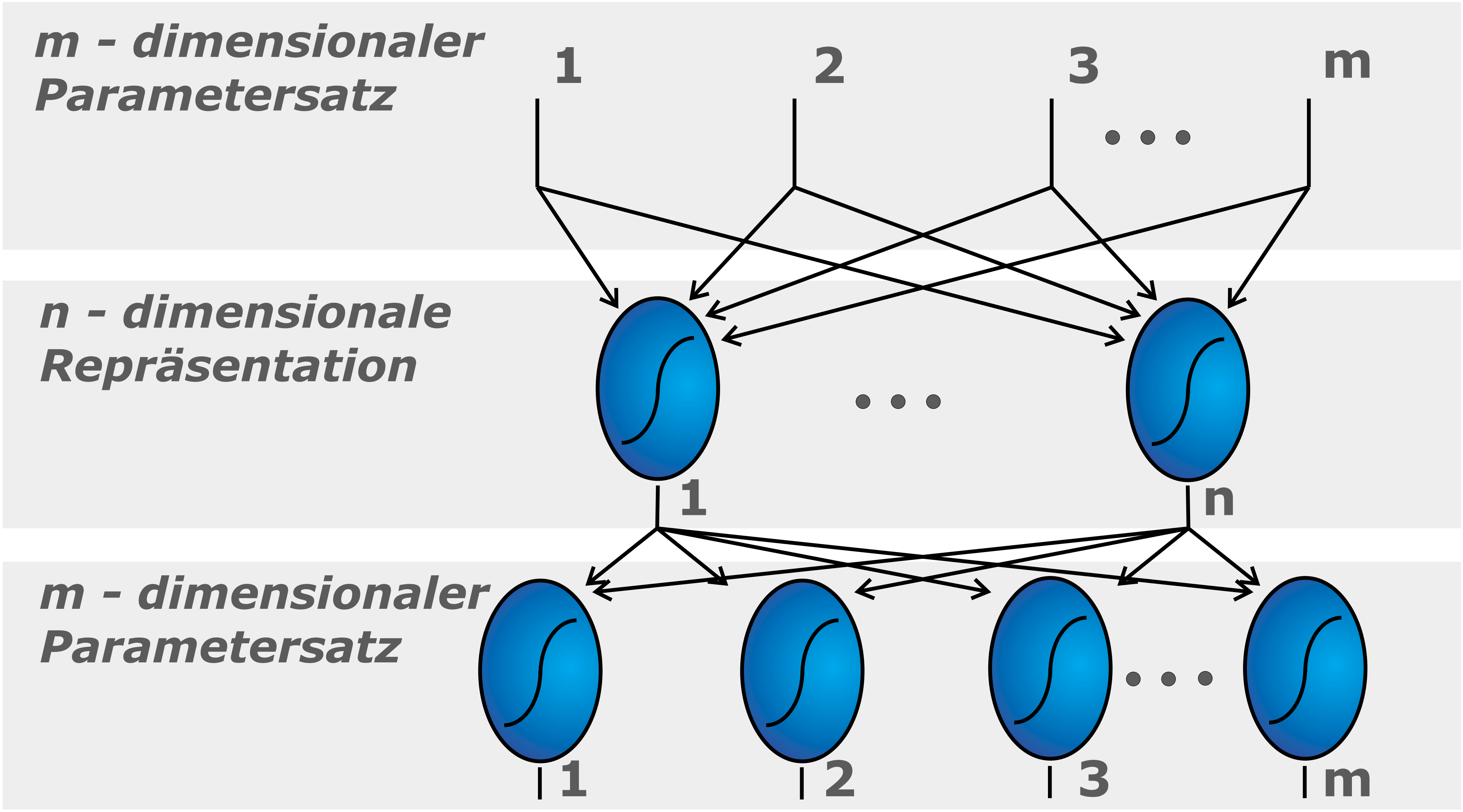


Abbildung 16: Symmetrisches neuronales Netz, wie es zur Reduktion der Dimension von Parametersätzen verwendet wird. Der *m*-dimensional Eingabevektor wird in der verdeckten Schicht auf *n* Dimensionen reduziert. Die dort beobachteten Zahlenwerte können als reduzierte Parameterrepräsentation verwandt werden, um Ähnlichkeitsanalysen durchzuführen oder als Eingabe für eine weitere Datenverarbeitung dienen.

Mittels der symmetrischen neuronalen Netzen gelingt es, aus fünf bzw. sieben dimensionalen, heterogenen Parametersätzen für die 20 proteinogenen Aminosäuren dreidimensionale Repräsentationen zu extrahieren, die die Information nahezu vollständig enthalten. Dazu werden Seitenketteneigenschaften wie Volumen, Hydrophobizität, Polarisierbarkeit, ein sterischer Parameter und der isoelektrische Punkt mit der oben erwähnten, statistisch abgeleiteten Häufigkeit, mit der eine Aminosäure in einer α-Helix bzw. in einem β-Faltblatt auftritt, kombiniert. Die niederdimensionalen Projektionen ermöglichen eine Visualisierung der Verhältnisse der Aminosäuren untereinander, wie in Abbildung 17 für eine und drei Dimensionen gezeigt ist. Aminosäuren mit ähnlichen Eigenschaften ergeben kurze Abstände, während unähnliche Aminosäuren weit voneinander entfernt sind. Gut lässt sich die Einteilung der Aminosäuren in Gruppen (z.B. basische, aromatische, aliphatische) nachvollziehen. Um die Vollständigkeit der erhaltenen Projektionen zu testen, werden neuronale Netze sowohl mit den vollständigen, als auch mit den reduzierten Parametersätzen trainiert, die Sekundärstruktur von Proteinen vorherzusagen. Die detaillierte Durchführung und die Ergebnisse dieser Analysen sind in der Literatur[134] ausführlich dargestellt. Sowohl mit den vollständigen Parametersätzen, als auch mit den abgeleiteten dreidimensionalen Repräsentationen, wird ein Q_3 wert von 66 % ± 1 % erreicht. Dieses Ergebnis zeigt an, dass in den reduzierten dreidimensionalen Repräsentationen die für die Sekundärstrukturberechnung notwendige Information enthalten ist. Das Ergebnis ist vergleichbar mit anderen Implementierungen[126]. Die teilweise erreichten Werte von Q_3 = 80 %[26] sind nur durch die Einbeziehung zusätzlicher Information möglich, die über die Primärsequenz der Proteins hinausgeht. Der Zeitgewinn beim Training des Netzes beträgt über 80% (*4h* statt *24h*) und bei der Anwendung des Netzes zur Berechnung der Sekundärstruktur immer noch mehr als 50%.

Mittels der symmetrischen neuronalen Netzen können aus fünf bzw. sieben dimensionalen, heterogenen Parametersätzen für die 20 proteinogenen Aminosäuren dreidimensionale Repräsentationen extrahiert werden, die die Information nahezu vollständig enthalten. Die niederdimensionalen Projektionen ermöglichen eine Visualisierung der Verhältnisse der Aminosäuren untereinander. Die reduzierten Parameterrepräsentationen sind geeignet, als Eingabe für ein neuronales Netz zu dienen, welches die Sekundärstruktur eines Proteins vorhersagt. Durch die Reduktion der Parametersätze erreicht man eine erhebliche Zeitersparnis, ohne Information zu verlieren. Die reduzierten Parametersätze, ein Programm zur Sekundärstrukturberechnung („Secondary“) sowie die genutzte Datenbank repräsentativer Proteinfaltungen (Kapitel 3.2) sind im WWW[86] zur wissenschaftlichen Nutzung zugänglich gemacht.

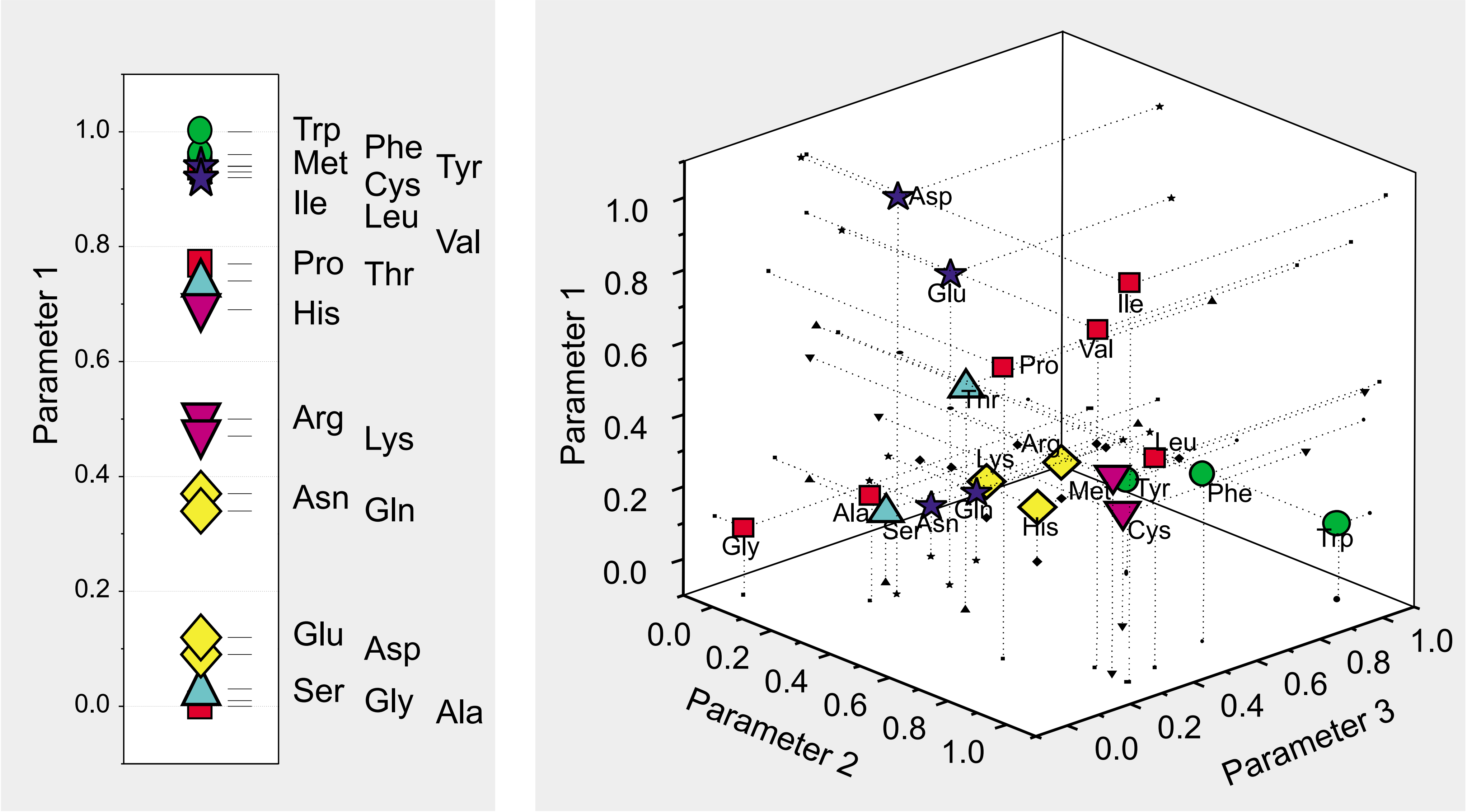


Abbildung 17: Ein- und Dreidimensionale reduzierte Parameterrepräsentationen für die 20 natürlichen proteinogenen Aminosäuren. Diese Parametersätze wurden aus dem Volumen, der Hydrophobizität, der Polarisierbarkeit, einem sterischen Parameter und dem isoelektrischen Punkt mit einem symmetrischen neuronalen Netz erhalten.

6 Wirkstoffoptimierung mittels neuronaler Netze am Beispiel des Epothilons

Die vielseitige Nutzung künstlicher neuronaler Netze für chemisch und biochemisch relevante Fragestellungen wurde bereits ausführlich in Kapitel 4.1 beschrieben[96]. Insbesondere werden neuronale Netze zur Analyse von Beziehungen zwischen der Struktur von Wirkstoffen und deren Aktivität genutzt[29]. Die hohe Eignung neuronaler Netze resultiert aus ihrer Fähigkeit, komplexe Wechselwirkungen zwischen verschiedenen Eingangsparametern und Ausgabeparametern zu erkennen und für die Erstellung eines QSPR- oder QSAR- Modells zu nutzen. Um neuronale Netze zur Etablierung eines solchen QSAR-Modells zu implementieren, sind lediglich eine numerische Codierung von Struktur und Aktivität sowie eine experimentelle Datengrundlage für den überdeckten Strukturraum notwendig. Die sich „selbst organisierenden" Netze stellen dabei einen weitgehend modellfreien Ansatz dar. Insbesondere für biologische Aktivitäten sind neuronale Netze aufgrund der genannten Vorteile geeignet, da hier häufig kein mathematisches Modell vorhanden ist, aber vielschichtige Wechselwirkungen und Abhängigkeiten eine Rolle spielen[135]. Ein Nachteil der Methode ist, dass das erstellte Modell im Anschluss nur schwer analysiert werden kann, da hierzu die Gesamtheit der oft großen Zahl der trainierten Gewichte betrachtet werden muss. Damit lässt sich durch ein neuronales Netz der bekannte Strukturraum sehr gut beschreiben, allgemeine Aussagen und Schlussfolgerungen für den gesamten Strukturraum und damit auch für bis dato nicht synthetisierte Moleküle, lassen sich aber bisher nur schwer ableiten.

Das erstellte Modell ist durch die Gesamtheit aller Gewichte vollständig beschrieben. Deshalb müssen sich aus den Gewichten auch verallgemeinerbare Aussagen über das beschriebene Problem ableiten lassen. Weiterhin sind die trainierten neuronalen Netze auf die Vorhersage eines Testdatensatzes optimiert (vgl. Kapitel 4.1). Sie können konsequenterweise auch zur Berechnung der Eigenschaften bzw. Aktivitäten bis dato nicht getesteter Substanzen genutzt werden. Im folgenden werden die Ergebnisse einer Analyse von über 200 Epothilonderivaten mit neuronalen Netzen zusammengefasst, die in der Literatur[136] detailliert beschrieben sind. Im besonderen werden Ansätze zur Lösung der genannten Nachteile vorgestellt.

Die Epothilone A and B wurden vom myxobacterium *Sorangium cellulosum* strain 90 von Höfle et. al.[46,47] isoliert. Die Entdeckung ihrer zelltoxischen Wirkung gegen Tumorzellen führte zur intensiven Erforschung ihrer Chemie und Biologie. Bollag et al.[3] entdeckten die induzierende Wirkung auf die Tubulin[48,49] Polymerisation dieser Substanzklasse, ähnlich dem Taxol[1]. Der Effekt der Stabilisierung der Mikrotubuli in Taxol resistenten Krebszellen[50] erhöhte ihr Potential für die Chemotherapie weiter[3,51,52].

Die komplette Struktur mit gelöster Stereochemie wurde von Höfle[53] publiziert (Schema 3). Wenig später wurden viele Synthesen von Vorstufen, den Epothilonen selbst und Analogen vorgeschlagen[137-141]. Viele dieser Derivate wurden auf ihre biologische Aktivität überprüft. Damit ist eine breite Datenbasis biologischer Aktivitäten bekannt und dient zu qualitativen SAR Untersuchungen[142-145]. Wang et. al.[146] stellten erste quantitative SAR durch Docking von 26 Derivaten an ein Rezeptormodell des Tubulin auf. Darüber hinaus existieren bis jetzt keine weiteren und im speziellen keine modellfreien QSAR Untersuchungen.

Nicolaou et. al.[144] publizierten für über 200 Epothilonderivate eine Konstante, die den Anteil des unter bestimmten experimentellen Bedingungen polymerisierten Tubulins beschreibt, sowie für fast 40 dieser Derivate IC_{50} Werte, die die Inhibierung des Krebszellenwachstum dreier Eierstockzelllinien unter dem Einfluss des jeweiligen Derivates beschreiben. Eine numerische Codierung dieser über 200 Epothilonderivate, erlaubt es, mittels neuronaler Netze, Modelle, sowohl zur Berechnung der Induktion der Tubulin Polymerisation, als auch direkt zur Inhibierung des Krebszellenwachstums, zu erstellen. Die eingeführten numerischen Parameter beschreiben variierte Atomtypen, Stereochemie und Substituenten. Abbildung 18 stellt das angewendete Prinzip dar: Das neuronale Netz wird dabei an die Stelle des Proteins Tubulin bzw. der gesamten Zelle gesetzt und mit den experimentellen Daten trainiert, sich gegenüber den angebotenen Epothilonderivaten genauso zu verhalten, wie Tubulin bzw. die Zelle in der Realität. Die trainierten neuronalen Netze sind danach in der Lage die Abhängigkeit der biologischen Aktivität von den definierten Strukturparametern zu beschreiben. Die Korrelation der berechneten Induktionskonstanten der Tubulinpolymerisation mit den experimentellen Werten beträgt $R = 0.73$ für den Testdatensatz. Im besonderen werden die *in vivo* bestimmten Inhibierungen des Krebszellwachstums mit sehr geringen Fehlern berechnet ($R = 0.94$ für den Testdatensatz).

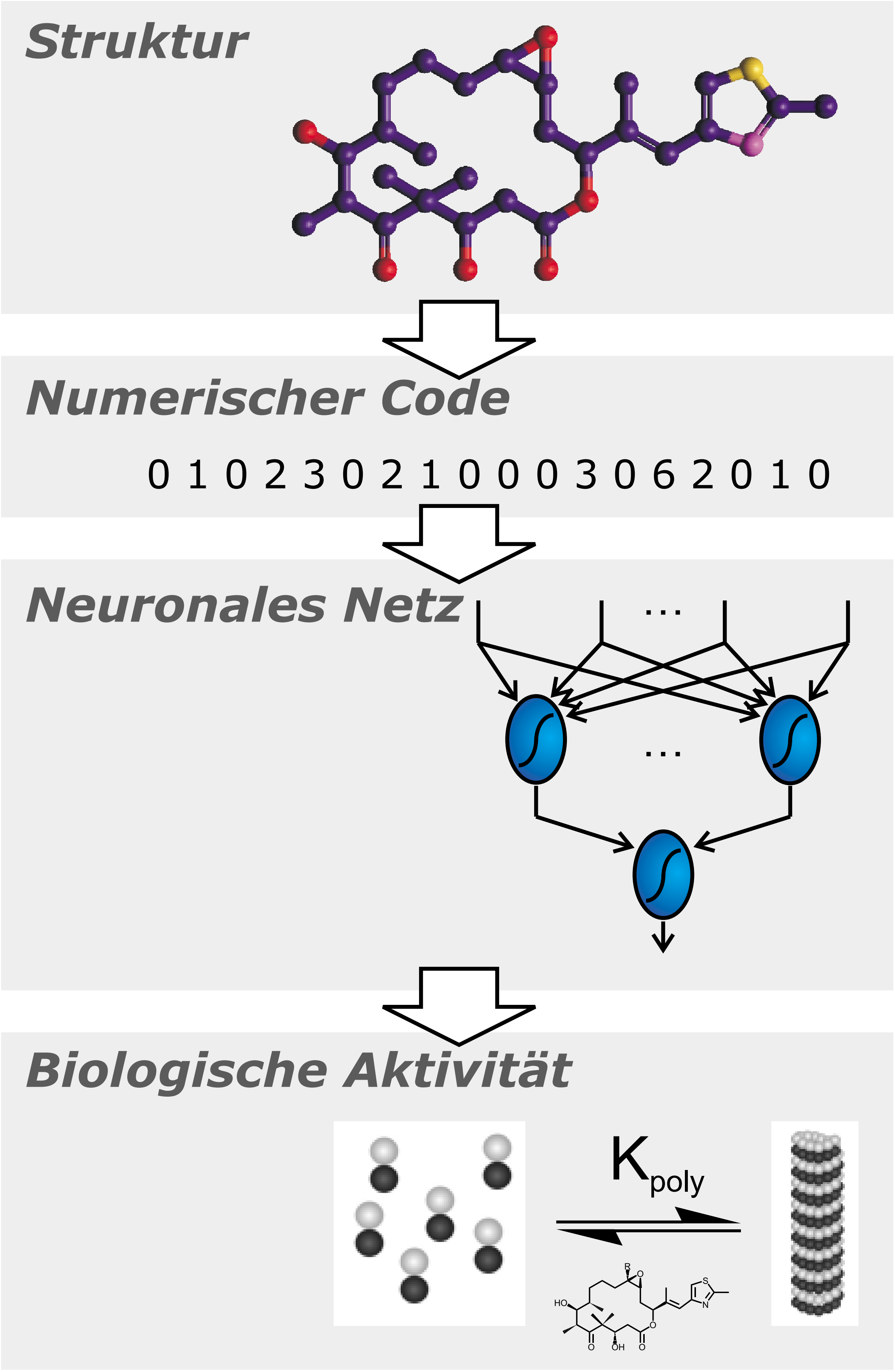


Abbildung 18: Prinzip der Aktivitätsvorhersage von Wirkstoffen mit neuronalen Netzen am Beispiel des Epothilons: Das Netz nimmt die Position des Tubulin ein und ist trainiert, sich gegenüber den dargebotenen Epothilon Derivaten genauso zu verhalten.

Die so trainierten neuronalen Netze können zunächst genutzt werden, um den Einfluss einzelner der eingeführten Strukturparameter auf die Aktivität zu prüfen. Dazu werden die Gewichte des Netzwerkes in einer Sensitivitätsanalyse untersucht. Bei diesem Verfahren wird nacheinander der Eingabewert jedes Eingangs variiert, wobei allen anderen Eingängen ein neutrales Signal (z.B. Null) zugeordnet wird. Der Wertebereich eines Ausgangs stellt nun die „Sensitivität“ dieses Ausgangsneurons gegenüber dem entsprechenden Eingang dar. Im vorliegenden Fall kann diese Information im Anschluss mit der 3D Struktur des Wirkstoffes verglichen werden, da jeder Eingang einem bestimmten Strukturelement zuzuordnen ist. Bereiche hoher Sensitivität stellen dabei wahrscheinliche Wechselwirkungsstellen mit dem Protein dar, da Veränderungen der Struktur an diesen Stellen die Wirkung stark beeinflussen. Modifiziert man diese, können die Eigenschaften des Wirkstoffes verändert und damit auch optimiert werden. Abbildung 19 vergleicht die hier erhaltenen größten Sensitivitätswerte mit der 3D Struktur der gebundenen und der freien Form des Epothilon A. Auffällig ist, dass in der gebundenen im Gegensatz zur freien Form ein Stickstoffatom, welches für die Wechselwirkung essentiell ist, frei zugänglich wird und auf einer Seite des Moleküls mit anderen potentiellen Wechselwirkungsstellen zu liegen kommt. Gegenwärtige Studien vergleichen diese Seite mit den bindenden Molekülteilen des Taxol und docken sie in ein 3D Strukturmodell des Tubulin.

Neben der Etablierung eines QSAR Modells ist mit diesen neuronalen Netzen auch die Vorhersage der biologischen Aktivität weiterer Substanzen möglich, welche durch die eingeführten numerischen Parameter beschrieben werden. Wird also das künstliche neuronale Netz mit einem vorgeschalteten Strukturgenerator kombiniert, der alle möglichen Strukturen des Parameterraumes durch Permutation der definierten Parameterwerte erzeugt (Abbildung 20), können alle möglichen Derivate gemäß ihrer berechneten Aktivität geordnet werden. Die Zahl der so erzeugten Strukturen erreicht 2.6 Milliarden, bezieht man alle 24 definierten Strukturparameter ein. Sie ist durch klassische und selbst kombinatorische Synthese kaum abzudecken. Auch wenn der Fehler der Berechnungsmethode in die Analyse einbezogen wird, lassen sich Vorschläge für aktivere Epothilonderivate finden oder zumindest verallgemeinerbare Regeln für eine gezieltere Optimierung der Wirkstoffstruktur ableiten. Im Falle der Epothilone kann weiterhin durch die parallele Berechnung der drei IC_{50} Werte die Sicherheit der Vorhersage erhöht werden. Zum einen erhöht sich die Güte des neuronalen Netzes durch diese parallele Berechnung in einem Netzwerk, da hier während des Trainings

Gemeinsamkeiten der Parameter für die Erstellung des Modells genutzt werden können. Zum anderen erhöht sich auch die Wahrscheinlichkeit, ein aktiveres Derivat zu detektieren, wenn es eine auffällige Reduktion in allen drei IC_{50} Werten zeigt. Damit hat dieses Verfahren das Potential, die Optimierung biologisch aktiver Substanzen essentiell zu beschleunigen.

Neuronale Netze sind geeignet Struktur – Aktivitäts – Beziehungen zwischen der Struktur und der biologischen Aktivität von Wirkstoffen zu beschreiben. Die *anti* – Tumor Aktivität von über 200 Derivaten des Epothilons lässt sich aus 24 Strukturparametern berechnen. Die trainierten neuronalen Netze können in einer Sensitivitätsanalyse genutzt werden, um die Bindungsstellen des Moleküls zu identifizieren und damit Hinweise auf weitere geeignete Modifikationen geben. Darüber hinaus können die trainierten neuronalen Netze in Kombination mit einem vorgeschalteten Strukturgenerator genutzt werden, um den definierten Parameterraum nach potentiell aktiveren Derivaten zu durchsuchen. Dabei werden Strukturen mit einer bis zu 1 000 mal erhöhten Wirkung gegenüber Epothilon A berechnet.

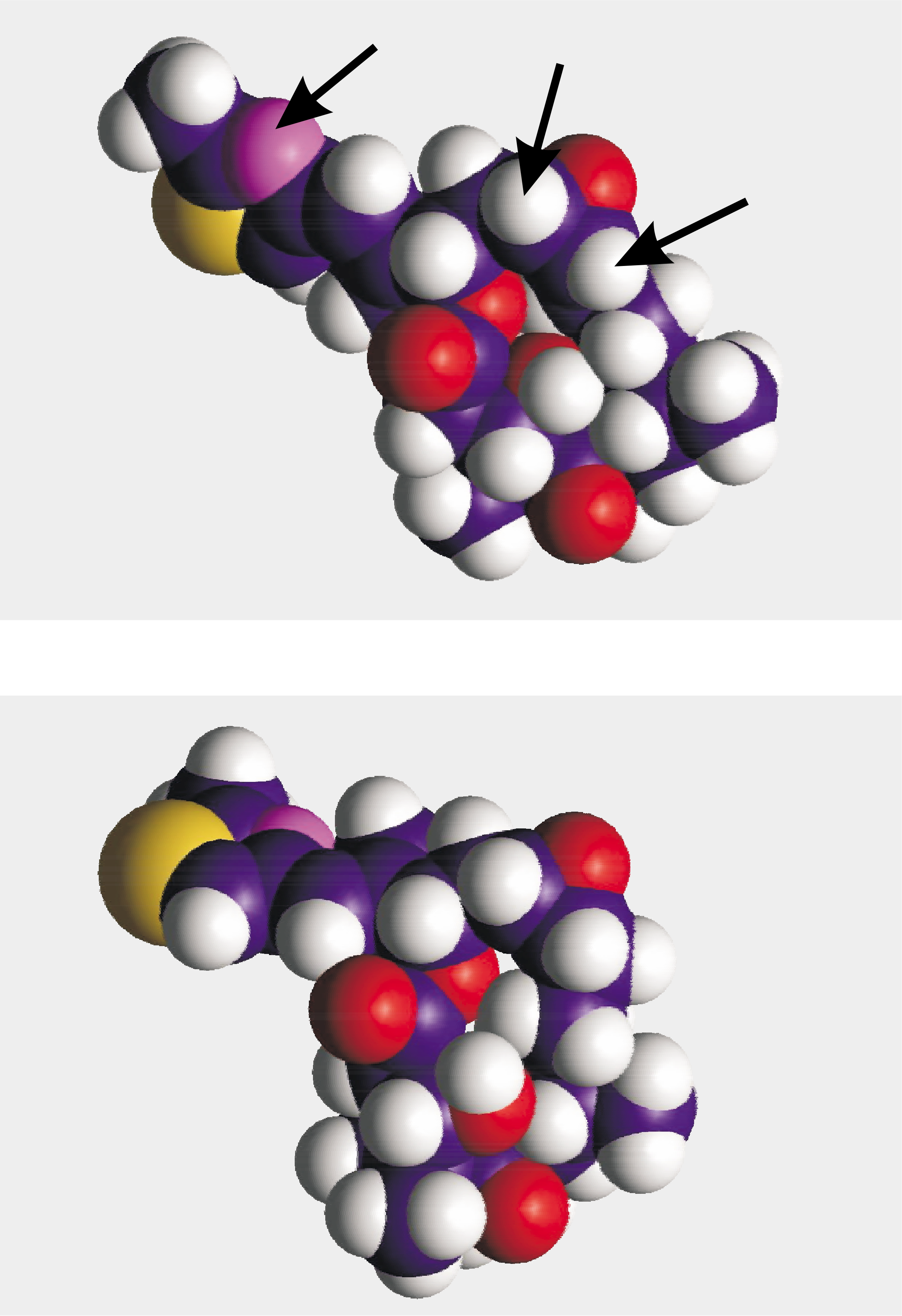


Abbildung 19: Kalottenmodell des Epothilon A in gebundener Form (oben) und in freier Form (unten). Die aus der Sensitivitätsanalyse des neuronalen Netzes für die Bindung mit Tubulin entscheidenden Strukturelemente (Stickstoff des Thiazolrings sowie Epoxid mit den möglichen Substituenten an C12 und C13, Pfeile) liegen in der gebundenen Form des Epothilon A auf einer Seite des Moleküls. In der freien Form liegt das Stickstoffatom auf der anderen Seite des Moleküls und ist zudem durch die beiden benachbarten Methylgruppen kaum zugänglich für Wechselwirkungen mit dem Protein.

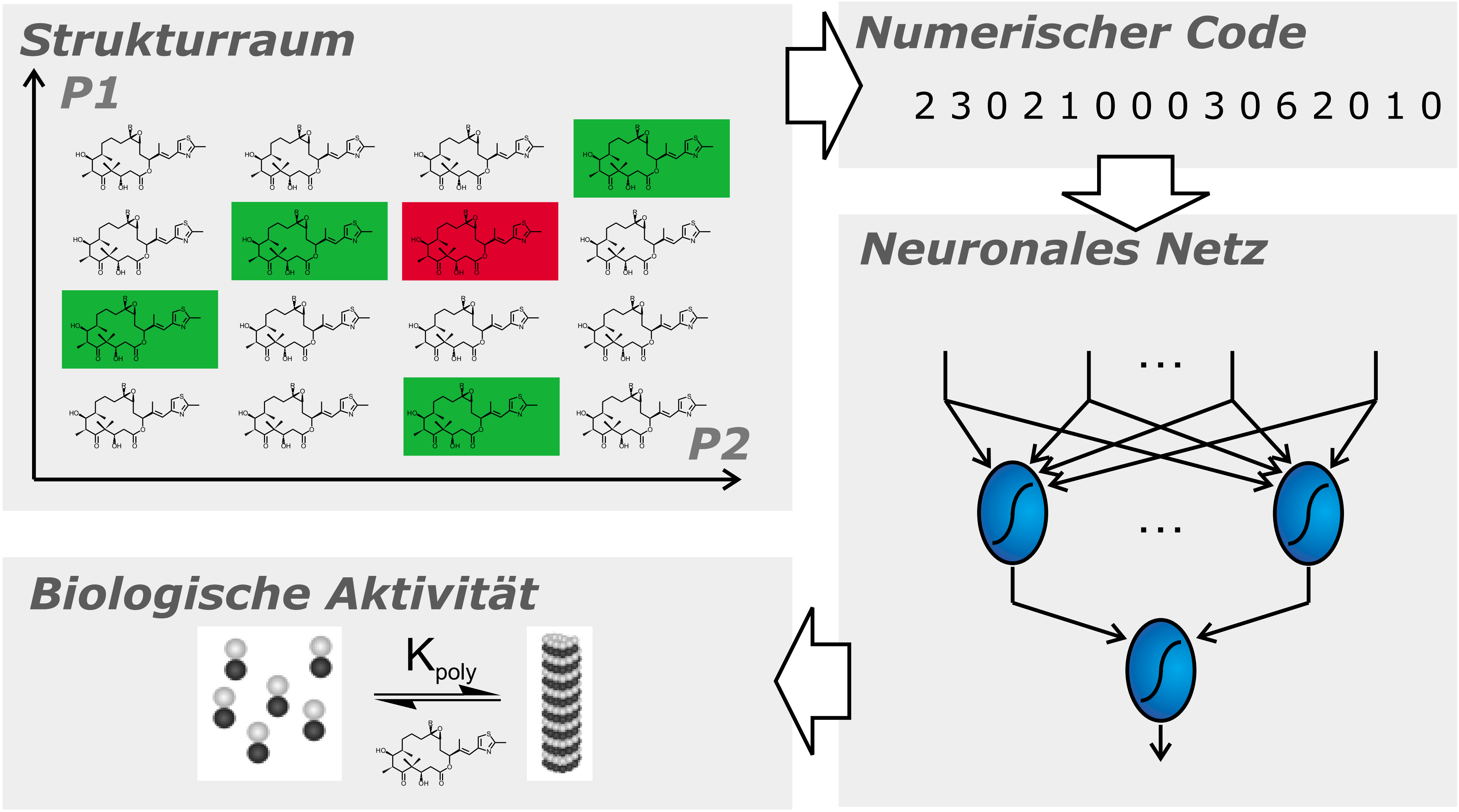


Abbildung 20: Prinzip des Absuchens eines Strukturraumes nach biologisch aktiven Verbindungen mittels eines neuronalen Netzes: Das neuronale Netz wird mit einem Unterraum des Strukturraums (hier beschrieben durch zwei Strukturparameter P1 und P2) trainiert (grün gekennzeichnete Verbindungen) und ist danach in der Lage, für alle Verbindungen die biologische Aktivität zu berechnen. Dabei können auch potentiell aktivere Substanzen gefunden werden (rot gekennzeichnete Verbindung).

7 Zusammenfassung

In der vorliegenden Arbeit werden Verfahren der Mathematik und Informatik entwickelt und eingesetzt, um Struktur, Dynamik und biologische Aktivität aus NMR spektroskopischen und empirischen Parametern zu bestimmen.

Dolastatin 10 und Epothilon A sind potentielle Wirkstoffe gegen Krebs, da sie durch Wechselwirkung mit Tubulin die Zellteilung unterbinden. Die 3D Struktur beider Wirkstoffe in Lösung und die Struktur von an Tubulin gebundenem Epothilon A wird aus NMR spektroskopischen Parametern bestimmt. Dolastatin 10 liegt in einem konformationellen Gleichgewicht zwischen der *cis* – und *trans* – Konformation in der ungewöhnlichen Aminosäure DAP vor. Beide Konformationen des flexiblen Pentapeptids können bestimmt werden mit RMSD = 1.423 Å für das *cis* – Konformer und RMSD = 1.488 Å für das *trans* – Konformer. Während das *trans* – Konformer gestreckt vorliegt, faltet das *cis* – Konformer am DAP zurück. Epothilone A ist durch einen Makrozyklus weniger flexibel und sowohl die an Tubulin gebundene Struktur (RMSD = 0.537 Å) als auch freie Form (RMSD = 0.497 Å) kann mit geringen RMSD – Werten bestimmt werden. Die Struktur der freien Form, welche in Lösung hauptsächlich vorliegt, ist mit der Röntgenstruktur weitgehend identisch. In der an Tubulin gebundenen Form wird eine essentielle Umorientierung der Seitenkette beobachtet, die für die Wechselwirkung mit Tubulin entscheidend ist.

Dipolare Kopplungen eines Proteins sind geeignet, eine 3D Homologiesuche in der PDB durchzuführen, da die relative Orientierung von Sekundärstrukturelementen und Domänen durch sie beschrieben wird[85]. Die frühe Erkennung 3D homologer Proteinfaltungen eröffnet die Möglichkeit, die Bestimmung von Proteinstrukturen zu beschleunigen. Eine Homolgiesuche unter Nutzung dipolarer Kopplungen ist in der Lage, Proteine oder zumindest Fragmente mit ähnlicher 3D Struktur zu finden, auch wenn die Primärsequenzhomologie gering ist. Darüber hinaus wird eine Transformation für experimentelle dipolare Kopplungen entwickelt, die die indirekte Orientierungsinformation eines Vektors relativ zu einem externen Tensor in den möglichen Bereich für den Projektionswinkel zwischen zwei Vektoren und somit in eine intramolekulare Strukturinformation übersetzt. Diese Einschränkungen können in der Strukturbestimmung von Proteinen mittels Molekulardynamik genutzt werden[92]. Im Gegensatz zu allen existierenden Implementierungen wird die Konvergenz der Rechnung durch die auf diese Weise eingeführten dipolare Kopplungsinformation kaum beeinflusst. Die dipolaren Kopplungen werden trotzdem von den errechneten Strukturen erfüllt. Auch ohne die Nutzung bereits bekannter Protein- oder Fragmentstrukturen kann so ein erheblicher Teil der NOE – Information substituiert werden. Die Dynamik des Vektors, der die beiden wechselwirkenden Dipole verbindet, beeinflusst den Messwert der dipolaren Kopplung. Dadurch wird Information über die Dynamik von Molekülen auf der µs-Zeitskala zugänglich, die bisher nur schwer untersucht werden konnte. Die Messung dipolarer Kopplungen für einen Vektor in verschiedenen Orientierungen erlaubt die Analyse seiner Bewegung[89]. Im besonderen ist die Ableitung eines modellfreien Ordnungsparameters S^2 möglich. Weiterhin lassen sich ebenso modellfrei eine mittlere Orientierung des Vektors, axialsymmetrische Anteile und nichtaxialsymmetrische Anteile der Dynamik ableiten und auswerten. Die Anwendung der so entwickelten Protokolle auf experimentelle Daten[90] lässt Proteine deutlich dynamischer erscheinen als auf der Zeitskala der Relaxationsexperimente zu erkennen ist. Der mittlere Ordnungsparameter sinkt von 0.8 auf 0.6. Dies entspricht einer Erhöhung des Öffnungswinkels der Bewegung von ca. 22 ° auf ca. 33°. Die Bewegungen weichen teilweise bis zu 40% und im Mittel 15% von der Axialsymmetrie ab.

Neuronale Netze erlauben eine schnelle (ca. 5000 chemische Verschiebungen pro Sekunde) und exakte (mittleren Abweichung von 1.6 ppm) Berechnung der ^{13}C NMR chemischen Verschiebung[115]. Dabei kombinieren sie die Vorteile bisher bekannter Datenbankabschätzungen (hohe Genauigkeit) und Inkrementverfahren (hohe Geschwindigkeit). Das ^{13}C NMR Spektrum einer organischen Verbindung stellt eine detaillierte Beschreibung seiner Struktur dar. Resultate des Strukturgenerators COCON können durch den Vergleich des experimentellen mit den berechneten ^{13}C NMR Spektren auf ca. 1 ‰ der vorgeschlagenen Strukturen eingeschränkt werden, die eine geringe Abweichung zum experimentellen Spektrum haben[122]. Die Kombination mit einer Substrukturanalyse erlaubt weiterhin die Erkennung wahrscheinlicher, geschlossener Ringsysteme und gibt einen Überblick über die Struktur des generierten Konstitutionssubraumes. Genetische Algorithmen können die Struktur organischer Moleküle ausgehend von derer Summenformel auf eine Übereinstimmung mit dem experimentellen ^{13}C NMR Spektrum optimieren. Die Konstitution von Molekülen wird dafür durch einen Vektor der Bindungszustände zwischen allen Atom – Atom Paaren beschrieben. Selbige Vektoren sind geeignet, in einem genetischen Algorithmus als genetischer Code von Konstitutionen betrachtet zu werden. Diese Methode erlaubt die automatisierte Bestimmung der Konstitution von Molekülen mit 10 bis 20 Nichtwasserstoffatomen[123].

Symmetrische neuronale Netze können fünf bzw. sieben dimensionale, heterogene Parameterrepräsentationen der 20 proteinogenen Aminosäuren unter Erhalt der wesentlichen Information in den dreidimensionalen Raum projizieren[134]. Die niederdimensionalen Projektionen ermöglichen eine Visualisierung der Beziehungen der Aminosäuren untereinander. Die reduzierten Parameterrepräsentationen sind geeignet, als Eingabe für ein neuronales Netz zu dienen, welches die Sekundärstruktur eines Proteins mit einer Genauigkeit von 66 % im Q_3 – Wert berechnet.

Neuronale Netzte sind aufgrund ihrer flexiblen Struktur besonders geeignet, quantitative Beziehungen zwischen Struktur und Aktivität zu beschreiben, da hier hochgradig nichtlineare, komplexe Zusammenhänge vorliegen. Eine numerische Codierung der über 200 in der Literatur beschriebenen Epothilonderivate erlaubt es, Modelle zur Berechnung der Induktion der Tubulin Polymerisation (R = 0.73) und der Inhibierung des Krebszellenwachstums (R = 0.94) zu erstellen[136]. Die trainierten neuronalen Netze können in einer Sensitivitätsanalyse genutzt werden, um die Bindungsstellen des Moleküls zu identifizieren. Aus der Berechnung der Aktivität für alle Moleküle des durch die Parameter definierten Strukturraums ergeben sich Vorschläge für Epothilonderivate, die bis zu 1 000 mal aktiver als die bisher synthetisierten sein könnten.

8 Abkürzungen

Abkürzung	*Beschreibung*
2D	zweidimensional
3D	dreidimensional
DAP	Dolaproin (Dreibuchstabencode)
DIL	Dolaisoleuin (Dreibuchstabencode)
DNA	**D**esoxy Ribo **N**ucleid **A**cid (Desoxyribonucleinsäure)
DOE	Dolaphenin (Dreibuchstabencode)
DOV	Dolavalin (Dreibuchstabencode)
NMR	**N**uclear **M**agnetic **R**esonance (Kernmagnetische Resonanz)
NOE	**N**uclear **O**verhauser **E**nhancement (Kern Overhauser Effekt)
NOESY	**N**uclear **O**verhauser and **E**xchange **S**pectrocop**y**
PDB	**P**rotein **D**aten **B**ase (Protein Datenbank)
QSAR	**q**uantitative **SAR** (quantitative ~)
QSPR	**q**uantitative **SPR** (quantitative ~)
R	Korrelationskoeffizient
RMSD	**R**oot **M**ean **S**quare **D**eviation (Wurzel der mittleren quadratischen Abweichung)
RNA	**R**ibo **N**ucleid **A**cid (Ribonucleinsäure)
SAR	**S**tructure **A**ctivity **R**elation (Struktur-Aktivitäts-Beziehung)
SPR	**S**tructure **P**roperty **R**elation (Struktur-Eigenschafts-Beziehung)
WWW	**W**orld **W**ide **W**eb (Internet)

9 Naturkonstanten

Konstante	*Wert*
π	≈ 3.14159265359
h	$\approx 6.626176 \cdot 10^{-34} J \cdot s$ (PLANKsches Wirkungsquantum)
μ_0	$= 4\pi \cdot 10^{-7} \frac{T^2 m^3}{J} \approx 1.256637 \cdot 10^{-6} \frac{T^2 m^3}{J}$
k	$\approx 1.380652 \cdot 10^{-23} \frac{J}{K}$ (BOLTZMANN Konstante)
γ_H	$\approx 2.6751 \cdot 10^{8} \frac{rad}{T \cdot s}$ (gyromagnetisches Verhältnis des Wasserstoff)
γ_C	$\approx 6.7263 \cdot 10^{7} \frac{rad}{T \cdot s}$ (gyromagnetisches Verhältnis des Kohlenstoffs)
γ_N	$\approx -2.7120 \cdot 10^{7} \frac{rad}{T \cdot s}$ (gyromagnetisches Verhältnis des Stickstoff)
γ_F	$\approx 2.5180 \cdot 10^{8} \frac{rad}{T \cdot s}$ (gyromagnetisches Verhältnis des Fluors)
γ_P	$\approx 1.0839 \cdot 10^{8} \frac{rad}{T \cdot s}$ (gyromagnetisches Verhältnis des Phosphors)

10 Schemata

11 Tabellen

12 Abbildungsverzeichnis

13 Literatur

[1] Schiff, P. B.; Fant, J.; Horwitz, S. B. "Promotion of microtubule assembly in vitro by taxol", *Nature* **1979**, *277*, 665-668.

[2] Pettit, G. R.; Kamano, Y.; Herald, C. L.; Tuinman, A. A.; Boettner, F. E.; Kizu, H.; Schmidt, J. M.; Baczynskyj, L.; Tomer, K. B.; Bontems, R. J. "The isolation and structure of a remarkable marine animal antineoplastic constituent: dolastatin 10", *J. Am. Chem. Soc.* **1987**, *109*, 6883-6885.

[3] Bollag, D. M.; McQueney, P. A.; Zhu, J.; Hensens, O.; Koupal, L.; Liesch, J.; Goetz, M.; Lazarides, E.; Woods, C. M. "Epothilones, a New Class of Microtubulue-stabilizing Agents with a Taxol-like Mechanism of Action", *Cancer Research* **1995**, *55*, 2325-2333.

[4] Benedetti, E.; Fraternali, F.; Hamada, Y.; Paolillo, L.; Shioiri, T. "Conformational Analysis of Dolastatin 10: An NMR amd Theoretical Approach", *Biopolymers* **1994**, *35*, 525-538.

[5] Alattia, T.; Roux, F.; Poncet, J.; Cavé, A.; Jouin, P. "Conformational Study of Dolastatin 10", *Tetrahedron Letters* **1995**, *51*, 2593-2604.

[6] Taylor, R. E.; Zajicek, J. "Conformational Properties of Epothilone", *J. Org. Chem.* **1999**, *64*, 7224-7228.

[7] Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. "Nuclear magnetic dipole interactions in filed-oriented proteins: Information for structure determination in solution", *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 9279-9283.

[8] Tjandra, N.; Bax, A. "Direct Measurment of Distances and Angles in Biomolecules by NMR in a Dilute Liquid Crystalline Medium", *Science* **1997**, *278*, 1111-1113.

[9] Losonczi, J. A.; Andrec, M.; Fischer, M. W. F.; Prestegard, J. H. "Order Matrix Analysis of Residual Dipolar Couplings Using Singular Value Decomposition", *J. Magn. Res.* **1999**, *138*, 334-342.

[10] Annila, A.; Aitio, H.; Thulin, E.; Drakenberg, T. "Recognition of protein folds via dipolar couplings", *J. Biomol. NMR* **1999**, *14*, 223-230.

[11] Hus, J.-C.; Marion, D.; Blackledge, M. "*De novo* Determination of Protein Structure by NMR using Orientational and Long-range Order Restraints", *J. Mol. Biol.* **2000**, *00*, 1-10.

[12] Delaglio, F.; Kontaxis, G.; Bax, A. "Protein Structure Determination Using Molecular Fragment Replacement and NMR Dipolar Couplings", *J. Am. Chem. Soc.* **2000**, *122*, 2142-2143.

[13] Tjandra, N.; Omichinski, J. G.; Gronenborn, A. M.; Clore, G. M.; Bax, A. "Use of dipolar 1H-15N and 1H-13C couplings in the structure determination of magnetically oriented macromolecules in solution", *Nature structural biology* **1997**, *4*, 732-738.

[14] Bruenger, A. T. "X-PLOR A System for X-Ray Christallography and NMR", *Yale University Press (New Haven)* **1992**.

[15] Tolman, J. R.; Al-Hashimi, H. M.; Kay, L. E.; Prestegard, J. H. "Structural and Dynamic Analysis of Residual Dipolar Coupling Data for Proteins", *J. Am. Chem. Soc.* **2000**.

[16] Bremser, W.; Ernst, L.; Franke, B.; Gerhards, R.; Hardt, A. *Carbon-13 NMR Spectral Data*; Verlag Chemie: Weinheim, 1981.

[17] Schindler, M.; Kutzelnigg, W. "Theory of magnetic susceptibilities and NMR chemical shifts in terms of localized quantities. II. Application to some simple molecules", *J. Chem. Phys.* **1982**, *76*, 1919-1933.

[18] Robien, W. "Das CSEARCH-NMR-Datenbanksystem", *Nachr. Chem. Tech. Lab.* **1998**, *46*, 74-77.

[19] *SpecInfo database*; Chemical Concepts: Karlsruhe, 2001.

[20] Bremser, W. "HOSE - A Novel Substructure Code", *Anal. Chim. Acta* **1978**, *103*, 355-365.

[21] Clerc, J.-T.; Sommerauer, H. "A Minicomputer Program Based On Additivity Rules For The Estimation Of 13C NMR Chemical Shifts", *Anal. Chim. Acta* **1977**, *95*, 33-40.

[22] Kvasnicka, V.; Sklenak, S.; Pospichal, J. "Application of Recurrent Neural Network in Chemistry. Prediction and Classification of 13C NMR Chemical Shiftsin a Series of Monosubstituted Benzenes", *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 742-747.

[23] Doucet, J.-P.; Panaye, A.; Feuilleaubois, E.; Ladd, P. "Neural networks and carbon-13 NMR shift prediction", *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 320-324.

[24] Meiler, J.; Meusinger, R.; Will, M. "Prediction of 13C-NMR Chemical Shifts of Substituted Benzenes by Means of a Neural Network", In *Software - Entwicklung in der Chemie*; Fels, G., Schubert, V., Eds.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1996; Vol. 11, pp 234-238.

[25] Fauchere, J.-L.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. "Amino Acid Side Chain Parameters for correlation Studies in Biology and Pharmacology", *Int. J. Peptide Protein* **1988**, *32*, 269-278.

[26] Rost, B. "PHD: predicting one-dimensional protein structure by profile-based neural networks", *Methods Enzymol.* **1996**, *266*, 525-539.

[27] Otto, M. "Chemometrie", *VCH Verlagsgesellschaft mbH, Weinheim* **1997**, *3.527-28837-6*.

[28] Livingstone, D. J.; Hesketh, G.; Clayworth, D. "Novel method for the display of multivariate data using neural networks", *J. Mol. Graphics* **1991**, *9*, 115-118.

[29] Devillers, J. *Neural networks in QSAR and drug design*; Acad. Press: London, 1996.

[30] Jeener, J.; Meier, B. H.; Bachmann, P.; Ernst, R. R. "Investigation of exchange processes by two-dimensional NMR spectroscopy", *J. Chem. Phys.* **1979**, *71*, 4546-4553.

[31] Bothner-By, A. A.; Stephens, R. L.; Lee, J.; Warren, C. D.; Jeanloz, R. W. "Structure determination of a tetrasaccharide: transient nuclear Overhauser effects in the rotating frame", *J. Am. Chem. Soc.* **1984**, *106*, 811-813.

[32] Rance, M. "Improved techniques for homonuclear rotating-frame and isotropic mixing experiments", *J. Magn. Reson.* **1987**, *74*, 557-564.

[33] Schleucher, J.; Quant, J.; Glaser, S. J.; Griesinger, C. "A theorem relating cross relaxation and Hartmann-Hahn transfer in multiple pulse sequences. Optimal

suppression of TOCSY transfer in ROESY", *J. Magn. Reson., Ser. A* **1995**, *112*, 144-151.

[34] Wüthrich, K. *NMR of Proteins and Nucleic Acids (1H-NMR shifts of amino acids)*; John Wiley & Sons: New York, Chichester, Brisbane, Toronto, Singapore, 1986; Vol. ISBN 0-471-82893-9.

[35] Reif, B.; Diener, A.; Hennig, M.; Maurer, M.; Griesinger, C. "Cross-Correlated Relaxation for the Measurement of Angles between Tensorial Interactions", *J. Magn. Reson.* **2000**, *143*, 45-68.

[36] Reif, B.; Hennig, M.; Griesinger, C. "Direct measurement of angles between bond vectors in high-resolution NMR", *Science (Washington, D. C.)* **1997**, *276*, 1230-1233.

[37] Diener, A. "Einführung neuer Strukturparameter in die Bestimmung der Struktur von Biomakromolekülen mit Hilfe von NMR-Spektroskopie und Molekulardynamik", *Dissertation (Universität Frankfurt)* **2000**.

[38] Crippen, G. M.; Havel, T. F. "Stable calculation of coordinates from distance information", *Acta Crystallogr., Sect. A* **1978**, *A34*, 282-284.

[39] Crippen, G. M.; Havel, T. F. *Distance Geometry and Molecular Conformation*; Research Studies Press: Taunton, Somerset, 1988.

[40] Crippen, G. M.; Havel, T. F. *Distance Geometry and Molecular Dynamics*; John Wiley & Sons: New York, Chichester, Brisbane, Toronto, Singapore, 1988.

[41] Haile, J. M. *Molecular Dynamics Simulation, Elementary Methods*; John Wiley & Sons: New York, Chichester, Brisbane, Toronto, Singapore, 1992.

[42] Evans, J. *Biomolecular NMR Spectroscopy*; Oxford University Press: Oxford, New York, Tokio, 1995.

[43] Pettit, G. R.; Singh, S. B.; Hogan, F.; Lloyd-Williams, P.; Herald, D. L.; Burkett, D. D.; Clewlow, P. J. "Antineoplastic agents. Part 189. The absolute configuration and synthesis of natural (-)-dolastatin 10", *J. Am. Chem. Soc.* **1989**, *111*, 5463-5465.

[44] Pettit, G. R.; Srirangam, J. K.; Herald, D. L.; Hamel, E. "The Dolastatins. 21. Synthesis, X-ray Crystal Structure, and Molecular Modeling of (6R)-Isodolastatin 10", *J. Org. Chem.* **1994**, *59*, 6127-6130.

[45] Quant, J. "Neue Experimente für homo- und heteronuclearen Kohärenztransfer im rotierenden Koordinatensystem und Anwendung auf Biomakromoleküle", *Dissertation (Universität Frankfurt)* **1996**.

[46] Höfle, G.; Bedorf, N.; Gerth, K. "Epothilone derivatives", *Chemical Abstracts* **1994**, *120*, 836.

[47] Gerth, K.; Bedorf, N.; Höfle, G.; Irschik, H.; Reichenbach, H. "Epothilons A and B: Antifungal and Cytotoxic Compounds from Sorangium cellulosum (Myxobacteria) - Production, Physico-chemical and Biological Properties", *J. Antibiothics* **1997**, *49*, 560-563.

[48] Nogales, E.; Wolf, S. G.; Khan, I. A.; Luduena, R. F.; Downing, K. H. "Structure of tubulin at 6.5A ad location of the taxol-binding site", *Nature* **1995**, *375*, 424-427.

[49] Nogales, E.; Wolf, S. G.; Downing, K. H. "Structure of the αβ tubulin dimer by electron crystallography", *Nature* **1998**, *39*, 199-203.

[50] Kowalski, R. J.; Giannakakous, P.; Hamel, E. "Activities of the Microtubule-stabilizing Agents Epothilone A and B with Purified and in Cells Resistant to Paclitaxel (Taxol)", *J. Biol. Chem.* **1997**, *272*, 2534-2541.

[51] Giannakakou, P.; Gussio, R.; Nogales, E.; Downing, K. H.; Zaharevitz, D.; Bollbuck, B.; Poy, G.; Sackett, D.; Nicolaou, K. C.; Fojo, T. "A common pharmacophore for epothilone and taxanes: molecular basis for drug resistance conferred by tubulin mutations in human cancer cells", *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 2904-2909.

[52] Martello, L. A.; McDaid, H. M.; Regl, D. L.; Yang, C.-P. H.; Meng, D.; Pettus, T. R. R.; Kaufman, M. D.; Arimoto, H.; Danishefsky, S. J.; Smith, A. B., III; Horwitz, S. B. "Taxol and discodermolide represent a synergistic drug combination in human carcinoma cell lines", *Clin. Cancer Res.* **2000**, *6*, 1978-1987.

[53] Höfle, G.; Bedorf, N.; Steinmetz, H.; Reichenbach, H.; Gerth, K. "Epthilon A und B - neuartige, 16gliedrige Makrolide mit cytotoxischer Wirkung: Isolierung, Struktur im Kristall und Konformation in Lösung", *Angewandte Chemie* **1996**, *108*, 1671-1673.

[54] Rhis, G.; Blommers, M. J. J., Röntgenkristallstruktur des Epothilon A.

[55] Blommers, M. J. J.; Carlomagno, T., Transfer NOE Intensitäten und kreuzkorrelierte Relaxationsraten von Epothilone A gebunden an Tubulin.

[56] Kamlowski, A.; Parac, T., Epothilone A NOE Intensitäten gemessen in DMSO (300ms, 500ms, 700ms).

[57] Sanders, C. R.; Schwonek, J. P. "bicellenarticel dmpcdhpc", *Biochemistry* **1992**, *31*, 8898-8905.

[58] Sanders II, C. R.; Hare, B. J.; Howard, K. P.; Prestegard, J. H. "Magnetically-oriented phosphorlipid micelles as a tool for the study of membrane-associated molecules", *Prog. NMR Spec.* **1994**, *26*, 421-444.

[59] Vold, R. R.; Prosser, R. S. "Magnetically Oriented Phospholipid Bilayers Micells for Structural Studies of Polypeptides. Does the ideal Bicelle exist?", *J. Magn. Res. B* **1996**, *113*, 267-271.

[60] Vold, R. R.; Prosser, R. S.; Deese, A. J. "Isotropic solutions of phospholipid bicells: A new membrane mimetic for high-resolution NMR studies of polypeptides", *J. Biomol. NMR* **1997**, *9*, 329-335.

[61] Losonczi, J. A.; Prestegard, J. H. "Improved dilute bicelle solutions for high-resolution NMR of biological macromolecules", *J. Biomol. NMR* **1998**, *12*, 447-451.

[62] Ottiger, M.; Bax, A. "Characterisation of magnetically oriented phospholpid micelles for measurement of dipolar couplings in macromolecules", *J. Biomol. NMR* **1998**, *12*, 361-372.

[63] Sanders, C. R.; Prosser, R. S. "Bicells: a model memnrane system for all seasons?", *Curr. Biol.* **1998**, *6*, 1227-1234.

[64] Boyd, J.; Redfield, C. "Characterization of 15N Chemical Shift Anisotropy from Orientation-Dependent Changes to 15N Chemical Shifts in Dilute Bicelle Solutions", *J. Am. Chem. Soc.* **1999**, *WWW*.

[65] Cavagnero, S.; Dyson, J. H.; Wright, P. E. "Improved low pH bicelle system for orienting macromolecules over a wide temperature range", *J. Biomol. NMR* **1999**, *13*, 387-391.

[66] Ottiger, M.; Bax, A. "Bicelle-based liquid crystals for NMR-measurement of dipolar couplings at acidic and basic pH values", *J. Biomol. NMR* **1999**, *13*, 187-191.

[67] Koenig, B. W.; Hu, J.-S.; Ottiger, M.; Bose, S.; Hendler, R. W.; Bax, A. "NMR Measurement of Dipolar Couplings in Proteins Aligned by Transient Binding to Purple Membrane Fragments", *J. Am. Chem. Soc.* **1999**, *121*, 1385-1386.

[68] Sass, J.; Cordier, F.; Hoffmann, A.; Cousin, A.; Grzesiek, S. "Purple Membrane Induced Alignment of Biological Macromolecules in the Magnetic Field", *J. Am. Chem. Soc.* **1999**, *121*, 2047-2055.

[69] Flemming, K.; Gray, D.; Prasannan, S.; Matthews, S. "Cellulose Crystallites: a new and robust liquid cyrystalline meium for the measurement of residual dipolar couplings", *J. Am. Chem. Soc.* **2000**, *122*, 5224-5225.

[70] Tycko, R.; Blanco, F. J.; Ishii , Y. "Alignment of Biopolymers in Strained Gels: A New Way To Create Detectable Dipole-Dipole Couplings in High-Resolution Biomolecular NMR", *J. Am. Chem. Soc.* **2000**, *122*, 9340 -9341.

[71] Hansen, M. R.; Mueller, L.; Pardi, A. "Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions", *Nat. Sruct. Biol.* **1998**, *5*, 1065-1074.

[72] Ojennus, D. D.; Mitton-Fry, R. M.; Wuttke, D. S. "Induced alignment and measurement of dipolar couplings of an SH2 domain through direct binding with filamentous phage", *J. Biomol. NMR* **1999**, *14*, 175-179.

[73] Kemple, M. D.; Ray, B. D.; Lipkowitz, K. B.; Prendergast, F. G.; Rao, B. D. N. "The Use of Lanthanides for Solution Structure Determination of Biomolecules by NMR: Evaluation of the Methodology with EDTA Derivatives as Model Systems", *J. Am. Chem. Soc.* **1988**, *110*, 8275-8287.

[74] Gochin, M. "Nuclear Magnetic Resonance Studies of a Paramagnetic Metallo DNA Complex", *J. Am. Chem. Soc.* **1997**, *119*, 3377-3378.

[75] Gochin, M.; Roder, H. "Use of the Pseudocontact Shift as a Structural Constraint for Macromolecules in Solution", *Bulletin of Magnetic Resonance* **1997**, *17*, 298-299.

[76] Beger, R.; Marathias, V. M.; Volkman, B. F.; Bolton, P. H. "Determination of Internuclear Angles of DNA Using Paramagnetic-Assisted Magnetic Alignment", *J. Magn. Res.* **1998**, *135*, 256-259.

[77] Boisbouvier, J.; Gans, P.; Blackedge, M.; Brutscher, B.; Marion, D. "Long Range Structural Information in NMR Studies of Paramagnetic Molecules from Electron Spin-Nuclear Spin Cross-Correlated Relaxation", *J. Am. Chem. Soc.* **1999**, *WWW*.

[78] Nguyen, B. D.; Xia, Z.; Yeh, D. C.; Deaguero, H.; Mar, G. N. L. "Solution NMR Determination of the Anisotropy an Orientation of the Paramagnetic Susceptibility Tensor as a Function of Temperature for Metmyoglobin Cyanide: Implication for the Population of Excited Electronic States", *J. Am. Chem. Soc.* **1999**, *121*, 208-217.

[79] Silver, B. L. *Irreducible Tensor Methods - An Introduction for Chemists* New York, San Franzisko, London, 1976.

[80] Sali, A. "100,000 protein structures for the biologist", *Nature structural biology* **1998**, *5*, 1029-1032.

[81] Fischer, D.; Eisenberg, D. "Predicting structures for genome proteins", *Curr. Opin. Struct. Biol.* **1999**, *9*, 208-211.

[82] Kay, L. E.; Gardner, K. H. "Solution NMR spectroscopy beyond 25 kDa", *Curr. Opin. Struct. Biol.* **1997**, *7*, 722-731.

[83] Rost, B.; Sander, C. "Combining evolutionary information and neural networks to predict protein secondary structure", *Proteins: Struct., Funct., Genet.* **1994**, *19*, 55-72.

[84] Holm, L.; Sander, C. "Mapping the protein universe", *Science* **1996**, *273*, 595-602.

[85] Meiler, J.; Peti, W.; Griesinger, C. "DipoCoup: A versatile program for 3D-structure homology comparision based on residual dipolar couplings and pseudocontact shifts", *J. Biomol. NMR* **2000**, *17*, 283-294.

[86] Meiler, J., *www.jens-meiler.de* **2001**.

[87] Peti, W. *Promotion*; Universität Frankfurt: Frankfurt, 2001.

[88] Hus, J.-C.; Marion, D.; Blackedge, M. "Ab initio determination of protein backbone structure using only residual dipolar couplings", **2000**.

[89] Meiler, J.; Peti, W.; Prompers, J.; Griesinger, C.; Brueschweiler, R. "Model-Free Approach to the Dynamic Interpretation of Residual Dipolar Couplings in Globular Proteins", *submitted* **2001**.

[90] Peti, W.; Meiler, J.; Prompers, J.; Brueschweiler, R.; Griesinger, C. "Influence of Molecular Motion on Residual Dipolar Spin-Spin Couplings in Proteins - A Practical Approach", *submitted* **2001**.

[91] Junker, J. "Use of Dipolar Couplings in X-Plor", *personal information* **2000**.

[92] Meiler, J.; Blomberg, N.; Nilges, M.; Griesinger, C. "A new Approach for Applying Residual Dipolar Couplings as Restraints in Structure Elucidation", *J. Biomol. NMR* **2000**, *16*, 245-252.

[93] Neubauer, H.; Meiler, J.; Peti, W.; Griesinger, C. "NMR Structure Determination of Saccharose and Raffinose by Means of Homo- and Heteronuclear Dipolar Couplings", *Hel. Chim. Acta* **2001**, *84*, 243-2858.

[94] Lipari, G.; Szabo, A. "Model-Free Approach to the Interpretation of Nucllear Resonance Relaxation in Macromolecules. 1. Theory and Range of Validity", *J. Am. Chem. Soc.* **1982**, *104*, 4546-4559.

[95] Lipari, G.; Szabo, A. "Model-Free Approach to the Interpretation of Nucllear Resonance Relaxation in Macromolecules. 2. Analysis of Experimental Results", *J. Am. Chem. Soc.* **1982**, *104*, 4559-4570.

[96] Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH Verlagsgesellschaft mbH: Weinheim, 1993.

[97] Devillers, J. "Genetic Algorithmus in Computer-Aided Molecular Design", In *Genetic Algorithm in Molecular Modeling*; Devillers, J., Ed.; Acad. Press.: London, 1996; pp 1-34.

[98] Ebraheem, K. A. K.; Webb, G. A. "Semi-Empirical Calculations of the Chemical Shifts of Nuclei other than Protons", *Progress in NMR Spectroscopy* **1977**, *11*, 149-181.

[99] Fürst, A.; Pretsch, E. "A computer program for the prediction of 13C NMR chemical shifts of organic compounds", *Anal. Chim. Acta* **1990**, *229*, 17-25.

[100] Hearmon, R. A.; Liu, H. M.; Laverick, S.; Tayler, P. "Microcomputer Prediction and Assessment of Substituted Benzene 13C NMR Chemical Shifts", *Mag. Res. Chem.* **1992**, *30*, 240-248.

[101] Cremer, D.; Olsson, L.; Reichel, F.; Kraka, E. "Calculation of NMR Chemical Shifts - the third Dimension of Quantum Chemistry", *Israel Journal of Chemistry* **1993**, *33*, 369-385.

[102] Gauss, J. "Accurate Calculation of NMR Chemical Shifts", *Berichte der Bunsen Gesellschaft für Physikalische Chemie - An International Journal of Physical Chemistry* **1995**, *99*, 1001-1008.

[103] Forsyth, D. A.; Sebag, A. B. "Computed 13C NMR Chemical Shifts via Empirically Scaled GIAO Shieldings and Molecular Mechanics Geometries. Conformation and Configuration from 13C Shifts", *J. Am. Chem. Soc.* **1997**, *119*, 9483-9494.

[104] Kvasnicka, V.; Sklenak, S.; Pospichal, J. "Application of recurrent neural networks in chemistry. Prediction and classification of carbon-13 NMR chemical shifts in a series of monosubstituted benzenes", *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 742-747.

[105] Kvasnicka, V.; Sklenak, S.; Pospichal, J. "Application of neural networks with feedback connections in chemistry: prediction of carbon-13 NMR chemical shifts in a series of monosubstituted benzenes", *Theochem* **1992**, *96*, 87-107.

[106] Sklenak, S.; Kvasnicka, V.; Pospichal, J. "Prediction of 13C NMR chemical shifts by neural networks in a series of monosubstituted benzenes", *Chem. Pap.* **1994**, *48*, 135-140.

[107] Svozil, D.; Pospichal, J.; Kvasnicka, V. "Neural Network Prediction of Carbon-13 NMR Chemical Shifts of Alkanes", *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 924-928.

[108] Thomas, S.; Kleinpeter, E. "Assignment of the 13C NMR chemical shifts of substituted naphthalenes from charge density with an artificial neural network", *J. Prakt. Chem./Chem.-Ztg.* **1995**, *337*, 504-507.

[109] Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J.-P. "13C NMR Chemical Shift Prediction of sp2 Carbon Atoms in Acyclic Alkenes Using Neural Networks", *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 644-653.

[110] Schweitzer, R. C.; Small, G. W. "Enhanced Structural Encoding for Database Retrievals of Carbon-13 Nuclear Magnetic Resonance Chemical Shifts", *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 310-322.

[111] Li, Z.; Huang, Y.; Hu, F.; Sheng, Q.; Peng, S. "Neural Networks in spectroscopy. Estimation and prediction of chemical shifts of 13C NMR in alkanes by using subgraphs", *Bopuxue Zazhi* **1997**, *14*, 507-514.

[112] Thomas, S.; Brühl, I.; Heilmann, D.; Kleinpeter, E. "13C NMR Chemical Shift Calculations for Some Substituted Pyridines: A Comparative Consideration", *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 726-730.

[113] Meiler, J.; Meusinger, R.; Will, M. "Neural Network Prediction of 13C NMR Chemical Shifts of Substituted Benzenes", *Monatshefte für Chemie* **1999**, *130*, 1089-1095.

[114] Le Bret, C. "A General 13C NMR Spectrim Predictor using Data Mining Techniques", *SAR and QSAR in Environmental Research* **2000**, *11*, 211-234.

[115] Meiler, J.; Will, M.; Meusinger, R. "Fast Determination of 13C-NMR Chemical Shifts Using Artificial Neural Networks", *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1169-1176.

[116] Benecke, C.; Grund, R.; Hohberger, R.; Kerber, A.; Laue, R.; Wieland, T. "MOLGEN+, a generator of connectivity isomers and stereoisomers for molecular structure elucidation", *Anal. Chim. Acta* **1995**, *314*, 141-147.

[117] Wieland, T.; Kerber, A.; Laue, R. "Principles of the Generation of Constitutional and Configurational Isomers", *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 413-419.

[118] Lindel, T.; Junker, J.; Köck, M. "COCON: From NMR Correlation Data to Molecular Constitution", *J. Mol. Model.* **1997**, *3*, 364-368.

[119] Junker, J.; Maier, W.; Lindel, T.; Köck, M. "Computer-Assisted Constitutional Assignment of Large Molecules: COCON Analysisi of Ascomycin", *Org. Letters* **1999**, *1*, 737-740.

[120] Köck, M.; Junker, J.; Maier, W.; Will, M.; Lindel, T. "A COCON Analysis of Proton-Poor Heterocycles - Application of Carbon Chemical Shift Predictions for the Evaluation of Structural Proposels", *Eur. J. Org. Chem.* **1999**, 579-586.

[121] Lindel, T.; Junker, J.; Köck, M. "2D-NMR-Guided Costitutional Analysis of Organic Compounds Employing the Computer Program COCON", *Eur. J. Org. Chem.* **1999**, 573-577.

[122] Meiler, J.; Köck, M. "Structure Elucidation by Automatic Generation and Analysis of Molecule Databases from NMR Connectivity Information Using Substructure Analysis and 13C-NMR Chemical Shift Prediction", *submitted* **2001**.

[123] Meiler, J.; Will, M. "Structure Elucidation from 13C-NMR Chemical Shifts by Genetic Algorithms", *submitted* **2001**.

[124] Livingstone, D. J. "Multivariate Data Display Using Neural Networks", In *Neural networks in QSAR and drug design*; Devillers, J., Ed.; Acad. Press: London, 1996.

[125] Kocjancic, R.; Zupan, J. "Application of a Feed-Forward Artificial Neural Network as a Mapping Device", *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 985-989.

[126] Rost, B.; Sander, C. "Prediction of protein secondary structure at better than 70% accuracy", *J. Mol. Biol.* **1993**, *232*, 584-599.

[127] Sasagawa, F.; Tajima, K. "Prediction of protein secondary structures by a neural network", *Comput. Appl. Biosci.* **1993**, *9*, 147-152.

[128] Rost, B.; Sander, C.; Schneider, R. "Redefining the Goals of Protein Secondary Structure Prediction", *J. Mol. Biol.* **1994**, *235*, 13-26.

[129] Chandonia, J.-M.; Karplus, M. "Neural networks for secondary structure and structural class predictions", *Protein Sci.* **1995**, *4*, 275-285.

[130] Salamov, A. A.; Solovyev, V. V. "Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments", *J. Mol. Biol.* **1995**, *247*, 11-15.

[131] Choy, W. Y.; Sanctuary, B. C.; Zhu, G. "Using neural network predicted secondary structure information in automatic protein NMR assignment", *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1086-1094.

[132] Baldi, P.; Brunak, S.; Frasconi, P.; Soda, G.; Pollastri, G. "Exploiting the past and future in protein secondary structure prediction", *Bioinformatics* **1999**, *15*, 937-946.

[133] Chandonia, J.-M.; Karplus, M. "New methods for accurate prediction of protein secondary structure", *Proteins: Struct., Funct., Genet.* **1999**, *35*, 293-306.

[134] Meiler, J.; Müller, M.; Zeidler, A.; Schmäschke, F. "Generation and Evaluation of Dimension Reduced Amino Acid Parameter Representations by Artificial Neural Networks", *submitted* **2001**.

[135] Meiler, J. "Untersuchung von Struktur-Eigenschafts-Beziehungen für die Spezifität von Serin-Proteasen gegenüber Polypeptiden mittels NMR-Spektroskopie und künstlicher neuronaler Netze" (Diplomarbeit)", *Universität Leipzig* **1998**, *diploma thesis*.

[136] Meiler, J.; Bleckmann, A. "Epothilones: QSAR Studies Performed by Artificial Neural Networks Leading to New Drug Proposals", *submitted* **2001**.

[137] Schinzer, D.; Limberg, A.; Böhm, O. M. "Studies Towards the Total Synthesis of Epothilones: Asymmetric Synthesis of the Key Fragments", *Chem. Eur. J.* **1996**, *2*, 1477-1482.

[138] Nicolaou, K. C.; Vourloumis, D.; Li, T.; Pastor, J.; Winssinger, N.; He, Y.; Ninkovic, S.; Sarabia, F.; Vallberg, H.; Roschanger, F.; King, P. N.; Finlay, M. R. V.; Giannakakous, P.; Verdier-Pinard, P.; Hamel, E. "Gezielt entworfene Epothilone: kombinatorische Synthese, Induktion der Tubulin-Polymerisation und cytotoxische Wirkung gegen taxolresistente Tumorzellen", *Angewandte Chemie* **1997**, *109*, 2181-2187.

[139] Höfle, G.; Glaser, N.; Kiffe, M.; Hecht, H.-J.; Sasse, F.; Reichenbach, H. "N-oxidation of epothilone A-C and O-acyl rearrangement to C-19- and C-21-substituted epothilones", *Angew. Chem., Int. Ed.* **1999**, *38*, 1971-1974.

[140] Mulzer, J. "Epothilone B and its derivatives as novel antitumor drugs: total and partial synthesis and biological evaluation", *Monatsh. Chem.* **2000**, *131*, 205-238.

[141] Altmann, K.-H.; Bold, G.; Caravatti, G.; End, N.; Florsheimer, A.; Guagnano, V.; O'Reilly, T.; Wartmann, M. "Epothilones and their analogs - potential new weapons in the fight against cancer", *Chimia* **2000**, *54*, 612-621.

[142] Winkler, J. D.; Axelsen, P. H. "A Model for the Taxol (Paclitaxel)/Epothilone Pharmacophore", *Bioorganic & Medicinal Chemistry Letters* **1996**, *6*, 2963-2966.

[143] Su, D.-S.; Balog, A.; Meng, D.; Bertinato, P.; Danishefsky, S. J.; Zheng, Y.-H.; Chou, T.-C.; He, L.; Horwitz, S. B. "Structure-activity relationships of the epothilones and the

first in vivo comparison with paclitaxel", *Angew. Chem., Int. Ed. Engl.* **1997**, *36*, 2093-2096.

[144] Nicolaou, K. C.; Roschangar, F.; Vourloumis, D. "Chemie und Biologie der Epothilone", *Angewandte Chemie* **1998**, *110*, 2120-2153.

[145] He, L.; Jagtap, P. G.; Kingston, D. G. I.; Shen, H.-J.; Orr, G. A.; Horwitz, S. B. "A Common Pharmacophore for Taxol and the Epothilones Based on the Biological Activity of a Taxane Molecule Lacking a C-13 Side Chain", *Biochemistry* **2000**, *39*, 3972-3978.

[146] Wang, M.; Xia, X.; Kim, Y.; Hwang, D.; Jansen, J. M.; Botta, M.; Liotta, D. C.; Snyder, J. P. "A Unified and Quantitative Receptor Model for the Microtubule Binding of Paclitaxel and Epothilone", *Org. Lett.* **1999**, *1*, 43-46.

14 Anhang A (Definition der Topologie und des Kraftfeldes für Dolastatin 10)

```
remark  file top_dola_jens.pro
remark  geomatric energy function topology for distance geometry and simulated annealing
remark
remark  written for dolastatin 10
remark
remark  written 270798
remark
remark  written by Jens Meiler, University Frankfurt
remark  charges unchecked

autogenerate
  angles     = true
  dihedrales = false
end

residue DMVJ
  group
    atom N      type = N_AM   charge =  0.00 end
    atom C1     type = C_QU   charge =  0.00 end
    atom H11    type = H_QU   charge =  0.00 excl =( H21  H22  H23)  end
    atom H12    type = H_QU   charge =  0.00 excl =( H21  H22  H23)  end
    atom H13    type = H_QU   charge =  0.00 excl =( H21  H22  H23)  end
    atom C2     type = C_QU   charge =  0.00 end
    atom H21    type = H_QU   charge =  0.00 excl =( H11  H12  H13)  end
    atom H22    type = H_QU   charge =  0.00 excl =( H11  H12  H13)  end
    atom H23    type = H_QU   charge =  0.00 excl =( H11  H12  H13)  end
    atom CA     type = C_QU   charge =  0.00 end
    atom HA     type = H_QU   charge =  0.00 end
    atom CB     type = C_QU   charge =  0.00 end
    atom HB     type = H_QU   charge =  0.00 end
    atom CG1    type = C_QU   charge =  0.00 end
    atom HG11   type = H_QU   charge =  0.00 excl =( HG21 HG22 HG23) end
    atom HG12   type = H_QU   charge =  0.00 excl =( HG21 HG22 HG23) end
    atom HG13   type = H_QU   charge =  0.00 excl =( HG21 HG22 HG23) end
    atom CG2    type = C_QU   charge =  0.00 end
    atom HG21   type = H_QU   charge =  0.00 excl =( HG11 HG12 HG13) end
    atom HG22   type = H_QU   charge =  0.00 excl =( HG11 HG12 HG13) end
    atom HG23   type = H_QU   charge =  0.00 excl =( HG11 HG12 HG13) end
    atom C      type = C_CA   charge =  0.00 end
    atom O      type = O_CA   charge =  0.00 end

    bond N   C1     bond C1  H11     bond C1  H12     bond C1  H13
    bond N   C2     bond C2  H21     bond C2  H22     bond C2  H23
    bond N   CA     bond CA  HA
    bond CA  CB     bond CB  HB
    bond CB  CG1    bond CG1 HG11    bond CG1 HG12    bond CG1 HG13
    bond CB  CG2    bond CG2 HG21    bond CG2 HG22    bond CG2 HG23
    bond CA  C
    bond C   O

    improper HA   N    C    CB      ! stereo CA
    improper HB   CA   CG1  CG2     ! stereo CB
    improper HG12 HG11 CB   HG13    ! stereo G1
    improper HG22 HG21 CB   HG23    ! stereo G2
    improper H12  H11  N    H13     ! stereo C1
    improper H22  H21  N    H23     ! stereo C2
end


residue VALJ
  group
    atom N      type = N_AM   charge =  0.00 end
    atom HN     type = H_AM   charge =  0.00 end
    atom CA     type = C_QU   charge =  0.00 end
    atom HA     type = H_QU   charge =  0.00 end
    atom CB     type = C_QU   charge =  0.00 end
    atom HB     type = H_QU   charge =  0.00 end
    atom CG1    type = C_QU   charge =  0.00 end
    atom HG11   type = H_QU   charge =  0.00 excl =( HG21 HG22 HG23) end
    atom HG12   type = H_QU   charge =  0.00 excl =( HG21 HG22 HG23) end
    atom HG13   type = H_QU   charge =  0.00 excl =( HG21 HG22 HG23) end
    atom CG2    type = C_QU   charge =  0.00 end
    atom HG21   type = H_QU   charge =  0.00 excl =( HG11 HG12 HG13) end
    atom HG22   type = H_QU   charge =  0.00 excl =( HG11 HG12 HG13) end
    atom HG23   type = H_QU   charge =  0.00 excl =( HG11 HG12 HG13) end
    atom C      type = C_CA   charge =  0.00 end
    atom O      type = O_CA   charge =  0.00 end

    bond N   HN
    bond N   CA     bond CA  HA
    bond CA  CB     bond CB  HB
    bond CB  CG1    bond CG1 HG11    bond CG1 HG12    bond CG1 HG13
    bond CB  CG2    bond CG2 HG21    bond CG2 HG22    bond CG2 HG23
    bond CA  C
    bond C   O

    improper HA   N    C    CB      ! stereo CA
    improper HB   CA   CG1  CG2     ! stereo CB
    improper HG12 HG11 CB   HG13    ! stereo G1
    improper HG22 HG21 CB   HG23    ! stereo G2
end


residue DILJ
  group
    atom N      type = N_AM   charge =  0.00 end
    atom CM     type = C_QU   charge =  0.00 end
    atom HM1    type = H_QU   charge =  0.00 end
    atom HM2    type = H_QU   charge =  0.00 end
    atom HM3    type = H_QU   charge =  0.00 end
    atom CA     type = C_QU   charge =  0.00 end
    atom HA     type = H_QU   charge =  0.00 end
    atom CB     type = C_QU   charge =  0.00 end
    atom HB     type = H_QU   charge =  0.00 end
    atom CG1    type = C_QU   charge =  0.00 end
    atom HG11   type = H_QU   charge =  0.00 excl =( HG21 HG22 HD1 HD2 HD3) end
    atom HG12   type = H_QU   charge =  0.00 excl =( HG21 HG22 HD1 HD2 HD3) end
    atom HG13   type = H_QU   charge =  0.00 excl =( HG21 HG22 HD1 HD2 HD3) end
    atom CG2    type = C_QU   charge =  0.00 end
    atom HG21   type = H_QU   charge =  0.00 excl =( HG11 HG12 HG13) end
    atom HG22   type = H_QU   charge =  0.00 excl =( HG11 HG12 HG13) end
    atom CD     type = C_QU   charge =  0.00 end
    atom HD1    type = H_QU   charge =  0.00 excl =( HG11 HG12 HG13) end
    atom HD2    type = H_QU   charge =  0.00 excl =( HG11 HG12 HG13) end
    atom HD3    type = H_QU   charge =  0.00 excl =( HG11 HG12 HG13) end
    atom CA1    type = C_QU   charge =  0.00 end
    atom HA1    type = H_QU   charge =  0.00 end
    atom O1     type = O_HY   charge =  0.00 end
    atom CP     type = C_QU   charge =  0.00 end
    atom HP1    type = H_QU   charge =  0.00 end
    atom HP2    type = H_QU   charge =  0.00 end
    atom HP3    type = H_QU   charge =  0.00 end
    atom CA2    type = C_QU   charge =  0.00 end
    atom HA21   type = H_QU   charge =  0.00 end
    atom HA22   type = H_QU   charge =  0.00 end
    atom C      type = C_CA   charge =  0.00 end
    atom O      type = O_CA   charge =  0.00 end

    bond N   CM     bond CM  HM1     bond CM  HM2     bond CM  HM3
    bond N   CA     bond CA  HA
    bond CA  CB     bond CB  HB
    bond CB  CG1    bond CG1 HG11    bond CG1 HG12    bond CG1 HG13
    bond CB  CG2    bond CG2 HG21    bond CG2 HG22
    bond CG2 CD     bond CD  HD1     bond CD  HD2     bond CD  HD3
    bond CA  CA1    bond CA1 HA1     bond CA1 O1
    bond O1  CP     bond CP  HP1     bond CP  HP2     bond CP  HP3
    bond CA1 CA2    bond CA2 HA21    bond CA2 HA22
    bond CA2 C
    bond C   O

    improper HA   N    CA1  CB      ! stereo CA
    improper HB   CA   CG2  CG1     ! stereo CB
    improper HG12 HG11 CB   HG13    ! stereo CG1
    improper HG21 HG22 CD   CB      ! stereo CG2
    improper HD2  HD1  CG2  HD3     ! stereo CD
    improper HM2  HM1  N    HM3     ! stereo CM
    improper HA1  O1   CA2  CA      ! stereo CA1
    improper HP2  HP1  O1   HP3     ! stereo CP
    improper HA21 HA22 C    CA1     ! stereo CA2
end


residue DAPJ
  group
    atom N      type = N_PR   charge =  0.00 end
    atom CA     type = C_PA   charge =  0.00 end
    atom CB     type = C_PB   charge =  0.00 end
    atom CG     type = C_PG   charge =  0.00 end
    atom CD     type = C_PD   charge =  0.00 end
    atom HA     type = H_QU   charge =  0.00 end
    atom HB1    type = H_QU   charge =  0.00 end
    atom HB2    type = H_QU   charge =  0.00 end

    atom HG1    type = H_QU   charge =  0.00 end
    atom HG2    type = H_QU   charge =  0.00 end
    atom HD1    type = H_QU   charge =  0.00 end
    atom HD2    type = H_QU   charge =  0.00 end
    atom CA1    type = C_QU   charge =  0.00 end
    atom HA1    type = H_QU   charge =  0.00 end
    atom O1     type = O_HY   charge =  0.00 end
    atom CP     type = C_QU   charge =  0.00 end
    atom HP1    type = H_QU   charge =  0.00 end
    atom HP2    type = H_QU   charge =  0.00 end
    atom HP3    type = H_QU   charge =  0.00 end
    atom CA2    type = C_QU   charge =  0.00 end
    atom HA2    type = H_QU   charge =  0.00 end
    atom CF     type = C_QU   charge =  0.00 end
    atom HF1    type = H_QU   charge =  0.00 end
    atom HF2    type = H_QU   charge =  0.00 end
    atom HF3    type = H_QU   charge =  0.00 end
    atom C      type = C_CA   charge =  0.00 end
    atom O      type = O_CA   charge =  0.00 end

    bond N   CA     bond CA  HA
    bond CA  CB     bond CB  HB1     bond CB  HB2
    bond CB  CG     bond CG  HG1     bond CG  HG2
    bond CG  CD     bond CD  HD1     bond CD  HD2
    bond CD  N
    bond CA  CA1    bond CA1 HA1     bond CA1 O1
    bond O1  CP     bond CP  HP1     bond CP  HP2     bond CP  HP3
    bond CA1 CA2    bond CA2 HA2
    bond CA2 CF     bond CF  HF1     bond CF  HF2     bond CF  HF3
    bond CA2 C
    bond C   O

    improper HA   N    CA1   CB     ! stereo CA
    improper HB1  HB2  CA    CG     ! stereo CB
    improper HG1  HG2  CB    CD     ! stereo CG
    improper HD1  HD2  CG    N      ! stereo CD
    improper HA1  O1   CA2   CA     ! stereo CA1
    improper HP2  HP1  O1    HP3    ! stereo CP
    improper HA2  CA1  C     CF     ! stereo CA2
    improper HF2  HF1  CA2   HF3    ! stereo CF

    improper N    CA   CB    CG     ! proline ring pucker
end


residue DOEJ
  group
    atom N      type = N_AM   charge =  0.00 end
    atom HN     type = H_AM   charge =  0.00 end
    atom CA     type = C_QU   charge =  0.00 end
    atom HA     type = H_QU   charge =  0.00 end
    atom CB     type = C_QU   charge =  0.00 end
    atom HB1    type = H_QU   charge =  0.00 end
    atom HB2    type = H_QU   charge =  0.00 end
    atom C      type = C_TA   charge =  0.00 end
    atom NTA    type = N_TH   charge =  0.00 end
    atom C1     type = C_TB   charge =  0.00 end
    atom H1     type = H_TB   charge =  0.00 end
    atom C2     type = C_TG   charge =  0.00 end
    atom H2     type = H_TG   charge =  0.00 end
    atom STA    type = S_TH   charge =  0.00 end

    atom CG     type = C_AR   charge =  0.00 exclude =(  CR3) end
    atom CR11   type = C_AR   charge =  0.00 exclude =( CR22) end
    atom HR11   type = H_AR   charge =  0.00 end
    atom CR12   type = C_AR   charge =  0.00 exclude =( CR21) end
    atom HR12   type = H_AR   charge =  0.00 end
    atom CR21   type = C_AR   charge =  0.00 exclude =( CR12) end
    atom HR21   type = H_AR   charge =  0.00 end
    atom CR22   type = C_AR   charge =  0.00 exclude =( CR11) end
    atom HR22   type = H_AR   charge =  0.00 end
    atom CR3    type = C_AR   charge =  0.00 exclude =(   CG) end
    atom HR3    type = H_AR   charge =  0.00 end

    bond N    HN
    bond N    CA     bond CA   HA
    bond CA   CB     bond CB   HB1   bond CB   HB2
    bond CA   C
    bond CB   CG

    bond CG   CR11   bond CR11 HR11
    bond CG   CR12   bond CR12 HR12
    bond CR11 CR21   bond CR21 HR21
    bond CR12 CR22   bond CR22 HR22
    bond CR21 CR3    bond CR3  HR3
    bond CR22 CR3

    bond C    NTA
    bond NTA  C1     bond C1  H1
    bond C1   C2     bond C2  H2
    bond C2   STA
    bond STA  C

    improper HA   N    C    CB       ! stereo CA
    improper HB2  HB1  CA   CG       ! stereo CB

    improper HR12 CR12 CR22 CR3
    improper HR22 CR22 CR3  CR21
    improper HR3  CR3  CR21 CR11
    improper HR21 CR21 CR11 CG
    improper HR11 CR11 CG   CR12
    improper CB   CG   CR12 CR22

    improper CG   CR11 CR21 CR3
    improper CR11 CR21 CR3  CR22
    improper CR21 CR3  CR22 CR12
    improper CR3  CR22 CR12 CG
    improper CR22 CR12 CG   CR11
    improper CR12 CG   CR11 CR21

    improper CA   C    NTA  C1
    improper C    NTA  C1   H1
    improper NTA  C1   C2   H2

    improper C    NTA  C1   C2
    improper NTA  C1   C2   STA
    improper C1   C2   STA  C
    improper C2   STA  C    NTA
    improper STA  C    NTA  C1
end

presidue PEPT          ! PEPTide bond link, for all except the  *(-) - (+)PRO link
                       ! "*(-) - PEPT - (+)*:
  add bond  -C   +N

  add angle -CA  -C   +N
  add angle -O   -C   +N
  add angle -C   +N   +CA
  add angle -C   +N   +HN

  add improper -O    -C    +N    -CA  ! planar (-C)
  add improper +HN   +N    -C    +CA  ! planar (+N)
  add improper -CA   -C    +N    +CA  ! planar peptide bond

! add dihedral -CA   -C    +N    +CA  ! planar peptide bond
end

presidue PED2          ! PEPTide bond link, for VAL - DIL
                       ! "*(-) - PED2 - (+)*:
  add bond  -C   +N

  add angle -CA  -C   +N
  add angle -O   -C   +N
  add angle -C   +N   +CA
  add angle -C   +N   +CM

  add improper -O    -C    +N    -CA  ! planar (-C)
  add improper +CM   +N    -C    +CA  ! planar (+N)
  add improper -CA   -C    +N    +CA  ! planar peptide bond

! add dihedral -CA   -C    +N    +CA  ! planar peptide bond
end

presidue PED3          ! PEPTide bond link, for DIL - DAP
                       ! "*(-) - PED3 - (+)*:
  add bond  -C   +N

  add angle -CA2 -C   +N
  add angle -O   -C   +N
  add angle -C   +N   +CA
  add angle -C   +N   +CD

  add improper -O    -C    +N    -CA2 ! planar (-C)
  add improper +CD   +N    -C    +CA  ! planar (+N)
  add improper -CA2  -C    +N    +CA  ! planar peptide bond


! add dihedral -CA2  -C    +N    +CA  ! planar peptide bond
end

presidue PED4          ! PEPTide bond link, for DAP - DOE
                       ! "*(-) - PED4 - (+)*:
  add bond  -C   +N

  add angle -CA2 -C   +N
  add angle -O   -C   +N
  add angle -C   +N   +CA
  add angle -C   +N   +HN

  add improper -O    -C    +N    -CA2 ! planar (-C)
  add improper +HN   +N    -C    +CA  ! planar (+N)
  add improper -CA2  -C    +N    +CA  ! planar peptide bond

! add dihedral -CA2  -C    +N    +CA  ! planar peptide bond
end
```

```
remark  file par_dola_jens.pro
remark  geomatric energy function parameters for distance geometry and simulated annealing
remark
remark  written for dolastatin 10
remark
remark  written 270798
remark
remark  written by Jens Meiler, University Frankfurt


! force constants

  evaluate( $kbon = 1000)
  evaluate( $kang =  500)
  evaluate( $kpla =  500)
  evaluate( $kchi =  500)
  evaluate( $kmet =  500)
  evaluate( $kmen =  500)
  evaluate( $kbac =  500)
  evaluate( $kdih =  500)

! bonds

  bond       H_AM  N_AM             $kbon     0.98
  bond       N_AM  C_QU             $kbon     1.49
  bond       C_QU  C_QU             $kbon     1.53
  bond       C_QU  H_QU             $kbon     1.08
  bond       C_QU  C_CA             $kbon     1.53
  bond       C_CA  O_CA             $kbon     1.215
  bond       C_QU  O_HY             $kbon     1.42
  bond       C_AR  C_AR             $kbon     1.40
  bond       C_AR  C_QU             $kbon     1.51
  bond       H_AR  C_AR             $kbon     1.08
  bond       C_CA  N_AM             $kbon     1.305

  bond       C_PA  H_QU             $kbon     1.08    ! PR ring hydrogene
  bond       C_PB  H_QU             $kbon     1.08    ! PR ring hydrogene
  bond       C_PG  H_QU             $kbon     1.08    ! PR ring hydrogene
  bond       C_PD  H_QU             $kbon     1.08    ! PR ring hydrogene
  bond       C_PA  C_PB             $kbon     1.53    ! PR ring inner ring
  bond       C_PB  C_PG             $kbon     1.49    ! PR ring inner ring
  bond       C_PG  C_PD             $kbon     1.50    ! PR ring inner ring
  bond       C_PD  N_PR             $kbon     1.47    ! PR ring inner ring
  bond       C_PA  N_PR             $kbon     1.49    ! PR ring inner ring

  bond       C_PA  C_QU             $kbon     1.53    ! PR ring connection
  bond       C_CA  N_PR             $kbon     1.305   ! PR ring connection

  bond       C_TB  H_TB             $kbon     1.08    ! TH ring hydrogene
  bond       C_TG  H_TG             $kbon     1.08    ! TH ring hydrogene
  bond       C_QU  C_TA             $kbon     1.504   ! TH ring connection
  bond       C_TA  N_TH             $kbon     1.42    ! TH ring inner ring
  bond       N_TH  C_TB             $kbon     1.38    ! TH ring inner ring
  bond       C_TB  C_TG             $kbon     1.49    ! TH ring inner ring
  bond       C_TG  S_TH             $kbon     1.77    ! TH ring inner ring
  bond       S_TH  C_TA             $kbon     1.80    ! TH ring inner ring


! angles

  angle      C_CA  N_AM  C_QU       $kang   120.0
  angle      C_CA  N_AM  H_AM       $kang   120.0
  angle      C_QU  N_AM  H_AM       $kang   120.0
  angle      C_QU  N_AM  C_QU       $kang   120.0

  angle      N_AM  C_QU  C_QU       $kang   109.5
  angle      N_AM  C_QU  H_QU       $kang   109.5
  angle      N_AM  C_QU  C_CA       $kang   109.5
  angle      C_QU  C_QU  H_QU       $kang   109.5
  angle      C_QU  C_QU  C_QU       $kang   109.5
  angle      H_QU  C_QU  H_QU       $kang   109.5
  angle      H_QU  C_QU  C_CA       $kang   109.5
  angle      C_QU  C_QU  C_CA       $kang   109.5
  angle      C_QU  C_QU  O_HY       $kang   109.5
  angle      H_QU  C_QU  O_HY        $kang   109.5
  angle      C_PA  C_QU  O_HY        $kang   109.5
  angle      C_PA  C_QU  H_QU        $kang   109.5
  angle      C_PA  C_QU  C_QU        $kang   109.5
  angle      N_AM  C_QU  C_TA        $kang   109.5
  angle      H_QU  C_QU  C_TA        $kang   109.5
  angle      C_QU  C_QU  C_TA        $kang   109.5
  angle      C_QU  C_QU  C_AR        $kang   109.5
  angle      H_QU  C_QU  C_AR        $kang   109.5

  angle      C_QU  C_CA  O_CA        $kang   120.0
  angle      C_QU  C_CA  N_AM        $kang   120.0
  angle      N_AM  C_CA  O_CA        $kang   120.0
  angle      C_QU  C_CA  N_PR        $kang   120.0
  angle      N_PR  C_CA  O_CA        $kang   120.0

  angle      C_QU  O_HY  C_QU        $kang   117.8

  angle      C_AR  C_AR  C_AR        $kang   120.0
  angle      C_AR  C_AR  H_AR        $kang   120.0
  angle      C_QU  C_AR  C_AR        $kang   120.0

  angle      H_QU  C_PA  N_PR        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PA  C_PB        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PA  C_QU        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PB  C_PA        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PB  C_PG        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PB  H_QU        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PG  C_PB        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PG  C_PD        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PG  H_QU        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PD  C_PG        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PD  N_PR        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PD  H_QU        $kang   109.5     ! PR ring hydrogene
  angle      C_PA  C_PB  C_PG        $kang   104.5     ! PR ring inner ring
  angle      C_PB  C_PG  C_PD        $kang   106.1     ! PR ring inner ring
  angle      C_PG  C_PD  N_PR        $kang   103.2     ! PR ring inner ring
  angle      C_PD  N_PR  C_PA        $kang   112.0     ! PR ring inner ring
  angle      N_PR  C_PA  C_PB        $kang   104.0     ! PR ring inner ring
  angle      N_PR  C_PA  C_QU        $kang   104.0     ! PR ring connection
  angle      C_PB  C_PA  C_QU        $kang   112.5     ! PR ring connection
  angle      C_CA  N_PR  C_PA        $kang   119.1     ! PR ring connection
  angle      C_CA  N_PR  C_PD        $kang   125.0     ! PR ring connection

  angle      H_TB  C_TB  C_TG        $kang   121.9     ! TH ring hydrogene
  angle      H_TB  C_TB  N_TH        $kang   121.7     ! TH ring hydrogene
  angle      H_TG  C_TG  C_TB        $kang   126.3     ! TH ring hydrogene
  angle      H_TG  C_TG  S_TH        $kang   121.1     ! TH ring hydrogene
  angle      C_TA  N_TH  C_TB        $kang   112.3     ! TH ring inner ring
  angle      N_TH  C_TB  C_TG        $kang   112.4     ! TH ring inner ring
  angle      C_TB  C_TG  S_TH        $kang   112.6     ! TH ring inner ring
  angle      C_TG  S_TH  C_TA        $kang    88.8     ! TH ring inner ring
  angle      S_TH  C_TA  N_TH        $kang   114.3     ! TH ring inner ring
  angle      C_QU  C_TA  N_TH        $kang   118.9     ! TH ring connection
  angle      C_QU  C_TA  S_TH        $kang   126.8     ! TH ring connection


! impropers

  improper   H_QU  H_QU  C_QU  H_QU  $kmet 0   66.514   ! methyl    group
  improper   H_QU  H_QU  O_HY  H_QU  $kmet 0   66.514   ! methyl    group
  improper   H_QU  H_QU  N_AM  H_QU  $kmet 0   66.514   ! methyl    group
  improper   H_QU  H_QU  C_QU  C_QU  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  C_QU  C_AR  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  C_CA  C_QU  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  C_PA  C_PG  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  C_PB  C_PD  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  C_PG  N_PR  $kmen 0   70.874   ! methylene group
  improper   H_QU  C_QU  C_QU  C_QU  $kchi 0   65.977   ! methine   group
  improper   H_QU  C_QU  C_CA  C_QU  $kchi 0   65.977   ! methine   group
  improper   H_QU  N_AM  C_QU  C_QU  $kchi 0   65.977   ! methine   group
  improper   H_QU  O_HY  C_QU  C_QU  $kchi 0   65.977   ! methine   group
  improper   H_QU  N_AM  C_CA  C_QU  $kchi 0   65.977   ! methine   group
  improper   H_QU  N_PR  C_QU  C_PB  $kchi 0   65.977   ! methine   group
  improper   H_QU  O_HY  C_QU  C_PA  $kchi 0   65.977   ! methine   group
  improper   H_QU  N_AM  C_TA  C_QU  $kchi 0   65.977   ! methine   group

  improper   O_CA  C_CA  N_AM  C_QU  $kbac 0  180.0     ! planar (-C)
  improper   H_AM  N_AM  C_CA  C_QU  $kbac 0  180.0     ! planar (+N)
  improper   C_QU  C_CA  N_AM  C_QU  $kbac 0  180.0     ! planar peptid bond trans and planar
(+N) (DIL/DOE)
  improper   C_QU  C_CA  C_QU  N_AM  $kbac 0  180.0     ! planar peptid bond cis
  improper   O_CA  C_CA  N_PR  C_QU  $kbac 0  180.0     ! PR planar (-C)
  improper   C_PD  N_PR  C_CA  C_PA  $kbac 0  180.0     ! PR planar (+N)
! improper   C_QU  C_CA  N_PR  C_PA  $kbac 0  180.0     ! PR planar peptid bond trans
  improper   C_QU  C_CA  N_PR  C_PA  $kbac 0    0.0     ! PR planar peptid bond cis

  dihedral   C_QU  C_CA  N_PR  C_PA  $kdih 2  180.0     ! PR planar peptid trans or cis
  dihedral   C_QU  C_CA  N_AM  C_QU  $kdih 2  180.0     ! planar peptid bond trans or cis

  improper   H_AR  C_AR  C_AR  C_AR  $kpla 0  180.0
  improper   C_QU  C_AR  C_AR  C_AR  $kpla 0  180.0
  improper   C_AR  C_AR  C_AR  C_AR  $kpla 0    0.0

  improper   C_QU  C_TA  N_TH  C_TB  $kpla 0  180.0     ! TH ring
  improper   C_TA  N_TH  C_TB  H_TB  $kpla 0  180.0     ! TH ring
  improper   N_TH  C_TB  C_TG  H_TG  $kpla 0  180.0     ! TH ring

  improper   C_TA  N_TH  C_TB  C_TG  $kpla 0    0.0     ! TH ring
  improper   N_TH  C_TB  C_TG  S_TH  $kpla 0    0.0     ! TH ring
  improper   C_TB  C_TG  S_TH  C_TA  $kpla 0    0.0     ! TH ring
  improper   C_TG  S_TH  C_TA  N_TH  $kpla 0    0.0     ! TH ring
  improper   S_TH  C_TA  N_TH  C_TB  $kpla 0    0.0     ! TH ring

  improper   N_PR  C_PA  C_PB  C_PG  $kbac 0   25.0     ! PR ring pucker


! the non-bonded parameters are approximately the same as
! DISMAN's large radii for repel=0.9, and as DISMAN's small
! radii for repel=0.75. (ECEPP2)
! the epsilon values are generally not used (and should not be)
! with this force field
! the radius is sigma*2^(-5/6)
!
!                      eps      sigma       eps(1:4) sigma(1:4)

  NONBonded  C_CA      0.0903   3.3409      0.0903   3.3409
  NONBonded  C_AR      0.120    3.3409      0.1200   3.3409
  NONBonded  C_TA      0.145    3.3409      0.1450   3.3409
  NONBonded  C_TB      0.1200   3.3409      0.1200   3.3409
  NONBonded  C_TG      0.1200   3.3409      0.1200   3.3409
  NONBonded  C_QU      0.0903   3.3409      0.0903   3.3409
  NONBonded  C_PA      0.0903   3.3409      0.0903   3.3409
  NONBonded  C_PB      0.0903   3.3409      0.0903   3.3409
  NONBonded  C_PG      0.1450   3.3409      0.1450   3.3409
  NONBonded  C_PD      0.1450   3.3409      0.1450   3.3409
  NONBonded  H_QU      0.0045   2.2272      0.0045   2.2272
  NONBonded  H_AM      0.0498   2.2272      0.0498   2.2272
  NONBonded  H_AR      0.0045   2.2272      0.0045   2.2272
  NONBonded  H_TB      0.0045   2.2272      0.0045   2.2272
  NONBonded  H_TG      0.0045   2.2272      0.0045   2.2272
  NONBonded  N_PR      0.1592   3.0068      0.1592   3.0068
  NONBonded  N_TH      0.1592   3.0068      0.1592   3.0068
  NONBonded  N_AM      0.1592   3.0068      0.1592   3.0068
  NONBonded  O_CA      0.2342   2.7755      0.2342   2.7755
  NONBonded  O_HY      0.2342   2.7755      0.2342   2.7755
  NONBonded  S_TH      0.0239   3.7458      0.0239   3.7458

! the following nbfixes allow hydrogen bonding
! the distance used is (2A/B)^(1/6)*repel  distances
!                             A      B   A1-4  B1-4

  nbfix      H_AM     O_CA     44.2  1.0  44.2  1.0
  nbfix      H_AM     O_HY     44.2  1.0  44.2  1.0

set message on echo on end
```

15 Anhang B (Resultate der Strukturrechnung für Dolastatin 10)

		cis – Konformation			*trans* – Konformation		
Atom 1	Atom 2	NOE	gerechnete Strukt.		NOE	gerechnete Strukt.	
HA(DOV)	HB(DOV)	2.830 Å	2.496 Å	±0.033 Å	2.454 Å	2.471 Å	±0.011 Å
HA(DOV)	HG1*(DOV)	4.299 Å	3.216 Å	±0.344 Å	3.871 Å	3.156 Å	±0.249 Å
HA(DOV)	HG2*(DOV)	4.489 Å	3.729 Å	±0.330 Å	4.042 Å	3.817 Å	±0.236 Å
HB(DOV)	HG1*(DOV)	3.996 Å	2.656 Å	±0.001 Å	3.496 Å	2.656 Å	±0.001 Å
HB(DOV)	HG2*(DOV)	4.083 Å	2.656 Å	±0.001 Å	3.677 Å	2.657 Å	±0.001 Å
HA(DOV)	HN(VAL)	3.090 Å	2.780 Å	±0.154 Å	2.721 Å	2.670 Å	±0.122 Å
HA(VAL)	HB(VAL)	3.062 Å	2.661 Å	±0.240 Å	2.574 Å	2.522 Å	±0.029 Å
HA(VAL)	HG1*(VAL)	4.518 Å	3.388 Å	±0.410 Å	4.110 Å	3.889 Å	±0.005 Å
HA(VAL)	HG2*(VAL)	4.389 Å	3.318 Å	±0.387 Å	3.740 Å	3.020 Å	±0.037 Å
HB(VAL)	HG1*(VAL)	4.258 Å	2.657 Å	±0.001 Å	–	–	–
HB(VAL)	HG2*(VAL)	3.877 Å	2.657 Å	±0.001 Å	3.598 Å	2.656 Å	±0.001 Å
HA(VAL)	HM*(DIL)	4.669 Å	2.662 Å	±0.052 Å	3.539 Å	2.912 Å	±0.011 Å
HB(VAL)	HM*(DIL)	5.026 Å	4.337 Å	±0.767 Å	4.608 Å	2.611 Å	±0.015 Å
HG1*(VAL)	HM*(DIL)	6.782 Å	4.763 Å	±0.802 Å	–	–	–
HG2*(VAL)	HM*(DIL)	7.189 Å	5.087 Å	±0.631 Å	5.760 Å	4.579 Å	±0.089 Å
HG1*(VAL)	HR*(DOE)	7.697 Å	6.928 Å	±1.015 Å	–	–	–
HA(DIL)	HD*(DIL)	–	–	–	4.082 Å	3.199 Å	±0.279 Å
HA1(DIL)	HA2*(DIL)	3.762 Å	2.779 Å	±0.027 Å	3.584 Å	2.823 Å	±0.027 Å
HA1(DIL)	HG2*(DIL)	3.774 Å	3.910 Å	±0.006 Å	3.507 Å	2.660 Å	±0.114 Å
HA1(DIL)	HM*(DIL)	4.932 Å	2.948 Å	±0.017 Å	4.549 Å	4.612 Å	±0.067 Å
HA1(DIL)	HP*(DIL)	4.182 Å	2.905 Å	±0.060 Å	3.865 Å	2.926 Å	±0.078 Å
HA2*(DIL)	HB(DIL)	3.650 Å	2.944 Å	±0.063 Å	–	–	–
HA2*(DIL)	HG2*(DIL)	5.281 Å	4.910 Å	±0.050 Å	–	–	–
HA2*(DIL)	HM*(DIL)	5.201 Å	5.185 Å	±0.005 Å	4.931 Å	3.322 Å	±0.142 Å
HB(DIL)	HD*(DIL)	4.644 Å	3.084 Å	±0.017 Å	–	–	–
HB(DIL)	HM*(DIL)	4.057 Å	4.305 Å	±0.007 Å	4.027 Å	2.879 Å	±0.027 Å
HD*(DIL)	HM*(DIL)	7.768 Å	4.967 Å	±0.009 Å	6.585 Å	5.497 Å	±0.027 Å
HG2*(DIL)	HM*(DIL)	6.601 Å	2.775 Å	±0.007 Å	–	–	–
HA2*(DIL)	HA(DAP)	3.407 Å	2.869 Å	±0.031 Å	–	–	–
HA2*(DIL)	HA1(DAP)	3.545 Å	2.504 Å	±0.075 Å	–	–	–
HA2*(DIL)	HD1(DAP)	–	–	–	5.101 Å	2.810 Å	±0.063 Å
HA2*(DIL)	HD2(DAP)	–	–	–	5.421 Å	2.883 Å	±0.014 Å
HP*(DIL)	HA1(DAP)	–	–	–	3.753 Å	3.625 Å	±0.122 Å
HP*(DIL)	HF*(DAP)	–	–	–	5.953 Å	5.979 Å	±0.072 Å
HM*(DIL)	HA(DOE)	5.467 Å	5.473 Å	±0.093 Å	–	–	–
HM*(DIL)	HR1*(DOE)	7.654 Å	5.933 Å	±0.350 Å	–	–	–
HP*(DIL)	HA(DOE)	5.185 Å	2.452 Å	±0.131 Å	–	–	–
HA(DAP)	HA1(DAP)	–	–	–	2.358 Å	2.454 Å	±0.007 Å
HA(DAP)	HB1(DAP)	4.774 Å	2.744 Å	±0.001 Å	4.614 Å	2.752 Å	±0.002 Å
HA(DAP)	HB2(DAP)	4.412 Å	2.331 Å	±0.001 Å	4.170 Å	2.335 Å	±0.001 Å
HA(DAP)	HF*(DAP)	–	–	–	4.785 Å	4.687 Å	±0.013 Å
HA1(DAP)	HA2(DAP)	3.225 Å	3.033 Å	±0.002 Å	2.626 Å	2.941 Å	±0.004 Å

		cis – Konformation			*trans* – Konformation		
Atom 1	Atom 2	NOE	gerechnete Strukt.		NOE	gerechnete Strukt.	
HA1(DAP)	HF*(DAP)	4.531 Å	3.088 Å	±0.039 Å	4.173 Å	3.361 Å	±0.015 Å
HA2(DAP)	HB1(DAP)	4.355 Å	2.027 Å	±0.079 Å	4.149 Å	1.936 Å	±0.029 Å
HA2(DAP)	HF*(DAP)	4.223 Å	2.659 Å	±0.001 Å	3.937 Å	2.655 Å	±0.001 Å
HB1(DAP)	HB2(DAP)	4.004 Å	1.774 Å	±0.000 Å	3.739 Å	1.775 Å	±0.001 Å
HB1(DAP)	HD2(DAP)	–	–	–	4.452 Å	3.965 Å	±0.008 Å
HB2(DAP)	HD1(DAP)	–	–	–	4.992 Å	3.909 Å	±0.004 Å
HB2(DAP)	HG1(DAP)	5.071 Å	2.957 Å	±0.006 Å	–	–	–
HD1(DAP)	HD2(DAP)	3.770 Å	1.776 Å	±0.000 Å	3.712 Å	1.778 Å	±0.000 Å
HD1(DAP)	HG1(DAP)	5.260 Å	2.255 Å	±0.005 Å	4.259 Å	2.278 Å	±0.003 Å
HD1(DAP)	HG2(DAP)	4.587 Å	2.814 Å	±0.018 Å	4.574 Å	2.746 Å	±0.006 Å
HD2(DAP)	HG2(DAP)	–	–	–	4.054 Å	2.289 Å	±0.003 Å
HF*(DAP)	HP*(DAP)	6.391 Å	3.989 Å	±0.277 Å	–	–	–
HG1(DAP)	HG2(DAP)	4.038 Å	1.768 Å	±0.001 Å	3.735 Å	1.770 Å	±0.001 Å
HA2(DAP)	HN(DOE)	2.892 Å	2.290 Å	±0.043 Å	2.521 Å	2.447 Å	±0.090 Å
HA(DOE)	HB1(DOE)	2.802 Å	2.608 Å	±0.117 Å	2.468 Å	2.437 Å	±0.043 Å
HA(DOE)	HB2(DOE)	3.379 Å	2.995 Å	±0.035 Å	–	–	–
HA(DOE)	HN(DOE)	3.675 Å	2.908 Å	±0.038 Å	3.344 Å	2.969 Å	±0.028 Å
HA(DOE)	HR1*(DOE)	4.889 Å	3.360 Å	±0.093 Å	–	–	–
HB1(DOE)	HB2(DOE)	1.883 Å	1.764 Å	±0.001 Å	1.714 Å	1.760 Å	±0.001 Å
HB1(DOE)	HR*(DOE)	5.242 Å	3.077 Å	±0.019 Å	4.912 Å	4.434 Å	±0.008 Å
HB2(DOE)	HN(DOE)	3.343 Å	2.319 Å	±0.161 Å	2.933 Å	2.687 Å	±0.140 Å
HB2(DOE)	HR*(DOE)	4.874 Å	3.059 Å	±0.020 Å	4.511 Å	4.421 Å	±0.005 Å

Tabelle 1: NOESY – Distanzen und berechnete Distanzen in Dolastatin 10

		cis – Konformation			*trans* – Konformation		
Atom 1	Atom 2	J (Exp.)	gerechnete Strukt.		J (Exp.)	gerechnete Strukt.	
HA(DIL)	HA1(DIL)	4.8 Hz	0.5 Hz	−88.1 °	4.0 Hz	0.5 Hz	88.7 °
HA(DAP)	HA1(DAP)	3.0 Hz	2.4 Hz	−63.1 °	3.0 Hz	3.2 Hz	−57.8 °
HA1(DAP)	HA2(DAP)	10.0 Hz	12.0 Hz	179.7 °	9.5 Hz	9.0 Hz	150.1 °
HN(DOE)	HA(DOE)	8.2 Hz	6.8 Hz	−136.3 °	8.4 Hz	11.9 Hz	−173.9 °

Tabelle 2: Skalare Kopplungen und gefundene Dihedralwinkel in Dolastatin 10

	cis – Konformation	*trans* – Konformation
RMSD (alle Atome)	1.423 Å	1.488 Å
RMSD (Nichtwasserstoffatome)	1.042 Å	1.220 Å
RMSD (Rückgrat)	0.463 Å	0.500 Å

Tabelle 3: RMSD – Werte für die zehn energieärmsten Dolastatin 10 Strukturen

Atom 1	Atom 2	Atom 3	Atom 4	*cis* – Konform.		*trans* – Konform.	
N(DOV)	CA(DOV)	C(DOV)	N(VAL)	30.5 °	±65.7 °	19.4 °	±99.3 °
	Konformer 1 ⇒ vgl. Kapitel 2.2			51.9 °	±14.4 °	57.4 °	±2.5 °
	Konformer 2 ⇒ vgl. Kapitel 2.2			–162.1 °	±0.0 °	170.8 °	±22.6 °
CA(DOV)	C(DOV)	N(VAL)	CA(VAL)	180.0 °	±0.1 °	180.0 °	±0.1 °
C(DOV)	N(VAL)	CA(VAL)	C(VAL)	–84.6 °	±56.9 °	–88.2 °	±25.8 °
	Konformer 1 ⇒ vgl. Kapitel 2.2			–59.0 °	±4.1 °	–61.1 °	±9.9 °
N(VAL)	CA(VAL)	C(VAL)	N(DIL)	127.5 °	±23.2 °	170.3 °	±0.7 °
CA(VAL)	C(VAL)	N(DIL)	CA(DIL)	180.0 °	±0.1 °	180.0 °	±0.1 °
C(VAL)	N(DIL)	CA(DIL)	CA1(DIL)	–140.6 °	±0.6 °	–75.1 °	±6.4 °
N(DIL)	CA(DIL)	CA1(DIL)	CA2(DIL)	151.2 °	±1.0 °	–33.1 °	±1.5 °
CA(DIL)	CA1(DIL)	CA2(DIL)	C(DIL)	–70.4 °	±12.9 °	–152.2 °	±11.5 °
CA1(DIL)	CA2(DIL)	C(DIL)	N(DAP)	–96.0 °	±6.4 °	–151.6 °	±11.9 °
CA2(DIL)	C(DIL)	N(DAP)	CA(DAP)	–0.2 °	±0.1 °	180.0 °	±0.1 °
C(DIL)	N(DAP)	CA(DAP)	CA1(DAP)	–80.7 °	±2.2 °	–71.4 °	±0.9 °
N(DAP)	CA(DAP)	CA1(DAP)	CA2(DAP)	175.3 °	±5.2 °	–179.1 °	±1.5 °
CA(DAP)	CA1(DAP)	CA2(DAP)	C(DAP)	–60.3 °	±4.1 °	–88.3 °	±1.2 °
CA1(DAP)	CA2(DAP)	C(DAP)	N(DOE)	126.7 °	±12.7 °	144.2 °	±29.4 °
CA2(DAP)	C(DAP)	N(DOE)	CA(DOE)	180.0 °	±0.1 °	180.0 °	±0.1 °
C(DAP)	N(DOE)	CA(DOE)	C(DOE)	–76.4 °	±11.1 °	–114.0 °	±23.0 °

Tabelle 4: Dihedralwinkel im Rückgrat der bestimmten Dolastatin 10 Struktur

16 Anhang C (Definition der Topologie und des Kraftfeldes für Epothilone A)

```
remark  file top_epothilone.pro
remark  geometric energy function topology for distance geometry and simulated annealing
remark
remark  written for epothilone
remark
remark  written 051099
remark
remark  written by Jens Meiler, University Frankfurt
remark  charges unchecked


autogenerate
  angles     = true
  dihedrales = false
end


residue EPOA
  group
    atom C01    type = C_CA   charge =  0.00 end
    atom C02    type = C_QU   charge =  0.00 end
    atom C03    type = C_QU   charge =  0.00 end
    atom C04    type = C_QU   charge =  0.00 end
    atom C05    type = C_CA   charge =  0.00 end
    atom C06    type = C_QU   charge =  0.00 end
    atom C07    type = C_QU   charge =  0.00 end
    atom C08    type = C_QU   charge =  0.00 end
    atom C09    type = C_QU   charge =  0.00 end
    atom C10    type = C_QU   charge =  0.00 end
    atom C11    type = C_QU   charge =  0.00 end
    atom C12    type = C_EP   charge =  0.00 end
    atom C13    type = C_EP   charge =  0.00 end
    atom C14    type = C_QU   charge =  0.00 end
    atom C15    type = C_QU   charge =  0.00 end
    atom C16    type = C_OL   charge =  0.00 end
    atom C17    type = C_OL   charge =  0.00 end
    atom C18    type = C_TB   charge =  0.00 end
    atom C19    type = C_TG   charge =  0.00 end
    atom C20    type = C_TA   charge =  0.00 end
    atom C21    type = C_QU   charge =  0.00 end
    atom C22    type = C_QU   charge =  0.00 end
    atom C23    type = C_QU   charge =  0.00 end
    atom C24    type = C_QU   charge =  0.00 end
    atom C25    type = C_QU   charge =  0.00 end
    atom C27    type = C_QU   charge =  0.00 end
    atom O011   type = O_HY   charge =  0.00 end
    atom O012   type = O_CA   charge =  0.00 end
    atom O03    type = O_HY   charge =  0.00 end
    atom O05    type = O_CA   charge =  0.00 end
    atom O07    type = O_HY   charge =  0.00 end
    atom O12    type = O_HY   charge =  0.00 end
    atom S19    type = S_TH   charge =  0.00 end
    atom N18    type = N_TH   charge =  0.00 end
    atom H021   type = H_QU   charge =  0.00 end
    atom H022   type = H_QU   charge =  0.00 end
    atom H03    type = H_QU   charge =  0.00 end
    atom HO03   type = H_HY   charge =  0.00 end
    atom H06    type = H_QU   charge =  0.00 end
    atom H07    type = H_QU   charge =  0.00 end
    atom HO07   type = H_HY   charge =  0.00 end
    atom H08    type = H_QU   charge =  0.00 end
    atom H091   type = H_QU   charge =  0.00 end
    atom H092   type = H_QU   charge =  0.00 end
    atom H101   type = H_QU   charge =  0.00 end
    atom H102   type = H_QU   charge =  0.00 end
    atom H111   type = H_QU   charge =  0.00 end
    atom H112   type = H_QU   charge =  0.00 end
    atom H12    type = H_QU   charge =  0.00 end
    atom H13    type = H_QU   charge =  0.00 end
    atom H141   type = H_QU   charge =  0.00 end
    atom H142   type = H_QU   charge =  0.00 end
    atom H15    type = H_QU   charge =  0.00 end
    atom H17    type = H_OL   charge =  0.00 end
    atom H19    type = H_TG   charge =  0.00 end
    atom H211   type = H_QU   charge =  0.00 excl =( H211 H212 H213) end
    atom H212   type = H_QU   charge =  0.00 excl =( H211 H212 H213) end
    atom H213   type = H_QU   charge =  0.00 excl =( H211 H212 H213) end
    atom H221   type = H_QU   charge =  0.00 excl =( H221 H222 H223) end
    atom H222   type = H_QU   charge =  0.00 excl =( H221 H222 H223) end
    atom H223   type = H_QU   charge =  0.00 excl =( H221 H222 H223) end
    atom H231   type = H_QU   charge =  0.00 excl =( H231 H232 H233) end
    atom H232   type = H_QU   charge =  0.00 excl =( H231 H232 H233) end
    atom H233   type = H_QU   charge =  0.00 excl =( H231 H232 H233) end
    atom H241   type = H_QU   charge =  0.00 excl =( H241 H242 H243) end
    atom H242   type = H_QU   charge =  0.00 excl =( H241 H242 H243) end
    atom H243   type = H_QU   charge =  0.00 excl =( H241 H242 H243) end
    atom H251   type = H_QU   charge =  0.00 excl =( H251 H252 H253) end
    atom H252   type = H_QU   charge =  0.00 excl =( H251 H252 H253) end
    atom H253   type = H_QU   charge =  0.00 excl =( H251 H252 H253) end
    atom H271   type = H_QU   charge =  0.00 excl =( H271 H272 H273) end
    atom H272   type = H_QU   charge =  0.00 excl =( H271 H272 H273) end
    atom H273   type = H_QU   charge =  0.00 excl =( H271 H272 H273) end

                    bond C01  O011  bond C01  O012
    bond C01  C02   bond C02  H021  bond C02  H022
    bond C02  C03   bond C03  H03   bond C03  O03   bond O03  HO03
    bond C03  C04
    bond C04  C22   bond C22  H221  bond C22  H222  bond C22  H223
    bond C04  C23   bond C23  H231  bond C23  H232  bond C23  H233
    bond C04  C05   bond C05  O05
    bond C05  C06   bond C06  H06
    bond C06  C24   bond C24  H241  bond C24  H242  bond C24  H243
    bond C06  C07   bond C07  H07   bond C07  O07   bond O07  HO07
    bond C07  C08   bond C08  H08
    bond C08  C25   bond C25  H251  bond C25  H252  bond C25  H253
    bond C08  C09   bond C09  H091  bond C09  H092
    bond C09  C10   bond C10  H101  bond C10  H102
    bond C10  C11   bond C11  H111  bond C11  H112
    bond C11  C12   bond C12  H12   bond C12  O12
    bond C12  C13   bond C13  H13   bond C13  O12
    bond C13  C14   bond C14  H141  bond C14  H142
    bond C14  C15   bond C15  O011  bond C15  H15
    bond C15  C16
    bond C16  C27   bond C27  H271  bond C27  H272  bond C27  H273
    bond C16  C17   bond C17  H17
    bond C17  C18   bond C18  N18   bond N18  C20
    bond C18  C19   bond C19  S19   bond S19  C20   bond C19  H19
    bond C20  C21   bond C21  H211  bond C21  H212  bond C21  H213

    improper H021 H022 C03  C01     ! stereo C02
    improper H03  O03  C04  C02     ! stereo C03
    improper C23  C22  C05  C03     ! stereo C04
    improper H06  C07  C05  C24     ! stereo C06
    improper H07  O07  C06  C08     ! stereo C07
    improper H08  C25  C07  C09     ! stereo C08
    improper H091 H092 C10  C08     ! stereo C09
    improper H101 H102 C11  C09     ! stereo C10
    improper H111 H112 C12  C10     ! stereo C11
    improper H12  O12  C11  C13     ! stereo C12
    improper H13  O12  C12  C14     ! stereo C13
    improper H141 H142 C15  C13     ! stereo C14
    improper H15  O011 C16  C14     ! stereo C15
    improper H211 H212 C20  H213    ! stereo C21
    improper H221 H222 C04  H223    ! stereo C22
    improper H231 H232 C04  H233    ! stereo C23
    improper H241 H242 C06  H243    ! stereo C24
    improper H251 H252 C08  H253    ! stereo C25
    improper H271 H272 C16  H273    ! stereo C27

    improper C20  N18  C18  C19
    improper N18  C18  C19  S19
    improper C18  C19  S19  C20
    improper C19  S19  C20  N18
    improper S19  C20  N18  C18
```

```
    improper C21  C20  N18  C18
    improper C20  N18  C18  C17
    improper N18  C18  C19  H19

    improper O011 C02  C01  O012
    improper C04  C06  C05  O05

    improper C18  C16  C17  H17
    improper C27  C15  C16  C17
    improper C18  C17  C16  C15
    improper N18  C18  C17  C16
!   improper C19  C18  C17  C16

end
```

```
remark  file par_epothilone.pro
remark  geomatric energy function parameters for distance geometry and simulated annealing
remark
remark  written for epothilone
remark
remark  written 051099
remark
remark  written by Jens Meiler, University Frankfurt


! force constants

  evaluate( $kbon = 1000)
  evaluate( $kang =  500)
  evaluate( $kpla =  500)
  evaluate( $kchi =  500)
  evaluate( $kmet =  500)
  evaluate( $kmen =  500)
  evaluate( $kbac =  500)
  evaluate( $kdih =  500)


! bonds

  bond       H_AM  N_AM              $kbon     0.98
  bond       N_AM  C_QU              $kbon     1.49
  bond       C_QU  C_QU              $kbon     1.53
  bond       C_QU  H_QU              $kbon     1.08
  bond       C_QU  C_CA              $kbon     1.53
  bond       C_PA  C_CA              $kbon     1.53
  bond       C_CA  O_CA              $kbon     1.215
  bond       C_QU  O_HY              $kbon     1.42
  bond       C_AR  C_AR              $kbon     1.40
  bond       C_AR  C_QU              $kbon     1.51
  bond       H_AR  C_AR              $kbon     1.08
  bond       C_CA  N_AM              $kbon     1.305
  bond       C_CA  O_HY              $kbon     1.42
  bond       O_HY  H_HY              $kbon     0.96
  bond       C_QU  C_OL              $kbon     1.504
  bond       C_OL  C_OL              $kbon     1.49
  bond       C_OL  H_OL              $kbon     1.08
  bond       C_QU  C_EP              $kbon     1.53
  bond       C_EP  C_EP              $kbon     1.53
  bond       C_EP  H_QU              $kbon     1.08
  bond       C_EP  O_HY              $kbon     1.42

  bond       C_PA  H_QU              $kbon     1.08    ! PR ring hydrogene
  bond       C_PB  H_QU              $kbon     1.08    ! PR ring hydrogene
  bond       C_PG  H_QU              $kbon     1.08    ! PR ring hydrogene
  bond       C_PD  H_QU              $kbon     1.08    ! PR ring hydrogene
  bond       C_PA  C_PB              $kbon     1.53    ! PR ring inner ring
  bond       C_PB  C_PG              $kbon     1.49    ! PR ring inner ring
  bond       C_PG  C_PD              $kbon     1.50    ! PR ring inner ring
  bond       C_PD  N_PR              $kbon     1.47    ! PR ring inner ring
  bond       C_PA  N_PR              $kbon     1.49    ! PR ring inner ring
  bond       C_PA  C_QU              $kbon     1.53    ! PR ring connection
  bond       C_CA  N_PR              $kbon     1.305   ! PR ring connection

  bond       C_TB  H_TB              $kbon     1.08    ! TH ring hydrogene
  bond       C_TG  H_TG              $kbon     1.08    ! TH ring hydrogene
  bond       C_QU  C_TA              $kbon     1.504   ! TH ring connection
  bond       C_OL  C_TB              $kbon     1.49    ! TH ring connection
  bond       C_TA  N_TH              $kbon     1.42    ! TH ring inner ring
  bond       N_TH  C_TB              $kbon     1.38    ! TH ring inner ring
  bond       C_TB  C_TG              $kbon     1.49    ! TH ring inner ring
  bond       C_TG  S_TH              $kbon     1.77    ! TH ring inner ring
  bond       S_TH  C_TA              $kbon     1.80    ! TH ring inner ring


! angles

  angle      C_CA  N_AM  C_QU        $kang   120.0
  angle      C_CA  N_AM  H_AM        $kang   120.0
  angle      C_QU  N_AM  H_AM        $kang   120.0
  angle      C_QU  N_AM  C_QU        $kang   120.0

  angle      N_AM  C_QU  C_QU        $kang   109.5
  angle      N_AM  C_QU  H_QU        $kang   109.5
  angle      N_AM  C_QU  C_CA        $kang   109.5
  angle      C_QU  C_QU  H_QU        $kang   109.5
  angle      C_QU  C_QU  C_QU        $kang   109.5
  angle      H_QU  C_QU  H_QU        $kang   109.5
  angle      H_QU  C_QU  C_CA        $kang   109.5
  angle      C_QU  C_QU  C_CA        $kang   109.5
  angle      C_QU  C_QU  O_HY        $kang   109.5
  angle      H_QU  C_QU  O_HY        $kang   109.5
  angle      C_PA  C_QU  O_HY        $kang   109.5
  angle      C_PA  C_QU  H_QU        $kang   109.5
  angle      C_PA  C_QU  C_QU        $kang   109.5
  angle      N_AM  C_QU  C_TA        $kang   109.5
  angle      H_QU  C_QU  C_TA        $kang   109.5
  angle      C_QU  C_QU  C_TA        $kang   109.5
  angle      C_QU  C_QU  C_AR        $kang   109.5
  angle      H_QU  C_QU  C_AR        $kang   109.5
  angle      C_QU  C_QU  C_OL        $kang   109.5
  angle      C_OL  C_QU  O_HY        $kang   109.5
  angle      H_QU  C_QU  C_OL        $kang   109.5
  angle      H_QU  C_QU  C_EP        $kang   109.5
  angle      C_QU  C_QU  C_EP        $kang   109.5

  angle      C_QU  C_CA  O_CA        $kang   120.0
  angle      C_QU  C_CA  N_AM        $kang   120.0
  angle      N_AM  C_CA  O_CA        $kang   120.0
  angle      C_QU  C_CA  N_PR        $kang   120.0
  angle      N_PR  C_CA  O_CA        $kang   120.0
  angle      C_PA  C_CA  O_CA        $kang   120.0
  angle      C_PA  C_CA  N_AM        $kang   120.0
  angle      O_HY  C_CA  O_CA        $kang   120.0
  angle      O_HY  C_CA  C_QU        $kang   120.0
  angle      C_QU  C_CA  C_QU        $kang   120.0


  angle      C_QU  C_OL  C_QU        $kang   120.0
  angle      C_QU  C_OL  C_OL        $kang   120.0
  angle      C_OL  C_OL  H_OL        $kang   120.0
  angle      C_OL  C_OL  C_TB        $kang   120.0
  angle      H_OL  C_OL  C_TB        $kang   120.0

  angle      C_EP  C_EP  O_HY        $kang    60.2
  angle      C_QU  C_EP  O_HY        $kang   112.5
  angle      C_QU  C_EP  C_EP        $kang   120.0
  angle      H_QU  C_EP  O_HY        $kang   112.5
  angle      H_QU  C_EP  C_EP        $kang   120.0
  angle      H_QU  C_EP  C_QU        $kang   109.5

  angle      C_EP  O_HY  C_EP        $kang    59.4
  angle      C_QU  O_HY  C_QU        $kang   117.8
  angle      C_CA  O_HY  C_QU        $kang   117.8
  angle      C_QU  O_HY  H_HY        $kang   108.0

  angle      C_AR  C_AR  C_AR        $kang   120.0
  angle      C_AR  C_AR  H_AR        $kang   120.0
  angle      C_QU  C_AR  C_AR        $kang   120.0

  angle      H_QU  C_PA  C_PB        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PA  N_PR        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PA  C_CA        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PA  C_QU        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PB  C_PA        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PB  C_PG        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PB  H_QU        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PG  C_PB        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PG  C_PD        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PG  H_QU        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PD  C_PG        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PD  N_PR        $kang   109.5     ! PR ring hydrogene
  angle      H_QU  C_PD  H_QU        $kang   109.5     ! PR ring hydrogene
  angle      C_PA  C_PB  C_PG        $kang   104.5     ! PR ring inner ring
```

```
  angle      C_PB  C_PG  C_PD        $kang   106.1     ! PR ring inner ring
  angle      C_PG  C_PD  N_PR        $kang   103.2     ! PR ring inner ring
  angle      C_PD  N_PR  C_PA        $kang   112.0     ! PR ring inner ring
  angle      N_PR  C_PA  C_PB        $kang   104.0     ! PR ring inner ring
  angle      N_PR  C_PA  C_CA        $kang   104.0     ! PR ring connection
  angle      C_PB  C_PA  C_CA        $kang   112.5     ! PR ring connection
  angle      C_CA  N_PR  C_PA        $kang   119.1     ! PR ring connection
  angle      C_CA  N_PR  C_PD        $kang   125.0     ! PR ring connection

  angle      H_TB  C_TB  C_TG        $kang   121.9     ! TH ring hydrogene
  angle      C_OL  C_TB  C_TG        $kang   121.9     ! TH ring hydrogene
  angle      H_TB  C_TB  N_TH        $kang   121.7     ! TH ring hydrogene
  angle      C_OL  C_TB  N_TH        $kang   121.7     ! TH ring hydrogene
  angle      H_TG  C_TG  C_TB        $kang   126.3     ! TH ring hydrogene
  angle      H_TG  C_TG  S_TH        $kang   121.1     ! TH ring hydrogene
  angle      C_TA  N_TH  C_TB        $kang   112.3     ! TH ring inner ring
  angle      N_TH  C_TB  C_TG        $kang   112.4     ! TH ring inner ring
  angle      C_TB  C_TG  S_TH        $kang   112.6     ! TH ring inner ring
  angle      C_TG  S_TH  C_TA        $kang    88.8     ! TH ring inner ring
  angle      S_TH  C_TA  N_TH        $kang   114.3     ! TH ring inner ring
  angle      C_QU  C_TA  N_TH        $kang   118.9     ! TH ring connection
  angle      C_QU  C_TA  S_TH        $kang   126.8     ! TH ring connection


! impropers

  improper   H_QU  H_QU  C_QU  H_QU  $kmet 0   66.514   ! methyl    group
  improper   H_QU  H_QU  O_HY  H_QU  $kmet 0   66.514   ! methyl    group
  improper   H_QU  H_QU  N_AM  H_QU  $kmet 0   66.514   ! methyl    group
  improper   H_QU  H_QU  C_TA  H_QU  $kmet 0   66.514   ! methyl    group
  improper   H_QU  H_QU  C_OL  H_QU  $kmet 0   66.514   ! methyl    group
  improper   H_QU  H_QU  C_QU  C_QU  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  C_EP  C_QU  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  C_QU  C_EP  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  C_QU  C_AR  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  C_CA  C_QU  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  C_PA  C_PG  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  C_PB  C_PD  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  C_PG  N_PR  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  N_AM  C_QU  $kmen 0   70.874   ! methylene group
  improper   H_QU  H_QU  C_QU  C_CA  $kmen 0   70.874   ! methylene group
  improper   H_QU  C_QU  C_QU  C_QU  $kchi 0   65.977   ! methine   group
  improper   H_QU  C_QU  C_CA  C_QU  $kchi 0   65.977   ! methine   group
  improper   H_QU  N_AM  C_QU  C_QU  $kchi 0   65.977   ! methine   group
  improper   H_QU  O_HY  C_QU  C_QU  $kchi 0   65.977   ! methine   group
  improper   H_QU  N_AM  C_CA  C_QU  $kchi 0   65.977   ! methine   group
  improper   H_QU  N_PR  C_CA  C_PB  $kchi 0   65.977   ! methine   group
  improper   H_QU  O_HY  C_QU  C_PA  $kchi 0   72.977   ! methine   group
  improper   H_QU  N_AM  C_TA  C_QU  $kchi 0   65.977   ! methine   group
  improper   H_QU  O_HY  C_OL  C_QU  $kchi 0   65.977   ! methine   group
  improper   H_QU  O_HY  C_EP  C_QU  $kchi 0   65.977   ! methine   group
  improper   H_QU  O_HY  C_QU  C_EP  $kchi 0   65.977   ! methine   group
  improper   C_QU  C_QU  C_CA  C_QU  $kchi 0   70.528   ! quartaer carbon

  improper   O_CA  C_CA  N_AM  C_QU  $kbac 0  180.0     ! planar (-C)
  improper   O_CA  C_CA  N_AM  C_PA  $kbac 0  180.0     ! planar (-C)
  improper   H_AM  N_AM  C_CA  C_QU  $kbac 0  180.0     ! planar (+N)
  improper   C_QU  C_CA  N_AM  C_QU  $kbac 0  180.0     ! planar peptid bond trans and planar
(+N) (DIL/DOE)
  improper   C_PA  C_CA  N_AM  C_QU  $kbac 0  180.0     ! planar peptid bond trans and planar
(+N)
  improper   O_CA  C_CA  N_PR  C_QU  $kbac 0  180.0     ! PR planar (-C)
  improper   C_PD  N_PR  C_CA  C_PA  $kbac 0  180.0     ! PR planar (+N)
! improper   C_QU  C_CA  N_PR  C_PA  $kbac 0  180.0     ! PR planar peptid bond trans
! improper   C_QU  C_CA  N_PR  C_PA  $kbac 0    0.0     ! PR planar peptid bond cis

  dihedral   C_QU  C_CA  N_PR  C_PA  $kdih 2  180.0     ! PR planar peptid trans or cis
  dihedral   C_QU  C_CA  N_AM  C_QU  $kdih 2  180.0     ! planar peptid bond trans or cis
  dihedral   C_PA  C_CA  N_AM  C_QU  $kdih 2  180.0     ! planar peptid bond trans or cis

  improper   H_AR  C_AR  C_AR  C_AR  $kpla 0  180.0
  improper   C_QU  C_AR  C_AR  C_AR  $kpla 0  180.0
  improper   C_AR  C_AR  C_AR  C_AR  $kpla 0    0.0
  improper   C_QU  C_TA  N_TH  C_TB  $kpla 0  180.0     ! TH ring
  improper   C_TA  N_TH  C_TB  H_TB  $kpla 0  180.0     ! TH ring
  improper   N_TH  C_TB  C_TG  H_TG  $kpla 0  180.0     ! TH ring
  improper   C_TA  N_TH  C_TB  C_OL  $kpla 0  180.0     ! TH ring

  improper   C_TA  N_TH  C_TB  C_TG  $kpla 0    0.0     ! TH ring
  improper   N_TH  C_TB  C_TG  S_TH  $kpla 0    0.0     ! TH ring
  improper   C_TB  C_TG  S_TH  C_TA  $kpla 0    0.0     ! TH ring
  improper   C_TG  S_TH  C_TA  N_TH  $kpla 0    0.0     ! TH ring
  improper   S_TH  C_TA  N_TH  C_TB  $kpla 0    0.0     ! TH ring

  improper   O_HY  C_QU  C_CA  O_CA  $kbac 0  180.0
  improper   C_QU  C_QU  C_CA  O_CA  $kbac 0  180.0

  improper   C_TB  C_OL  C_OL  H_OL  $kbac 0  180.0
  improper   C_QU  C_QU  C_OL  C_OL  $kbac 0  180.0
  improper   C_TB  C_OL  C_OL  C_QU  $kbac 0  180.0
  improper   N_TH  C_TB  C_OL  C_OL  $kbac 0    0.0

  improper   N_PR  C_PA  C_PB  C_PG  $kbac 0   25.0     ! PR ring pucker

! the non-bonded parameters are approximately the same as
! DISMAN's large radii for repel=0.9, and as DISMAN's small
! radii for repel=0.75. (ECEPP2)
! the epsilon values are generally not used (and should not be)
! with this force field
! the radius is sigma*2^(-5/6)
!
!                      eps      sigma       eps(1:4) sigma(1:4)

  NONBonded  C_CA      0.0903   3.3409      0.0903   3.3409
  NONBonded  C_AR      0.120    3.3409      0.1200   3.3409
  NONBonded  C_TA      0.145    3.3409      0.1450   3.3409
  NONBonded  C_TB      0.1200   3.3409      0.1200   3.3409
  NONBonded  C_TG      0.1200   3.3409      0.1200   3.3409
  NONBonded  C_OL      0.1200   3.3409      0.1200   3.3409
  NONBonded  C_EP      0.1200   3.3409      0.1200   3.3409
  NONBonded  C_QU      0.0903   3.3409      0.0903   3.3409
  NONBonded  C_PA      0.0903   3.3409      0.0903   3.3409
  NONBonded  C_PB      0.0903   3.3409      0.0903   3.3409
  NONBonded  C_PG      0.1450   3.3409      0.1450   3.3409
  NONBonded  C_PD      0.1450   3.3409      0.1450   3.3409
  NONBonded  H_QU      0.0045   2.2272      0.0045   2.2272
  NONBonded  H_AM      0.0498   2.2272      0.0498   2.2272
  NONBonded  H_HY      0.0498   2.2272      0.0498   2.2272 ! to be checked
  NONBonded  H_AR      0.0045   2.2272      0.0045   2.2272
  NONBonded  H_TB      0.0045   2.2272      0.0045   2.2272
  NONBonded  H_TG      0.0045   2.2272      0.0045   2.2272
  NONBonded  H_OL      0.0045   2.2272      0.0045   2.2272
  NONBonded  N_PR      0.1592   3.0068      0.1592   3.0068
  NONBonded  N_TH      0.1592   3.0068      0.1592   3.0068
  NONBonded  N_AM      0.1592   3.0068      0.1592   3.0068
  NONBonded  O_CA      0.2342   2.7755      0.2342   2.7755
  NONBonded  O_HY      0.2342   2.7755      0.2342   2.7755
  NONBonded  S_TH      0.0239   3.7458      0.0239   3.7458

! the following nbfixes allow hydrogen bonding
! the distance used is (2A/B)^(1/6)*repel  distances
!                              A     B    A1-4  B1-4

  nbfix      H_AM     O_CA     44.2  1.0  44.2  1.0
  nbfix      H_AM     O_HY     44.2  1.0  44.2  1.0

set message on echo on end
```

17 Anhang D (Resultate der Strukturrechnung für Epothilon A)

		gebundene Form			freie Form			Röntgen-
Atom 1	Atom 2	NOE	gerechnete Strukt.		NOE	gerechnete Strukt.		struktur
H02*	H03	–	–	–	2.500 Å	2.749 Å	±0.002 Å	2.615 Å
H02*	H06	–	–	–	3.500 Å	2.863 Å	±0.016 Å	2.391 Å
H02*	H08	5.000 Å	5.196 Å	±0.044 Å	–	–	–	3.515 Å
H02*	H091	–	–	–	3.500 Å	3.266 Å	±0.022 Å	2.886 Å
H02*	H092	–	–	–	5.000 Å	5.017 Å	±0.022 Å	4.397 Å
H02*	H17	–	–	–	3.500 Å	4.250 Å	±0.013 Å	5.705 Å
H02*	H22*	4.200 Å	3.292 Å	±0.005 Å	3.000 Å	2.882 Å	±0.005 Å	3.233 Å
H02*	H23*	4.200 Å	3.209 Å	±0.029 Å	4.000 Å	4.462 Å	±0.005 Å	4.544 Å
H02*	HO03	–	–	–	3.500 Å	3.307 Å	±0.017 Å	3.532 Å
H03	H06	–	–	–	2.500 Å	4.104 Å	±0.012 Å	3.960 Å
H03	H07	–	–	–	3.500 Å	5.896 Å	±0.007 Å	5.785 Å
H03	H08	5.000 Å	3.235 Å	±0.045 Å	–	–	–	4.334 Å
H03	H102	–	–	–	3.500 Å	3.929 Å	±0.022 Å	4.398 Å
H03	H17	–	–	–	5.000 Å	4.965 Å	±0.012 Å	5.754 Å
H03	H22*	4.200 Å	3.906 Å	±0.002 Å	3.000 Å	3.558 Å	±0.004 Å	3.065 Å
H03	H23*	4.200 Å	3.048 Å	±0.017 Å	4.200 Å	2.788 Å	±0.005 Å	2.835 Å
H03	H24*	6.000 Å	5.615 Å	±0.005 Å	–	–	–	5.609 Å
H03	HO03	–	–	–	2.500 Å	2.151 Å	±0.011 Å	2.061 Å
H06	H08	3.500 Å	3.099 Å	±0.055 Å	–	–	–	2.864 Å
H06	H091	–	–	–	2.500 Å	2.702 Å	±0.011 Å	2.422 Å
H06	H22*	3.000 Å	2.568 Å	±0.014 Å	–	–	–	3.138 Å
H06	H23*	–	–	–	4.000 Å	4.484 Å	±0.006 Å	4.581 Å
H06	H24*	3.000 Å	2.658 Å	±0.001 Å	3.000 Å	2.618 Å	±0.003 Å	2.518 Å
H06	H25*	6.000 Å	4.606 Å	±0.010 Å	3.000 Å	4.819 Å	±0.003 Å	4.652 Å
H06	HO07	–	–	–	3.500 Å	3.303 Å	±0.180 Å	2.677 Å
H07	H08	3.500 Å	2.466 Å	±0.017 Å	2.500 Å	2.707 Å	±0.002 Å	2.461 Å
H07	H092	–	–	–	3.500 Å	3.502 Å	±0.007 Å	3.586 Å
H07	H22*	6.000 Å	5.164 Å	±0.011 Å	–	–	–	5.220 Å
H07	H23*	–	–	–	6.000 Å	5.682 Å	±0.010 Å	5.325 Å
H07	H24*	3.000 Å	3.058 Å	±0.019 Å	3.000 Å	2.906 Å	±0.006 Å	2.989 Å
H07	H25*	4.200 Å	3.084 Å	±0.028 Å	3.000 Å	2.792 Å	±0.009 Å	2.811 Å
H07	HO07	–	–	–	2.500 Å	2.231 Å	±0.081 Å	2.276 Å
H08	H25*	–	–	–	3.000 Å	2.648 Å	±0.003 Å	2.502 Å
H091	H092	–	–	–	3.000 Å	1.758 Å	±0.002 Å	1.565 Å
H091	H25*	–	–	–	3.500 Å	3.847 Å	±0.005 Å	3.708 Å
H091	HO07	–	–	–	2.500 Å	2.947 Å	±0.259 Å	3.159 Å
H092	HO07	–	–	–	3.500 Å	2.341 Å	±0.437 Å	3.324 Å
H101	H102	–	–	–	2.500 Å	1.765 Å	±0.002 Å	1.570 Å
H101	H111	–	–	–	3.500 Å	3.003 Å	±0.003 Å	2.814 Å
H102	H111	–	–	–	2.500 Å	2.621 Å	±0.002 Å	2.375 Å

		gebundene Form			freie Form			Röntgen-
Atom 1	Atom 2	NOE	gerechnete Strukt.		NOE	gerechnete Strukt.		struktur
H102	H112	–	–	–	3.500 Å	3.004 Å	±0.002 Å	2.816 Å
H102	H12	–	–	–	3.500 Å	2.781 Å	±0.009 Å	3.125 Å
H102	H142	–	–	–	4.000 Å	4.388 Å	±0.013 Å	4.894 Å
H111	H112	–	–	–	3.500 Å	1.762 Å	±0.002 Å	1.566 Å
H111	H12	–	–	–	2.500 Å	3.031 Å	±0.001 Å	3.352 Å
H112	H12	–	–	–	3.500 Å	2.461 Å	±0.005 Å	2.854 Å
H112	H141	2.500 Å	2.717 Å	±0.038 Å	–	–	–	3.459 Å
H112	H142	–	–	–	4.200 Å	4.850 Å	±0.008 Å	4.621 Å
H12	H13	–	–	–	2.500 Å	2.559 Å	±0.004 Å	2.570 Å
H13	H141	3.500 Å	3.060 Å	±0.004 Å	2.500 Å	3.042 Å	±0.002 Å	2.849 Å
H13	H142	3.500 Å	2.429 Å	±0.012 Å	3.000 Å	2.530 Å	±0.003 Å	2.386 Å
H13	H15	2.500 Å	2.690 Å	±0.036 Å	2.500 Å	2.338 Å	±0.010 Å	2.557 Å
H13	H17	5.000 Å	4.274 Å	±0.102 Å	5.000 Å	4.803 Å	±0.010 Å	4.276 Å
H141	H142	–	–	–	2.500 Å	1.762 Å	±0.001 Å	1.567 Å
H141	H15	3.500 Å	3.026 Å	±0.002 Å	3.000 Å	2.954 Å	±0.001 Å	2.800 Å
H141	H17	3.500 Å	3.730 Å	±0.081 Å	–	–	–	4.415 Å
H142	H15	3.500 Å	2.456 Å	±0.019 Å	4.200 Å	2.598 Å	±0.004 Å	2.500 Å
H142	H17	5.000 Å	2.536 Å	±0.130 Å	4.200 Å	4.383 Å	±0.005 Å	3.584 Å
H15	H17	2.500 Å	2.598 Å	±0.050 Å	2.500 Å	2.526 Å	±0.005 Å	2.234 Å
H17	H19	5.000 Å	4.107 Å	±0.001 Å	3.500 Å	2.792 Å	±0.005 Å	2.428 Å
H17	H27*	–	–	–	4.200 Å	4.036 Å	±0.003 Å	3.892 Å
H17	HO03	–	–	–	5.000 Å	5.135 Å	±0.026 Å	6.617 Å
H19	H27*	3.200 Å	2.363 Å	±0.001 Å	4.200 Å	5.549 Å	±0.005 Å	5.351 Å
H22*	H23*	–	–	–	3.000 Å	3.388 Å	±0.005 Å	3.235 Å
H22*	HO03	–	–	–	4.200 Å	4.853 Å	±0.005 Å	4.464 Å
H23*	HO03	–	–	–	4.200 Å	4.245 Å	±0.016 Å	2.945 Å
H24*	HO07	–	–	–	4.200 Å	3.969 Å	±0.365 Å	2.656 Å
H25*	HO07	–	–	–	4.200 Å	3.340 Å	±0.313 Å	4.099 Å

Tabelle 5: NOESY – Distanzen und berechnete Distanzen in Epothilon A

			gebundene Form		Röntgenkristallstruktur-	
Vektor 1	Vektor 2	Rate (Exp.)	gerechnete Struktur		analyse	
C02 H021	C03 H03	3.0 Hz	3.6 Hz	−148.5 °	5.3 Hz	165.3 °
C02 H022	C03 H03	2.1 Hz	1.0 Hz	−28.5 °	−2.8 Hz	−74.7 °
C03 H03	C04 C22	−1.3 Hz	−1.9 Hz	176.9 °	0.9 Hz	71.2 °
C03 H03	C04 C23	0.0 Hz	0.5 Hz	56.9 °	0.3 Hz	−48.8 °
C07 H07	H08 C08	−2.0 Hz	−1.9 Hz	−60.0 °	−2.9 Hz	−80.7 °

Tabelle 6: Kreuzkorrelierte Raten und gefundene Dihedralwinkel in Epothilon A

			freie Form		Röntgenkristallstruktur-	
Atom 1	Atom 2	J (Exp.)	gerechnete Struktur		analyse	
H021	H03	11.0 Hz	12.0 Hz	177.6 °	11.2 Hz	165.7 °
H022	H03	3.4 Hz	2.6 Hz	−61.9 °	0.9 Hz	−76.7 °
H06	H07	8.3 Hz	10.1 Hz	−154.9 °	12.0 Hz	−179.5 °
H07	H08	1.0 Hz	0.7 Hz	−95.3 °	0.6 Hz	−82.4 °
H091	H101	8.5 Hz	8.0 Hz	144.5 °	11.9 Hz	176.0 °
H092	H102	9.4 Hz	8.1 Hz	144.9 °	11.7 Hz	171.2 °
H102	H112	11.3 Hz	10.0 Hz	156.4 °	11.9 Hz	176.4 °
H13	H141	9.8 Hz	12.0 Hz	−175.3 °	11.9 Hz	177.1 °
H13	H142	3.4 Hz	2.7 Hz	62.9 °	4.2 Hz	54.0 °
H141	H15	9.3 Hz	10.0 Hz	156.1 °	9.3 Hz	151.8 °
H142	H15	2.0 Hz	0.6 Hz	−82.3 °	0.5 Hz	−85.0 °
H15	C13	5.0 Hz	4.0 Hz	38.0 °	4.5 Hz	33.5 °
H15	C01	1.5 Hz	1.8 Hz	56.7 °	3.0 Hz	46.0 °
H15	C27	2.0 Hz	5.1 Hz	140.0 °	8.5 Hz	−178.1 °
H15	C17	4.0 Hz	4.4 Hz	−34.5 °	6.6 Hz	5.8 °
H03	C22	2.0 Hz	0.6 Hz	103.1 °	0.5 Hz	70.8 °
H03	C23	1.0 Hz	5.8 Hz	−20.9 °	2.4 Hz	−51.3 °
H03	C05	1.0 Hz	4.2 Hz	−133.5 °	8.1 Hz	−167.9 °
H03	C01	1.0 Hz	1.6 Hz	58.5 °	3.2 Hz	44.4 °
H07	C24	2.0 Hz	3.8 Hz	−39.6 °	1.5 Hz	−59.1 °
H07	C25	4.0 Hz	5.5 Hz	24.6 °	4.3 Hz	35.8 °
H07	C09	3.0 Hz	5.3 Hz	141.1 °	7.4 Hz	158.5 °
H07	C05	0.5 Hz	0.0 Hz	84.0 °	1.5 Hz	59.7 °
H17	C19	3.7 Hz	6.7 Hz	1.4 °	6.2 Hz	−15.1 °
H17	C15	7.5 Hz	6.7 Hz	−1.5 °	6.7 Hz	−1.6 °
H17	C27	9.1 Hz	8.5 Hz	−176.1 °	8.5 Hz	−177.1 °

Tabelle 7: Skalare Kopplungen und gefundene Dihedralwinkel in Epothilon A

	gebundene Form	freie Form
RMSD (alle Atome)	0.537 Å	0.497 Å
RMSD (Nichtwasserstoffatome)	0.103 Å	0.016 Å

Tabelle 8: RMSD – Werte für die zehn energieärmsten Epothilon A Strukturen

Atom 1	Atom 2	Atom 3	Atom 4	gebundene Form		freie Form		Röntgen
C01	C02	C03	C04	−148.5 °	±0.7 °	179.4 °	±0.5 °	165.3 °
C02	C03	C04	C05	176.9 °	±1.0 °	104.7 °	±0.2 °	71.2 °
C03	C04	C05	C06	−115.3 °	±2.1 °	−86.4 °	±0.6 °	−76.6 °
C04	C05	C06	C07	137.5 °	±1.0 °	136.0 °	±0.7 °	145.5 °
C05	C06	C07	C08	−57.2 °	±2.0 °	−38.0 °	±0.4 °	−61.0 °

Atom 1	Atom 2	Atom 3	Atom 4	gebundene Form		freie Form		Röntgen
C06	C07	C08	C09	−60.0 °	±2.7 °	−96.2 °	±0.3 °	−80.7 °
C07	C08	C09	C10	179.2 °	±1.6 °	167.9 °	±0.8 °	158.6 °
C08	C09	C10	C11	−151.7 °	±7.1 °	145.4 °	±0.4 °	173.6 °
C09	C10	C11	C12	174.0 °	±1.1 °	155.8 °	±0.3 °	173.7 °
C10	C11	C12	C13	−166.7 °	±4.0 °	−93.7 °	±0.5 °	−112.1 °
C11	C12	C13	C14	6.7 °	±0.4 °	3.3 °	±0.3 °	0.7 °
C12	C13	C14	C15	81.8 °	±1.8 °	100.5 °	±0.4 °	93.6 °
C13	C14	C15	O011	−59.4 °	±2.6 °	−81.2 °	±0.2 °	−84.0 °
C14	C15	O011	C01	169.4 °	±2.9 °	175.6 °	±0.7 °	163.4 °
C15	O011	C01	C02	133.8 °	±3.8 °	149.8 °	±0.2 °	174.1 °
O011	C01	C02	C03	69.8 °	±2.0 °	102.8 °	±0.8 °	152.4 °
C14	C15	C16	C17	−73.7 °	±5.1 °	−154.7 °	±0.1 °	−116.6 °

Tabelle 9: Dihedralwinkel im Rückgrat der bestimmten Epothilon A Struktur

18 Anhang E (Implementierung der Projektionswinkelbeschränkungen in X-Plor)

```
C===============
C
C vectangl.fcm
C
C by Jens Meiler / Michael Nilges Jan 1999
C===============
      INTEGER MAXVEANCLASSES
      PARAMETER (MAXVEANCLASSES = 10)
C
C arrays that hold vectorangle info
C vectassndx tells ending index of the vect_x arrays (below)
C for each class.
C vectforces holds k1 and k2 for each class
C vectclasstypes holds the type of each class
C
C these are small enough not to bother with the heap.
C
      INTEGER VEANASSNDX    (MAXVEANCLASSES)
      DOUBLE PRECISION VEANFORCES(2,MAXVEANCLASSES)
      CHARACTER*8 VEANCLASSNAMES (MAXVEANCLASSES)
      LOGICAL PRINTCLASS(MAXVEANCLASSES)
C
C maxvects = number of slots set aside for vectorangle constant
C           assignments
C nvects = total number of vectorangle constants entered
C
      INTEGER MAXVEANS, NVEANS, NCLASSES, CURCLASS
C
C pointers to arrays to hold atom numbers, J-obs, and
C error for each vectorangle assignment
C
      INTEGER VEANIPTR, VEANJPTR, VEANKPTR, VEANLPTR, VEANDATYPTR,
     &     VEANVOBSPTR, VEANVERRPTR, VEANVTENPTR, CALCVEANPTR
C
C  cross-validation
C  by Jens Meiler / Michael Nilges Jan 1999
C
C cross-validation array
      INTEGER VEANCV
C
C cross-validation test number
      INTEGER VICV
C
C input modes
C

      INTEGER MODE, NEW, UPDATE
      PARAMETER (NEW = 1)
      PARAMETER (UPDATE = 2)
C
C===>parameter MCONST truncates the reciprocals and makes sure
C===>we don't devide by zero
      DOUBLE PRECISION MCONST
      PARAMETER (MCONST=0.0001D0)
C

C common blocks
C
      COMMON /CVEAN/ VEANCLASSNAMES
      COMMON /IVEAN/ VEANASSNDX,
     &     MAXVEANS, NVEANS, CURCLASS, NCLASSES,
     &     VEANIPTR, VEANJPTR, VEANKPTR, VEANLPTR, VEANDATYPTR,
     &     VEANVOBSPTR, VEANVERRPTR, VEANVTENPTR, CALCVEANPTR, MODE,
     &     VICV, VEANCV
      COMMON /RVEAN/ VEANFORCES
      COMMON /LVEAN/ PRINTCLASS
      SAVE /CVEAN/
      SAVE /IVEAN/
      SAVE /RVEAN/
      SAVE /LVEAN/
```

```
C===============
      SUBROUTINE EVEAN (EV, WHICH)
C
C Calls EVEAN2, which does the actual energy calculation
C
C by Jens Meiler / Michael Nilges Jan 1999
C===============
      IMPLICIT NONE
C include files
      INCLUDE 'VECTANGL.FCM'
      INCLUDE 'HEAP.FCM'
C i/o
      DOUBLE PRECISION EV
      CHARACTER*7 WHICH
C begin
C
C
      CALL EVEAN2(EV, HEAP(VEANIPTR), HEAP(VEANJPTR), HEAP(VEANKPTR),
     &     HEAP(VEANLPTR), HEAP(VEANDATYPTR),
     &     HEAP(VEANVOBSPTR), HEAP(VEANVERRPTR), HEAP(VEANVTENPTR),
     &     HEAP(CALCVEANPTR), WHICH, HEAP(VEANCV))
      RETURN
      END
C===============
      SUBROUTINE EVEAN2 (EV, ATOMI, ATOMJ, ATOMK, ATOML, DATY,
     &     VOBS, VERR, VTEN, VCALC, WHICH, VCV)
C
C Calculates vectorangle constant energies
C
C energies are of the form
C      E = kbord*deltaV**2 or E*kcent*(1+cos(deltaV / delta))
C where
C      kbord = energy constant for angles at border,
C      kcent = energy constant for angles in center,
C
C by Jens Meiler / Michael Nilges Jan 1999
C
C===============
      IMPLICIT NONE
C include files
      INCLUDE 'COORD.FCM'
      INCLUDE 'NUMBERS.FCM'
      INCLUDE 'DERIV.FCM'
      INCLUDE 'VECTANGL.FCM'
      INCLUDE 'CONSTA.FCM'
      INCLUDE 'COMAND.FCM'
C
      INCLUDE 'PSF.FCM'
C i/o
      INTEGER ATOMI(*), ATOMJ(*), ATOMK(*), ATOML(*), VCV(*), DATY(*)
      DOUBLE PRECISION VOBS(2,*), VERR(2,*), VTEN(2,*), VCALC(2,*)
      DOUBLE PRECISION EV
      CHARACTER*7 WHICH
C local variables
      INTEGER COUNT, CLASS, LOOPIJ, LOOPKL
      DOUBLE PRECISION Kveancent, Kveanbord, Ealpha, Dalpha
      DOUBLE PRECISION XI, XJ, XK, XL, YI, YJ, YK, YL, ZI, ZJ, ZK, ZL
      DOUBLE PRECISION XIJ, XKL, YIJ, YKL, ZIJ, ZKL, LIJ, LKL
      DOUBLE PRECISION COSALPHA, SINALPHA, RECSIN, ALPHA, VOBSA, VOBSB, VOBSC, VOBSD, DV
      DOUBLE PRECISION DCIJ, DCKL, PHIIJMIN, PHIIJMAX, PHIKLMIN, PHIKLMAX, PHIIJ, PHIKL
      DOUBLE PRECISION TEMPFIJ, TEMPFKL, ALPHAPLUS, ALPHAMINUS, SQRTNORM, SQRTINVERT,
DEVIATION
      DOUBLE PRECISION DPRIJX, DPRIJY, DPRIJZ, DPRKLX, DPRKLY, DPRKLZ
C begin
C
C following Axel's code in ETOR,
C
C zero out partial energy
C
      EV = ZERO
C
      CLASS = 1
      Kveancent = VEANFORCES(2,CLASS)
      Kveanbord = VEANFORCES(1,CLASS)
      DO COUNT = 1, NVEANS
          IF (VEANASSNDX(CLASS).LT.COUNT) THEN
               CLASS = CLASS + 1
               Kveancent = VEANFORCES(1,CLASS)
               Kveanbord = VEANFORCES(2,CLASS)
          END IF
C
          XI = X(ATOMI(COUNT))
          XJ = X(ATOMJ(COUNT))
          XK = X(ATOMK(COUNT))
          XL = X(ATOML(COUNT))
C
          YI = Y(ATOMI(COUNT))
          YJ = Y(ATOMJ(COUNT))
          YK = Y(ATOMK(COUNT))
          YL = Y(ATOML(COUNT))
C
          ZI = Z(ATOMI(COUNT))
          ZJ = Z(ATOMJ(COUNT))
          ZK = Z(ATOMK(COUNT))
          ZL = Z(ATOML(COUNT))
C
C now calculate diffs RIJ=RI-RJ, RJK=RJ-RK, RKL=RK-RL
C
          XIJ = XI - XJ
          XKL = XK - XL
C
          YIJ = YI - YJ
          YKL = YK - YL
C
          ZIJ = ZI - ZJ
          ZKL = ZK - ZL
C
C calculate the norm of A, B, & C & set to MCONST if it's too small
C
          LIJ = ONE/SQRT(MAX(MCONST, XIJ**2+YIJ**2+ZIJ**2))
          LKL = ONE/SQRT(MAX(MCONST, XKL**2+YKL**2+ZKL**2))
C
C normalize A, B, & C
C
          XIJ = XIJ * LIJ
          YIJ = YIJ * LIJ
          ZIJ = ZIJ * LIJ
          XKL = XKL * LKL
          YKL = YKL * LKL
          ZKL = ZKL * LKL
C
C calculate cos(alpha)
C
          COSALPHA = XIJ*XKL+YIJ*YKL+ZIJ*ZKL
          SINALPHA = SQRT(MAX(ZERO,ONE-COSALPHA**2))
C
C calculate alpha (make sure cos is within bounds and get sign from
C sin) and keep it it radians
C
          ALPHA = ACOS(MIN(ONE,MAX(-ONE,COSALPHA)))
C
C if type not 0 calculate VOBSA VOBSB VOBSC VOBSD with actual distances
C otherwise set it from prdefined values
C
          IF(DATY(COUNT).eq.0) then
               VOBSA = (VOBS(1,COUNT) - VERR(1,COUNT)) * PI / 180
               VOBSB = (VOBS(1,COUNT) + VERR(1,COUNT)) * PI / 180
               VOBSC = (VOBS(2,COUNT) - VERR(2,COUNT)) * PI / 180
               VOBSD = (VOBS(2,COUNT) + VERR(2,COUNT)) * PI / 180
               IF(VOBSC.lt.VOBSB) then
                    VOBSC = PI/2
                    VOBSB = PI/2

               END IF
          ELSE
C
C scale couplings with respect to actual distance
C
```

```
               DCIJ = VOBS(1,COUNT) / ((LIJ)**3)
               DCKL = VOBS(2,COUNT) / ((LKL)**3)
C
C set ranges for angle aplha to maximum
C
               VOBSA = PI
               VOBSB = 0
               VOBSC = PI
               VOBSD = 0
C
C calculate possible ranges for angle phi
C
               PHIIJMIN = ((DCIJ / VTEN(1,COUNT) + 1) * 2 / 3 / VTEN(2,COUNT))
               PHIKLMIN = ((DCKL / VTEN(1,COUNT) + 1) * 2 / 3 / VTEN(2,COUNT))
               PHIIJMIN = ACOS(MIN(ONE,MAX(-ONE,PHIIJMIN))) / 2
               PHIKLMIN = ACOS(MIN(ONE,MAX(-ONE,PHIKLMIN))) / 2
               PHIIJMAX = PI - PHIIJMIN
               PHIKLMAX = PI - PHIKLMIN
C
C check all possible combinations for largest range
C
               PHIIJ = PHIIJMIN
               DO LOOPIJ = 1, 3
                    PHIKL = PHIKLMIN
                    DO LOOPKL = 1, 3
                         TEMPFIJ = MIN(ONE,MAX(0,(DCIJ / VTEN(1,COUNT) - 2) / (VTEN(2,COUNT)*
COS(2 * PHIIJ) - 2) * 2 / 3))
                         TEMPFKL = MIN(ONE,MAX(0,(DCKL / VTEN(1,COUNT) - 2) / (VTEN(2,COUNT)*
COS(2 * PHIKL) - 2) * 2 / 3))
                         SQRTNORM   = SQRT(TEMPFIJ*TEMPFKL)
                         SQRTINVERT = SQRT((1-TEMPFIJ)*(1-TEMPFKL))
                         ALPHAPLUS  = ACOS(MIN(ONE,MAX(-ONE,SQRTNORM * COS(PHIIJ - PHIKL)+
SQRTINVERT)))
                         ALPHAMINUS = ACOS(MIN(ONE,MAX(-ONE,SQRTNORM * COS(PHIIJ - PHIKL)-
SQRTINVERT)))
                         IF(ALPHAPLUS.lt.VOBSA)  VOBSA = ALPHAPLUS
                         IF(ALPHAMINUS.lt.VOBSC) VOBSC = ALPHAMINUS
                         ALPHAPLUS  = ACOS(MIN(ONE,MAX(-ONE,SQRTNORM * COS(PHIIJ + PHIKL)+
SQRTINVERT)))
                         ALPHAMINUS = ACOS(MIN(ONE,MAX(-ONE,SQRTNORM * COS(PHIIJ + PHIKL)-
SQRTINVERT)))
                         IF(ALPHAPLUS.gt.VOBSB)  VOBSB = ALPHAPLUS
                         IF(ALPHAMINUS.gt.VOBSD) VOBSD = ALPHAMINUS
                    PHIKL = PHIKL + (PHIKLMAX - PHIKLMIN) / 2
                    END DO
               PHIIJ = PHIIJ + (PHIIJMAX - PHIIJMIN) / 2
               END DO
C
               IF(VOBSC.lt.VOBSB) THEN
                    VOBSC = PI / 2
                    VOBSB = PI / 2
               END IF
C
C end calculation
C
          END IF
C
C calculate energy and forces for normal entries
C
          if (ALPHA.lt.VOBSA) then
               Ealpha = Kveanbord*((ALPHA-VOBSA)**2)
               Dalpha = Kveanbord* (ALPHA-VOBSA) *2
               DEVIATION = VOBSA - ALPHA
          elseif (ALPHA.lt.VOBSB) then
               Ealpha = 0
               Dalpha = 0
               DEVIATION = 0
          elseif (ALPHA.lt.VOBSC.AND.VOBSB.lt.VOBSC) then
               Ealpha = 0.5*Kveancent*(1+COS(2*PI*(ALPHA-VOBSB)/(VOBSC-VOBSB)-PI))
               Dalpha = Kveancent*PI/(VOBSB-VOBSC)*SIN(2*PI*(ALPHA-VOBSB)/(VOBSC-VOBSB)-PI)
               DEVIATION = MIN(VOBSC - ALPHA, ALPHA-VOBSB)
          elseif (ALPHA.lt.VOBSD) then
               Ealpha = 0
               Dalpha = 0
               DEVIATION = 0
          else
               Ealpha = Kveanbord*((ALPHA-VOBSD)**2)
               Dalpha = Kveanbord* (ALPHA-VOBSD) *2
               DEVIATION = ALPHA - VOBSD
          end if
C
C If we're printing out the couplings, then we need to
C try both possible assignments of observed with the
C the angle and pick the one with the lower total violation.
C
          IF (WHICH.EQ.'ANALYZE') THEN
                VCALC(1,COUNT) = ALPHA
                VCALC(2,COUNT) = DEVIATION
          END IF
C
C accumulate energy (only for working set)
C
          EV = Ealpha + EV
          DV = Dalpha
C
C compute reciprocal of SP multiplied by the derivative of the
C function DF
          RECSIN = DV/MAX(MCONST,SINALPHA)
C
C compute:
C    d (alpha)
C DF ---------,  etc.
C    d rij
          DPRIJX = RECSIN*LIJ*(COSALPHA*XIJ-XKL)
          DPRIJY = RECSIN*LIJ*(COSALPHA*YIJ-YKL)
          DPRIJZ = RECSIN*LIJ*(COSALPHA*ZIJ-ZKL)
          DPRKLX = RECSIN*LKL*(COSALPHA*XKL-XIJ)
          DPRKLY = RECSIN*LKL*(COSALPHA*YKL-YIJ)
          DPRKLZ = RECSIN*LKL*(COSALPHA*ZKL-ZIJ)
C
C now update forces if in energy & force mode
C
          IF (WHICH.NE.'ANALYZE') THEN
            IF (VCV(COUNT).NE.VICV) THEN
               DX(ATOMI(COUNT))=DX(ATOMI(COUNT))+DPRIJX
               DY(ATOMI(COUNT))=DY(ATOMI(COUNT))+DPRIJY
               DZ(ATOMI(COUNT))=DZ(ATOMI(COUNT))+DPRIJZ
               DX(ATOMJ(COUNT))=DX(ATOMJ(COUNT))-DPRIJX
               DY(ATOMJ(COUNT))=DY(ATOMJ(COUNT))-DPRIJY
               DZ(ATOMJ(COUNT))=DZ(ATOMJ(COUNT))-DPRIJZ
               DX(ATOMK(COUNT))=DX(ATOMK(COUNT))+DPRKLX
               DY(ATOMK(COUNT))=DY(ATOMK(COUNT))+DPRKLY
               DZ(ATOMK(COUNT))=DZ(ATOMK(COUNT))+DPRKLZ
               DX(ATOML(COUNT))=DX(ATOML(COUNT))-DPRKLX
               DY(ATOML(COUNT))=DY(ATOML(COUNT))-DPRKLY
               DZ(ATOML(COUNT))=DZ(ATOML(COUNT))-DPRKLZ
           END IF
         END IF
      END DO
      RETURN
      END
C================
      SUBROUTINE READVEAN
C
C reads in vectorangle constant information
C
C by Jens Meiler / Michael Nilges Jan 1999
C================
      IMPLICIT NONE
C include files
      INCLUDE 'COMAND.FCM'
      INCLUDE 'VECTANGL.FCM'
      INCLUDE 'FUNCT.FCM'
      INCLUDE 'PSF.FCM'
      INCLUDE 'HEAP.FCM'
      INCLUDE 'NUMBERS.FCM'
      INCLUDE 'CTITLA.FCM'
C i/o
```

```
C local variables
      INTEGER COUNT, SPTR, OLDCLASS, OLDMAXVEANS, TEMP
      DOUBLE PRECISION K1, K2, CUTOFF
      CHARACTER*4 THENAME
C begin
C
C this is used by READVEAN2 to hold the selection
C
      SPTR=ALLHP(INTEG4(NATOM))
C
C reset database only if no vectorangles have been entered
C ie., this is the first time in the xplor script that
C vectorangles has appeared
C
      IF (VEANIPTR.EQ.0) THEN
           CALL VEANDEFAULTS
           CALL ALLOCVEANS(0, MAXVEANS)
      END IF
C
C now read input
C
      CALL PUSEND('VEANORANGLES>')
      DO UNTIL (DONE)
           CALL NEXTWD('VEANORANGLES>')
           CALL MISCOM('VEANORANGLES>',USED)
           IF (.NOT.USED) THEN
C
           IF (WD(1:4).EQ.'HELP') THEN
C-DOCUMENTATION-SOURCE-BEGIN
                WRITE(DUNIT,'(5X,A)')
     &'VEAN E<vectorangles-statement>L END ',
     &'<VEANORANGLES-statement>:== ',
     &'   ASSIgn <slctn> <slctn> <slctn> <slctn>',
     &'          <real>  <real>  <real>  <real>'
                WRITE(DUNIT,'(5X,A)')
     &'       E* atom i   atom j   atom k  atom l  angle1  error1  angle2  error2*L',
     &'   CLASs <name> ! Starts a new class. Applies to all ',
     &'          ! ASSIgn , TYPE, FORCe, and FLAT entries until '
                WRITE(DUNIT,'(5X,A)')
     &'          ! another CLASS entry is issued. '
                WRITE(DUNIT,'(5X,A)')
     &'   CV=<integer> ! select partition number for ',
     &'                !  cross-validation '
                WRITE(DUNIT,'(5X,A)')
     &'   FORCe <real> <real> ! force constants for all ',
     &'          ! assignments in the ',
     &'          ! current class. Edefault = 50, 10L ',
     &'   NREStraints <integer> ! number of slots for vectorangles ',
     &'          ! restraints to allocate in memory Edefault = 200L '
                WRITE(DUNIT,'(5X,A)')
     &'   PARTition=<integer> ! number of partitions for complete ',
     &'                       ! cross-validation ',
     &'   PRINt THREshold <real> <ALL | CLASs <name>> ! prints ',
     &'          ! vectorangle violations '
                WRITE(DUNIT,'(3X,A)')
     &'          ! greater than the specified value (in Hz) ',
     &'   RESEt  ! erases the vectorangle assignment table, ',
     &'          ! but keeps NREStraints the same. '
C-DOCUMENTATION-SOURCE-END
C
C Get class name.  Determine if it's an already-defined class.
C Insert a new class if it's not.
C
           ELSE IF (WD(1:4).EQ.'CLAS') THEN
                OLDCLASS = CURCLASS
                CALL NEXTA4('class name =', THENAME)
                MODE = NEW
                DO COUNT = 1, NCLASSES
                     IF (VEANCLASSNAMES(COUNT).EQ.THENAME) THEN
                          MODE = UPDATE
                          CURCLASS = COUNT
                     END IF
                END DO
                IF (MODE.EQ.NEW) THEN
```

```
C
C make sure you can't add more than the maximum
C number of classes
C
                     IF (OLDCLASS.EQ.MAXVEANCLASSES) THEN
                          CALL DSPERR('VEAN','Too many classes.')
                          CALL DSPERR('VEAN',
     &                      'Increase MAXVEANCLASSES and recompile.')
                          CALL WRNDIE(-5, 'READVEAN',
     &                                'Too many vectorangle classes.')
                     END IF
                     NCLASSES = NCLASSES + 1
                     CURCLASS = NCLASSES
                     VEANCLASSNAMES(CURCLASS) = THENAME
C
C If this isn't the first class, close off the old class
C
                    IF (NCLASSES.GT.1) THEN
                         VEANASSNDX(OLDCLASS) = NVEANS
                    END IF
                END IF
C
C set force constant for current class
C
           ELSE IF (WD(1:4).EQ.'FORC') THEN
                CALL NEXTF('border force constant =', K1)
                CALL NEXTF('center force constant =', K2)
C
C start a default class if there isn't one defined
C
                IF (CURCLASS.EQ.0) THEN
                     NCLASSES = 1
                     CURCLASS = 1
                END IF
                WRITE(DUNIT, '(A, A, A, F8.3, F8.3)')
     &              'Setting force consts for class ',
     &               VEANCLASSNAMES(CURCLASS), ' to ', K1, K2
                VEANFORCES(1,CURCLASS) = K1
                VEANFORCES(2,CURCLASS) = K2
C
C reset vectorangle database
C
           ELSE IF (WD(1:4).EQ.'RESE') THEN
                CALL VEANDEFAULTS
                CALL ALLOCVEANS(MAXVEANS, MAXVEANS)
C
C change number of assignment slots
C
           ELSE IF (WD(1:4).EQ.'NRES') THEN
                OLDMAXVEANS = MAXVEANS
                CALL NEXTI('number of slots =', MAXVEANS)
                CALL ALLOCVEANS(OLDMAXVEANS, MAXVEANS)
C
C read in an assignment
C
           ELSE IF (WD(1:4).EQ.'ASSI') THEN
C
C make sure you can't add more vectorangle assignments
C than you have slots for
C
                IF (NVEANS.EQ.MAXVEANS) THEN
                     CALL DSPERR('VEAN','Too many assignments.')
                     CALL DSPERR('VEAN',
     &               'Increase NREStraints and run again.')
                     CALL WRNDIE(-1,'VEAN>',
     &               'exceeded allocation for vectorangle restraints')
                END IF
C
C if there isn't a class specified,
C start a default class
C
                IF (CURCLASS.EQ.0) THEN
                     NCLASSES = 1
                     CURCLASS = 1
                END IF
                CALL READVEAN2(HEAP(VEANIPTR), HEAP(VEANJPTR),
     &               HEAP(VEANKPTR), HEAP(VEANLPTR),
     &               HEAP(VEANDATYPTR), HEAP(VEANVOBSPTR),
     &               HEAP(VEANVERRPTR), HEAP(VEANVTENPTR),
     &               HEAP(SPTR), HEAP(VEANCV))
C
C print violations
C
           ELSE IF (WD(1:4).EQ.'PRIN') THEN
                CALL NEXTWD('PRINt>')
                IF (WD(1:4).NE.'THRE') THEN
                     CALL DSPERR('VEANORANGLES',
     &                           'print expects THREshold parameter.')
                ELSE
                     CALL NEXTF('THREshold =', CUTOFF)
                     IF (CUTOFF.LT.ZERO) THEN
                          CALL DSPERR('VEANORANGLES',
     &                                'cutoff must be positive.')
                          CUTOFF = ABS(CUTOFF)
                     END IF
                     CALL NEXTA4('ALL or CLASs>', THENAME)
                     IF (THENAME(1:3).EQ.'ALL') THEN
                          DO COUNT = 1,NCLASSES
                               PRINTCLASS(COUNT) = .TRUE.
                          END DO
                     ELSE IF (THENAME(1:4).EQ.'CLAS') THEN
                          CALL NEXTA4('class name =', THENAME)
                          DO COUNT = 1,NCLASSES
                               IF (VEANCLASSNAMES(COUNT).EQ.
     &                              THENAME) THEN
                                    PRINTCLASS(COUNT) = .TRUE.
                               ELSE
                                    PRINTCLASS(COUNT) = .FALSE.
                               END IF
                          END DO
                     ELSE
                          CALL DSPERR('VEANORANGLES',
     &                         'not understood.  Printing all classes')
                          DO COUNT = 1,NCLASSES
                               PRINTCLASS(COUNT) = .TRUE.
                          END DO
                     END IF
C
                     CALL PRINTVEANS(CUTOFF, HEAP(CALCVEANPTR),
     &                    HEAP(VEANVOBSPTR), HEAP(VEANVERRPTR),
     &                    HEAP(VEANVTENPTR), HEAP(VEANDATYPTR),
     &                    HEAP(VEANIPTR), HEAP(VEANJPTR),
     &                    HEAP(VEANKPTR), HEAP(VEANLPTR),
     &                    HEAP(VEANCV), 0)
                     IF (VICV.GT.0) THEN
                       CALL PRINTVEANS(CUTOFF, HEAP(CALCVEANPTR),
     &                      HEAP(VEANVOBSPTR), HEAP(VEANVERRPTR),
     &                      HEAP(VEANIPTR), HEAP(VEANJPTR),
     &                      HEAP(VEANKPTR), HEAP(VEANLPTR),
     &                      HEAP(VEANCV), 1)
                     END IF
                END IF
C=====================================================================
      ELSE IF (WD(1:2).EQ.'CV') THEN
      CALL NEXTI('CV excluded partition number:',VICV)
C=====================================================================
      ELSE IF (WD(1:4).EQ.'PART') THEN
      CALL NEXTI('number of PARTitions:',TEMP)
      TEMP=MAX(0,TEMP)
      CALL VCVS(NVEANS,HEAP(VEANCV),TEMP)
      IF (TEMP.EQ.0) THEN
      VICV=0
      END IF
C
C check for END statement
C
           ELSE
                CALL CHKEND('VEANANGL>', DONE)
           END IF
           END IF
      END DO
      DONE = .FALSE.
      CALL FREHP(SPTR,INTEG4(NATOM))
      RETURN
      END
C===============
      SUBROUTINE ALLOCVEANS (OLDSIZE, NEWSIZE)
C
C resets vectorangle constant arrays to hold SIZE entries
C
C by Jens Meiler / Michael Nilges Jan 1999
C===============
      IMPLICIT NONE
C include files
      INCLUDE 'FUNCT.FCM'
      INCLUDE 'VECTANGL.FCM'
C i/o
      INTEGER OLDSIZE, NEWSIZE
C begin
      IF (OLDSIZE.NE.0) THEN
           CALL FREHP(VEANIPTR, INTEG4(OLDSIZE))
           CALL FREHP(VEANJPTR, INTEG4(OLDSIZE))
           CALL FREHP(VEANKPTR, INTEG4(OLDSIZE))
           CALL FREHP(VEANLPTR, INTEG4(OLDSIZE))
           CALL FREHP(VEANCV  , INTEG4(OLDSIZE))
           CALL FREHP(VEANVOBSPTR, IREAL8(2*OLDSIZE))
           CALL FREHP(VEANVERRPTR, IREAL8(2*OLDSIZE))
           CALL FREHP(CALCVEANPTR, IREAL8(2*OLDSIZE))
           CALL FREHP(VEANVTENPTR, IREAL8(2*OLDSIZE))
           CALL FREHP(VEANDATYPTR, INTEG4(OLDSIZE))
       END IF
C
      VEANIPTR = ALLHP(INTEG4(NEWSIZE))
      VEANJPTR = ALLHP(INTEG4(NEWSIZE))
      VEANKPTR = ALLHP(INTEG4(NEWSIZE))
      VEANLPTR = ALLHP(INTEG4(NEWSIZE))
      VEANCV   = ALLHP(INTEG4(NEWSIZE))
      VEANDATYPTR = ALLHP(INTEG4(NEWSIZE))
      VEANVOBSPTR = ALLHP(IREAL8(2*NEWSIZE))
      VEANVERRPTR = ALLHP(IREAL8(2*NEWSIZE))
      VEANVTENPTR = ALLHP(IREAL8(2*NEWSIZE))
      CALCVEANPTR = ALLHP(IREAL8(2*NEWSIZE))
      RETURN
      END
C==============
      SUBROUTINE VEANDEFAULTS
C
C sets up defaults
C
C by Jens Meiler / Michael Nilges Jan 1999
C==============
      IMPLICIT NONE
C include files
      INCLUDE 'VECTANGL.FCM'
C local variables
      INTEGER COUNT
      DOUBLE PRECISION P
C begin
      MODE = NEW
      MAXVEANS = 1000
      NVEANS = 0
      NCLASSES = 0
      CURCLASS = 0
      VICV = 0
      DO COUNT = 1, MAXVEANCLASSES
           VEANCLASSNAMES(COUNT) = 'DEFAULT'
           VEANASSNDX(COUNT) = 0
           VEANFORCES (1,COUNT) =  40
           VEANFORCES (2,COUNT) =  10
      END DO
      RETURN
      END
```

```
C==============
      SUBROUTINE READVEAN2 (ATOMI, ATOMJ, ATOMK, ATOML,
     &     DATY, VOBS, VERR, VTEN, SEL, VCV)
C
C reads actual vectorangle assignments into arrays
C
C by Jens Meiler / Michael Nilges Jan 1999
C==============
      IMPLICIT NONE
C include files
      INCLUDE 'COORD.FCM'
      INCLUDE 'VECTANGL.FCM'
      INCLUDE 'NUMBERS.FCM'
      INCLUDE 'PSF.FCM'
      INCLUDE 'CONSTA.FCM'
      INCLUDE 'COMAND.FCM'
C i/o
      INTEGER ATOMI(*), ATOMJ(*), ATOMK(*), ATOML(*), DATY(*),
     &        SEL(*), VCV(*)
      DOUBLE PRECISION VOBS(2,*), VERR(2,*), VTEN(2,*)
C local variables
      INTEGER NSEL, INSERTPOS, COUNT, CURSTOP, OTHERSTOP, LOOPIJ, LOOPKL
      DOUBLE PRECISION JO, JE, SI, RH,
     &                 DCIJ, DCKL,
     &                 PHIIJMIN, PHIIJMAX, PHIKLMIN, PHIKLMAX,
     &                 PHIIJ, PHIKL, TEMPFIJ, TEMPFKL,
     &                 TEMPFIJ, TEMPFKL, ALPHAPLUS, ALPHAMINUS,
     &                 SQRTNORM, SQRTINVERT, DEVIATION,
     &                 VOBSA, VOBSB, VOBSC, VOBSD
C begin
C
C if we're in update mode, make a space for the new line
C
      IF (MODE.EQ.UPDATE) THEN
           DO COUNT = NVEANS+1, VEANASSNDX(CURCLASS)+1, -1
                ATOMI(COUNT) = ATOMI(COUNT-1)
                ATOMJ(COUNT) = ATOMJ(COUNT-1)
                ATOMK(COUNT) = ATOMK(COUNT-1)
                ATOML(COUNT) = ATOML(COUNT-1)
                DATY(COUNT) = DATY(COUNT-1)
                VOBS(1,COUNT) = VOBS(1,COUNT-1)
                VOBS(2,COUNT) = VOBS(2,COUNT-1)
                VERR(1,COUNT) = VERR(1,COUNT-1)
                VERR(2,COUNT) = VERR(2,COUNT-1)
                VTEN(1,COUNT) = VTEN(1,COUNT-1)
                VTEN(2,COUNT) = VTEN(2,COUNT-1)
                VCV(COUNT) = VCV(COUNT-1)
           END DO
           CURSTOP = VEANASSNDX(CURCLASS)
           DO COUNT = 1, NCLASSES
                OTHERSTOP = VEANASSNDX(COUNT)
                IF (OTHERSTOP.GT.CURSTOP) THEN
                     VEANASSNDX(COUNT) = OTHERSTOP + 1
                END IF
           END DO
           VEANASSNDX(CURCLASS) = CURSTOP + 1
           INSERTPOS = CURSTOP
           NVEANS = NVEANS + 1
      ELSE
           NVEANS = NVEANS + 1
           INSERTPOS = NVEANS
           VEANASSNDX(CURCLASS) = INSERTPOS
      END IF
C
      CALL SELCTA(SEL, NSEL, X, Y, Z,.TRUE.)
      IF (NSEL.GT.1) THEN
           CALL DSPERR('VEAN',
     &     'more than 1 atom in selection for atom i. Using first')
      END IF
      CALL MAKIND(SEL, NATOM, NSEL)
      ATOMI(INSERTPOS) = SEL(1)
C
      CALL SELCTA(SEL, NSEL, X, Y, Z,.TRUE.)
      IF (NSEL.GT.1) THEN
```

```
           CALL DSPERR('VEAN',
     &     'more than 1 atom in selection for atom j. Using first')
      END IF
      CALL MAKIND(SEL, NATOM, NSEL)
      ATOMJ(INSERTPOS) = SEL(1)
C
      CALL SELCTA(SEL, NSEL, X, Y, Z,.TRUE.)
      IF (NSEL.GT.1) THEN
           CALL DSPERR('VEAN',
     &     'more than 1 atom in selection for atom k. Using first')
      END IF
      CALL MAKIND(SEL, NATOM, NSEL)
      ATOMK(INSERTPOS) = SEL(1)
C
      CALL SELCTA(SEL, NSEL, X, Y, Z,.TRUE.)
      IF (NSEL.GT.1) THEN
           CALL DSPERR('VEAN',
     &     'more than 1 atom in selection for atom l. Using first')
      END IF
      CALL MAKIND(SEL, NATOM, NSEL)
      ATOML(INSERTPOS) = SEL(1)
C
C read type of data
C 0 = angles are given
C 1 = dc are given with hard distances
C 2 = dc are given with weak distances
C
      CALL NEXTI('type of given data (0,1,2) =', NSEL)
      DATY(INSERTPOS) = NSEL
      CALL NEXTF('observed vectorangle =', JO)
      CALL NEXTF('error in vectorangle =', JE)
      VOBS(1,INSERTPOS) = JO
      VERR(1,INSERTPOS) = JE
      CALL NEXTF('observed vectorangle =', JO)
      CALL NEXTF('error in vectorangle =', JE)
      VOBS(2,INSERTPOS) = JO
      VERR(2,INSERTPOS) = JE
      CALL NEXTF('tensor axial =', SI)
      VTEN(1,INSERTPOS) = SI
      CALL NEXTF('tensor rhombicity =', RH)
      VTEN(2,INSERTPOS) = RH
C
C check if type = 1 and calculate angles
C
      IF(NSEL.eq.1) THEN
C
           DATY(INSERTPOS) = 0
C
C scale couplings with respect to actual distance
C
           DCIJ = VOBS(1,INSERTPOS) * (VERR(1,INSERTPOS)**3)
           DCKL = VOBS(2,INSERTPOS) * (VERR(2,INSERTPOS)**3)
C
C set ranges for angle aplha to maximum
C
           VOBSA = PI
           VOBSB = 0
           VOBSC = PI
           VOBSD = 0
C
C calculate possible ranges for angles phi
C
           PHIIJMIN = ((DCIJ / VTEN(1,INSERTPOS) + 1) * 2 / 3 / VTEN(2,INSERTPOS))
           PHIKLMIN = ((DCKL / VTEN(1,INSERTPOS) + 1) * 2 / 3 / VTEN(2,INSERTPOS))
           PHIIJMIN = ACOS(MIN(ONE,MAX(-ONE,PHIIJMIN))) / 2
           PHIKLMIN = ACOS(MIN(ONE,MAX(-ONE,PHIKLMIN))) / 2
           PHIIJMAX = PI - PHIIJMIN
           PHIKLMAX = PI - PHIKLMIN
C
C check all possible combinations for largest range
C
           PHIIJ = PHIIJMIN
           DO LOOPIJ = 1, 3
                PHIKL = PHIKLMIN
                DO LOOPKL = 1, 3
                     TEMPFIJ = MIN(ONE,MAX(0,(DCIJ / VTEN(1,INSERTPOS) - 2) /
(VTEN(2,INSERTPOS)* COS(2 * PHIIJ) - 2) * 2 / 3))
                     TEMPFKL = MIN(ONE,MAX(0,(DCKL / VTEN(1,INSERTPOS) - 2) /
(VTEN(2,INSERTPOS)* COS(2 * PHIKL) - 2) * 2 / 3))
                     SQRTNORM   = SQRT(TEMPFIJ*TEMPFKL)
                     SQRTINVERT = SQRT((1-TEMPFIJ)*(1-TEMPFKL))
                     ALPHAPLUS  = ACOS(MIN(ONE,MAX(-ONE,SQRTNORM * COS(PHIIJ - PHIKL)+
SQRTINVERT)))
                     ALPHAMINUS = ACOS(MIN(ONE,MAX(-ONE,SQRTNORM * COS(PHIIJ - PHIKL)-
SQRTINVERT)))
                     IF(ALPHAPLUS.lt.VOBSA)  VOBSA = ALPHAPLUS
                     IF(ALPHAMINUS.lt.VOBSC) VOBSC = ALPHAMINUS
                     ALPHAPLUS  = ACOS(MIN(ONE,MAX(-ONE,SQRTNORM * COS(PHIIJ + PHIKL)+
SQRTINVERT)))
                     ALPHAMINUS = ACOS(MIN(ONE,MAX(-ONE,SQRTNORM * COS(PHIIJ + PHIKL)-
SQRTINVERT)))
                     IF(ALPHAPLUS.gt.VOBSB)  VOBSB = ALPHAPLUS
                     IF(ALPHAMINUS.gt.VOBSD) VOBSD = ALPHAMINUS
                PHIKL = PHIKL + (PHIKLMAX - PHIKLMIN) / 2
                END DO
           PHIIJ = PHIIJ + (PHIIJMAX - PHIIJMIN) / 2
           END DO
C
           IF(VOBSC.lt.VOBSB) THEN
                VOBSC = PI / 2
                VOBSB = PI / 2
           END IF
C
           VOBS(1,INSERTPOS) = 90.0 / PI * (VOBSA + VOBSB)
           VOBS(2,INSERTPOS) = 90.0 / PI * (VOBSC + VOBSD)
           VERR(1,INSERTPOS) = 90.0 / PI * (VOBSB - VOBSA)
           VERR(2,INSERTPOS) = 90.0 / PI * (VOBSD - VOBSC)
C
      END IF
C
C end
C
      VCV(INSERTPOS)  = -1
      RETURN
      END
C===============
      SUBROUTINE VEANINIT
C
C initializes vectorangles
C
C by Jens Meiler / Michael Nilges Jan 1999
C==============
      IMPLICIT NONE
C include files
      INCLUDE 'VECTANGL.FCM'
C begin
      CALL VEANDEFAULTS
      CALL ALLOCVEANS(0, MAXVEANS)
      RETURN
      END
C===============
      SUBROUTINE VEANHP
C
C deallocates vectorangles
C
C by Jens Meiler / Michael Nilges Jan 1999
C================
      IMPLICIT NONE
C include files
      INCLUDE 'VECTANGL.FCM'
C begin
      CALL ALLOCVEANS(MAXVEANS,0)
      RETURN
      END
C=================
      SUBROUTINE PRINTVEANS (CUTOFF, VCALC,
     &     VOBS, VERR, VTEN, DATY,
     &     ATOMI, ATOMJ, ATOMK, ATOML,
     &     VCV, ITEST)
C
C prints vectorangles with delta alpha greater than cutoff
C calculates RMS deviation and puts it into $RESULT
C
C by Jens Meiler / Michael Nilges Jan 1999
C=================
      IMPLICIT NONE
C include files
      INCLUDE 'VECTANGL.FCM'
      INCLUDE 'COMAND.FCM'
      INCLUDE 'PSF.FCM'
      INCLUDE 'NUMBERS.FCM'
      INCLUDE 'CONSTA.FCM'
C i/o
      DOUBLE PRECISION CUTOFF, VCALC(2,*), VOBS(2,*), VERR(2,*), VTEN(2,*)
      INTEGER ATOMI(*), ATOMJ(*), ATOMK(*), ATOML(*), DATY(*)
      INTEGER VCV(*), ITEST
C local variables
      DOUBLE PRECISION RMS, VIOLS, VENERGY, DELTA, CALCV, VOBSA, VOBSB, VOBSC, VOBSD
      INTEGER NASSIGNSINCLUDED, COUNT, CLASS, I, J, K, L, LOOPIJ, LOOPKL
      LOGICAL PRINTTHISCLASS
      DOUBLE COMPLEX DUMMY2
C begin
      RMS = ZERO
      VIOLS = ZERO
      NASSIGNSINCLUDED = ZERO
C
C make sure that the calcJ array is up to date
C
      CALL EVEAN(VENERGY, 'ANALYZE')
      IF (VICV.GT.0) THEN
        IF (ITEST.EQ.0) THEN
           WRITE(PUNIT,'(A)')
     & ' $$$$$$$$$$$$$$$$$$ working set $$$$$$$$$$$$$$$$$$$$$$$$$$$$$$'
        ELSE
           WRITE(PUNIT,'(A,I5,A)')
     & ' $$$$$$$$$$$$$$$$$$$$ test set (TEST=',VICV,
     & ')  $$$$$$$$$$$$$$$$$$$$$$'
        END IF
      END IF
      WRITE (PUNIT, '(A)') 'The following vectorangles have delta'
      WRITE (PUNIT, '(A)') 'greater than the cutoff:'
      WRITE (PUNIT, '(A)') '(calculated)(observed borderA borderB borderC borderD)(delta)'
C
C write out first class heading
C
      CLASS = 1
      PRINTTHISCLASS = PRINTCLASS(CLASS)
      IF (PRINTTHISCLASS) THEN
           WRITE (PUNIT, '(A, A)') 'class ', VEANCLASSNAMES(1)
      END IF
C
C for every vectorangle entry,
C
      DO COUNT = 1, NVEANS
C
C is this the start of a new class?
C
        IF (VEANASSNDX(CLASS).LT.COUNT) THEN
               CLASS = CLASS + 1
               PRINTTHISCLASS = PRINTCLASS(CLASS)
               IF (PRINTTHISCLASS) THEN
               WRITE (PUNIT, '(A, A)') 'class ', VEANCLASSNAMES(CLASS)
               END IF
        END IF
C
C check if in test set or not (cross-validation)
C
        IF ((ITEST.EQ.0.AND.VCV(COUNT).NE.VICV).OR.
     &      (ITEST.EQ.1.AND.VCV(COUNT).EQ.VICV)) THEN
C       IF (1.EQ.1) THEN
C
C if this ASSIgnment is in a class that will be printed,
C
        IF (PRINTTHISCLASS) THEN
C
C update RMS delta J
C
               CALCV = VCALC(1,COUNT) * 180 / PI
               DELTA = VCALC(2,COUNT) * 180 / PI
C
C if type not 0 calculate VOBSA VOBSB VOBSC VOBSD with actual distances
C otherwise set it from prdefined values
C
               IF(DATY(COUNT).eq.0) THEN
                     VOBSA = (VOBS(1,COUNT) - VERR(1,COUNT))
                     VOBSB = (VOBS(1,COUNT) + VERR(1,COUNT))
                     VOBSC = (VOBS(2,COUNT) - VERR(2,COUNT))
                     VOBSD = (VOBS(2,COUNT) + VERR(2,COUNT))
                     IF(VOBSC.lt.VOBSB) THEN
                           VOBSC = 90.0
                           VOBSB = 90.0
                     END IF
               END IF
C
               RMS = RMS + DELTA**2
               NASSIGNSINCLUDED = NASSIGNSINCLUDED + 1
C
C print out delta Js greater than cutoff
C and update number of violations
C
               IF (ABS(DELTA).GT.CUTOFF) THEN
                     I = ATOMI(COUNT)
                     J = ATOMJ(COUNT)
                     K = ATOMK(COUNT)
                     L = ATOML(COUNT)
                     WRITE(PUNIT,'(9X,16(1X,A), 6(F8.3))')
     &                   SEGID(I),RESID(I),RES(I),DATY(I),
     &                   SEGID(J),RESID(J),RES(J),DATY(J),
     &                   SEGID(K),RESID(K),RES(K),DATY(K),
     &                   SEGID(L),RESID(L),RES(L),DATY(L),
     &                   CALCV, VOBSA, VOBSB, VOBSC, VOBSD, DELTA
                     VIOLS = VIOLS + ONE
               END IF
          END IF
        END IF
      END DO
C
      IF (NASSIGNSINCLUDED.GT.ZERO) THEN
           RMS = SQRT(RMS / NASSIGNSINCLUDED)
      ELSE
           RMS = ZERO
      END IF
      WRITE(PUNIT,'(A,F8.3,A,F5.2,A,F6.0,A,I6,A)')
     & '  RMS diff. =',RMS,
     & ', #(violat.>',CUTOFF,')=',VIOLS,
     & ' of ',NASSIGNSINCLUDED,' vectorangles'
      IF (ITEST.EQ.1) THEN
        CALL DECLAR('RESULT', 'DP', ' ', DUMMY2, RMS)
        CALL DECLAR('TEST_RMS', 'DP', ' ', DUMMY2, RMS)
        CALL DECLAR('TEST_VIOLATIONS', 'DP', ' ', DUMMY2, VIOLS)
      ELSE
        CALL DECLAR('RESULT', 'DP', ' ', DUMMY2, RMS)
        CALL DECLAR('RMS', 'DP', ' ', DUMMY2, RMS)
        CALL DECLAR('VIOLATIONS', 'DP', ' ', DUMMY2, VIOLS)
      END IF
      RETURN
      END
C
C=====================================================================
      SUBROUTINE VCVS(VNUM,VCV,PART)
C
C Routine partitions vectorangle data into PART sets.
C VCV will contain integer numbers between 1 and PART.
C
C Author: Axel T. Brunger
C Modifed for vectorangles: Jens Meiler / Michael Nilges Jan 1999
```

```
C
C
C     IMPLICIT NONE
C I/O
      INTEGER VNUM, VCV(*)
      INTEGER PART
C local
      INTEGER I, P, NP, NRETRY, NTRYTOT
      DOUBLE PRECISION RNUM
      NRETRY = 0
      NTRYTOT = 0
C begin
  100 CONTINUE
      IF (PART.GT.0) THEN
      DO I=1,VNUM
      CALL GGUBFS(RNUM)
      VCV(I)=MAX(1,MIN(PART,INT(RNUM*PART)+1))
      END DO
C
      IF (PART.EQ.VNUM) THEN
      DO I=1,VNUM
      VCV(I)=I
      END DO
      END IF
C
      DO P=1,PART
      NP=0
      DO I=1,VNUM
      IF (VCV(I).EQ.P) THEN
      NP=NP+1
      END IF
      END DO
      IF (NP .EQ. 0) THEN
        NRETRY = 1
      ENDIF
      WRITE(6,'(A,I3,A,I5,A)') ' For set ',P,
     & ' there are ',NP,' vectorangle restraints.'
      END DO
      ELSE
      WRITE(6,'(A)')
     & ' Data are not partitioned or partitioning removed.'
      DO I=1,VNUM
      VCV(I)=-1
      END DO
      END IF
      NTRYTOT = NTRYTOT + NRETRY
      IF (NTRYTOT .GT. 0 .AND. NTRYTOT .LE. 10) THEN
      WRITE(6,'(A)')
     & ' Test set with 0 constraints! New trial...'
      GOTO 100
      ELSE IF (NTRYTOT .GT. 10) THEN
      CALL WRNDIE(-1,'VCVS',
     & 'Unable to partition the vectorangle data within ten trials')
      ENDIF
C
      RETURN
      END
C
```

19 Anhang F (Lebenslauf)

geboren am 31.08.74 in Leipzig

Vater: Dr. Wolfgang Meiler, Diplom Physiker
Mutter: Dr. Monika Meiler, Diplom Mathematiker

1981	Einschulung
1989-1993	Gymnasium „Wilhelm Ostwald“ in Leipzig
1993	Abschluss der Schule mit der allgemeinen Hochschulreife („sehr gut“)
1993/94	Bundeswehrdienst
1994-1998	Stipendium der „Studienstiftung des Deutschen Volkes“
1994-1998	Chemiestudium an der Universität Leipzig
1995	Diplom Vorexamen („sehr gut“)
1998	Diplom Hauptexamen („sehr gut”)
1998-2000	„Kekulé” Stipendium des Fonds der chemischen Industrie
1998	Beginn der Dissertation an der Universität Frankfurt bei Herrn Prof. Griesinger

20 Anhang G (akademische Lehrer)

Prof. Dr. S. Berger, Prof. Dr. L. Beyer, Prof. Dr. J. Borsdorf, Prof. Dr. K. Burger, Prof. Dr. F. Dietz, Prof. Dr. W. Engewald, Prof. Dr. C. Griesinger, Prof. Dr. H. Hennig, Prof. Dr. R. Herzschuh, Prof. Dr. E. Hey-Hawkins, P. D. Dr. M. Köck, Dr. M. Meiler, Doz. Dr. W. Meiler, P. D. Dr. R. Meusinger, Prof. Dr. H. Papp, Prof. Dr. K. Quitsch, Prof. Dr. J. Reinhold, Prof. Dr. J. Sieler, Prof. Dr. R. Szargan, Prof. Dr. P. Welzel, Prof. Dr. G. Werner

21 Anhang H (eigene Publikationen)

[1] Meiler, J. "Untersuchung von Reaktionsprodukten bei einigen Reaktionen von Derivaten der Abietinsäure", *Junge Wissenschaft* **1993**, *32*, 40-45.

[2] Meiler, J.; Meusinger, R. "Use of Neural Networks to Determine Properties of Alkanes from their 13C-NMR Spectra", In *Software - Entwicklung in der Chemie*; Gasteiger, J., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1995; Vol. 10, pp 259-263.

[3] Meiler, J.; Meusinger, R.; Will, M. "Prediction of 13C-NMR Chemical Shifts of Substituted Benzenes by Means of a Neural Network", In *Software - Entwicklung in der Chemie*; Fels, G., Schubert, V., Eds.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1996; Vol. 11, pp 234-238.

[4] Meiler, J. "Künstliche Intelligenz zu Aufdeckung von Spektren-Eigenschafts-Beziehungen", *Junge Wissenschaft* **1996**, *43*, 33-41.

[5] Meiler, J.; Seidel, P. "Reaktionen in der Kerzenflamme", *Junge Wissenschaft* **1996**, *41*, 10-22.

[6] Meiler, J. "Untersuchung von Struktur-Eigenschafts-Beziehungen für die Spezifität von Serin-Proteasen gegenüber Polypeptiden mittels NMR-Spektroskopie und künstlicher neuronaler Netze" (Diplomarbeit)", *Universität Leipzig* **1998**, *diploma thesis*.

[7] Wendt, M. A.; Meiler, J.; Weinhold, F.; Farrar, T. C. "Solvent And Concentration Dependence Of The Hydroxyl Chemical Shift Of Methanol", *Molecular Physics* **1998**, *93*, 145-151.

[8] Meiler, J.; Meusinger, R.; Will, M. "Neural Network Prediction of 13C NMR Chemical Shifts of Substituted Benzenes", *Monatshefte für Chemie* **1999**, *130*, 1089-1095.

[9] Meiler, J.; Will, M.; Meusinger, R. "Fast Determination of 13C-NMR Chemical Shifts Using Artificial Neural Networks", *J. Chem. Inf. Comput. Sci.* 2000, *40*, 1169-1176.

[10] Meiler, J.; Peti, W.; Griesinger, C. "DipoCoup: A versatile program for 3D-structure homology comparision based on residual dipolar couplings and pseudocontact shifts", *J. Biomol. NMR* 2000, *17*, 283-294.

[11] Meiler, J.; Blomberg, N.; Nilges, M.; Griesinger, C. "A new Approach for Applying Residual Dipolar Couplings as Restraints in Structure Elucidation", *J. Biomol. NMR* 2000, *16*, 245-252.

[12] Meiler, J.; Müller, M.; Zeidler, A.; Schmäschke, F. "Generation and Evaluation of Dimension Reduced Amino Acid Parameter Representations by Artificial Neural Networks", *accepted* 2001.

[13] Meiler, J.; Peti, W.; Prompers, J.; Griesinger, C.; Brueschweiler, R. "Model-Free Approach to the Dynamic Interpretation of Residual Dipolar Couplings in Globular Proteins ", *in press* 2001.

[14] **Meiler, J.; Köck, M. "Structure Elucidation by Automatic Generation and Analysis of Molecule Databases from NMR Connectivity Information Using Substructure Analysis and 13C-NMR Chemical Shift Prediction", *accepted* 2001.**

[15] **Meiler, J.; Will, M. "Structure Elucidation from 13C-NMR Chemical Shifts by Genetic Algorithms", *submitted* 2001.**

[16] **Meiler, J.; Bleckmann, A. "Epothilones: QSAR Studies Performed by Artificial Neural Networks Leading to New Drug Proposals", *submitted* 2001.**

[17] Neubauer, H.; Meiler, J.; Peti, W.; Griesinger, C. "NMR Structure Determination of Saccharose and Raffinose by Means of Homo- and Heteronuclear Dipolar Couplings", *Hel. Chim. Acta* **2001**, *84*, 243-2858.

[18] Peti, W.; Meiler, J.; Prompers, J.; Brueschweiler, R.; Griesinger, C. "Influence of Molecular Motion on Residual Dipolar Spin-Spin Couplings in Proteins - A Practical Approach", *in prep.* **2001.**

Jegliches hat seine Zeit,
Steine sammeln, Steine zerstreun,
Bäume pflanzen, Bäume abhaun,
Leben und Sterben und Frieden und Streit.

Ulrich Plenzdorf

Journal of Biomolecular NMR, **17:** 283–294, 2000.
KLUWER/ESCOM
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

DipoCoup: A versatile program for 3D-structure homology comparison based on residual dipolar couplings and pseudocontact shifts

Jens Meiler[a], Wolfgang Peti[a] & Christian Griesinger[a,b,*]
[a]*Universität Frankfurt, Institut für Organische Chemie, Marie-Curie-Str. 11, D-60439 Frankfurt am Main, Germany*
[b]*Max-Planck Institute for Biophysical Chemistry, Am Faßberg 11, D-37077 Göttingen, Germany*

Received 22 Februuary 2000; Accepted 28 April 2000

Key words: bioinformatics, dipolar couplings, postgenomic research, protein data base, structure elucidation, three dimensional homology search

Abstract

A program, *DipoCoup*, is presented that allows to search the protein data bank for proteins which have a three dimensional fold that is at least partially homologous to a protein under investigation. The three dimensional homology search uses secondary structure alignment based on chemical shifts and dipolar couplings or pseudocontact shifts for the three dimensional orientation of secondary structure elements. Moreover, the program offers additional tools for handling and analyzing dipolar couplings.

Introduction

One goal of post genomic research is to determine all protein folds. The number of folds is expected to be limited (Sali, 1998; Fischer and Eisenberg, 1999). Sequence profile methods nowadays have a big impact in fold recognition. *Ab initio* structure prediction works up to 40–60 amino acids and may emerge as a powerful tool for structure prediction in the future (Moult, 1999). To obtain a complete coverage of folds most effectively, it is important to focus on the elucidation of structures with novel folds rather than rediscovering known folds on new proteins. Blast threading and *ab initio* approaches rely on the analysis of primary and secondary structure in the context of a three dimensional structure database. We will present experimental tools that allow to compare the 3D fold of a new protein to all known folds in an early stage of NMR based structure determination. This approach has the potential to predict folds of a new protein with little homology to proteins with known folds. By the same token, structure elucidation of a new protein with a structure homologous to a known fold will be accelerated. There is so far only one example of using experimental NMR parameters in an early stage for 3D homology searches (Annila et al., 1999). Recently the possibility for using protein fragments generated from PDB and chosen by aligning similar dipolar couplings and chemical shifts for structure determination was shown (Delaglio et al., 2000). The availability of orientation information from NMR experiments in terms of residual dipolar couplings (Tolman et al., 1995; Tjandra and Bax, 1997; Bax and Tjandra, 1997; Clore et al., 1998a; Fischer et al., 1999; Peti and Griesinger, 2000; Meiler et al., 2000) offers new possibilities in this field. In this paper, we present a versatile program, *DipoCoup*, that uses chemical shifts for the alignment of secondary structure elements and tertiary structure alignment from dipolar couplings and pseudocontact shifts for the homology search in the PDB. We will show, using examples, that the program is fast enough to search through a large number of pdb files.

[*] To whom correspondence should be addressed. E-mail: cigr@org.chemie.uni-frankfurt.de

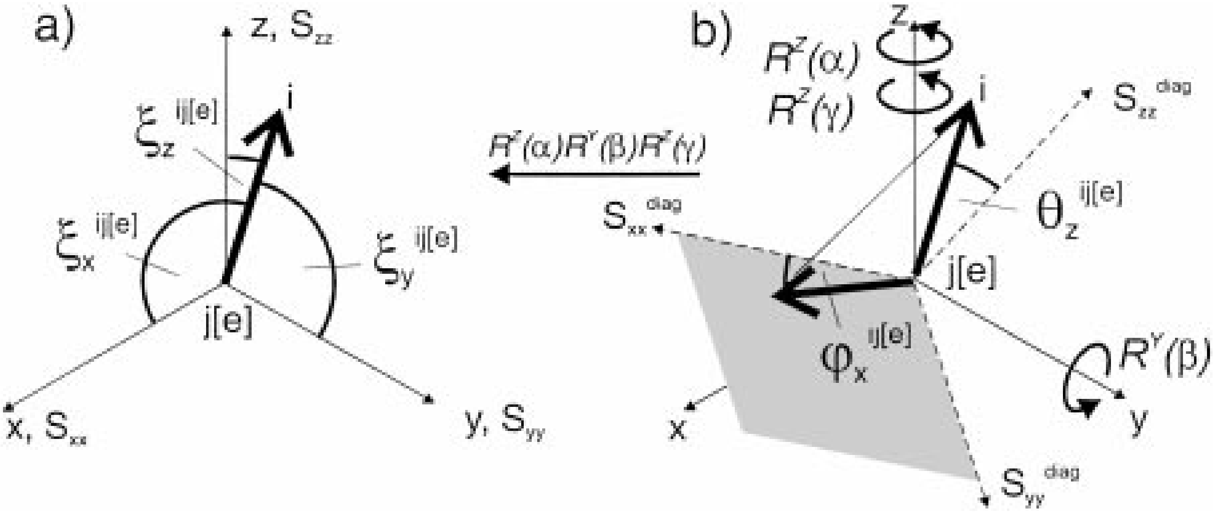


Figure 1. (a) Coordinate system of the molecule *(x, y, z)* with a bond vector between the two nuclei *i* and *j* or a nucleus *i* and an electron *e*. The projection angles of the vector onto the *x, y, z* axes are ξ_x^{ij}, ξ_y^{ij} and ξ_z^{ij}, respectively. (b) Representation of the vector in the frame of the tensor S_{xx}^{diag}, S_{yy}^{diag}, S_{zz}^{diag}. The Euler rotation transforms the tensor into the coordinate system of the molecule. The orientation of the bond vector $\vec{r}_{ij}$ is defined by the angles θ_z^{ij} and φ_x^{ij}.

Theory

Experimental dipolar couplings between nuclei i and $j\,(D^{ij})$ and pseudocontact shifts between nucleus i and electron $e(\delta_{PC}^{ie})$ are related to the alignment tensor (principal components: A_{xx}, A_{yy}, A_{zz}) or to the magnetic susceptibility tensor (principal components: χ_{xx}, χ_{yy}, χ_{zz}) and to the orientation of a specific vector with respect to the alignment tensor expressed by the projection angles θ_z^{ij} and φ_x^{ij} according to Equation 1. The vector is either the vector $\vec{r}_{ij}$ between the two coupled atoms i and j in case of dipolar coupling (Equation 1a) or the vector $\vec{r}_{ie}$ between the electron spin e (paramagnetic center) and the active nucleus i (Equation 1b).

$$D^{ij}(\theta_z^{ij},\varphi_x^{ij}) = \frac{-\mu_0 hS\gamma_i\gamma_j}{8\pi^3 r_{ij}^3}\left[\frac{1}{6}(2A_{zz}-A_{xx} - A_{yy})(3\cos^2\theta_z^{ij}-1) + \frac{1}{2}(A_{xx}-A_{yy})\cos 2\varphi_x^{ij}\sin^2\theta_z^{ij}\right], \tag{1a}$$

$$\delta_{PC}^{ie}(\theta_z^{ie},\varphi_x^{ie}) = \frac{10^6}{6\pi r_{ie}^3}\left[\frac{1}{6}(2\chi_{zz}-\chi_{xx} - \chi_{yy})(3\cos^2\theta_z^{ie}-1) + \frac{1}{2}(\chi_{xx}-\chi_{yy})\cos 2\varphi_x^{ie}\sin^2\theta_z^{ie}\right] \tag{1b}$$

In case of paramagnetic alignment the susceptibility tensor $\hat{\chi}$ is related to the alignment tensor $\hat{A}$ by $\hat{\chi} = \hat{A}(15\mu_0 kT/4B_0\pi)$. Equation 1 uses the alignment tensor as the frame of reference (Figure 1). The formal dependence of dipolar couplings and pseudo contact shifts is the same while the prefactors differ. The prefactor is constant for dipolar couplings (Equation 1a) if the distance r_{ij} is constant. For pseudocontact shifts and also for dipolar couplings between nuclei whose distance is not fixed a priori in the bonding network they vary because the distance r_{ie} or r_{ij} cannot be regarded as constant in this case (Ghose and Prestegard, 1997; Clore and Garrett, 1999).

However, the measured orientation value cannot be translated directly in a combination of θ_z^{ij} and φ_x^{ij}. An infinite number of combinations of θ_z^{ij} and φ_x^{ij} exist, that fulfill an experimental value. Still if one pair of angles θ_z^{ij} and φ_x^{ij} can be found to be correct due to the alignment of a whole molecule, four orientations of the molecule fulfill all experimental values, since the signs of angles θ_z^{ij} and φ_x^{ij} can be reversed independently in Equation 1 without the change of either dipolar couplings or pseudo contact shifts.

In the context of a 3D homology search, the coordinate system of a protein in the 3D structure data file (e.g. PDB) is the natural frame of reference. Therefore we express Equation 1 in this coordinate system which is rotated by three Euler angles α, β, γ with respect to the alignment tensor (Figure 1). Equation 2 expresses the dipolar couplings in the molecular frame (an identical equation is obtained for pseudocontact shifts δ_{PC}^{ie} by replacing j with e, all following equations are only given for dipolar couplings):

$$D^{ij}(\xi_x^{ij},\xi_y^{ij},\xi_z^{ij}) =$$

$$= F_{ij}\begin{pmatrix}\cos\xi_x^{ij}\\ \cos\xi_y^{ij}\\ \cos\xi_z^{ij}\end{pmatrix}^T \begin{pmatrix}-S_{yy}-S_{zz} & S_{xy} & S_{xz}\\ S_{xy} & S_{yy} & S_{yz}\\ S_{xz} & S_{yz} & S_{zz}\end{pmatrix}\begin{pmatrix}\cos\xi_x^{ij}\\ \cos\xi_y^{ij}\\ \cos\xi_z^{ij}\end{pmatrix}$$

$$= F_{ij}\begin{pmatrix}(\cos^2\xi_y^{ij}-\cos^2\xi_x^{ij})S_{yy}+(\cos^2\xi_z^{ij}-\cos^2\xi_x^{ij})S_{zz}\\ +(2\cos\xi_x^{ij}\cos\xi_y^{ij})S_{xy}+(2\cos\xi_x^{ij}\cos\xi_z^{ij})S_{xz}\\ +(2\cos\xi_y^{ij}\cos\xi_z^{ij})S_{yz}\end{pmatrix} \tag{2}$$

with

$$F_{ij} = \frac{-\mu_0 h S\gamma_i\gamma_j}{8\pi^3 r_{ij}^3}$$

In this molecular frame the alignment tensor is no longer diagonal and can be expressed by a symmetric three by three traceless matrix holding five independent elements S_{xx}, S_{zz}, S_{xy}, S_{xz} and S_{yz}, the elements of the Saupe matrix (Saupe, 1968). The eigenvalues of this matrix S_{xx}^{diag}, S_{yy}^{diag}, S_{zz}^{diag} are identical to the principal components of the alignment tensor A_{xx}, A_{yy}, A_{zz}. The angles ξ_x^{ij}, ξ_y^{ij}, ξ_z^{ij} define the projection angles of the bond vector $\vec{r}_{ij}$ or the vector between the nucleus and the electron $\vec{r}_{ie}$ using pseudocontact shifts onto the molecular frame. For a given structure and experimental dipolar couplings D_{exp}^{ij}, the five independent tensor contributions can be determined directly by solving the linear system of equations given from Equation 3 for a set of experimental dipolar couplings for n pairs of nuclei i and j requiring $D_{\mathrm{exp}}^{ij} = D_{\mathrm{theor}}^{ij}$ (Losonczi et al., 1999).

$$\begin{pmatrix}D_{\mathrm{exp}}^{ij1}/F_{ij}\\ \vdots\\ D_{\mathrm{exp}}^{ijn}/F_{ij}\end{pmatrix} \overset{!}{=} \begin{pmatrix}D_{\mathrm{theor}}^{ij1}/F_{ij}\\ \vdots\\ D_{\mathrm{theor}}^{ijn}/F_{ij}\end{pmatrix} =$$

$$\begin{pmatrix}\cos^2\xi_y^{ij1}-\cos^2\xi_x^{ij1} & \cos^2\xi_y^{ij1}-\cos^2\xi_x^{ij1} & 2\cos\xi_x^{ij1}\cos\xi_y^{ij1} & \cdots & 2\cos\xi_x^{ij1}\cos\xi_y^{ij1} & 2\cos\xi_x^{ij1}\cos\xi_y^{ij1}\\ \vdots & \vdots & \vdots & & \vdots & \vdots\\ \cos^2\xi_y^{ijn}-\cos^2\xi_x^{ijn} & \cos^2\xi_y^{ijn}-\cos^2\xi_x^{ijn} & 2\cos\xi_x^{ijn}\cos\xi_y^{ijn} & \cdots & 2\cos\xi_x^{ijn}\cos\xi_y^{ijn} & 2\cos\xi_x^{ijn}\cos\xi_y^{ijn}\end{pmatrix}\begin{pmatrix}S_{yy}\\ S_{zz}\\ S_{xy}\\ S_{xz}\\ S_{yz}\end{pmatrix} = \mathbf{C}\vec{S} \tag{3}$$

This system of equations can be solved by multiplication of the pseudo inverse of the rectangular matrix **C**, i.e., by calculating the Moore-Penrose-Inverse of the matrix yielding the vector $\vec{S}$. Rebuilding the Saupe matrix from these values and analyzing its eigensystem yields the eigenvalues of the tensor S_{xx}^{diag}, S_{yy}^{diag}, S_{zz}^{diag} as well as its orientation given by the eigenvectors. It can be expressed in terms of three Euler angles in α, β, and γ.

$$\begin{pmatrix}-S_{yy}-S_{zz} & S_{xy} & S_{xz}\\ S_{xy} & S_{yy} & S_{yz}\\ S_{xz} & S_{yz} & S_{zz}\end{pmatrix} = \left(R^Z(\alpha)R^Y(\beta)R^Z(\gamma)\right)^T$$

$$\begin{pmatrix}S_{xx}^{\mathrm{diag}} & 0 & 0\\ 0 & S_{yy}^{\mathrm{diag}} & 0\\ 0 & 0 & S_{zz}^{\mathrm{diag}}\end{pmatrix} \tag{4}$$

$$R^Z(\alpha)R^Y(\beta)R^Z(\gamma)$$

The solution of the Moore Penrose inversion problem is equivalent to finding a solution D_{theor}^{ij} with the least square deviation for a given experimental set of D_{exp}^{ij}. Experimental errors cannot be directly taken into consideration during this approach. Therefore a careful analysis afterwards is necessary according to Losonczi et al. (1999).

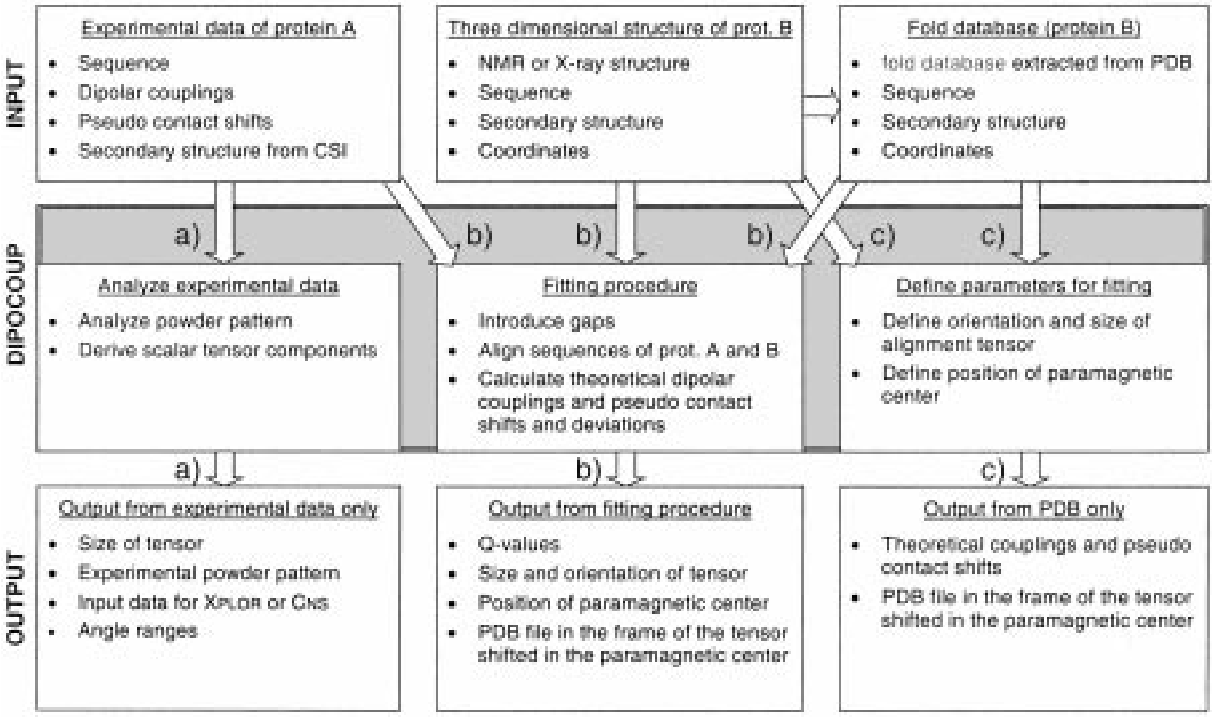

INPUT
Experimental data of protein A
Sequence
Dipolar couplings
Pseudo contact shifts
Secondary structure from CSI
Three dimensional structure of prot. B
NMR or X-ray structure
Sequence
Secondary structure
Coordinates
Fold database (protein B)
fold database extracted from PDB
Sequence
Secondary structure
Coordinates
DIPOCOUP
a)
b)
b)
b)
c)
c)
Analyze experimental data
Analyze powder pattern
Derive scalar tensor components
Fitting procedure
Introduce gaps
Align sequences of prot. A and B
Calculate theoretical dipolar couplings and pseudo contact shifts and deviations
Define parameters for fitting
Define orientation and size of alignment tensor
Define position of paramagnetic center
a)
b)
c)
OUTPUT
Output from experimental data only
Size of tensor
Experimental powder pattern
Input data for XPLOR or CNS
Angle ranges
Output from fitting procedure
Q-values
Size and orientation of tensor
Position of paramagnetic center
PDB file in the frame of the tensor shifted in the paramagnetic center
Output from PDB only
Theoretical couplings and pseudo contact shifts
PDB file in the frame of the tensor shifted in the paramagnetic center

Figure 2. Schematic features of the program *DipoCoup*. Arrows (a) describe the analysis of experimental dipolar couplings and/or pseudocontact shifts from protein **A** without the knowledge of its three dimensional structure and without the use of the database. Arrows (c) describe the analysis of the three dimensional structure of protein **B** for calculating theoretical dipolar couplings or pseudocontact shifts. Arrows (b) indicate the fitting procedure of protein **A** to the known three dimensional structure of protein **B**. This is used to obtain the orientation of the alignment tensor derived from the experimental data for **A** in the molecular frame of protein **B**. The quality of the fit is measured by the *Q*-value. Alternatively to a single molecule **B** a whole database can be searched finding homologous structures or structure fragments.

Materials and methods

The 3D homology search program *DipoCoup* was written in C^{++} and can be run on every standard PC working with either Windows95/98 or WindowsNT. The program offers three general means of analyzing dipolar couplings and pseudocontact shifts (Figure 2) of the protein **A** under investigation by comparing it to one or several selected proteins **B** from the database. In procedure (a) one can analyze the experimental dipolar couplings and/or pseudocontact shifts of protein **A** as well as obtain secondary structure information from ^{13}C chemical shift index data, CSI (Spera and Bax, 1991; Wishart et al., 1992, 1995; Wishart and Sykes, 1994). The program is able to handle different sets of dipolar couplings in combination with pseudocontact shifts for one alignment tensor. Dipolar couplings for atom pairs with defined distances (e.g., N-H^N, Cα-Hα, Cα-CO) in protein **A** scaled with F_{ij}^{-1} can be visualized in a histogram yielding a powder pattern. The eigenvalues S_{xx}^{diag}, S_{yy}^{diag}, S_{zz}^{diag} of the tensor can then be determined from the histogram (Clore et al., 1998b). With this information the program can generate input files for XPLOR- or CNS-annealing protocols which use residual dipolar couplings as restraints. It also calculates the angle projection ranges that allow to use dipolar couplings in XPLOR- or CNS-calculations without the necessity to define the orientation of the alignment tensor (Meiler et al., 2000).

3D homology searching is prepared in (c) by calculating NMR properties of a potentially homologous protein **B** which is extracted from a structure data base. From the given three dimensional structure of protein **B** a set of dipolar couplings and pseudocontact shifts can be generated. To do this, the program first adds hydrogen atoms that may be missing in the structure and corrects the bond lengths between all heavy atoms and their bound hydrogen atoms according to Bax and Ottiger (1998). For a given alignment tensor and paramagnetic center theoretical dipolar couplings and pseudocontact shifts can be calculated, visualized and exported to disk, respectively. Also the three dimensional structure can be exported to disk oriented in the frame of reference of the alignment tensor and shifted to the appropriate position with respect to the paramagnetic center.

Finally in procedure (b), both the experimental data from the protein **A** under investigation and the three dimensional structure from the protein **B** can be checked for matching 3D folds. One or several proteins **B** from the protein data bank (PDB) can be used, allowing one to compose the experimental data to a database of proteins. Hydrogen atoms are added or corrected for proteins **B** as already described in (c). The secondary structure elements of proteins **B** are calculated from the coordinates by analyzing hydrogen bonds and φ- and ψ-angles. Then the alignment of residues *a* of **A** and *b* of **B** is done such that $D_{\mathrm{exp}}(\mathrm{a})$ is assigned to the respective atoms of residue *b* of protein **B**. This set of 'experimental' dipolar couplings is used to calculate the alignment tensor and its orientation according to Equation 3. In this case no analysis of the histogram needs to be performed. As a quality measure the *Q*-value of the dipolar couplings (analogous for pseudocontact shifts) is used: $Q = \sqrt{\sum_{ij}(D^{ij}_{\mathrm{exp}} - D^{ij}_{\mathrm{theor}})^2 / \sum_{ij}(D^{ij}_{\mathrm{exp}})^2}$ (Cornilescu et al., 1998). *Q* is a normalized square deviation and is equivalent to $\sqrt{2}$ times the *R*-factor (Clore and Garrett, 1999). Moreover, the program calculates the correlation coefficient R (not to be mixed up with the *R*-factor) and offers therefore a second quality value.

The alignment of the residues *a* of protein **A** and *b* of protein **B** is not based on primary sequence homology. Rather, the sequences will be aligned to have a minimum *Q*-value. The program aligns first all amino acids of protein **A** over the amino acids of protein **B** starting with the first for both proteins, respectively. After calculation of the Q factor for this alignment the sequence of protein **A** is shifted by one residue and the procedure is repeated until the last amino acid of **A** is aligned with the last amino acid of **B**. This ensures that terminal secondary structure elements of protein **A** are fully used in the alignment process. This procedure avoids to find false positive hits due to a changing number of dipolar couplings and CSI data used. A check for matches of the secondary structure elements is performed. Secondary structure elements are derived from CSI for protein **A** and by analysis of H-bonds and ϕ and ψ (procedure (c) in Figure 2) for protein **B**. To achieve optimal alignment, the secondary structure elements of **A** can be disconnected and aligned individually with matching secondary structure elements of **B**. By default, disconnection of secondary structure elements in **A** occurs at boundaries of secondary structure elements, e.g. from β-sheet to random coil. However, the user may also suggest other positions for disconnecting the sequence, if additional information has to be used or other ideas have to be tesed. If no secondary structure alignment is possible the alignment with minimal *Q* without the use of secondary structural information is presented. The program allows for a search over the whole or part of the PDB database, as will be described subsequently.

If pseudocontact shifts are given, the position of the paramagnetic center either can be explicitly defined in the three dimensional structure or can be optimized by an interactive grid search protocol. For optimization to proceed, a starting position, a starting step, and the size of the cube to be searched has to be supplied. The program searches this given cube using the starting step size and restarts this search with the best point of the previous search and a decreased step size and size of the cube, until the step size is smaller than a predefined target value (e.g., 0.1 Å).

The program can be downloaded together with an example and two databases of 125 (Rost and Sander, 1994) and 500 representative folds out of the PDB from: http://krypton.org.chemie.uni-frankfurt.de/∼mj/software.html

$^1J_{\mathrm{NH}}$ and $^1D_{\mathrm{NH}}$ couplings were measured for the protein *Hgi*CIC (C46 → S) using the direct measurement of the $^1J_{\mathrm{NH}}$ splitting in the ^{15}N dimension of 2D ^{1}H-^{15}N HSQC spectra and $^1J_{\mathrm{NH}}$ modulated spectra (Tjandra et al., 1996). To measure the dipolar couplings, two ^{15}N labeled samples of *Hgi*CIC (C46 → S) were prepared: One for measuring isotropic $^1J_{\mathrm{NH}}$ couplings and one sample where the weak alignment to the magnetic field is induced using CHAPSO/DLPC lipid bicelles (Wang et al., 1998). Both samples contained 2.5 mM protein, 10 mM phosphate buffer at pH 6.5, 0.03% NaN_3, 0.1 mM Pefabloc SC, 600 mM NaCl, and 500 μl of 95% H_2O/5% D_2O in an 5 mm NMR tube.

The cyclophilin A sample was approximately 0.7 mM in 100 mM potassium phosphate buffer at pH 6.5 and 0.03% NaN_3. Solutions of 250 μl (95% H_2O/5% D_2O) were measured in Shigemi microcell tubes. Alignment was achieved by CHAPSO/DLPC/CTAB bicelles (5% total lipid conc.: 1: 5: 0.1; Losonczi and Prestegard, 1998).

All measurement were carried out on Bruker DRX-600 or Bruker DRX-800 (Bruker, Rheinstetten, Germany) spectrometers equipped with standard 5 mm triple-resonance, z-gradient probes. The temperature for all measurements was 303 K. The measurements of the $^1J_{NH}$ splitting in the ^{15}N dimension of 2D ^1H-^{15}N HSQC spectra were collected with 512 (t_1) × 2048 (t_2) complex data points. $^1J_{NH}$ modulated spectra were collected with 128 (t_1) × 2048 (t_2) complex data points. Data processing and analysis were performed using either XWinNMR 2.6 (Bruker, Karlsruhe, Germany) or Felix98.0 (MSI, San Diego, CA, USA).

Results and discussion

We have applied the program to three different protein structures: For rhodniin (Friedrich et al., 1993; van de Locht et al., 1995) we calculated a theoretical set of dipolar couplings and pseudocontact shifts using an NMR structure (Maurer and Griesinger, personal communication). A 3D homology search is performed on a restricted database of proteins according to procedure (b) of Figure 2. For cyclophilin A we recorded experimental dipolar couplings and procedure (a) of Figure 2 is used to analyze the experimental data. The dipolar couplings are fitted against the known NMR and X-ray structures, and the orientation of the alignment tensor is determined. The third example is the protein *Hgi*CIC which is currently under investigation in our laboratory. This protein contains a helix-turn-helix motif. Using experimentally derived dipolar couplings a 3D homology search on a restricted set of the PDB was performed.

Rhodniin consists of 103 amino acid and contains two similarly folded domains of 45 amino acids connected by a flexible linker of 10 amino acids. A set of $^1D_{NH}$ dipolar couplings and pseudocontact shifts for amide hydrogens was calculated from the known NMR structure of the protein for the N-terminal domain assuming a specific size and orientation of the alignment tensor and a specific position of a paramagnetic center. Only 36 couplings in rigid parts of the domain were used for the following calculations. The eigenvalues were set to be $S_{zz}^{\text{diag}} = 4.58 \times 10^{-4}$, $S_{yy}^{\text{diag}} = -2.96 \times 10^{-4}$ and $S_{xx}^{\text{diag}} = -1.62 \times 10^{-4}$, amounting to a rhombicity of 0.2. This set of dipolar couplings and pseudocontact shifts is used as an 'experimental' test set.

Measured dipolar couplings were fitted to the NMR structure of rhodniin, by omitting and including pseudocontact shifts. As expected, the dipolar couplings are reproduced in the first case (Figure 3, Table 1) when pseudocontact shifts were omitted. With a normalized square deviation of $Q = 0.00$ the tensor size and orientation exactly reproduce the predefined values. The Q-value is found to be 0.08 in the second case when the tensor and the position of the paramagnetic center were recalculated. The paramagnetic center is found with a deviation of 0.786 Å to its original position. This deviation is caused by the grid search step size of 0.5 Å yielding a maximum deviation of $\frac{1}{2}\sqrt{3}$ Å ≈ 0.866 Å. This deviation is also the reason for $Q > 0.00$. Additionally, deletion of one, two or three amino acids after residues 15 and 39, as well as the addition of amino acids at the same positions do not influence the result of the calculation. Sequence alignment is always found correctly, irrespective of the usage of pseudocontact shifts.

The 'experimental' set of dipolar couplings was fitted to the X-ray structure of ovomucoid (a homologous protein to the N-terminal domain of rhodniin). The 'experimental' values as well as the values calculated for the best fit are given in Figure 3, together with the visualization of both structures in the frame of the resulting alignment tensor. The program finds an eight amino acid shift in the sequence alignment (Table 1) which agrees with the primary sequence alignment for rhodniin and ovomucoid. In this case, the normalized square deviation was found to be $Q = 0.30$.

To speed up the process of three dimensional homology search, a subset of 125 folds was extracted from the PDB with a diverse set of folds according to Rost and Sander (1994). Loading the data and calculating secondary structure elements for all proteins in the fold database takes about 5 min on a 450 MHz Pentium II processor. The search itself takes only below 1 s for the whole database, if no gaps are introduced. This time increases to be 48 s if disconnecting of protein parts as explained above with a gap size of up to 5 amino acids is allowed.

The search over this database using the earlier mentioned theoretical set of dipolar couplings for the N terminal domain of rhodniin (a typical Kazal inhibitor) yields ovomucoid (1ovo_a) as 2nd best hit with a Q-value of 0.45 and porcine pancreatic secretory trypsin inhibitor (1tgs_i) as 16th best hit with a Q-value of 0.53. Both proteins are known as Kazal inhibitors and are homologous to rhodniin. In 9 out of these best 16 examples the α-helix of the rhodniin

Table 1. Results of fitting the experimental set of dipolar couplings of the N-terminal domain of rhodniin to ovomucoid. Identical amino acids in both sequences are labeled by | and similar amino acids are labeled by *

rhodniin	:	12	L	H	R	V	C	G	S	D	G	E	T	Y	S	N	P	C	T	L	N	C	A	K	F	N	G	K	P	E	L	V	L	V	H	D	G	C	47
				*		*	|	|	|	|			|	|		|	|	|					|					*	|		|		|		|		|	|	
ovomucoid	:	20	T	R	P	L	C	G	S	D	N	K	T	Y	G	N	P	C	N	F	C	N	A	V	V	E	S	N	P	T	L	T	L	S	H	F	G	C	55

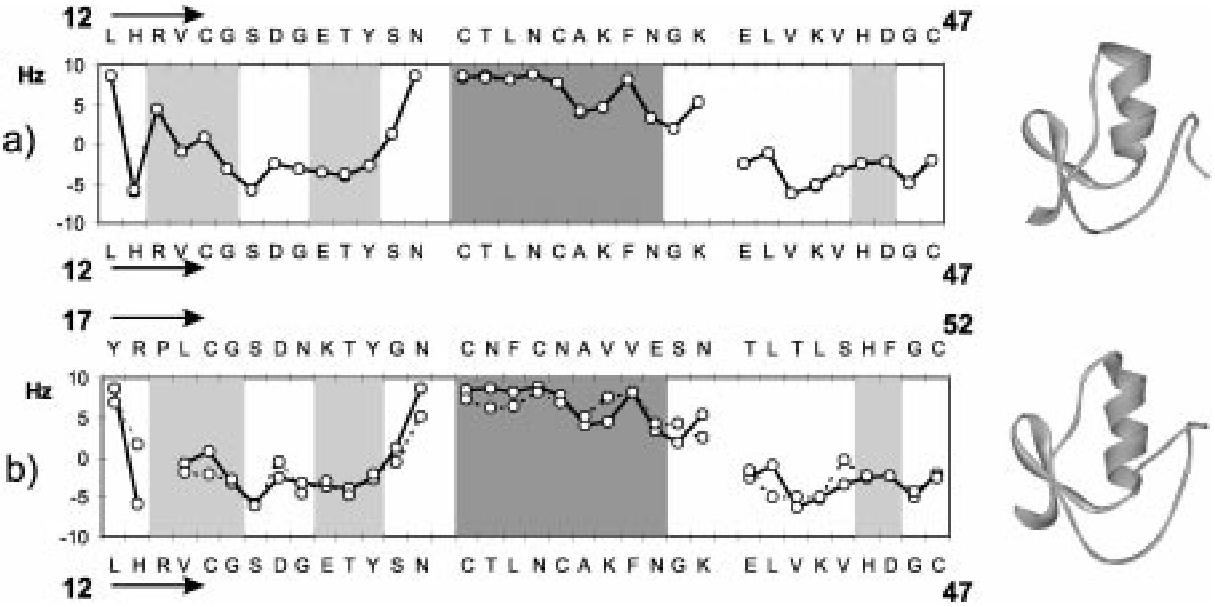


Figure 3. Results for fitting a theoretical set of dipolar couplings for the N-terminal domain of rhodniin to the rhodniin structure (protein **A**) itself (a) and to ovomucoid (protein **B**, a Kazal inhibitor), which is homologous in sequence and structure (b). The black lines indicate the theoretical calculated coupling values, the dotted lines indicate dipolar couplings calculated for the final fit. On the upper x-axis the amino acid number of protein **B**, on the lower x-axis the amino acid number of protein **A** is found. Secondary structure elements are shown by light gray areas (β-sheet) and dark gray areas (α-helix). The three dimensional structures are given in the coordinate system of the tensor (*y*- and *z*-axis are in the paper plane, the *x*-axis is perpendicular to the paper plane).

Table 2. Results of fitting the experimental set of dipolar couplings of the N-terminal domain of rhodniin to rhodniin itself and to an ensemble of eight Kazal inhibitors, some of which are in complex with serine proteases. For 1tbq the data of the N-terminal domain are fitted to the homologous C-terminal domain of rhodniin

Protein name	pdb code	Fit range	*Q*
Rhodniin	–	12–47	0.00
Rhodniin in complex with thrombin (Res.: 2.6 Å)	1tbr	12–47	0.27
Rhodniin in complex with thrombin (Res.: 3.1 Å)	1tbq	65–101	0.27
Ovomucoid	1ovo	20–55	0.30
Human pancreatic secretory inhibitor in complex with trypsin	1cgi	20–55	0.30
Procine pancreatic secretory inhibitor in complex with trypsin	1tgs	19–54	0.32
Human pancreatic secretory trypsin inhibitor	1hpt	20–55	0.36
Pig proteinase inhibitor (Kazal type)	1pce	24–59	0.38
Leech-derived inhibitor with procine in complex with trypsin	1ldt	10–45	0.49

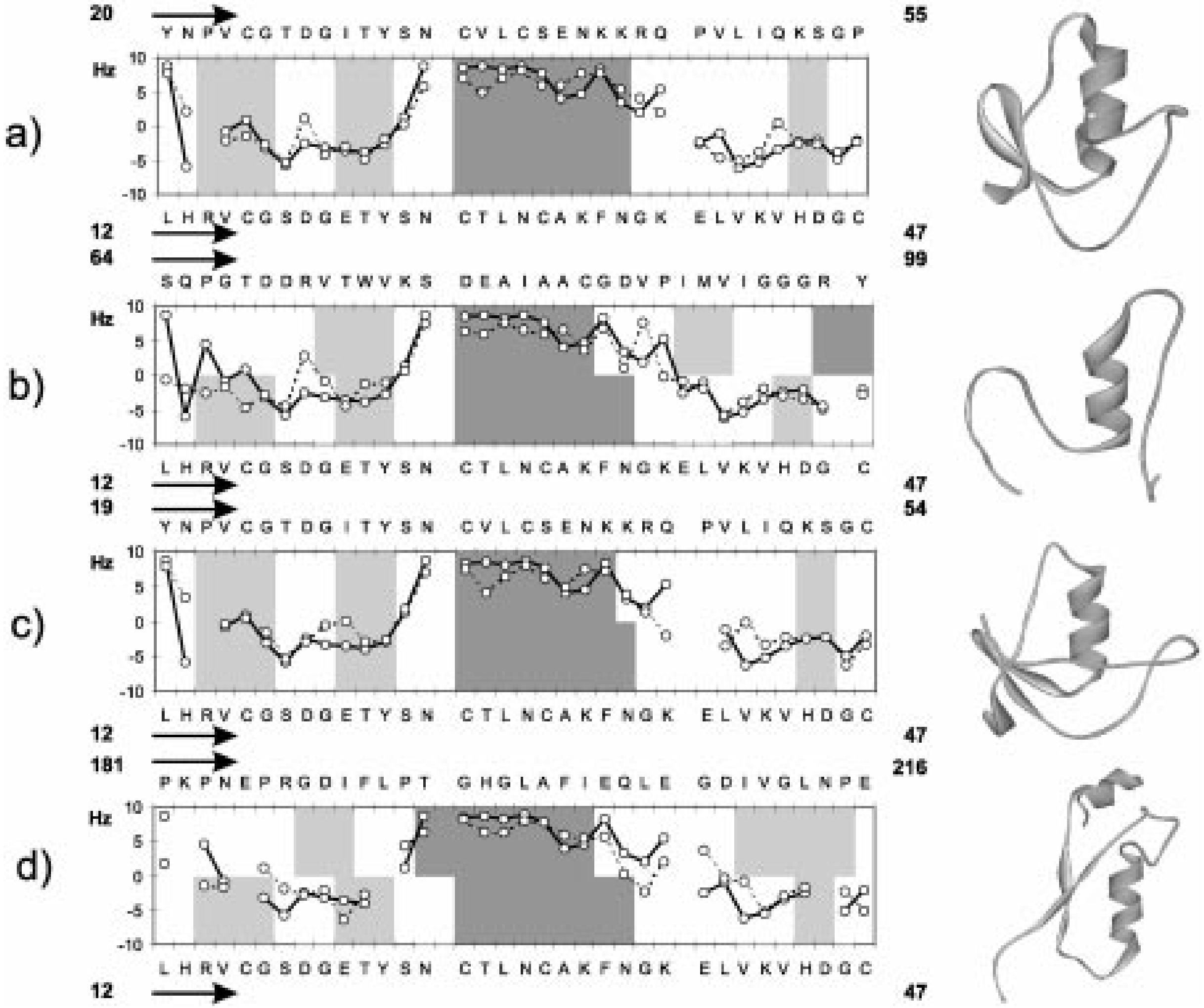


Figure 4. Results for a search in a database of 125 folds extracted from PDB for the theoretical set of dipolar couplings for the N-terminal domain of rhodniin. The black lines indicate the experimental coupling values (protein **A**, rhodniin), dotted lines indicate dipolar couplings calculated for protein **B** from the database. The upper and lower *x*-axes show the amino acid number of protein **B** and protein **A**, respectively. Secondary structure elements are represented similar to Figure 3. The results are ordered by increasing normalized square deviations (*Q*-values). (a) ovomucoid (1ovo_a) with a *Q*-value of 0.30, (b) fragment of an oxidoreductase (6fdr residues 64–99) with a *Q*-value of 0.31, (c) is again a proteinase inhibitor (1tgs_i) with a *Q*-value of 0.32 and (d) is a part of an intramolecular oxidoreductase (4xia_a residues 181–216) with a *Q*-value of 0.35. Subsequent hits have considerably worse matches with *Q*-values above 0.40.

domain is fitted over a β-strand of the protein from the PDB. This observation can be explained by the parallel orientation of N-H^N bond vectors in both secondary structure elements. Dipolar couplings are therefore of the same size in both secondary structure elements which makes a distinction difficult.

Much more significant results with less false positive answers and lower *Q*-values are obtained when secondary structure information from CSI is utilized by two simple rules: first, the alignment of β-strands over α-helices is excluded and second, only residues in well defined secondary structure regions are used for the calculation of *Q*-values. Using these rules, the two Kazal inhibitors of our database are ranked 1st (ovomucoid, 1ovo_a, $Q = 0.30$) and 3rd (porcine pancreatic secretory trypsin inhibitor, 1tgs_i, $Q = 0.32$). Figure 4 presents the first four hits of this search for which structures are displayed in the coordinate system of the tensor. The 2nd result is part of dihydrofolate reductase (6dfr) with a *Q*-value of 0.31 and the fourth result is part of D-xylose isomerase (4xia_a) with a *Q*-value of 0.35. Results (b) and (d) have a

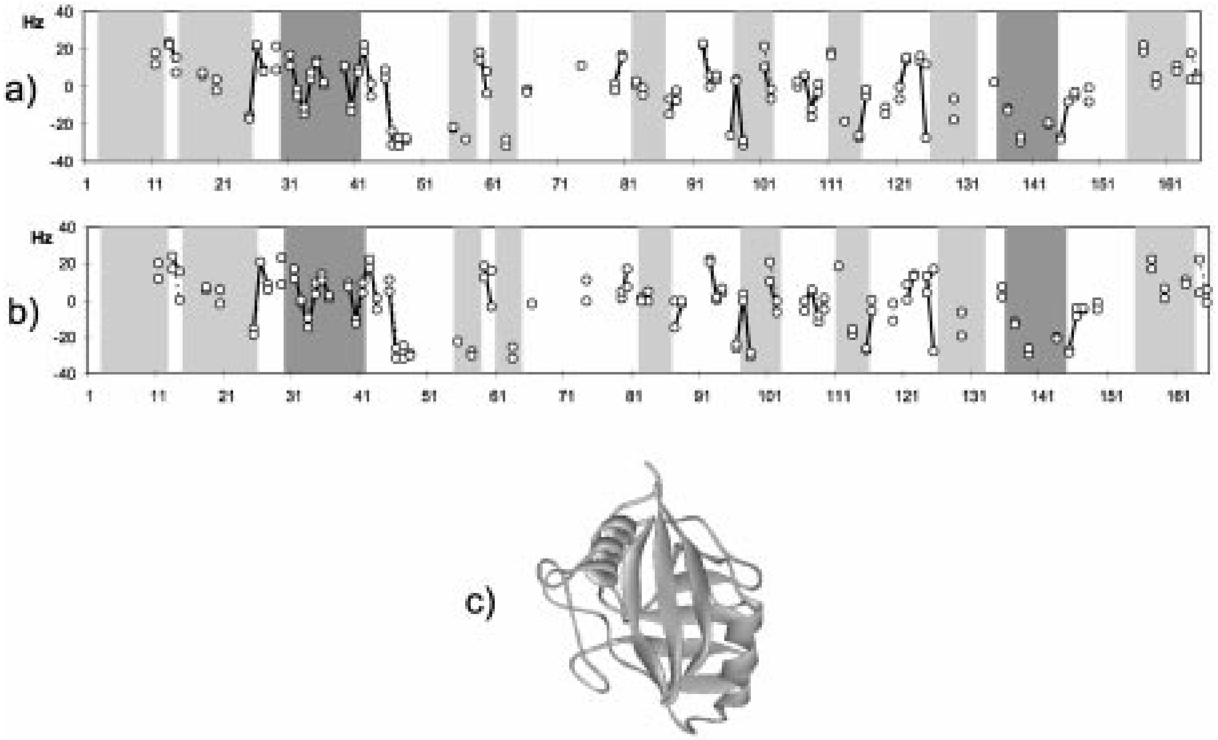


Figure 5. Results of fitting an experimental set of dipolar couplings for cylophilin A (protein **A**) to the NMR structure (a) and to the X-ray structure (b) (protein **B**). Definition of lines and shaded areas is like in Figure 3. *Q*-values are 0.28 and 0.21 for (a) and (b) respectively. The three dimensional structure is given in the coordinate frame of the tensor extracted from the fitting procedure (c).

similarly oriented α-helix and at least one of the three β-strands present in rhodniin with a similar orientation with respect to each other. Since the three β-strands are very short (three residues per strand) and nearly parallel, all dipolar couplings within them are of the same size. Therefore this matches very well with one larger β-strand or an extended region when all N-H^N bonds are parallel (d). Matches (b) and (d) have a primary sequence homology of only 11% and 5%. Thus the program finds 3D homology irrespective of sequence homology.

The result of this first homology search suggests rhodniin to be homologous to other Kazal inhibitors. Therefore a more thorough search for Kazal type inhibitors was performed in the PDB and a subset of such inhibitors was extracted. The *Q*-values of all eight structures range from 0.27 to 0.49 (Table 2).

The second example is cyclophilin for which only 69 fast and easily determinable dipolar couplings were extracted and fitted to the NMR structure (Ottiger et al., 1997) and X-ray (Weber et al., 1982) structures. Results are given in Figure 5 together with the three dimensional structures in the alignment tensor frame of reference. *Q*-values are 0.28 and 0.21 for NMR- and X-ray-structure, respectively. The good agreement of both structures with the experimental data proves that it is not necessary to determine all couplings for fitting. Moreover, the possibility of calculating dipolar couplings for other residues allows to accelerate further interpretation of spectra. While we detect 3D homology to other known cyclophilins, searching in a data bank of 125 folds only finds small parts of the whole sequence, in particular helix-strand-strand motives. It appears that cyclophilin has a rather unique 3D fold.

In the soilbacteria *Herpetosiphon giganteus* many restriction modification systems could be characterized. One of these systems is the *Hgi*CI system of which the C-protein (Controll protein) *Hgi*CIC (expressed with a His$_6$ tag and a C46 → S mutation) of 10 kDa molecular mass is currently under investigation in our laboratory and was used as a test system for *DipoCoup*. A total of 62 ^{1}D$_{NH}$ dipolar couplings could be extracted for the 88 residue protein *Hgi*CIC. The dipolar couplings range from −7.5 to 7.1 Hz. To establish weak alignment we used CHAPSO/DLPC (1:5) bicelles with a total lipid concentration of 5%.

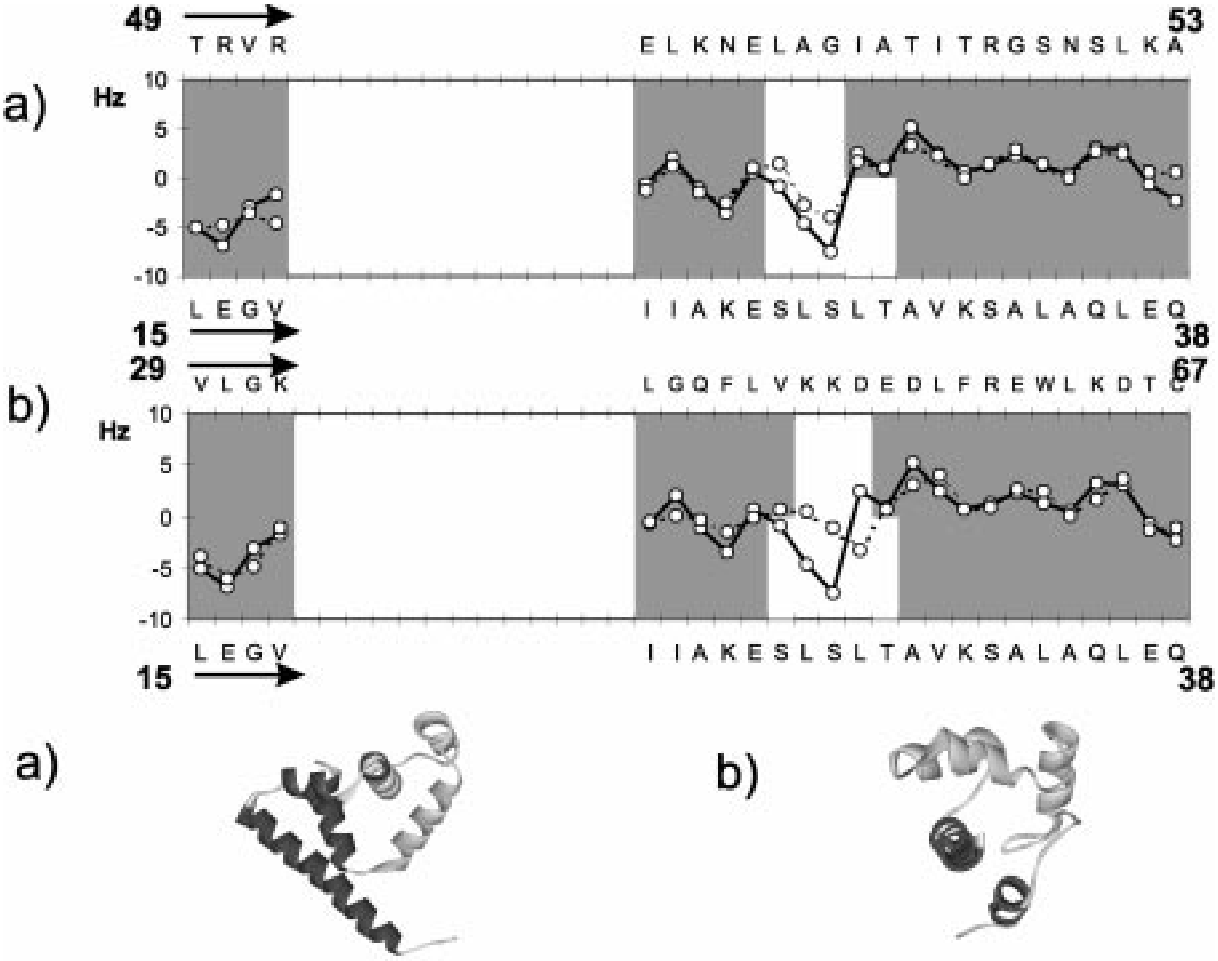


Figure 6. Result of the alignment of residues 15 to 53 of the protein *Hgi*CIC (C46S) to a database especially designed for helix-turn-helix proteins. The results with lowest *Q*-value are trp Repressor (a, $Q = 0.42$) and BAF (b, $Q = 0.43$). Definition of lines and shaded areas is as in Figure 3. The upper and lower x-axes show the amino acid number of protein **B** and protein **A**, respectively. A comparison of the dipolar couplings and corresponding structures in the alignment tensor coordinate system is shown for the two best fits. Light gray parts represent the fitted parts and dark gray parts are not fitted.

Secondary structure alignment indicated the protein might be a typical representative of the helix-turn-helix (HTH) fold family (Brennan and Matthews, 1989; Patto and Sauer, 1992; Harrison, 1999). Therefore we searched for known representatives of the HTH family in the DPInteract (http://arep.med.harvard.edu/dpinteract) database. There are two groups of known HTH proteins. One comprises all α-helical proteins and the other α+β proteins with a HTH motive. With this information we built a database with 19 helix-turn-helix proteins (also at http://krypton.org.chemie.uni-frankfurt.de/~mj/software.html). From the experimental $^1D_{NH}$ dipolar couplings of *Hgi*CIC the alignment tensor was calculated and the alignment search according to Figure 2b was performed with *DipoCoup*, including CSI data. The whole *Hgi*CIC (C46 → S) protein proved to be too large for an alignment with the structures of the database. We therefore partitioned the protein into two overlapping parts. The first part contained the residues 15 to 53 and the second one residues 32 to 71. Both regions can be aligned with parts of proteins in the HTH database. Alignment of the first part (residues 15 to 53) shows good match with the trp repressor (*Q*-value: 0.42) and the cellular factor BAF (*Q*-value of 0.43, Figure 6). The second part (residues 32–71), which includes also the HTH motif, does not match as well as the first stretch of amino acids. We find a best match with the structures of LexA (*Q*-value: 0.64) and with the struc-

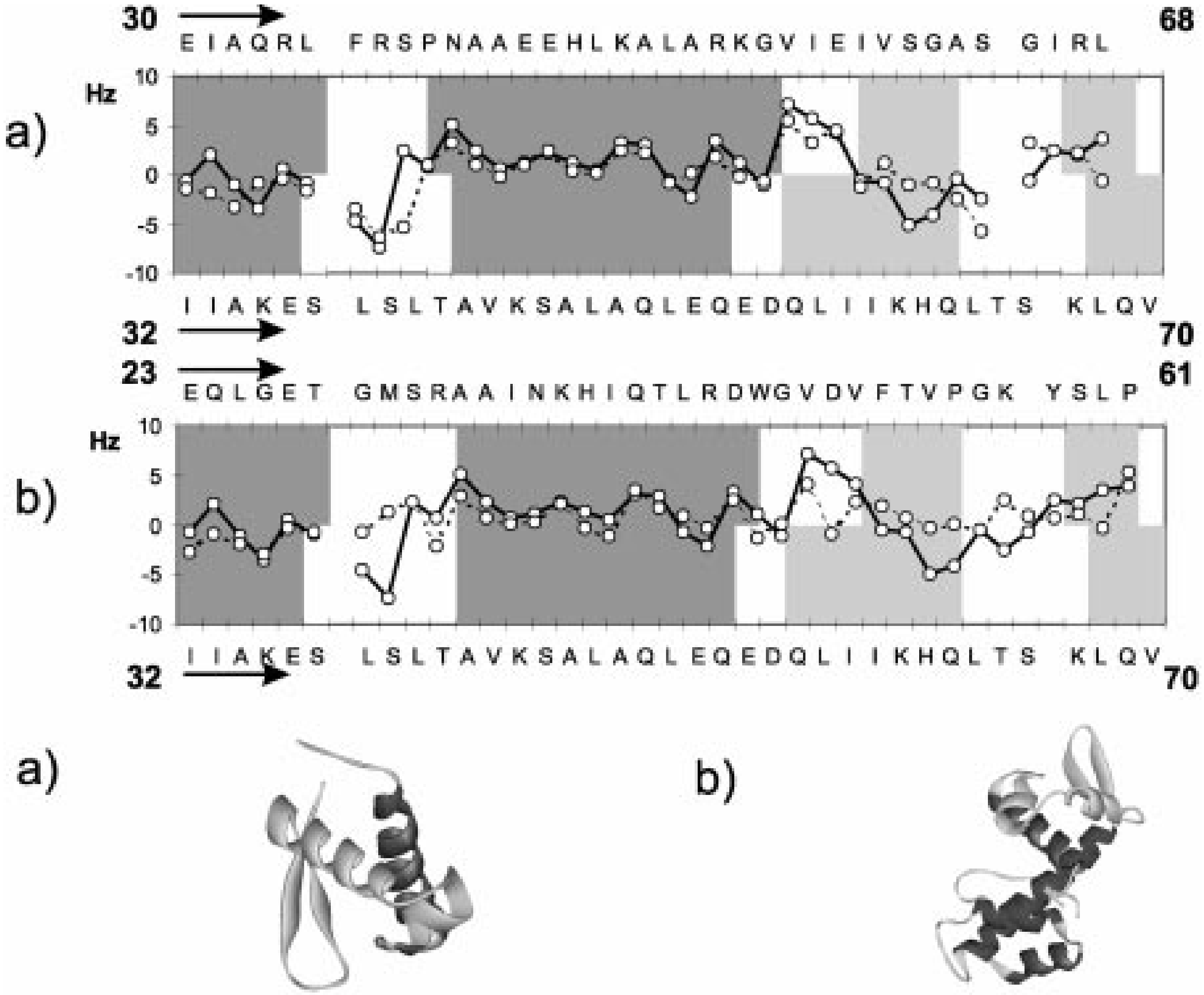


Figure 7. The best match of residues 32 to 71 of the protein *Hgi*CIC (C46S) with HTH protein database is shown. Measured dipolar couplings of the protein are plotted against the calculated dipolar couplings of the two best fitting proteins LexA, and diphtheria toxin repressor (b, $Q = 0.71$). The upper and lower *x*-axes show the amino acid number of protein **B** and protein **A**, respectively. The residue by residue match of the dipolar couplings is much better than the rather high *Q* value would suggest. Light gray parts represent the fitted parts and dark gray parts are not fitted.

ture of the diphtheria toxin repressor (*Q*-value: 0.71). Even though the *Q*-values for the alignment are quite high, the experimental and the calculated dipolar couplings match rather well on a residue by residue basis (Figure 7). A few large deviations can cause large *Q*-values, since *Q* depends quadratically on the deviation of dipolar couplings. Although *HgI*CIC is not very similar to any of the already known HTH-proteins in total, two parts of its structure match known protein folds from which a 3D model of the protein can be derived.

C proteins are also known to bind DNA. The HTH motif in *Hgi*CIC is consistent with the finding that *Hgi*CIC binds DNA as observed by band shift assays. The observed shifts upon DNA titration are most prominent for the amino acids in the helix-turn-helix motif.

Conclusions

We have demonstrated the possibility to use residual dipolar couplings and pseudocontact shifts together with secondary structure information to perform 3D structure homology searches in representative sub-databases of the PDB. We present the program *DipoCoup* which performs this homology search in a fast, accurate and user friendly way. Moreover, *DipoCoup* can be used to perform additional analysis of experimentally determined orientation data or 3D structures of proteins. The program is free for academic use, and can be downloaded from http://krypton.org.chemie.uni-frankfurt.de/∼mj/software.html.

Acknowledgements

This work was supported by the Fonds der Chemischen Industrie. It was further supported by the DFG through (Gr1211/4-4) and the MPG. We thank J.A. Losonczi for making her paper available for us prior to publication. We are grateful to M. Maurer for the NMR structure of rhodniin and for extracting the dipolar couplings of cyclophilin. We are thankful for enlightening discussions with Dr. K. Fiebig. Moreover, we thank A. Zeidler, M. Müller, B. Coligaev and F. Schmäschke for downloading, analyzing and preparing the subset of the PDB. We thank Prof. Dr. G. Fischer and Dr. H. Bang (MPI, Halle) for the preparation of cyclophilin A and Prof. Dr. M. Kröger and Dr. K. Meyer-Rogge (University Gießen) for the preparation of *Hgi*CIC (His$_6$) [C46 → S]. All spectra were recorded at the Large Scale Facility for Biomolecular NMR at the University of Frankfurt. W.P. and J.M. are both supported by a Kekulé stipend of the Fonds der Chemischen Industrie.

References

Annila, A., Aitio, H., Thulin, E. and Drakenberg, T. (1999) *J. Biomol. NMR*, **14**, 223–230.

Bax, A. and Ottiger, M. (1998) *J. Am. Chem. Soc.*, **120**, 12334–12341.

Bax, A. and Tjandra, N. (1997) *J. Biomol. NMR*, **10**, 289–292.

Brennan, R.G. and Matthews, B.W. (1989) *J. Biol. Chem.*, **264**, 1903–1906.

Clore, G.M. and Garrett, D.S. (1999) *J. Am. Chem. Soc.*, **121**, 9008–9012.

Clore, G.M., Gronenborn, A.M. and Bax, A. (1998) *J. Magn. Reson.*, **133**, 216–221.

Clore, G.M., Gronenborn, A.M. and Tjandra, N. (1998) *J. Magn. Reson.*, **131**, 159–162.

Cornilescu, G., Marquardt, J.L., Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 6836–6837.

Delaglio, F., Kontaxis, G. and Bax, A. (2000) *J. Am. Chem. Soc.*, **122**, 2142–2143.

Fischer, D. and Eisenberg, D. (1999) *Curr. Opin. Struct. Biol.*, **9**, 208–211.

Fischer, M.W.F., Losonczi, J.A., Weaver, J.L. and Prestegard, J.H. (1999) *Biochemistry*, **38**, 9013–9022.

Friedrich, T., Kröger, B., Bialojan, S., Lemaire, H.G., Höffken, H.W., Reuschenbach, P., Otte, M. and Dodt, J. (1993) *J. Biol. Chem.*, **268**, 16216–16220.

Ghose, R. and Prestegard, J.H. (1997) *J. Magn. Reson.*, **128**, 138–143.

Harrison, S.C. (1991) *Nature*, **353**, 715–719.

Losonczi, J.A., Andrec, M., Fischer, M.W.F. and Prestegard, J.H. (1999) *J. Magn. Reson.*, **138**, 334–342.

Losonczi, J.A. and Prestegard, J.H. (1998) *J. Biomol. NMR*, **12**, 447–451.

Meiler, J., Blomberg, N., Nilges, M. and Griesinger, C. (2000) *J. Biomol. NMR*, **16**, 245–252.

Moult, J. (1999) *Curr. Opin. Biotechnol.*, **10**, 583–588.

Ojennus, D.D., Mitton-Fry, R.M. and Wuttke, D.S. (1999) *J. Biomol. NMR*, **14**, 175–179.

Ottiger, M., Zerbe, O., Güntert, P. and Wüthrich, K. (1997) *J. Mol. Biol.*, **272**, 64–81.

Patto, C.O. and Saurer, R.T. (1992) *Annu. Rev. Biochem.*, **61**, 1053–1095.

Peti, W. and Griesinger, C. (2000), *J. Am. Chem. Soc.*, **122**, 3975–3976.

Rost, B. and Sander, C. (1994) *Proteins Struct. Funct. Genet.*, **19**, 55–72.

Sali, A. (1998) *Nat. Struct. Biol.*, **5**, 1029–1032.

Saupe, A. (1968) *Angew. Chem. Int. Ed. Engl.*, **7**, 97–102.

Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.

Tjandra, N. and Bax, A. (1997) *Science*, **278**, 1111–1113.

Tjandra, N., Grzesiek, S. and Bax, A. (1996) *J. Am. Chem. Soc.*, **118**, 6264–6272.

Tolman, J.R., Flanagan, J.M., Kennedy, M.A. and Prestegard, J.H. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 9279–9283.

van de Locht, A., Lambda, D., Bauer, M., Hubert, R., Friedrich, T., Kroeger, B., Hoffken, W. and Bode, W. (1995) *EMBO J.*, **14**, 5149–5155.

Wang, H., Eberstadt, M., Olejniczak, E.T., Meadows, R.P. and Fesik, S.W. (1998) *J. Biomol. NMR*, **12**, 443–446.

Weber, E., Papamokos, E., Bode, W., Huber, R., Kato, I. and Laskowski, M. (1982) *J. Mol. Biol.*, **158**, 515–520.

Wishart, D.S., Bigam, C.G., Holm, A., Hodges, R.S. and Sykes, B.D. (1995) *J. Biomol. NMR*, **5**, 67–81.

Wishart, D.S. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1992) *Biochemistry*, **31**, 1647–1651.

Journal of Biomolecular NMR, **16:** 245–252, 2000.
KLUWER/ESCOM
© 2000 *Kluwer Academic Publishers. Printed in the Netherlands.*

A new approach for applying residual dipolar couplings as restraints in structure elucidation

Jens Meiler[a], Niklas Blomberg[b], Michael Nilges[b] & Christian Griesinger[a]
[a]*Universität Frankfurt, Institut für Organische Chemie, Marie-Curie-Strasse 11, D-60439 Frankfurt am Main, Germany*
[b]*EMBL, Meyerhofstrasse 1, D-69012 Heidelberg, Germany*

Received 27 October 1999; Accepted 27 December 1999

Key words: protein structure determination, residual dipolar coupling

Abstract

Residual dipolar couplings are useful global structural restraints. The dipolar couplings define the orientation of a vector with respect to the alignment tensor. Although the size of the alignment tensor can be derived from the distribution of the experimental dipolar couplings, its orientation with respect to the coordinate system of the molecule is unknown at the beginning of structure determination. This causes convergence problems in the simulated annealing process. We therefore propose a protocol that translates dipolar couplings into intervector projection angles, which are independent of the orientation of the alignment tensor with respect to the molecule. These restraints can be used during the whole simulated annealing protocol.

Introduction

Residual dipolar couplings can be used to determine the orientation of an intermolecular vector in the coordinate system of the alignment tensor effected either by paramagnetic moieties of the molecule or by using diluted liquid crystals (Tolman et al., 1995; Hong et al., 1996; Bax and Tjandra, 1997; Tjandra and Bax, 1997; Wang et al., 1998; Cordier et al., 1999; Ojennus et al., 1999). Residual dipolar couplings are used in structure calculations optimizing the orientation of bond vectors with respect to the orientation of the external alignment or susceptibility tensor (Clore et al., 1998b, 1999; Ottiger et al., 1998; Bayer et al., 1999; Fischer et al., 1999; Olejniczak et al., 1999). From the distribution of the dipolar couplings it is possible to extract the size of the protein alignment tensor (Clore et al., 1998a); however, the orientation of the tensor with respect to the molecule that is defined by three Euler angles cannot be determined without structure calculation. In addition, the translation of the dipolar coupling into an orientation in the frame of the alignment tensor is ambiguous due to the fact that there is a continuum of α, β pairs for each dipolar coupling and the mirror reflection symmetry along each of the axes of the alignment tensor. A restraint based on the orientation of a vector in an alignment tensor therefore corresponds to a complicated energy hypersurface ensuing slow convergence properties of the algorithm.

To avoid convergence problems, we propose here to transform the dipolar coupling restraints into purely intramolecular projection restraints that do not require knowledge of the orientation of the molecule with respect to the alignment tensor. We demonstrate with two examples that with this implementation the convergence of the simulated annealing protocol with and without dipolar couplings is almost identical. On a model system it is shown that the precision of the structures is enhanced by the use of the dipolar couplings.

*To whom correspondence should be addressed. E-mail: cigr@org.chemie.uni-frankfurt.de

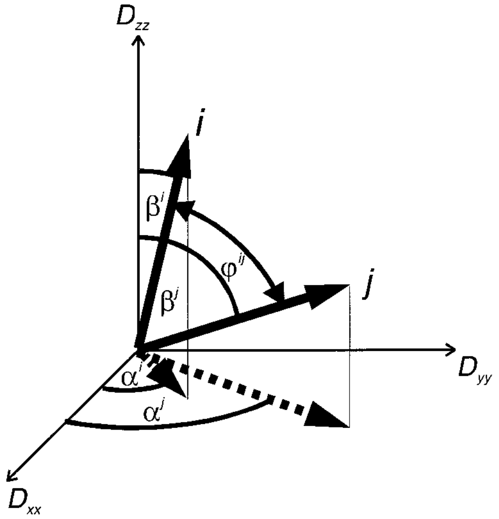


Figure 1. Two bond vectors *i* and *j* in the coordinate system of the alignment tensor *(D_{xx},D_{yy},D_{zz})*. The angles $\alpha^{i/j}$ and $\beta^{i/j}$ determine the projection of the vector onto the *z*-axis and of the *x,y*-component onto the *x*-axis of the alignment tensor, respectively. φ^{ij} determines the angle between the two vectors *i* and *j*.

Theory

In order to derive equations for the projection angles of interatomic vectors as a substitute for the orientation restraints, we introduce the coordinate system of Figure 1, in which the two vectors *i* and *j* are defined in the coordinate system of the alignment tensor with the values D_{xx}, D_{yy} and D_{zz}. Defining an axially symmetric $D_{\|}$ and a rhombic part $D_{\perp}$ of the tensor according to:

$$\begin{aligned} D_{\|} &= \frac{1}{3}\left(D_{zz} - \frac{D_{xx}+D_{yy}}{2}\right) \\ D_{\perp} &= \frac{2}{3}\left(\frac{D_{xx}-D_{yy}}{2}\right) \end{aligned} \quad (1)$$

one obtains:

$$\begin{aligned} D^i\left(\beta^i\alpha^i\right) &= D_{\|}\left(3\cos^2\beta^i - 1\right) \\ &+\frac{3}{2}D_{\perp}\left(\cos 2\alpha^i \sin^2\beta^i\right) \end{aligned} \quad (2)$$

With an experimental set of dipolar couplings the eigenvalues of the tensor D_{xx}, D_{yy}, D_{zz} can be determined from the powder pattern of experimental couplings (Clore et al., 1998a). We now derive equations that allow the use of the dipolar couplings as restraints without the need for defining the orientation of the alignment tensor. This is done by calculation of the projection angle φ^{ij} between all pairs of internuclear vectors *i* and *j* for which dipolar couplings have been measured.

$$\cos\varphi^{ij} = \begin{pmatrix} \cos\alpha^i \sin\beta^i \\ \cos\alpha^i \sin\beta^i \\ \cos\beta^i \end{pmatrix}^T \begin{pmatrix} \cos\alpha^i \sin\beta^j \\ \cos\alpha^j \sin\beta^j \\ \cos\beta^j \end{pmatrix} \quad (3)$$

Applying Equation 2 to Equation 3 one can eliminate two angles β^i, β^j and arrive at:

$$\begin{aligned} &\cos\varphi^{ij} = \\ &\sqrt{\frac{2(3D^i - 2D_{zz} + D_{xx} + D_{yy})}{3(\frac{3}{2}(D_{xx} - D_{yy})\cos\alpha^i - 2D_{zz} + D_{xx} + D_{yy})}} \\ &\sqrt{\frac{2(3D^j - 2D_{zz} + D_{xx} + D_{yy})}{3(\frac{3}{2}(D_{xx} - D_{yy})\cos\alpha^j - 2D_{zz} + D_{xx} + D_{yy})}} \\ &\cos(\alpha_2 \pm \alpha_1) \qquad (4) \\ &\pm\sqrt{1 - \frac{2(3D^i - 2D_{zz} + D_{xx} + D_{yy})}{3(\frac{3}{2}(D_{xx} - D_{yy})\cos\alpha^i - 2D_{zz} + D_{xx} + D_{yy})}} \\ &\sqrt{1 - \frac{2(3D^j - 2D_{zz} + D_{xx} + D_{yy})}{3(\frac{3}{2}(D_{xx} - D_{yy})\cos\alpha^j - 2D_{zz} + D_{xx} + D_{yy})}} \end{aligned}$$

In addition, the possible range for angle α^i is sometimes reduced by the measured coupling values:

$$if \ \left|\frac{6D^i + 2D_{zz} - D_{xx} - D_{yy}}{3\left(D_{xx} - D_{yy}\right)}\right| \leq 1$$

$$\left\{\begin{array}{l} \alpha^i_{\text{min}} = \frac{1}{2}\arccos\left(\frac{6D^i+2D_{zz}-D_{xx}-D_{yy}}{3(D_{xx}-D_{yy})}\right) \\ \alpha^i_{\text{max}} = \pi - \frac{1}{2}\arccos\left(\frac{6D^i+2D_{zz}-D_{xx}-D_{yy}}{3(D_{xx}-D_{yy}}\right) \end{array}\right\} \quad (5)$$

$$else \left\{\begin{array}{c} \alpha^i_{\text{min}} = 0; \\ \alpha^i_{\text{max}} = \pi \end{array}\right\}$$

Depending on the size of the measured coupling values, the angle φ^{ij} is no longer allowed to vary in the whole interval from 0 to π. Equation 5 states that the extreme values of φ^{ij} will always be found at the extreme values for α^i and α^j. The allowed range for φ^{ij} is in addition always symmetric about $\varphi^{ij} = \pi/2$ (Figure 2). Two general possibilities are found:

one allowed range: $\varphi^{ij} \in [\epsilon_1, \pi - \epsilon_1]$
two allowed ranges: $\varphi^{ij} \in [\epsilon_1, \pi/2 - \epsilon_2]$
or $\varphi^{ij} \in [\pi/2 + \epsilon_2, \pi - \epsilon_1]$

with $\epsilon_{1,2} \in [0,\pi/2]$. The symmetry of the allowed φ^{ij} ranges is directly related to the geometric symmetries of the dipolar couplings according to Equation 2. Figure 2 shows a range for one angle φ^{ij} that has been calculated in this way.

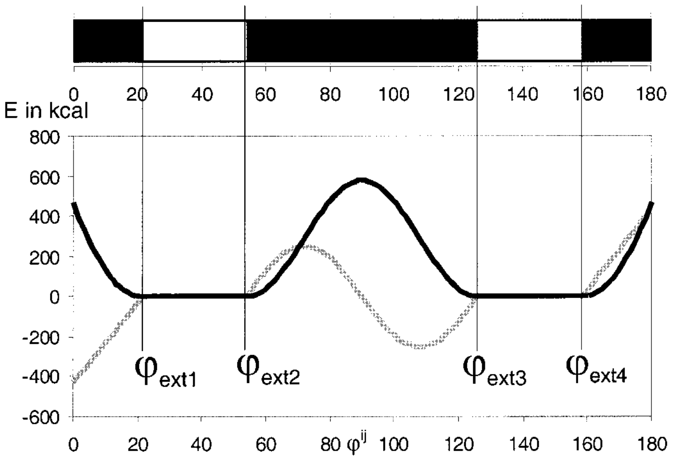


Figure 2. Potential employed to confine φ^{ij} within the allowed range (white) and exclude it from the forbidden range (black). A flat bottom potential is used for the allowed region, a parabolic potential for the margins close to *0* and π and a $\cos^2 \Delta\varphi^{ij}$ function for the inner forbidden part. The energy term used is given by the black line and its derivative (negative force) by the gray line.

Using this approach one can calculate from *n* dipolar couplings *n(n* − 1*)/2* ranges for angles φ^{ij}. These ranges are now free from information about tensor orientation. The translation of the dipolar couplings into intervector projection angles, i.e., the scalar product is the simplest pairwise relation between two vectors and any further more complicated vector relation yielding a scalar value can be derived from it. At the same time, the scalar product conserves all information and more complicated intervector relations are not required. This can be inferred from the following argument: If for a set of N vectors, the length of the vectors and all the mutual projection angles are known, the orientations of the vectors are uniquely defined. N vectors can be represented by N+1 atoms, where the N vectors all start from the (N+1st) atom. This set of atoms has 3(N+1) degrees of freedom. Since we know the length of the vectors (N+1) degrees of freedom and the projection angles: N(N+1)/2 degrees of freedom and subtracting 3 degrees of freedom each for rotation and translation, we find: 3(N+1)−(N+1)N/2−6−(N+1) ≤ 0 for all N. Thus, the projection angles would even overdetermine the coordinates of the vectors if they were exactly known.

To use these angle ranges φ^{ij} for structure determination, a new restraint was introduced in the X-PLOR program (Brünger, 1992). The energy function used is displayed in Figure 2 and given in Equation 6:

$$\begin{aligned}
E^{ij}_{0\rightarrow\varphi_{ext1}} &= k_1\left(\varphi^{ij} - \varphi^{ij}_{ext1}\right)^2\\
E^{ij}_{\varphi_{ext1}\rightarrow\varphi_{ext2}} &= 0\\
E^{ij}_{\varphi_{ext2}\rightarrow\varphi_{ext3}} &= k_2\cos^2\left(\pi\left(\frac{\varphi^{ij}-\varphi^{ij}_{ext2}}{\varphi^{ij}_{ext3}-\varphi^{ij}_{ext2}} - \frac{1}{2}\right)\right) \qquad (6)\\
E^{ij}_{\varphi_{ext3}\rightarrow\varphi_{ext4}} &= 0\\
E^{ij}_{\varphi_{ext4}\rightarrow 180^\circ} &= k_1\left(\varphi^{ij} - \varphi^{ij}_{ext4}\right)^2
\end{aligned}$$

k_1 and k_2 are the energy constants used in these implementations. Note that k_2 gives directly the height of the barriers between the two allowed φ^{ij} ranges while k_1 scales the square of the deviation from the extreme values. The two energy constants can be scaled separately during the simulated annealing protocol, which turned out to be essential as discussed below.

Results and discussion

The protocol has been applied to the protein Rhodniin, which has 103 amino acids and contains two similar folded domains of 45 amino acids and a flexible linker of 10 amino acids. A set of dipolar couplings between amide nitrogen and amide hydrogen was calculated from the known NMR structure (M. Maurer and C. Griesinger, in preparation) of the protein assuming a specific orientation of the alignment tensor. The eigenvalues were set to be $D_{zz} = 20.0$ Hz, $D_{yy} = -17.5$ Hz and $D_{xx} = -2.5$ Hz, amounting to a rhombicity of 0.5. This set of dipolar couplings is used as an *'experimental'* test set of data and the Rhodniin structure they were calculated from is called *'target structure'*.

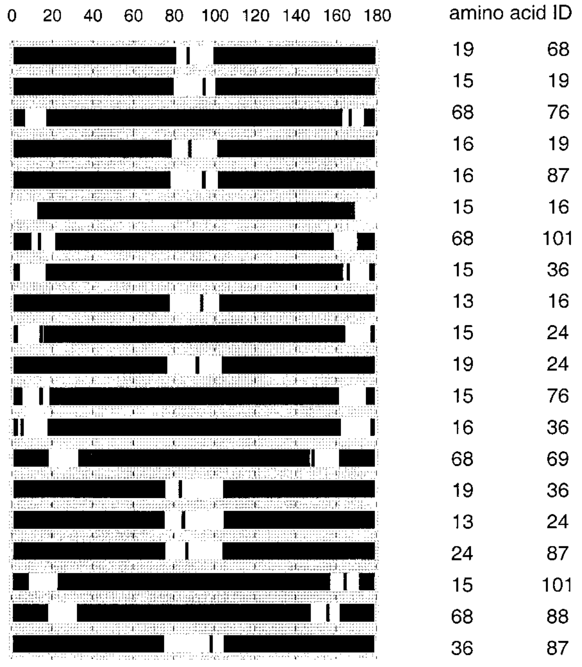


Figure 3. Allowed and forbidden ranges for the angles φ^{ij} between those N-H^N vectors that are most restricted by the dipolar couplings derived from one alignment tensor. The black bars are the φ^{ij} values of the 'target' structure of Rhodniin.

Table 1. Number of restraints that exclude a defined percentage of the possible φ^{ij} ranges

Percentage of φ^{ij} range excluded (%)	Number of restraints	Percentage of restraints (%)
0–10	802	26.18
10–20	781	22.75
20–30	736	18.92
30–40	624	14.25
40–50	494	8.51
50–60	304	5.41
60–70	213	2.86
70–80	163	0.91
80–90	67	0.38
90–100	2	0.00
Sum	4186	100.00

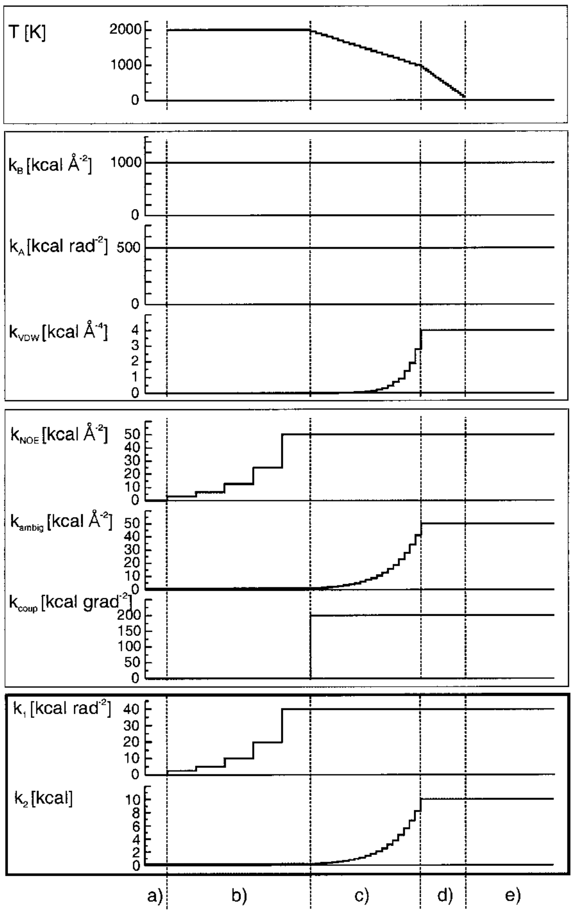


Figure 4. Time scale, temperature and size of the energy constants during the simulated annealing as used for the example Rhodniin in X-PLOR (Brünger, 1992). (a) Initial energy minimization of 50 steps is performed. (b) High temperature phase, lasting for 32.5 ps. The energy constants used for dipolar couplings k_1 are switched on together with the unambiguous NOEs. (c) First cooling phase, lasting for 25.0 ps: k_2 is switched on together with the rest of the NOE. Temperature is decreased from 2000 K to 1000 K. (d) Second cooling phase, lasting 10.0 ps. Temperature is decreased from 1000 K to 100 K. (e) Final energy minimization of 200 steps is performed using all restraints and their final force constants.

All NOEs used in the following calculations were put in three groups: strong NOEs < 2.5 Å, medium NOEs < 3.5 Å and weak NOEs < 5.0 Å. A flat bottom potential was used, being zero for all values smaller than the mentioned distances and quadratically increasing for larger distances.

The eigenvalues of the alignment tensor were back calculated from the powder pattern of the dipolar couplings (Clore et al., 1998a) yielding $D_{zz} = 19.6$ Hz, $D_{yy} = -17.4$ Hz and $D_{xx} = -2.2$ Hz. This tensor has a slight deviation from predefined values, caused by the fact that values are extracted from a powder pattern. From the dipolar couplings more than 4000 possible φ^{ij}–angle ranges were calculated using these eigenvalues. Figure 3 shows the 20 most restricted ranges. Table 1 reports the number of φ^{ij}–angle restraints that exclude a given percentage of the possible

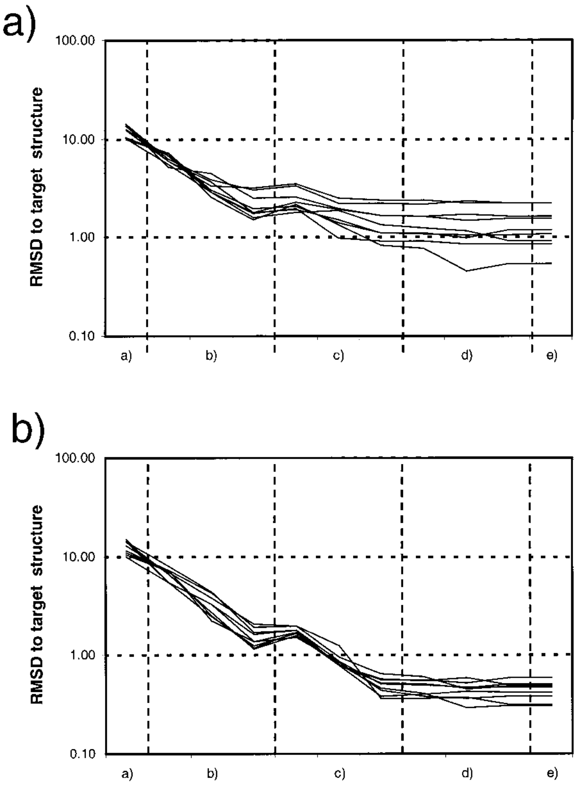


Figure 5. Logarithmic representation of rmsd to target structure during simulated annealing without (a) and with (b) the use of dipolar couplings for the N-domain of Rhodniin as a function of the sections of the SA explained in Figure 4. Best 10 out of an ensemble of 100 calculated structures. The precision and accuracy of the structure with dipolar couplings (both 0.5 Å) are higher than without (both 1.2 Å).

range from [0,π]. Since only a small part of the restraints contains most of the structural information, we find that only those 30% to 50% (depending on the data) most restricted ranges are necessary in the calculations.

k_1 and k_2 were varied between 0 kcal/mol and 200 kcal/mol and gave best results with $k_1 = 40$ kcal/mol rad^{-2} and $k_2 = 10$ kcal/mol. Figure 4 shows the scaling of the force constants used during the simulated annealing X-PLOR protocol. In order to keep the energy surface simple, only k_1 is ramped up, together with the force constant for the unambiguous NOEs in the high temperature phase. k_2 is ramped up in the first cooling phase, together with the force constant of the ambiguous NOEs. k_2 defines the height of the barrier between the two possible ranges φ^{ij}. It turns out to be essential to choose the right period in the simulated annealing protocol, when k_2 is ramped up. We find that the convergence deteriorates when k_2 is switched on too early, namely already in the high temperature phase. By contrast, the number of violated angle restraints increases when k_2 is ramped up too late, namely during the second cooling phase.

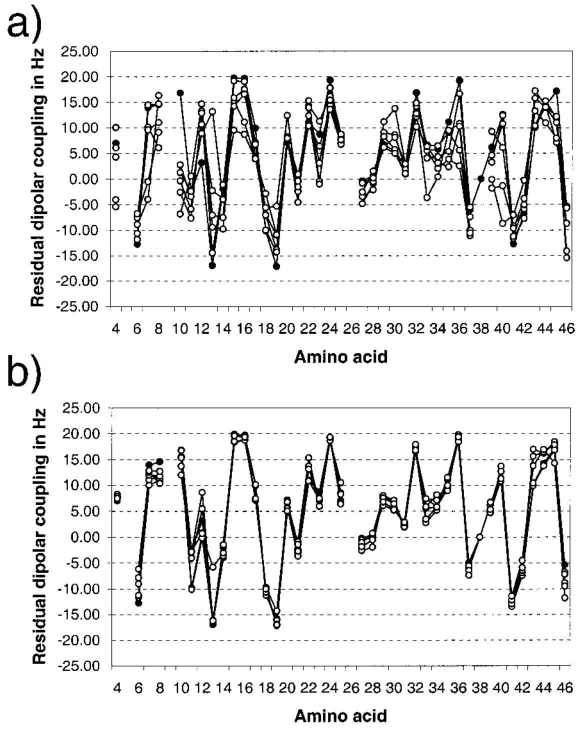


Figure 6. Couplings recalculated for the resulting structures without (a) and with (b) the restriction of the angle ranges. The normalized standard deviation (Q) decreases from 0.66 without (a) the use of the angle restraints to 0.19 with (b) the use of the angle restraints.

With the scaling of force constants as depicted in Figure 4, two ensembles of 100 structures were calculated with and without all angle restraints, all experimentally determined NOEs and J-coupling restraints. For the convergence, we focused on the N-terminal domain of Rhodniin. Figure 5 shows the rmsd of the heavy atoms in the N-terminal domain of the best 10 out of the 100 calculated structures to the target structure during the simulated annealing protocol with and without the angle restraints, respectively. The rmsd decreases from 1.2 Å to 0.5 Å by introducing the information derived from dipolar couplings, which is in line with previous results obtained for ubiquitin (Bax and Tjandra, 1997). In the best 10 of these structures no NOEs and between zero and three of the angle ranges are violated. Table 2 reports the energy distribution of the 100 structures. The introduction of the φ^{ij} restraints increases the mean energy of the ensemble. However, only 5 out of the 100 struc-

Table 2. Energy distribution of an ensemble of 100 structures after the simulated annealing protocol without angle restraints, with restraints and with restraints for two different tensors

Energy range (kcal)	Number of structures		
	Without dipolar restraints	With one set of dipolar restraints	With two sets of dipolar restraints
350–450	47	31	12
450–550	43	38	30
550–650	6	14	28
650–750	3	6	13
750–850	0	2	9
850–950	0	3	2
> 950	1	$5 + 1$	$5 + 1$

With the use of the dipolar restraints five structures do not converge at all. One converges, however, at 990 kcal/mol and 1100 kcal/mol, respectively.

Table 3. Rmsd to target structure with and without the use of angle restraints

Percentage of NOEs used (%)	Rmsd to target structure (in Å)	
	With dipolar restraints	Without dipolar restraints
20	5.39	5.75
40	4.20	4.40
60	2.90	4.13
80	1.78	3.50
100	1.73	2.26

20%, 40%, 60%, 80% and 100% of the NOEs including at least one H^N-atom are used. All other NOEs, including ambiguous NOEs, are excluded.

tures do not fold to a meaningful NMR structure and have a very high energy (beyond 5000 kcal/mol) but 83 have an energy lower than 650 kcal/mol. Without the use of the couplings 96 structures have an energy below 650 kcal/mol. The agreement between the dipolar couplings and the ‘experimental’ restraints is very good, as is reported in Figure 6 for 5 out of the 100 structures with the lowest energy. After the simulated annealing protocol, the orientation of the tensor was optimized (Losonczi et al., 1999) using the program DipoCoup (J. Meiler, W. Peti and C. Griesinger, submitted; Meiler, 1999). All deviations are given as dipolar *Q*-factor (normalized standard deviation, Cornilescu et al., 1998). This factor is bigger by $\sqrt{2}$ than the *R*-factor (Clore and Garrett, 1999). The dipolar *Q*-factor is 0.19 with the use of the restraints but 0.66 without angle restraints. Small deviations can be explained with the slightly incorrect alignment tensor eigenvalues used for calculating the restraints as well as with the violation of angle ranges in some special cases. The orientation of the tensor is reproduced within ±5°.

During a second type of experiment only part of the NOEs was used to test the possibility of replacing NOE information by dipolar couplings. All NOEs containing at least one amide hydrogen were selected first. Out of this subgroup of NOEs a varying percentage of 20% to 100% was randomly selected to be used in the calculation. Amide hydrogens were chosen, since their NOEs are the first and easiest to obtain during evaluation of the spectra and are the only NOEs available in deuterated proteins. Table 3 shows the rmsd values to the target structure with respect to the amount of NOEs used for the N-domain of the protein. The structures can be determined more accurately using the angle restraints. The largest improvement is observed when 60% to 80% of the amide-to-all-proton NOEs are used. Differences are smaller if more or less than 60% to 80% of the NOEs are used. This shows that, based on a minimal amount of NOEs that define the fold, dipolar couplings improve the structure. If the number of NOEs exceeds a certain number, they define the structure so well that dipolar couplings do not improve the accuracy of the structure considerably any more. If the number of NOEs is too low the fold is no longer determined and the dipolar couplings cannot remedy this situation.

A third calculation was performed to test the ability of this implementation to determine the global structure of Rhodniin. Therefore 50 structures were generated, and the orientation of the two Rhodniin domains with respect to each other was investigated. For this calculation the 500 most restricted interdomain φ^{ij} ranges as well as the 500 most restricted intradomain φ^{ij} ranges for the N-domain and the C-domain, respectively, are used. Figure 7 shows the five resulting structures with lowest energy. Three out of these five structures show the same relative orientation of the domains identical to the target structure. In the other two cases one domain is rotated by 180° around the x-axis of the alignment tensor. Due to the symmetry of the tensor, this is a valid solution. In all cases not more than five φ^{ij} ranges are slightly outside of the allowed range. Recalculating dipolar couplings from all five structures with an optimized tensor orientation shows that all five structures fulfill the dipolar coupling values.

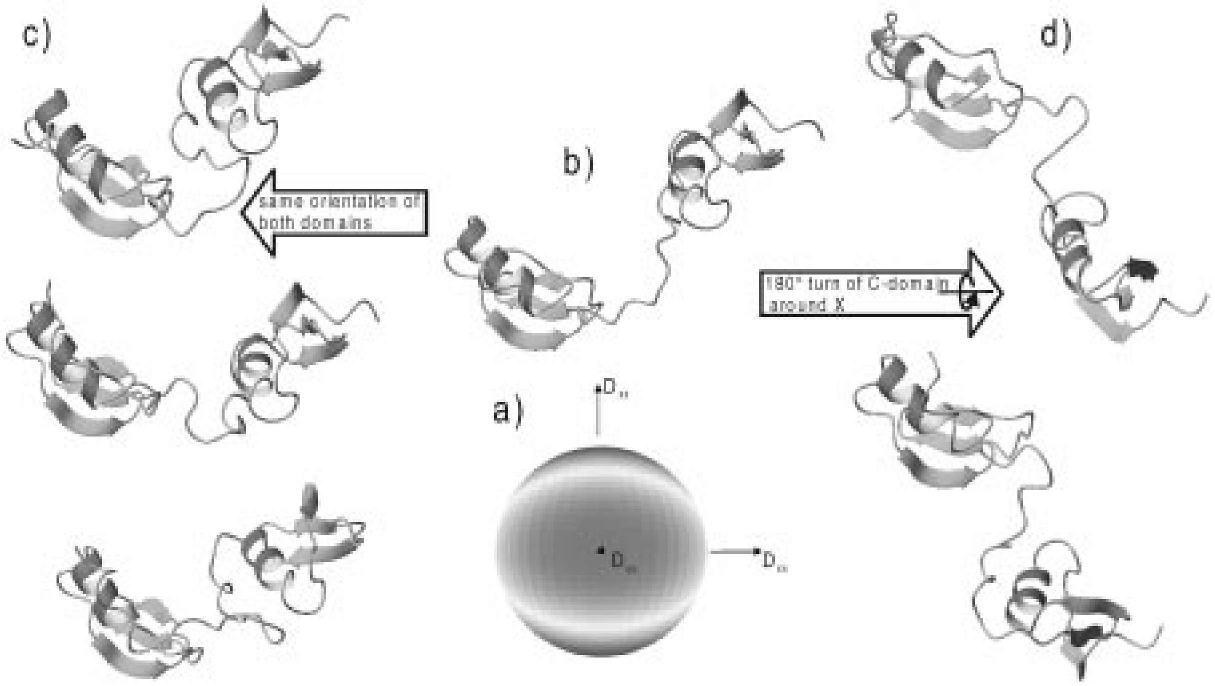

c)
same orientation of both domains
b)
180° turn of C-domain around X
d)
a)
D_{yy}
D_{zz}
D_{xx}

Figure 7. The relative orientation of the N- and C-domain of Rhodniin with respect to each other. The tensor (a) and the target structure (b) in the coordinate system of the tensor are shown. (c) and (d) display the best 5 out of 50 structures from the simulated annealing protocol of Figure 4. (c) The three structures with the same orientation of the domains as in the target structure are shown. (d) Two structures are obtained in which one domain is rotated by 180° about the D_{xx} axis with respect to the target structure.

To check the ability of the program to use two sets of dipolar couplings obtained with two different alignment tensors (Ramirez and Bax, 1998), a second tensor was defined, $D_{zz} = 10.0$ Hz, $D_{yy} = -6.5$ Hz and $D_{xx} = -3.5$ Hz, amounting to a rhombicity of 0.2. The orientation of the tensor with respect to the original PDB file of the target structure was changed from $(\alpha,\beta,\gamma) = (60°,135°,0°)$ for the first alignment tensor to $(\alpha,\beta,\gamma) = (120°,90°,133°)$ for the second alignment tensor. A second set of dipolar couplings was calculated and from these values derived φ^{ij} ranges were used together with the ranges depicted from the first set of dipolar couplings to recalculate the N-domain of Rhodniin.

With the use of both sets of angle restraints and the same protocol as used before, still 70 out of the 100 structures calculated have an energy lower than 650 kcal/mol. For the 10 structures with lowest energy no NOE is violated and only 0 to 10 of the angle restraints are slightly violated. The dipolar *Q*-factor is 0.16 for the first set of dipolar couplings and 0.14 for the second set of dipolar couplings. The rmsd of the structures to the target structure decreases also to be 0.3 Å.

In order to test the protocol with an experimental set of dipolar couplings, it was applied on a second protein. The PH domain of the protein Unc89 from *C. elegans* (residues 341 to 458) was expressed in *E. coli* and purified as described elsewhere (Blomberg et al., 2000). The domain has the native sequence with a single methionine residue added to the N-terminus. Deuterated dimethyl sulphoxide (DMSO; 15%) was added to the samples to suppress aggregation and precipitation. All NMR experiments were recorded at 303 K on a Bruker DRX600 or DRX500 spectrometer equipped with pulse field gradient triple resonance probes. The resonance-frequency assignment of the UNC-89 PH domain is deposited in the BioMagResBank (http://www.bmrb.wisc.edu, accession no. 4373). All experimental data as well as the structure determination of the protein is described and discussed elsewhere (Blomberg et al., 2000).

Residual dipolar couplings were measured for ^{1}J-NH-couplings from standard ^{15}N-HSQC experiments without decoupling in the indirect dimension. Alignment of the PH domain was achieved using DMPC/DLPC/SDS lipid bicelles (ratio 3.2:1:0.1; 5% w/v total lipid) (Losonczi and Prestegard, 1998; Ottiger et al., 1998). Bicelle formation, evidenced by the observed splitting of the D_2O signal, was achieved at 298 K and the measurements of dipolar couplings were performed at 303 K. A reference spectrum was recorded with the sample in the isotropic phase at 295 K. Residues with severe overlap in the ^{15}N-HSQC spectra and those showing evidence of large mobility (elevated ^{15}N T_2/T_1 ratio or short T_2's) were excluded, yielding a total of 41 long range angular restraints.

The magnitude of the tensor components was estimated from a histogram (Clore et al., 1998a) to be $D_{zz} = 8.32$ Hz, $D_{yy} = -6.34$ Hz and $D_{xx} = -1.98$ Hz, amounting to a rhombicity of 0.35. From this data 861 φ^{ij} ranges were derived and used in the simulated annealing process. It was necessary to increase the force constants to be $k_1 = 80$ kcal rad^{-2} and $k_2 = 20$ kcal to get no more than five angle restraints slightly violated. Two ensembles with 100 structures each were calculated, one without and the second with the use of the angle restraints. The dipolar Q-factor of the experimental coupling values to those calculated from the final structures is 0.20 with the use of the restraints, but 0.79 without angle restraints. As already obtained in the calculations of Rhodniin, the energies of the structures increase about 10% using angle restraints. The converged amount of structures with and without restraints is similar to the result for Rhodniin. The rmsd for the backbone of the protein decreases from 2.70 Å without the use of the angle restraints to 2.15 Å with the use of the restraints.

Conclusions

The use of dipolar couplings as φ^{ij} restraints allows to include these experimental data from the start of the simulated annealing protocol. Moreover, with this implementation the dipolar couplings can for the first time be directly translated into intramolecular restraints without the need to orient the alignment tensor during the simulated annealing protocol. The convergence of the simulated annealing protocol is almost unchanged as compared to calculations without dipolar couplings. The calculation time increases only slightly (below 5% for these examples) by taking dipolar couplings into account.

Acknowledgements

This work was supported by the 'Fond der chemischen Industrie', the DFG and the MPG. J.M. thanks the 'Fond der chemischen Industrie' for a Kekulé-Stipendium. We thank M. Maurer, Frankfurt, for making available all the Rhodniin data. We thank W. van Gunsteren, ETH Zürich, and S. Glaser, TU München, for enlightening discussions.

References

Bax, A. and Tjandra, N. (1997) *J. Biomol. NMR*, **10**, 289–292.

Bayer, P., Varani, L. and Varani, G. (1999) *J. Biomol. NMR*, **14**, 149–155.

Blomberg, N., Sattler, M. and Nilges, M. (1999) *J. Biomol. NMR*, **15**, 269–270.

Blomberg, N., Sattler, M., Baraldi, E., Saraste, M. and Nilges, M. (2000) submitted.

Bruenger, A.T. (1992) *X-PLOR, A System for X-Ray Crystallography and NMR*, Yale University Press, New Haven, CT.

Clore, G.M., Gronenborn, A.M. and Bax, A. (1998a) *J. Magn. Reson.*, **133**, 216–221.

Clore, G.M., Gronenborn, A.M. and Tjandra, N. (1998b) *J. Magn. Reson.*, **131**, 159–162.

Clore, G.M. and Garrett, D.S. (1999) *J. Am. Chem. Soc.*, **121**, 9008–9012.

Clore, G.M., Starich, M.R., Bewley, C.A., Cai, M. and Kuszewski, J. (1999) *J. Am. Chem. Soc.*, **121**, 6513–6514.

Cordier, F., Dingley, A.J. and Grzesiek, S. (1999) *J. Biomol. NMR*, **13**, 175–180.

Cornilescu, G., Marquardt, J.L., Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 6836–6837.

Fischer, M.W.F., Losonczi, J.A., Weaver, J.L. and Prestegard, J.H. (1999) *Biochemistry*, **38**, 9013–9022.

Hong, M., Schmidt-Rohr, K. and Zimmermann, H. (1996) *Biochemistry*, **35**, 8335–8341.

Losonczi, J.A., Andrec, M., Fischer, M.W.F. and Prestegard, J.H. (1999) *J. Magn. Reson.*, **138**, 334–342.

Losonczi, J.A. and Prestegard, J.H. (1998) *J. Biomol. NMR*, **12**, 447–451.

Meiler, J. (1999) http://krypton.org.chemie.uni-frankfurt.de/~mj/

Ojennus, D.D., Mitton-Fry, R.M. and Wuttke, D.S. (1999) *J. Biomol. NMR*, **14**, 175–179.

Olejniczak, E.T., Meadows, R.P., Wang, H., Cai, M. and Fesik, S.W. (1999) *J. Am. Chem. Soc.*, **121**, 9249–9250.

Ottiger, M., Delaglio, F., Marquardt, J.L., Tjandra, N. and Bax, A. (1998) *J. Magn. Reson.*, **134**, 365–369.

Ramirez, B. and Bax, A. (1998) *J. Am. Chem. Soc.*, **129**, 9106–9107.

Tjandra, N. and Bax, A. (1997) *Science*, **278**, 1111–1113.

Tolman, J.R., Flanagan, J.M., Kennedy, M.A. and Prestegard, J.H. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 9279–9283.

Wang, H., Eberstadt, M., Olejniczak, E.T., Meadows, R.P. and Fesik, S.W. (1998) *J. Biomol. NMR*, **12**, 443–446.

Model-Free Approach to the Dynamic Interpretation of Residual Dipolar Couplings in Globular Proteins

Jens Meiler,[†] Jeanine J. Prompers,[‡] Wolfgang Peti,[†] Christian Griesinger,*[,†,§] and Rafael Brüschweiler*[,‡]

Contribution from the Institut für Organische Chemie, Universität Frankfurt, Marie-Curie-Strasse 11, D-60439 Frankfurt am Main, Germany, Gustaf H. Carlson School of Chemistry and Biochemistry, Clark University, 950 Main Street, Worcester, Massachusetts 01610-1477, and Max-Planck Institute for Biophysical Chemistry, Am Fassberg 11, D-37077 Göttingen, Germany

Received January 2, 2001. Revised Manuscript Received March 28, 2001

Abstract: The effects of internal motions on residual dipolar NMR couplings of proteins partially aligned in a liquid-crystalline environment are analyzed using a 10 ns molecular dynamics (MD) computer simulation of ubiquitin. For a set of alignment tensors with different orientations and rhombicities, MD-averaged dipolar couplings are determined and subsequently interpreted for different scenarios in terms of effective alignment tensors, average orientations of dipolar vectors, and intramolecular reorientational vector distributions. Analytical relationships are derived that reflect similarities and differences between motional scaling of dipolar couplings and scaling of dipolar relaxation data (NMR order parameters). Application of the self-consistent procedure presented here to dipolar coupling measurements of biomolecules aligned in different liquid-crystalline media should allow one to extract in a "model-free" way average orientations of dipolar vectors and specific aspects of their motions.

1. Introduction

Since the first measurements of nuclear dipolar spin−spin couplings in proteins caused by the partial alignment of the proteins with respect to the external magnetic field,[1−5] these parameters have become widely used for the determination and refinement of structures of biomolecules in solution.[6−11] While in most applications residual dipolar couplings (rdc) are interpreted in the context of a static structure, it has been suggested from early on that these couplings also probe protein dynamics.[12] In multimodular systems, such as multidomain proteins and complex sugars, differences in alignment tensors determined for individual domains were attributed to differential motions between the domains.[13−17]

In the context of biomolecular structure determination, dipolar couplings are used to refine structures by optimizing agreement between experimental couplings, D_j^{exp}, and dipolar couplings predicted from the structural model, D_j^{calc}. A commonly used measure for the agreement is the Q value, defined by[18]

$$Q = \frac{\sum_j (D_j^{\text{exp}} - D_j^{\text{calc}})^2}{\sum_j (D_j^{\text{exp}})^2} \qquad (1)$$

The smaller Q, the better is the agreement between the structural model and the experimental data. In case of perfect agreement ($Q = 0$), $D_j^{\text{exp}} = D_j^{\text{calc}}$ for all j (values of Q larger than 1 are of little interest, since $Q = 1$ can always be achieved by setting $D_j^{\text{calc}} = 0$). In our experience, Q values for dipolar couplings determined directly from X-ray structures and NMR structures determined without the use of dipolar couplings typically lie between 0.2 and 0.5.[18,19] Possible reasons for Q values deviating from zero are experimental uncertainties, dynamic and exchange

* To whom correspondence should be addressed. C.G.: phone +49 69 7982-9130, fax +49 69 7982 9128, E-mail cigr@mpibpc.mpg.de. R.B.: phone (508) 793-7220, fax (508) 793-8861, E-mail bruschweiler@nmr.clarku.edu.

[†] Universität Frankfurt.

[‡] Clark University.

[§] Max-Planck Institute for Biophysical Chemistry.

(1) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9279−9283.

(2) Kung, H. C.; Wang, K. Y.; Goljer, I.; Bolton, P. H. *J. Magn. Reson. B* **1995**, *109*, 323−325.

(3) Tjandra, N.; Grzesiek, S.; Bax, A. *J. Am. Chem. Soc.* **1996**, *118*, 6264−6272.

(4) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111−1114.

(5) Bax, A.; Tjandra, N. *J. Biomol. NMR* **1997**, *10*, 289−292.

(6) Bewley, C. A.; Gustafson, K. R.; Boyd, M. R.; Covell, D. G.; Bax, A.; Clore, G. M.; Gronenborn, A. M. *Nat. Struct. Biol.* **1998**, *5*, 571−578.

(7) Cai, M.; Huang, Y.; Zheng, R.; Wei, S.-Q.; Ghirlando, R.; Lee, M. S.; Craigie, R.; Gronenborn, A. M.; Clore, G. M. *Nat. Struct. Biol.* **1998**, *5*, 903−909.

(8) Clore, G. M.; Starich, M. R.; Bewley, C. A.; Cai, M.; Kuszewski, J. *J. Am. Chem. Soc.* **1999**, *121*, 6513−6514.

(9) Drohat, A. C.; Tjandra, N.; Baldisseri, D. M.; Weber, D. J. *Protein Sci.* **1999**, *8*, 800−809.

(10) Mollova, E. T.; Hansen, M. R.; Pardi, A. *J. Am. Chem. Soc.* **2000**, *122*, 11561−11562.

(11) Hus, J. C.; Marion, D.; Blackledge, M. *J. Am. Chem. Soc.* **2001**, *123*, 1541−1542.

(12) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Nat. Struct. Biol.* **1997**, *4*, 292−297.

(13) Weaver, J. L.; Prestegard, J. H. *Biochemistry* **1998**, *37*, 116−128.

(14) Fischer, M. W. F.; Losconczi, J. A.; Weaver, J. L.; Prestegard, J. H. *Biochemistry* **1999**, *38*, 9013−9022.

(15) Skrynnikov, N.; Goto, N. K.; Yang, D.; Choy, W.-Y.; Tolman, J. R.; Mueller, G. A.; Kay, L. E. *J. Mol. Biol.* **2000**, *295*, 1265−1273.

(16) Neubauer, H.; Meiler, J.; Peti, W.; Griesinger, C. *Helv. Chim. Acta* **2001**, *84*, 243−258.

(17) Tian, F.; Al-Hashimi, H. M.; Craighead, J. L.; Prestegard, J. H. *J. Am. Chem. Soc.* **2001**, *123*, 485−492.

(18) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836−6837.

(19) Meiler, J.; Peti, W.; Griesinger, C. *J. Biomol. NMR* **2000**, *17*, 283−294.

10.1021/ja010002z CCC: $20.00 © xxxx American Chemical Society
Published on Web 00/00/0000 PAGE EST: 9.4

effects, and errors in the 3D structures, for example due to crystal packing in X-ray structures. NMR structures that are refined using dipolar couplings typically exhibit Q values between 0.05 and 0.3, depending also on the quality of the experimental data.[19,20]

In this work we investigate the effects of motions on Q values of backbone dipolar couplings using a 10 ns molecular dynamics (MD) simulation of ubiquitin and discuss several different scenarios for the structural and dynamic interpretation of dipolar couplings that explicitly take dynamics contributions into account. A practical procedure for including certain aspects of dynamics is the division of each dipolar coupling value by the corresponding Lipari−Szabo order parameter, S_{LS}, obtained from NMR relaxation experiments.[21] This procedure is valid only to a first-order approximation.[22] The quality of this approximation is quantitatively assessed here using the MD simulation as a reference from which averaged dipolar couplings as well as S_{LS} order parameters are computed and compared with each other. The inverse problem is then addressed to directly extract information on biomolecular structure and motions from dipolar couplings measured in multiple liquid-crystalline environments that give rise to different alignments. The proposed treatment is based on the assumptions that the structure and intramolecular motion are not significantly altered by the liquid-crystalline environment and that the alignment process is not affected by intramolecular motions. The treatment allows the determination of residual dipolar coupling order parameters, S_{rdc}, that probe motion up to the millisecond range and thus are complementary to the relaxation-derived Lipari−Szabo order parameters S_{LS}. We do not discuss here larger scale dynamics of (partially) unfolded proteins or interdomain dynamics of multidomain proteins. In the following section, the theoretical background of motional averaging effects on dipolar couplings is developed. In subsequent sections the theory is applied to the MD trajectory.

2. Motional Averaging of Dipolar Couplings

The residual dipolar couplings, which give rise to resonance splittings, result from the secular part of the magnetic dipole−dipole interactions between nuclear spins of molecules that are partially aligned in an anisotropic liquid. The dipolar splitting $\langle D\rangle$ (in units of hertz) between directly bonded heteronuclei X and H can be expressed in the laboratory frame as

$$\langle D\rangle = -\frac{\mu_0}{4\pi^2}\gamma_{\mathrm{X}}\gamma_{\mathrm{H}}\frac{h}{2\pi}(4\pi/5)^{1/2}\langle r_{\mathrm{XH}}{}^{-3}\rangle\,\langle P_2(\cos\chi)\rangle \qquad (2)$$

where $P_2(\cos\chi) = (3\cos^2\chi - 1)/2$, χ is the angle of the internuclear vector to the external B_0 field, $\mu_0/4\pi = 10^{-7}$ V•s/A•m, γ_{X}, γ_{H} are the gyromagnetic ratios, and r_{XH} is the distance between the two spins. The angular brackets denote an ensemble average over orientations χ and distances r_{XH} or, assuming that the system is ergodic, a time average over a single molecule. In eq 2, it is assumed that radial and angular averaging are statistically separable, as is the case for directly bonded N−H and C−H atom pairs. Furthermore, the radial part $\langle r_{\mathrm{XH}}{}^{-3}\rangle$ can often be considered to be identical for nuclear X−H pairs of the same kind.

For an internally static molecule, the dipolar couplings can alternatively be expressed in a molecular fixed frame in terms of a traceless reduced alignment tensor $\mathbf{D}$ (in units of hertz), with eigenvalues D_{xx}, D_{yy}, and D_{zz}, where $|D_{zz}| \geq |D_{yy}| \geq |D_{xx}|$.[22] In the eigenframe of this tensor, the dipolar coupling between two nuclei connected by an internuclear vector with orientation $\Omega = (\theta,\varphi)$, where θ,φ denote the polar angles in the eigenframe of $\mathbf{D}$, is given by D_{stat}:[22]

$$D_{\mathrm{stat}} = D_{\mathrm{a}}\left\{3\cos^2\theta - 1 + \frac{3}{2}R\sin^2\theta\cos 2\varphi\right\} \qquad (3)$$

where $D_{\mathrm{a}} = D_{zz}/2$ is the axial component and $R = {}^{2}/_{3}(D_{xx} - D_{yy})/D_{zz}$ is the rhombicity of $\mathbf{D}$ with $0 \leq R \leq {}^{2}/_{3}$. If $\mathbf{D}$ is symmetric, D_{zz} corresponds to the principal axis value along the symmetry axis of $\mathbf{D}$. For a given alignment tensor $\mathbf{D}$, D_{zz} is the largest coupling possible for the considered type of X−H spin pairs.

In the presence of intramolecular molecular dynamics, the experimental dipolar coupling corresponds to a conformational average, denoted by angular brackets, relative to the alignment tensor frame:

$$\langle D\rangle = D_{\mathrm{a}}\left\{\langle 3\cos^2\theta - 1\rangle + \frac{3}{2}R\langle\sin^2\theta\cos 2\varphi\rangle\right\} \qquad (4)$$

Equation 4 assumes that intramolecular motion does not interfere with the alignment process, i.e., that the alignment process is not significantly affected by internal motions. In the case of an alignment process due to steric effects,[23] this condition is fulfilled for motions that do not much alter the shape of the molecule. For small-amplitude, short-range motions, which can have a local or a concerted character,[24] eq 4 is expected to be more accurate than for larger amplitude motions of loops and termini, for example.

It is useful to express eq 4 in terms of normalized second-order spherical harmonic functions $Y_{2M}(\theta,\varphi)$:[15]

$$\frac{\langle D\rangle}{D_{zz}} = \sqrt{\frac{4\pi}{5}}\left(\langle Y_{20}(\theta,\varphi)\rangle + \sqrt{\frac{3}{8}}R(\langle Y_{22}(\theta,\varphi)\rangle + \langle Y^{*}_{22}(\theta,\varphi)\rangle)\right) \qquad (5)$$

where $Y_{20}(\theta,\varphi) = \sqrt{5/(16\pi)}(3\cos^2\theta - 1)$, $Y_{2\pm 2}(\theta,\varphi) = \sqrt{15/(32\pi)}\,e^{\pm 2i\varphi}\sin^2\theta$.[25] In what follows, $Y_{2\pm 1}(\theta,\varphi) = \mp\sqrt{15/(8\pi)}\,e^{\pm i\varphi}\cos\theta\sin\theta$ will also be used.

In analogy to eq 3, D_{stat} can be defined for reference purposes as the dipolar coupling expected from a static internuclear vector pointing along the average orientation $(\theta_{\mathrm{av}},\varphi_{\mathrm{av}}) = (\langle\theta\rangle,\langle\varphi\rangle)$:

$$\frac{D_{\mathrm{stat}}}{D_{zz}} = \sqrt{\frac{4\pi}{5}}\left(Y_{20}(\theta_{\mathrm{av}},\varphi_{\mathrm{av}}) + \sqrt{\frac{3}{8}}R(Y_{22}(\theta_{\mathrm{av}},\varphi_{\mathrm{av}}) + Y^{*}_{22}(\theta_{\mathrm{av}},\varphi_{\mathrm{av}}))\right) \qquad (6)$$

The effect of intramolecular reorientational motion on the dipolar coupling can be expressed by the *dipolar scaling factor*, λ_{rdc}:

$$\lambda_{\mathrm{rdc}} = \langle D\rangle/D_{\mathrm{stat}} \qquad (7)$$

In the absence of motion, $\lambda_{\mathrm{rdc}} = 1$, and in the presence of motion, $-\infty < \lambda_{\mathrm{rdc}} < \infty$.

[20] Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 12334−12341.

[21] Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546−4559.

[22] Clore, G. M.; Gronenborn, A. M.; Bax, A. *J. Magn. Reson.* **1998**, *133*, 216−221.

[23] Zweckstetter, M.; Bax, A. *J. Am. Chem. Soc.* **2000**, *122*, 3791−3792.

[24] Brüschweiler, R. *J. Chem. Phys.* **1995**, *102*, 3396−3403.

[25] Zare, R. N. *Angular Momentum*; John-Wiley & Sons: New York, 1988.

A benefit of eq 5 is that it can be easily transformed into a new reference frame related to the old reference frame by a three-dimensional rotation $\mathbf{R}(\alpha,\beta,\gamma)$ using the well-known transformation properties of the spherical harmonics $Y_{2M}(\theta,\varphi)$ under a three-dimensional rotation specified by the Euler angles α, β, and γ:[25]

$$\mathbf{R}(\alpha,\beta,\gamma)Y_{2M}(\theta,\varphi) = \sum_{M'=-2}^{2} \mathrm{e}^{-i\alpha M'} d^{(2)}_{M'M}(\beta)\, \mathrm{e}^{-i\gamma M}\, Y_{2M'}(\theta,\varphi) \tag{8}$$

Provided that the average protein structure and the dynamics do not vary with different alignment media, it follows that the dipolar coupling $\langle D\rangle$ measured in a new alignment frame i with axial component $D_{zz}{}^{(i)}$ and rhombicity $R^{(i)}$ that is related to the old frame by the rotation $\mathbf{R}(\alpha^{(i)},\beta^{(i)},\gamma^{(i)})$ can be expressed as

$$\begin{aligned}\frac{\langle D^{(i)}\rangle}{D_{zz}{}^{(i)}} = {} & \sqrt{\frac{4\pi}{5}}\left(\sum_{M'=-2}^{2} \mathrm{e}^{-iM'\alpha^{(i)}}\, d^{(2)}_{M'0}(\beta^{(i)})\langle Y_{2M'}\rangle\right) + \\ & \sqrt{\frac{4\pi}{5}}\sqrt{\frac{3}{8}}R\left(\sum_{M'=-2}^{2} \mathrm{e}^{-iM'\alpha^{(i)}}\, d^{(2)}_{M'2}(\beta^{(i)})\, \mathrm{e}^{-2i\gamma^{(i)}}\langle Y_{2M'}\rangle + \right. \\ & \left. \mathrm{e}^{-iM'\alpha^{(i)}}\, d^{(2)}_{M'-2}(\beta^{(i)})\, \mathrm{e}^{2i\gamma^{(i)}}\langle Y_{2M'}\rangle\right)\end{aligned} \tag{9}$$

Note that eq 9 is linear in the five motionally averaged spherical harmonics $\langle Y_{2M'}(\theta,\varphi)\rangle$. If the couplings $\langle D^{(i)}\rangle$ belonging to a certain dipolar interaction are measured for five (or more) known alignments tensors $\{D_{zz}{}^{(i)}, R^{(i)}, \alpha^{(i)}, \beta^{(i)}, \gamma^{(i)}\}$, the five quantities $\langle Y_{2M'}\rangle$, $M' = -2, -1, 0, 1, 2$ belonging to this dipolar interaction can be determined by solving the linear system of equations of eq 9 using, for example, singular-value decomposition or Moore−Penrose inversion.[26] The average vector orientation $(\theta_{\mathrm{av}},\varphi_{\mathrm{av}})$ can be approximated by the effective orientation $(\theta_{\mathrm{eff}},\varphi_{\mathrm{eff}})$ that is found by minimizing the sum

$$\sum_{M=-2}^{2}\Big(\langle Y_{2M}(\theta,\varphi)\rangle - Y_{2M}(\theta_{\mathrm{eff}},\varphi_{\mathrm{eff}})\Big)^2 \tag{10}$$

To discuss the effects of symmetry in the motional distributions of an internuclear vector, it is useful to describe the distribution in a frame with the z axis pointing along the average orientation of the vector. In this new frame the instantaneous orientation of a vector is denoted by (θ',φ'). If $\langle \mathrm{e}^{\pm i\varphi'}\rangle = \langle \mathrm{e}^{\pm 2i\varphi'}\rangle = 0$, as is the case for axially symmetric reorientational motion, it follows $\langle Y_{2M'}\rangle = 0$ except for $\langle Y_{20}\rangle$. To calculate the dipolar coupling, a coordinate transformation into the alignment frame is necessary, which is achieved by the rotation $\mathbf{R}'(\alpha' = 0, \beta' = -\theta_{\mathrm{av}}, \gamma' = -\varphi_{\mathrm{av}})$.

The extent of *nonaxial symmetry* of the motion can be quantified by the *motional asymmetry parameter* η fulfilling $0 \le \eta \le 1$:

$$\eta = \left(\frac{\sum_{M=\pm1,\pm2}\langle Y_{2M}(\theta',\varphi')\rangle\langle Y^*_{2M}(\theta',\varphi')\rangle}{\sum_{M=0,\pm1,\pm2}\langle Y_{2M}(\theta',\varphi')\rangle\langle Y^*_{2M}(\theta',\varphi')\rangle}\right)^{1/2}$$

$$= \frac{(S^2 - \langle P_2(\cos\theta')\rangle^2)^{1/2}}{S} \tag{11}$$

where the generalized S^2 order parameter has been introduced, which plays a key role in the "model-free" interpretation of heteronuclear NMR spin relaxation data by Lipari and Szabo,[21]

$$S^2 = \frac{4\pi}{5}\sum_{M=-2}^{2}\langle Y_{2M}(\theta,\varphi)\rangle\langle Y^*_{2M}(\theta,\varphi)\rangle \tag{12}$$

The S^2 order parameter extracted from spin relaxation data is sensitive to motions faster than the overall tumbling correlation time and is denoted here as S^2_{LS}. In contrast, an S^2 order parameter can be determined from residual dipolar couplings using eqs 9 and 12, which probes the much wider submillisecond time scale range and which is denoted as S^2_{rdc}. Therefore, S^2_{LS} is an upper limit for S^2_{rdc}, $S^2_{\mathrm{rdc}} \le S^2_{\mathrm{LS}}$. Note that when using the 10 ns MD trajectory for calculating S^2_{LS} and S^2_{rdc}, the two parameters probe the same time scales and are therefore identical.

For *axially symmetric motion* with respect to the average orientation $(\theta_{\mathrm{av}},\varphi_{\mathrm{av}})$, for which $\langle \mathrm{e}^{\pm i\varphi'}\rangle$ and $\langle \mathrm{e}^{\pm 2i\varphi'}\rangle$ vanish, the average dipolar coupling $\langle D\rangle$ can be expressed in a more compact way. Using eq 9 with $\langle Y_{2M'}(\theta',\varphi')\rangle = 0$ for $M' = \pm1$, ±2 it follows

$$\begin{aligned}\frac{\langle D\rangle_{\mathrm{sym}}}{D_{zz}} &= \sqrt{\frac{4\pi}{5}}\langle Y_{20}(\theta',\varphi')\rangle\left(d^{(2)}_{00}(\beta') + \sqrt{\frac{3}{8}}R(d^{(2)}_{02}(\beta')\, \mathrm{e}^{-2i\gamma'} + d^{(2)}_{0-2}(\beta')\, \mathrm{e}^{2i\gamma'})\right) \\ &= \left\langle\frac{1}{2}(3\cos^2\theta' - 1)\right\rangle\left(\frac{1}{2}(3\cos^2\beta' - 1) + \frac{3}{4}R\sin^2\beta'\cos 2\gamma'\right)\end{aligned} \tag{13}$$

where $\beta' = -\theta_{\mathrm{av}}$, $\gamma' = -\varphi_{\mathrm{av}}$. Consequently, the dipolar coupling of an internuclear vector is scaled under axially symmetric motion as compared to a static vector pointing along the average direction by

$$\lambda_{\mathrm{rdc,sym}} = \frac{\langle D\rangle_{\mathrm{sym}}}{D_{\mathrm{stat}}} = \langle P_2(\cos\theta')\rangle \tag{14}$$

where θ' is the angle between an instantaneous internuclear vector and the average vector orientation $(\theta_{\mathrm{av}},\varphi_{\mathrm{av}})$. Note that $\lambda_{\mathrm{rdc,sym}}$ does not depend on the relative orientation $(\theta_{\mathrm{av}},\varphi_{\mathrm{av}})$ with respect to the alignment frame. In the case of axially symmetric motion, S^2_{rdc} simplifies to $S^2_{\mathrm{rdc,sym}} = (4\pi/5)\langle Y_{20}(\theta',\varphi')\rangle^2 = \langle P_2(\cos\theta')\rangle^2$ and thus

$$\lambda_{\mathrm{rdc,sym}} = S_{\mathrm{rdc,sym}} \tag{15}$$

From eq 14 follows that $-0.5 \le \lambda_{\mathrm{rdc,sym}}, S_{\mathrm{rdc,sym}} \le 1$.

Knowledge of λ_{rdc} is useful for the determination of an average 3D protein structure using residual dipolar couplings. D_{stat} values, which are directly related to $(\theta_{\mathrm{av}},\varphi_{\mathrm{av}})$ by eq 6, could then be obtained by dividing experimental couplings $\langle D\rangle$ by λ_{rdc} according to eq 7. Since in practice λ_{rdc} values are not readily available, they sometimes are approximated by their respective S_{LS} values extracted from spin relaxation experiments.[18] $\lambda_{\mathrm{rdc}} = S_{\mathrm{LS}}$ holds if (i) internal reorientational motion is axially symmetric and (ii) all relevant motions take place on nanosecond and subnanosecond time scales. Condition (i) can be tested by

[26] Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in C*; Cambridge University Press: Cambridge, 1988.

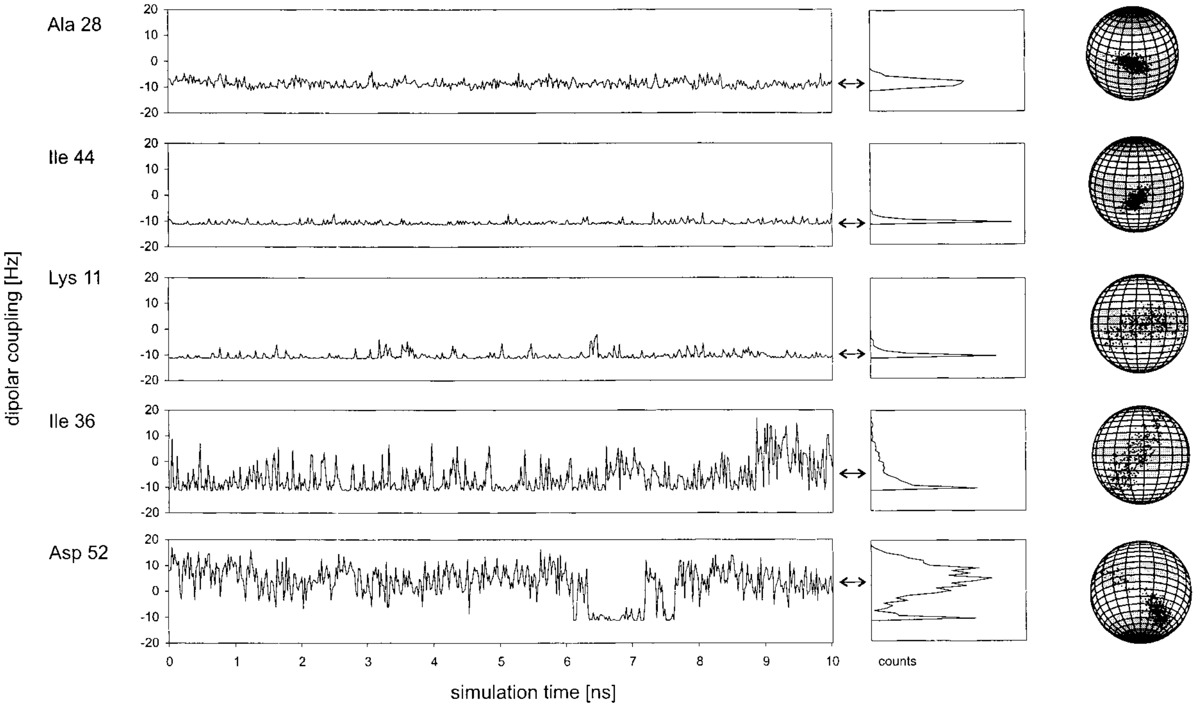


Figure 1. Time dependence of $^{15}N-^{1}H^{N}$ residual dipolar coupling values for selected amino acids of ubiquitin extracted from a 10 ns molecular dynamics (MD) simulation. Ala 28 belongs to the α helix, Ile 44 to a β strand, and Lys 11, Ile 36, Asp 52 to loop regions. In the middle, the distributions of the couplings are plotted vertically, with the horizontal arrows indicating average dipolar coupling values that would be observed experimentally. The dots plotted on the surface of the spheres (right) correspond to the $N-H^{N}$ orientations sampled during the MD trajectory (500 snapshots).

using a molecular dynamics (MD) simulation, which is done in the following section. Presently, condition (ii) can be assessed only by comparison with experimental data. If condition (i) is fulfilled but (ii) is not fulfilled, $S_{LS,sym}$ represents an upper limit, $\lambda_{rdc,sym} = S_{rdc,sym} \leq S_{LS,sym}$.

In addition, it was assumed here that the alignment tensor **D** is a priori known. In practice, however, **D** is iteratively adjusted during structure refinement based on residual dipolar couplings. Thus, the best fitting alignment tensor **D** implicitly includes certain motional contributions. In the following, a 10 ns MD simulation of ubiquitin is used to elucidate the influence of molecular motion on the interpretation of residual dipolar couplings.

3. Dipolar Couplings Calculated from MD Trajectory

A MD simulation of native ubiquitin was carried out under periodic boundary conditions using the program CHARMM 24.[27,28] An energy-minimized all-atom representation of the X-ray structure of ubiquitin[29] was embedded in a cubic box with a side length of 46.65 Å, containing a total of 2909 explicit water molecules. The simulation was performed at a temperature of 300 K with an integration time step of 1 fs. Details of this simulation have been reported elsewhere.[30] During a simulation time of 11 ns, snapshots were stored every 500 fs. From a 10 ns section of the trajectory, covering the range between 1 and 11 ns, 500 snapshots were selected with an increment of 20 ps for calculating dipolar couplings.

The 500 snapshots were reoriented and translated with respect to the snapshot at 6 ns by a least-squares superposition of their backbone atoms belonging to regular secondary structures. An average structure was constructed from the 500 reoriented snapshots by averaging over the Cartesian coordinates of all heavy atoms. The average positions of hydrogen atoms were determined by adding averaged X$-$H vectors (X = N or C atoms), which were rescaled to their standard lengths (1.02 Å for N$-$H and 1.09 Å for C$-$H), to the position of the corresponding X atom.

The shape of ubiquitin undergoes only small changes during the trajectory, as was assessed by computing inertia tensors for the 500 snapshots. The standard deviations of the moments of inertia tensor lie between 1% and 2%, which supports the validity of the assumptions underlying eq 4. It is assumed in the following that the MD trajectory represents a realistic description of the internal dynamics of ubiquitin, and thus slower time scale motions, which are not represented by the 10 ns simulation, are ignored.

To characterize the effect of dynamics, dipolar couplings were calculated from the 500 snapshots for a fixed alignment tensor **D** with $D_{zz} = 20$ Hz (with respect to $^{15}N-^{1}H^{N}$ couplings), $R = 0$, $\alpha = \beta = \gamma = 0$. The time dependence of backbone $^{15}N-^{1}H^{N}$ couplings is depicted in Figure 1 for a selection of five amino acids that experience variable amounts of motion: Ala 28 (α helix), Ile 44 (β sheet), Lys 11 (loop), Ile 36 (loop), and Asp 52 (loop). Also given in Figure 1 are the distributions of the dipolar couplings over the trajectory. Most of the displayed distributions, which also depend on the size and orientation of the alignment tensor, show quasi-singularities and are unimodal

[27] Brooks, R. B.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

[28] MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

[29] Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531–544.

[30] Lienin, S. F.; Bremi, T.; Brutscher, B.; Brüschweiler, R.; Ernst, R. R. *J. Am. Chem. Soc.* **1998**, *120*, 9870–9879.

Table 1. Back-Calculated Alignment Tensors and Q Values for Ubiquitin According to Scenario I (No Scaling)

nuclei[a]	$(\tilde{D}_{zz}/D_{zz})$[b]	R[c]	α[d] (deg)	β[e] (deg)	γ[f] (deg)	Q_{all}[g]	Q_{sec}[h]	Q_{loop}[i]
true	1.00	0.00	0	0	0	0.00	0.00	0.00
N H$^{\mathrm{N}}$	1.00	0.00	0	0	0	0.15	0.08	0.24
C$^{\alpha}$ H$^{\alpha}$	1.00	0.00	0	0	0	0.10	0.08	0.12
H$^{\mathrm{N}}$ H$^{\alpha}$	1.00	0.00	0	0	0	0.07	0.04	0.09
true	1.00	0.33	0	0	0	0.00	0.00	0.00
N H$^{\mathrm{N}}$	1.00	0.33	0	0	0	0.16	0.08	0.22
C$^{\alpha}$ H$^{\alpha}$	1.00	0.33	0	0	0	0.09	0.07	0.11
H$^{\mathrm{N}}$ H$^{\alpha}$	1.00	0.33	0	0	0	0.07	0.04	0.09
true	1.00	0.67	0	0	0	0.00	0.00	0.00
N H$^{\mathrm{N}}$	1.00	0.67	0	0	0	0.16	0.09	0.23
C$^{\alpha}$ H$^{\alpha}$	1.00	0.67	0	0	0	0.10	0.08	0.11
H$^{\mathrm{N}}$ H$^{\alpha}$	1.00	0.67	0	0	0	0.07	0.04	0.09
N H$^{\mathrm{N}}$	1.00	0.00	0	0	0	0.15	0.08	0.24
	1.00	0.00	0	45	0	0.14	0.07	0.19
	1.00	0.00	0	45	90	0.12	0.09	0.15
	1.00	0.00	0	45	180	0.10	0.07	0.13
	1.00	0.00	0	45	270	0.14	0.08	0.18
	1.00	0.00	0	90	0	0.12	0.06	0.24
	1.00	0.00	0	90	45	0.09	0.07	0.14
	1.00	0.00	0	90	90	0.14	0.06	0.20
	1.00	0.00	0	90	135	0.16	0.08	0.26

[a] Pairs of nuclei for which dipolar couplings are computed. [b] Ratio of the best fitting and the predefined tensor size ($D_{zz} = 20$ Hz for N$-$H$^{\mathrm{N}}$ couplings). [c] Rhombicities and [d,e,f] orientations of the predefined and back-calculated tensors. [g,h,i] Q_{all}, Q_{sec}, and Q_{loop} are the Q values (eq 1) for the whole protein, the secondary structural elements, and the loops, respectively.

except for Asp 52, where larger scale backbone modulations lead to a bimodal distribution.

4. Influence of Motion on Q Values

In practice, experimental dipolar couplings are commonly refined toward a single static structure. It is investigated here what level of agreement can be expected between experimental couplings and couplings calculated from the average structure in the presence of molecular motion occurring during the 10 ns MD trajectory of ubiquitin.

For the following analyses, sets of dipolar couplings belonging to 11 alignment tensors with different orientations and rhombicities were constructed from the 500 snapshots taken from the trajectory (see Table 1). Here and in the following, it is assumed that changes in the alignment tensor leave intramolecular motions unaffected. For three alignment tensors with $D_{zz} = 20$ Hz (for N$-$H$^{\mathrm{N}}$ dipolar couplings) and R values set to 0, $^{1}/_{3}$, and $^{2}/_{3}$, respectively, dipolar couplings were computed for N$-$H$^{\mathrm{N}}$, C$^{\alpha}-$H$^{\alpha}$, and H$^{\mathrm{N}}-$H$^{\alpha}$ spin pairs. In addition, eight more alignment tensors with $R = 0$ and $D_{zz} = 20$ Hz were defined by reorienting the original tensor using rotation matrixes $\mathbf{R}(\alpha,\beta,\gamma)$ with the following Euler angles to sample a representative distribution of tensor orientations:

$$(\alpha,\beta,\gamma) = \{(0°,45°,0°), (0°,45°,90°), (0°,45°,180°), (0°,45°,270°), (0°,90°,0°), (0°,90°,45°), (0°,90°,90°), (0°,90°,135°)\}$$

For these eight alignment tensors, only N$-$H$^{\mathrm{N}}$ couplings were computed.

On the basis of the MD simulation of ubiquitin, the effect of motion on Q was analyzed for three scenarios, I, II, and III, that involve different treatments of the data:

I. In this scenario, for a given alignment tensor, dipolar couplings were averaged over the 500 MD snapshots and compared with the dipolar couplings calculated from the average structure described in the previous section using the same alignment tensor.

II. In this scenario, for a given alignment tensor, dipolar couplings were averaged over the 500 MD snapshots and compared with the dipolar couplings calculated from the average structure using an optimized alignment tensor that was varied in size and orientation to minimize Q.

III. In this scenario, for a given alignment tensor, dipolar couplings were averaged over the 500 MD snapshots and subsequently divided by their respective S_{LS} order parameters calculated from the same snapshots. These rescaled dipolar couplings were then compared with the dipolar couplings calculated from the average structure using an optimized alignment tensor that was varied in size and orientation to minimize Q.

Scenario I corresponds to a situation where the "true" alignment tensor is known from external sources, for example theoretical calculations[23,31] or paramagnetic alignment. In this case, Q values become largest and motional effects are strongest, since they are not included in the form of a scaled alignment tensor. For scenario II, which is equivalent to overall scaling of all dipolar couplings combined with reorientation of the alignment tensor, readjustment of the alignment tensor can partially absorb internal motional effects. For example, if intramolecular motion reduces all dipolar couplings by 10% (compared to the couplings of the average structure), a new alignment tensor for the average structure that is 10% smaller would still yield $Q = 0$. This approach is equivalent to scaling of all couplings by a uniform λ_{rdc} value. Since the amplitudes of intramolecular motion generally vary between different protein sites, there will be no uniform scaling of dipolar couplings. Instead, individual motional scaling of dipolar couplings must be explicitly taken into account, which is the approach followed in scenario III. In the absence of any other information, a commonly used guess for the scaling factors are the S_{LS} order parameters of eq 12 obtained from spin relaxation measurements. As was shown in section 2 (eq 15), the scaling by S_{LS} values is adequate if all intramolecular motions are axially symmetric and take place on nanosecond and subnanosecond time scales that are accessed by spin relaxation experiments.

All three scenarios were analyzed for all 11 alignment tensors, and the results are compiled in Tables 1$-$3. For scenarios II and III, which involve fitting of the alignment tenors, the program DipoCoup[19] was used, performing a Moore$-$Penrose inversion, also known as singular-value decomposition,[26,32] previously used for the analysis of NMR relaxation data for anisotropic tumbling.[33] Since the results significantly differ between protein backbone parts with a well-defined secondary structure and loop regions, Q values were calculated for these different parts as well as for the whole protein backbone. Amino acids that belong to either a helix or a β sheet have residue numbers 2$-$7, 12$-$17, 23$-$34, 41$-$45, 49$-$50, and 65$-$72.

For scenario I, the Q values vary between 0.04 and 0.26 (see Table 1). They clearly depend on the type of vectors: N$-$H$^{\mathrm{N}}$ vectors show Q values that are larger than those of C$^{\alpha}-$H$^{\alpha}$ vectors, which in turn have Q values that are larger than those of H$^{\mathrm{N}}-$H$^{\alpha}$ vectors. This is not surprising since the H$^{\mathrm{N}}-$H$^{\alpha}$ distances are longer than the one-bond distances, and thus a displacement of the H$^{\mathrm{N}}$ or H$^{\alpha}$ atom causes only a minor change

(31) Ferrarini, A.; Moro, G. J.; Nordio, P. L. *Mol. Phys.* **1992**, *77*, 1$-$15.

(32) Losonczi, J. A.; Andrec, M.; Fischer, M. W. F.; Prestegard, J. H. *J. Magn. Res.* **1999**, *138*, 334$-$342.

(33) Brüschweiler, R.; Liao, X.; Wright, P. E. *Science* **1995**, *268*, 886$-$889.

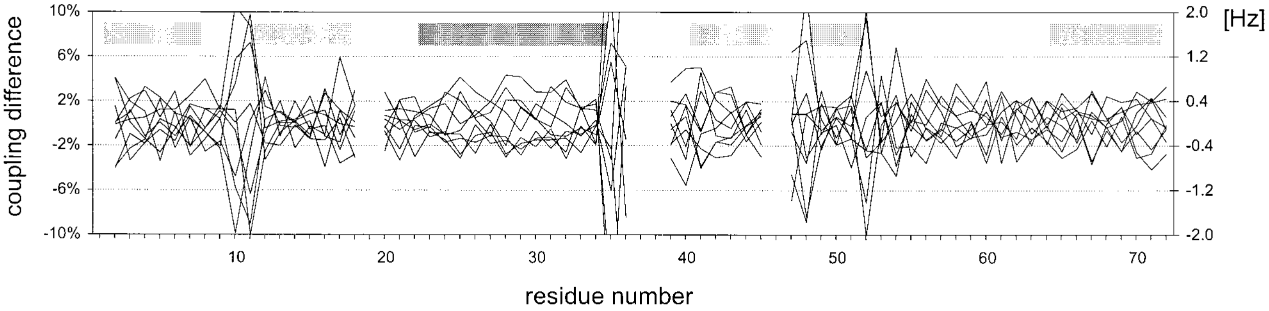


Figure 2. Differences between dipolar N−H^N couplings averaged over the 500 MD snapshots of ubiquitin and the back-calculated couplings determined for the (static) N−H^N vectors of the average structure as a function of residue number. The alignment tensor was optimized according to scenario II (see text). The calculation was done for each of the nine axially symmetric alignment tensors ($R = 0$) given in Table 2. The light gray bars on top of this figure (and also of Figures 3 and 6) indicates the β strands and the dark bar the α helix.

Table 2. Back-Calculated Alignment Tensors and Q Values for Ubiquitin According to Scenario II (Uniform Scaling)

	($\tilde{D}_{zz}$/		α[d]	β[e]	γ[f]				
nuclei[a]	D_{zz})[b]	R[c]	(deg)	(deg)	(deg)	Q_{all}[g]	Q_{sec}[h]	Q_{loop}[i]	S_{LS}[j]
true	1.00	0.00	0	0	0	0.00	0.00	0.00	
N H^N	0.90	0.04	0	0	0	0.10	0.05	0.17	0.91
C^α H^α	0.92	0.01	0	1	0	0.06	0.04	0.07	0.94
H^N H^α	0.95	0.01	0	−1	0	0.04	0.03	0.05	0.95
true	1.00	0.33	0	0	0	0.00	0.00	0.00	
N H^N	0.89	0.33	0	0	−2	0.11	0.06	0.15	0.91
C^α H^α	0.92	0.35	0	1	0	0.05	0.04	0.06	0.94
H^N H^α	0.95	0.33	0	−1	0	0.04	0.03	0.05	0.95
true	1.00	0.67	0	0	0	0.00	0.00	0.00	
N H^N	0.91	0.64	1	0	1	0.08	0.04	0.14	0.91
C^α H^α	0.95	0.65	0	89	0	0.06	0.03	0.07	0.94
H^N H^α	0.97	0.65	0	0	0	0.04	0.02	0.07	0.95
N H^N	0.90	0.04	0	0	0	0.10	0.05	0.17	0.91
	0.91	0.03	0	45	0	0.09	0.05	0.13	0.91
	0.90	0.02	0	45	91	0.06	0.05	0.07	0.91
	0.92	0.01	0	45	180	0.06	0.03	0.08	0.91
	0.91	0.01	0	45	270	0.09	0.06	0.11	0.91
	0.91	0.02	0	90	1	0.08	0.04	0.16	0.91
	0.93	0.01	0	90	45	0.05	0.03	0.09	0.91
	0.90	0.02	0	90	90	0.10	0.06	0.13	0.91
	0.89	0.02	0	90	135	0.11	0.05	0.18	0.91

[a] Pairs of nuclei for which dipolar couplings are computed. [b] Ratio of the best fitting and the predefined tensor size ($D_{zz} = 20$ Hz for N−H^N couplings). [c] Rhombicities and [d,e,f] orientations of the predefined and back-calculated tensors. [g,h,i] Q_{all}, Q_{sec}, and Q_{loop} are the Q values (eq 1) for the whole protein, the seconday structural elements, and the loops, respectively. [j] S_{LS} is the average order parameter for these vectors calculated from the MD trajectory according to eq 12.

in the vector orientation. The Q values depend on the orientation of the alignment tensor but they are nearly independent of R. Significant differences in Q are observed between regular secondary structures and loop regions.

The results of scenario II, which are summarized in Table 2, demonstrate the effect on Q values if the alignment tensor is allowed to vary. As compared to Table 1, the Q values drop by about 30%. The motional effects are contained in modified alignment tensors **D**. The directions of the principal axes change typically by less than 1°, and the rhombicity changes by 0.04 or less. The largest effects are seen in the new $\tilde{D}_{zz}$ values, which are scaled relative to the original D_{zz} values by factors between 0.89 and 0.95. Table 2 contains also average S_{LS} and S_{rdc} order parameters calculated from the 500 snapshots according to eq 12. The S_{LS} and S_{rdc} values vary between 0.91 and 0.95, which is comparable to the scaling factor variations $\tilde{D}_{zz}/D_{zz}$. The Q values depend on the details of the motional distributions and average orientations of the internuclear vectors relative to the alignment tensors. The Q values vary for the chosen alignment tensors by as much as a factor of 2. In Figure 2, the differences in back-calculated and "true" N−H^N dipolar couplings are plotted as a function of the residue number for the nine alignment tensors with different orientations defined in the lower part of Table 2. For individual N−H^N couplings, the motional influences characteristically depend on the directions sampled by the N−H^N vector relative to the alignment tensor. In the absence of rhombicity, $R = 0$, motion has the strongest influence for the average directions $\theta_{\mathrm{av}} = 0°,\pm 90°,180°$, for which $P_2(\cos\theta)$ has maximal curvature. The differences between back-calculated and "true" couplings, which are distributed around zero, are largest for the loop regions that do not belong to regular secondary structure. N−H^N vectors of these regions have calculated S^2 order parameters lower than 0.8 (see Supporting Information). For some but not all of these vectors, the orientational distributions have not converged during the 10 ns MD trajectory, which can also be seen for some of the examples shown in Figure 1. Figure 2 illustrates that the effect of dynamics on the observable dipolar coupling value depends on the orientation of the alignment tensor. The alignment tensor defines the projection along which motion is observable. The possibility to reconstruct characteristic motional features from dipolar couplings collected for different alignment tensors is discussed below.

For scenario III (Table 3), where the average dipolar couplings are individually divided by their S_{LS}, S_{rdc} values, the fitted alignment tensors almost identically reproduce the "true" alignment tensors with changes in R smaller than 0.03 and $\tilde{D}_{zz}$ values lying within 1% of D_{zz}. All Q values are further decreased as compared to the values in scenario II, with the largest reductions found for the mobile loop regions, where Q drops between 0.02 and 0.10. Figure 3 demonstrates the improved agreement for the individual N−H^N pairs as compared to the case in Figure 2. However, the Q values can still significantly differ from zero (see Table 3): for N−H^N dipolar couplings they vary between 0.04 and 0.07. This behavior is indicative of non-axially symmetric reorientational local motions of these internuclear vectors. Thus, the order parameter S_{LS}, S_{rdc} does not always accurately represent the motional scaling λ_{rdc} of dipolar couplings during the MD simulation. The residual discrepancies shown in Figure 3 are smallest for N−H^N vectors belonging to regular secondary structures, where dynamics is smaller and more closely matches axial symmetry than in the loop regions, where more complicated motion occurs that is generally more asymmetric. Analogous analyses carried out for alignment tensors with increasing rhombicities R indicate that changes in R can also have non-negligible effects on dipolar couplings.

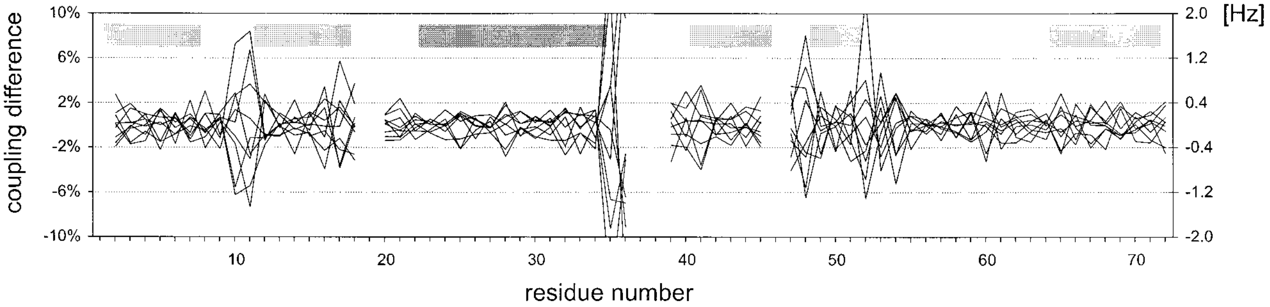


Figure 3. Differences between scaled dipolar N−H^N couplings averaged over the 500 MD snapshots of ubiquitin and the back-calculated couplings determined for the (static) N−H^N vectors of the average structure as a function of residue number. The dipolar couplings determined by averaging over the MD snapshots were scaled with their Lipari−Szabo order parameter S_{LS} according to scenario III. The calculation was done for each of the nine axially symmetric alignment tensors ($R = 0$) given in Table 3.

Table 3. Back-Calculated Alignment Tensors and Q Values for Ubiquitin According to Scenario III (Individual Scaling)

nuclei[a]	($\tilde{D}_{zz}$/ D_{zz})[b]	R[c]	α[d] (deg)	β[e] (deg)	γ[f] (deg)	Q_{all}[g]	Q_{sec}[h]	Q_{loop}[i]
true	1.00	0.00	0	0	0	0.00	0.00	0.00
N H^N	0.99	0.03	0	0	0	0.07	0.04	0.11
C^α H^α	1.00	0.00	0	0	0	0.02	0.02	0.02
H^N H^α	1.00	0.01	0	−1	0	0.03	0.02	0.04
true	1.00	0.33	0	0	0	0.00	0.00	0.00
N H^N	1.00	0.35	0	1	2	0.08	0.05	0.11
C^α H^α	1.00	0.34	0	0	0	0.03	0.03	0.03
H^N H^α	1.00	0.34	0	1	0	0.03	0.02	0.04
true	1.00	0.67	0	0	0	0.00	0.00	0.00
N H^N	1.01	0.65	0	0	0	0.04	0.02	0.07
C^α H^α	1.01	0.66	0	0	0	0.04	0.02	0.04
H^N H^α	1.01	0.66	−1	0	0	0.04	0.02	0.07
N H^N	0.99	0.03	0	0	0	0.07	0.04	0.11
	1.00	0.00	0	45	1	0.07	0.03	0.10
	1.00	0.01	0	46	90	0.05	0.03	0.07
	1.00	0.03	0	45	180	0.06	0.03	0.08
	1.00	0.01	0	45	270	0.06	0.04	0.07
	0.99	0.01	0	90	1	0.04	0.02	0.08
	1.01	0.01	0	90	45	0.04	0.02	0.06
	1.01	0.01	0	90	90	0.06	0.04	0.08
	0.99	0.01	0	90	135	0.05	0.03	0.08

[a] Pairs of nuclei for which dipolar couplings are computed. [b] Ratio of the best fitting and the predefined tensor size ($D_{zz} = 20$ Hz for N−H^N couplings). [c] Rhombicities and [d,e,f] orientations of the predefined and back-calculated tensors. [g,h,i] Q_{all}, Q_{sec}, and Q_{loop} are the Q values (eq 1) for the whole protein, the secondary structural elements, and the loops, respectively.

5. Reconstructing Motional Distributions from Multiple Alignment Data

From the previous analysis, it becomes clear that static and motional contributions to a dipolar coupling measured for a single alignment cannot readily be separated. The questions are addressed here of how this task can be accomplished by combining dipolar couplings measured for different alignment tensors and what aspects of the motional distributions of the internuclear vectors can be reconstructed.

For this purpose, five N−H^N pairs were selected in ubiquitin that show differential motional properties: Ala 28 (α helix) and Ile 44 (β sheet), which are part of regular secondary structures, and Lys 11, Ile 36, and Asp 52, which belong to the more mobile loop regions. The distinct motional behavior of these residues in the MD trajectory is reflected in their φ,ψ dihedral angle fluctuations. The right column in Figure 4 shows the φ,ψ distributions for the 500 MD snapshots: Ala 28 and Ile 44 show quite narrow φ,ψ distributions characteristic of α helix and β sheet structures, while Lys 11 and Ile 36 exhibit significantly wider distributions, in particular in their φ dihedral angle. Asp 52 exhibits a less regular behavior, indicative of a multimodal distribution. Dipolar couplings are sensitive to reorientations related to fluctuations of nearby dihedral angles as well as to longer range motions related to fluctuations of dihedral angles that are farther away.

In the left and middle panels of Figure 4, the orientations of the above-mentioned N−H^N vectors are displayed for the 500 MD snapshots indicated as dots as a function of the polar angles (θ,φ). The orientational distributions of the 500 snapshots are in all cases elongated (approximately elliptical for Ala 28 and Ile 44); i.e., they do not exhibit axial symmetry. The large filled circle in the center of each panel represents the orientation of the N−H^N vector in the average structure. The superimposed solid lines represent N−H^N vector orientations that are consistent with the dipolar couplings averaged over the 500 snapshots for the nine different alignment tensors with $R = 0$ described in the previous section (see Tables 2 and 3). Thus, any static N−H^N vector that points along a (θ,φ) direction belonging to a certain line could accurately reproduce the (scaled) dipolar coupling averaged over the trajectory for the alignment tensor associated with this line.

The panels in the left column correspond to scenario II, with the fitted alignment tensors given in the lower part of Table 2, while the panels in the middle column correspond to scenario III, where the couplings were divided by their individual S_{LS} order parameter. If all nine lines intersect at a single point, then a static N−H^N vector pointing along the intersection can simultaneously reproduce all MD-averaged couplings for the nine alignments. For the regular secondary structural residues Ala 28 and Ile 44, this behavior is approximately found for scenario II (left column of Figure 4), while it is not fulfilled for the three other residues, Lys 11, Ile 36, and Asp 52. For the latter residues, MD-averaged dipolar couplings measured for multiple alignments cannot be quantitatively reproduced by a static structural model.

Individual S_{LS} scaling of dipolar couplings (scenario III, middle column of Figure 4) improves the situation, in particular for Ala 28 and Ile 44, which belong to regular secondary structures although the reorientational distributions of these vectors are not axially symmetric. For these vectors, the intersections coincide with the dipolar coupling predicted from the average structure (filled circle). Thus, the MD-averaged dipolar couplings obtained in multiple alignment media scaled by their respective S_{LS} values allow for these residues the reconstruction of highly accurate average orientations. For a set of dipolar coupling measurements performed for a sufficiently large number of different alignments (five or more), it is conceivable to use an effective dipolar scaling factor $\lambda_{rdc,eff}$ as a fitting parameter. Since $\lambda_{rdc,eff}$ covers besides the relaxation-

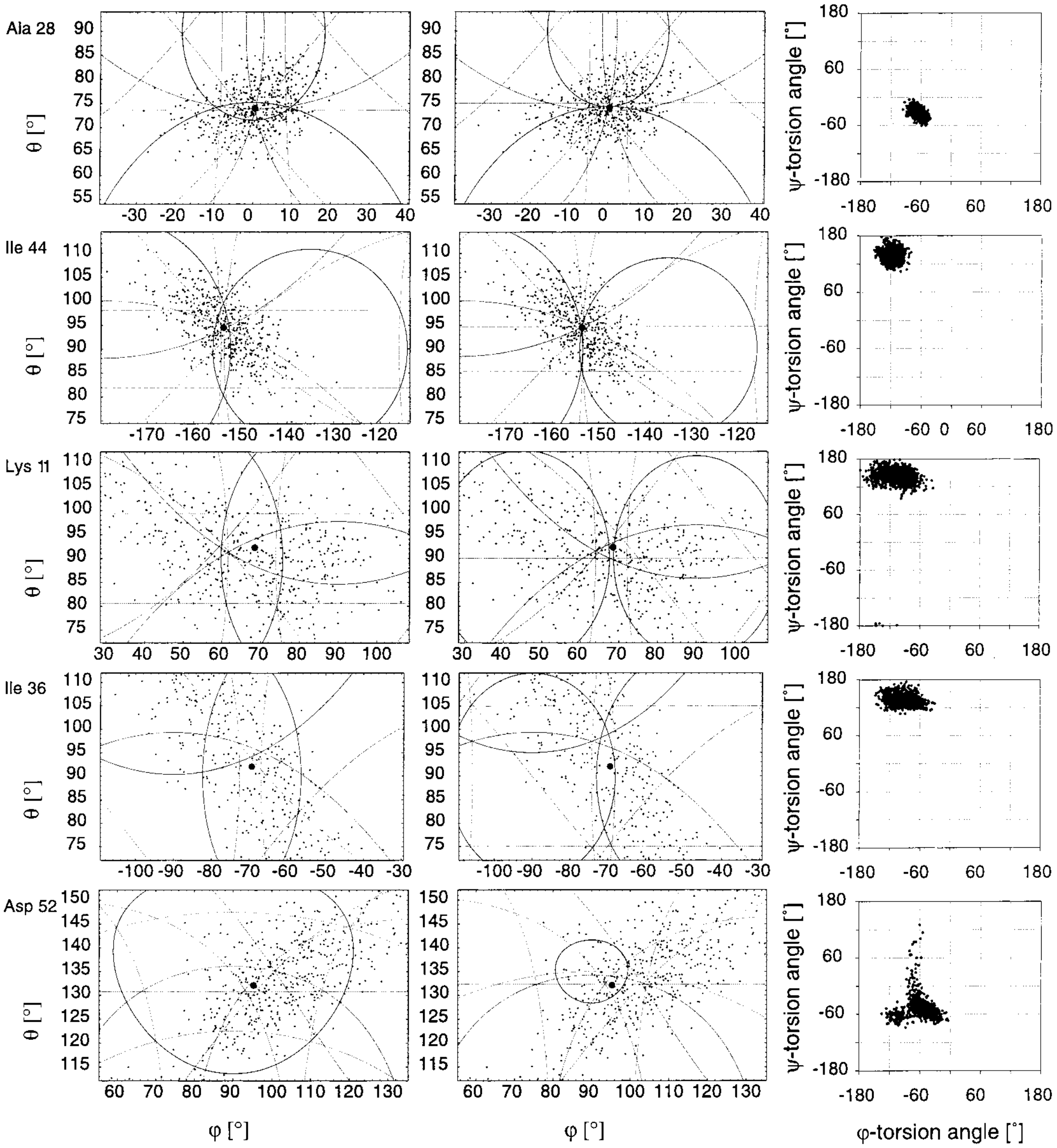


Figure 4. Determination of average N−H^N directions from dipolar couplings measured for nine different alignment tensors exemplified for residues Ala 28, Ile 44, Lys 11, Ile 36, and Asp 52. The dots in the panels in the left and the middle columns correspond to the orientations of the N−H^N vectors of 500 snapshots in the selected angular ranges. The full distributions are displayed on the spheres of Figure 1. The solid lines represent all static orientations that reproduce the MD-averaged couplings for the corresponding alignment tensors. The left panel corresponds to scenario II (no individual scaling of couplings), while the middle panel corresponds to scenario III (each coupling is scaled by its Lipari−Szabo order parameter S_{LS}). The panels in the right column show the φ,ψ dihedral angle distributions for these residues.

active motional time scales also slower time scales, comparison of $\lambda_{rdc,eff}$ with experimentally determined S_{LS} parameters should allow one to gain important insight into intramolecular motions occurring between nanosecond and millisecond time scales. It is expected that generally $\lambda_{rdc,eff} \leq S_{LS}$.

In contrast, for the mobile residues Lys 11, Ile 36, and Asp 52, none of the scenarios yields satisfactory results for the average orientations (left and middle columns of Figure 4). For scenario III (middle column), scaling by S_{LS} somewhat narrows down the range of possible average orientations, but obviously simple scaling remains insufficient for a quantitative determination of the average orientations because of the mathematical inequivalence of S_{LS} and λ_{rdc} for non-axially symmetric orientational distributions (cf. eqs 7, 9, and 12). Since according to eqs 7 and 9 λ_{rdc} directly depends on the average orientation of the dipolar vector, extraction of the average orientation and of the motional averaging effects becomes more complicated.

Model-Free Extraction of $\langle Y_{2M}\rangle$ and $(\theta_{eff},\varphi_{eff})$ Quantities. The following two-step procedure is proposed using experimental dipolar couplings and S_{LS} order parameters:

1. Absolute alignment tensors **D** for multiple liquid-crystalline media are determined from experimental dipolar couplings using S_{LS} order parameters obtained from relaxation experiments for residues belonging to well-defined secondary structures.

2. The average orientation of a dipolar vector belonging to a more mobile region is extracted by fitting the averaged spherical harmonics $\langle Y_{2M}\rangle$, $M = -2, -1, 0, 1, 2$, to the dipolar couplings collected in all available alignment media using eq 9 with **D** of 1 and by determining the orientations $(\theta_{eff},\varphi_{eff})$ by a least-squares fit according to eq 10.

The averaged spherical harmonics $\langle Y_{2M}\rangle$ quantities provide a “model-free” representation of motional effects on dipolar couplings in analogy to S^2_{LS} order parameters in spin relaxation studies.[21] In fact, the $\langle Y_{2M}\rangle$ quantities contain information about

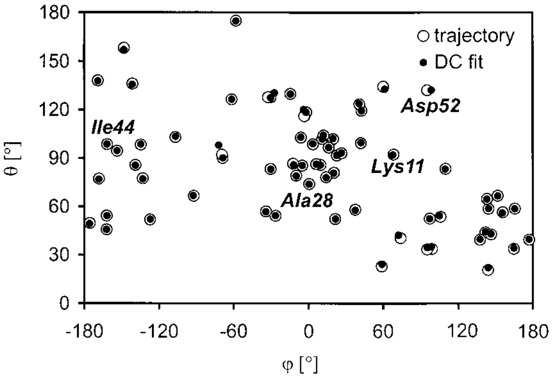


Figure 5. Comparison between average N$-$H^N directions (θ_{av},φ_{av}) determined from the MD trajectory and estimates (θ_{eff},φ_{eff}) determined by solving the linear system of equations (eq 9) followed by the minimization of the sum of eq 10. The figure shows that the estimate is generally within 2° of the exact average.

motional asymmetry, which has become lost in S^2_{LS}. The above procedure was applied to the MD-averaged dipolar couplings, with results shown in Figures 5 and 6. Comparison between average N$-$H^N directions (θ_{av},φ_{av}) computed from the MD trajectory and estimates (θ_{eff},φ_{eff}) determined by solving the overdetermined linear system of equations (eq 9), including the residual dipolar couplings determined for all nine alignment media, followed by the minimization of the sum of eq 10, yields differences that are less than 2° for residues in secondary structural elements (Figure 5). Although the deviation can be larger for loop regions (up to 5°), the (θ_{eff},φ_{eff}) values provide, on average, a much better and more reliable estimate for (θ_{av},φ_{av}) than the (θ,φ) values that are consistent with a single dipolar coupling value. From the extracted $\langle Y_{2M}\rangle$ quantities S^2_{rdc} values were determined according to eq 12. As expected, they turn out to be identical with the S^2_{LS} values determined directly from the trajectory.

Furthermore, asymmetry parameters η, which reflect the amount of asymmetry in reorientational motion (eq 11), were determined from the extracted $\langle Y_{2M}\rangle$ quantities, and they are shown in Figure 6 as a function of the amino acid number. The largest asymmetry is found in ubiquitin for residues in mobile loop regions with η values exceeding 10% (see Figure 6), while in secondary structural elements the asymmetry is typically well below 5%.

6. Conclusion

Intramolecular motions affect residual dipolar couplings in the form of a scaling by a factor λ_{rdc}, which generally also depends on the average orientation of the internuclear vector with respect to the alignment frame. Using a MD simulation as a reference, motional averaging effects of dipolar couplings have been described in detail, and a solution to the inverse problem has been presented that used theoretical dipolar couplings, assuming an optimal set of different alignment tensors. The proposed self-consistent analysis of dipolar couplings should allow the extraction of accurate structural information in terms of average orientations also when applied to experimental data.

Alignment tensors that are fitted to dipolar couplings tend to absorb a significant amount of intramolecular motional effects. If no information on S_{LS} order parameters is available, refinement of a static structural model should be "stopped" at Q values of about 0.05 for secondary structural parts and of about 0.1 for more mobile loop regions. If S_{LS} values are available, refinement to smaller Q values is conceivable, provided that no slower time scale motions are present.

Information on such slower time scale motions that are not reflected in spin relaxation data can be obtained from dipolar couplings measured in different liquid-crystalline media. The results presented here suggest that the combined use of dipolar coupling data sets measured in five or more different environments allows the accurate reconstruction of average positions and the retrieval of unique information on motional averaging of spherical harmonic functions of rank 2, $\langle Y_{2M}\rangle$, that is not readily accessible by S^2_{LS} order parameters obtained from spin relaxation measurements. Besides the longer time scales probed by dipolar couplings, also direct information about motional asymmetry of individual internuclear vectors, expressed by the parameter η, is available. For rapid axially symmetric reorientational motion of an internuclear vector, λ_{rdc} becomes equal to S_{rdc}. The $\langle Y_{2M}\rangle$ quantities have a "model-free" character similar to the model-free order parameters S^2_{LS} extracted from NMR spin relaxation experiments.[21] In analogy to the NMR relaxation field, interpretation of the $\langle Y_{2M}\rangle$ quantities in terms of concrete motional models, such as the 3D GAF model,[30] is possible as a subsequent step of data interpretation.

The basic assumption made here is that the liquid-crystalline environment does not affect biomolecular structure and dynamics. This assumption can be experimentally tested to some extent by verifying that chemical shifts, line widths, and homo- and heteronuclear relaxation parameters do not significantly change with the liquid-crystalline environment. In the case that the average protein structure varies for different alignment media, such variations would be reflected also in the $\langle Y_{2M}\rangle$ quantities.

At present, the requirement of five different liquid-crystalline environments may seem demanding. Moreover, the different alignment tensors should significantly differ with respect to each other in order to minimize the influence of experimental uncertainties in the residual dipolar couplings. Rapid progress in the development and understanding of aligning tools, however, makes it likely that soon a sufficient number of different alignment media will become available that lead to different alignment tensors.[4,5,34-52] Application of the presented protocol to experimental data is currently under way.

After submission of this work, a paper by Tolman et al.[53] appeared, in which the effects of protein motions on dipolar

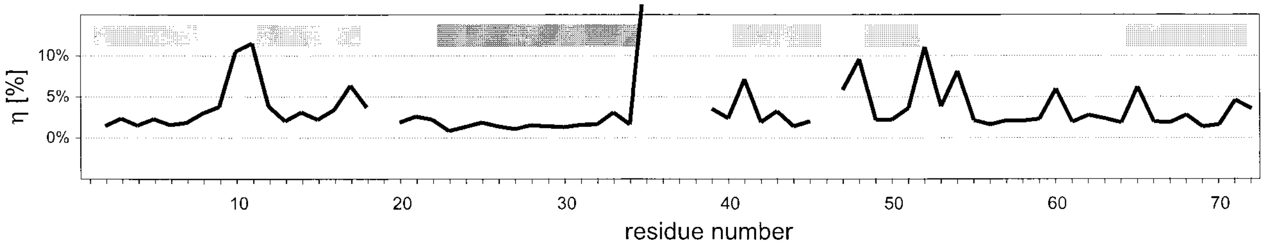


Figure 6. Motional asymmetry parameter η defined in eq 11 for N$-$H^N vectors as a function of the residue number. In regular secondary structure, η varies between 1% and 6%, while in more mobile loop regions the asymmetry can exceed 10%.

couplings measured for a single alignment tensor in ubiquitin were discussed. It differs from the one presented here in the following way. In the paper by Tolman et al., experimental dipolar couplings of different vectors in the peptide plane measured for one alignment medium were interpreted using analytical motional models, whereas in the present work MD-generated dipolar couplings of a single vector measured in multiple alignments were interpreted in a “model-free” way.

Acknowledgment. The authors thank reviewers for useful comments. This work was initiated during J.M.’s stay at Clark University. J.M. and W.P. are recipients of a Kekulé doctoral fellowship of the Fonds der chemischen Industrie, and J.J.P. is a recipient of a Human Frontier Science Program postdoctoral fellowship. This work was supported by the DFG, MPG, the Fond der chemischen Industrie (to C.G.), and NSF Grant MCB-9904875 (to R.B.).

Supporting Information Available: Two tables with structural and dipolar coupling averaging information determined from the 10 ns trajectory of ubiquitin; figure exemplifying the effect of errors in the dipolar couplings on the extraction of motional and structural information (PDF). This material is available free of charge via the Internet at http://pubs.acs.org.

JA010002Z

(34) Prosser, S. R.; Losonczi, J. A.; Shiyanovskaya, I. V. *J. Am. Chem. Soc.* **1998**, *120*, 11010−11011.

(35) Barrientos, L. G.; Dolan, C.; Gronenborn, A. M. *J. Biomol. NMR* **2000**, *16*, 329−337.

(36) Clore, G. M.; Starich, M. R.; Gronenborn, A. M. *J. Am. Chem. Soc.* **1998**, *120*, 10571−10572.

(37) Hansen, M. R.; Mueller, L.; Pardi, A. *Nat. Struct. Biol.* **1998**, *5*, 1065−1074.

(38) Hansen, M. R.; Rance, M.; Pardi, A. *J. Am. Chem. Soc.* **1998**, *120*, 11210−11211.

(39) Ojennus, D. D.; Mitton-Fry, R. M.; Wuttke, D. S. *J. Biomol. NMR* **1999**, *14*, 175−179.

(40) Ottiger, M.; Bax, A. *J. Biomol. NMR* **1998**, *12*, 361−372.

(41) Ottiger, M.; Bax, A. *J. Biomol. NMR* **1999**, *13*, 187−191.

(42) Sanders, C. R.; Schwonek, J. P. *Biochemistry* **1992**, *31*, 8898−8905.

(43) Sanders, C. R.; Hare, B. J.; Howard, K. P.; Prestegard, J. H. *Prog. NMR Spectrosc.* **1994**, *26*, 421−444.

(44) Wang, H.; Eberstadt, M.; Olejniczak, T.; Meadows, R. P.; Fesik, S. W. *J. Biomol. NMR* **1998**, *12*, 443−446.

(45) Cavagnero, S.; Dyson, J. H.; Wright, P. E. *J. Biomol. NMR* **1999**, *13*, 387−391.

(46) Losonczi, J. A.; Prestegard, J. H. *J. Biomol. NMR* **1998**, *12*, 447−451.

(47) Ruckert, M.; Otting, G. *J. Am. Chem. Soc*. **2000**, *122*, 7793−7797.

(48) Flemming, K.; Gray, D.; Prasannan, S.; Matthews, S. *J. Am. Chem. Soc*. **2000**, *122*, 5224−5225.

(49) Tycko, R.; Blanco, F. J.; Ishii, Y. *J. Am. Chem. Soc.* **2000**, *122*, 9340−9341.

(50) Sass, J.; Musco, G.; Stahl, S. J.; Wingfield, P. T.; Grzesiek, S. *J. Biomol. NMR* **2000**, *18*, 303−309.

(51) Koenig, B. W.; Hu, J.-S.; Ottiger, M.; Bose, S.; Hendler, R. W.; Bax, A. *J. Am. Chem. Soc*. **1999**, *121*, 1385−1386.

(52) Sass, J.; Cordier, F.; Hoffmann, A.; Rogowski, M.; Cousin, A.; Omichinski, J. G.; Lowen, H.; Grzesiek, S. *J. Am. Chem. Soc*. **1999**, *121*, 2047−2055.

(53) Tolman, J. R.; Al-Hashimi, H. M.; Kay, L. E.; Prestegard, J. H. *J. Am. Chem. Soc*. **2001**, *123*, 1416−1424.

Fast Determination of ^{13}C NMR Chemical Shifts Using Artificial Neural Networks

J. Meiler,*,† R. Meusinger,‡ and M. Will§

Institute of Organic Chemistry, Marie - Curie - Strasse 11, University of Frankfurt,
D-60439 Frankfurt, Germany, Institute of Organic Chemistry, University of Mainz,
D-55099 Mainz, Germany, and BASF AG Ludwigshafen, D-67056 Ludwigshafen, Germany

Received March 15, 2000

Nine different artificial neural networks were trained with the spherically encoded chemical environments of more than 500 000 carbon atoms to predict their ^{13}C NMR chemical shifts. Based on these results the PC-program “C_shift” was developed which allows the calculation of the ^{13}C NMR spectra of any proposed molecular structure consisting of the covalently bonded elements C, H, N, O, P, S and the halogens. Results were obtained with a mean deviation as low as 1.8 ppm; this accuracy is equivalent to a determination on the basis of a large database but, in a time as short as known from increment calculations, was demonstrated exemplary using the natural agent epothilone A. The artificial neural networks allow simultaneously a precise and fast prediction of a large number of ^{13}C NMR spectra, as needed for high throughput NMR and screening of a substance or spectra libraries.

INTRODUCTION

NMR spectroscopy is undoubtedly one of the most important methods used for structure determination of chemical compounds. In recent years the power of NMR methods and the sophistication of spectrometers increased clearly. This was achieved by a number of new techniques.[1] Only a few of them should be named here. The measurement time was decreased drastically by pulsed field gradients, double or single quantum coherence methods, and finally by the so-called “tubeless NMR”. This is the fitting of conventional high-resolution NMR spectrometers with flow-probes or special micro sample probes. Shorter NMR measuring times are required above all by the high throughput methods developed in combinatorial chemistry. With increasing amounts of spectral data available a new bottleneck has emerged: data analysis. Precise and fast computer programs are necessary to enhance the productivity here. Munk gave recently a vivid presentation of the evolution of computer enhanced structure elucidation exemplary by the structure determination of the antibiotic actinobolin.[2] In the 1960s the computer assisted elucidation of unknown structures required several man years using the structure generator ASSEMBLE. Forty years later, with both, more sophisticated NMR spectroscopic methods and computer software, the time required to determine the structure has been reduced to several days (time for data collection included). Now the program SESAMI generated four candidate structures in 5 min CPU time using only the available 1D and 2D NMR data. Lindel et al. also use both the NMR spectroscopic detectable connections between nuclei and their chemical shifts,[3] in their program COCON (*co*nstitutions from *con*nectivities) which was developed for the generation of all possible constitutions for complex natural products. The efforts which are spent for the development of efficient structure elucidation programs shall be presented here by two other current examples. CISOC-SES[4] is a computer assisted expert system that utilizes 1D and 2D NMR data. Recently the NMR assignment of a biologically active triterpenoid was shown by Peng et al.[5] With the program LSD (*L*ogic for *S*tructure *D*etermination) Nuzillard demonstrated impressively the potential of systematic structure elucidation of small molecules combining modern NMR spectroscopy with artificial intelligence at the example of gibberellic acid.[6] However, in most practical cases an elucidation of a completely unknown structure is not required. The more common type of structure determination is the structure verification. In this case, enough information is available perhaps on the basis of well-known synthetic reaction paths to propose a probable structure. The structure information which is achieved via the chemical shift is usually sufficient here.

NMR Chemical Shift Prediction. Atomic nuclei of one isotope located within one molecule in different chemical environments are shielded differently by their electron cloud. As a result, different resonance frequencies are observed during an NMR experiment exciting these isotopes. If these frequencies are measured as differences to the resonance frequency of an inner standard, they are designated as “chemical shifts”. The chemical shift value combines two advantages for structural analysis. It is an easily obtainable spectral parameter, and its dependence on chemical structure is well-known.[7] The chemical shift of a carbon is, in addition to its state of hybridization, mainly influenced by the kind and number of the bond atoms and by their distances to the observed carbon. The chemical shift of a carbon atom can be influenced by another atom in two different ways: electron interaction over covalent bonds or through space. In solution the second effect appears possibly as a “solvent effect”. However, electron interaction through space is only important if the distance between the observed and influencing atom is small. It has to be considered specially during

* Corresponding author phone: ++49 69 798 29 798; fax: ++49 69 798 29 128; e-mail: mj@org.chemie.uni-frankfurt.de.

† University of Frankfurt.

‡University of Mainz.

§ BASF AG Ludwigshafen.

10.1021/ci000021c CCC: $19.00
Published on Web 08/25/2000

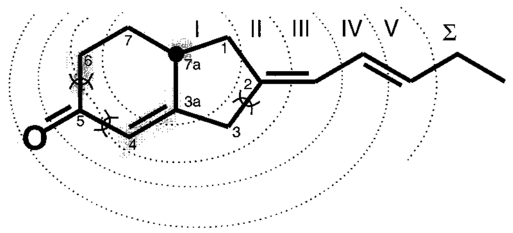


Figure 1. Schematic representation of the spherical division of the chemical environment of an carbon atom in a molecule. The carbon 7a of the substituted 1,2,3,6,7,7a-hexahydroinden-5-one is selected as focus (●). Five spheres starting from this focus are shown (I to V). The beyond environment is summarized in a "sum-sphere" (Σ). The π-contact areas of the focus carbon were marked with a gray background and the ring closure elements with)((see text).

stereochemical analyses. The stronger effect is transmitted via the orbitals following covalent bonds. It depends on the number and the σ- or π-type of the connecting bonds. Focusing on a randomly chosen atom in a molecule, one can consider all other atoms of this molecule as members of spheres. These spheres surround the focused atom, and their number is identical to the number of bonds between the focused atom and the atoms combined in this sphere. In Figure 1 the subdivision of a molecule into five different spheres (Roman numerals) with respect to one carbon atom (focus) is shown. In principle this procedure was already described in 1973 by Bremser with his well-known HOSE code (*h*ierarchically *o*rdered *s*pherical description of *e*nvironment).[8] The influence of the substituents on the chemical shift of the focused carbon normally decreases with an increasing number of bonds between them and therefore also with the increase of the sphere. By σ-bonded atoms the electrons are kept in the molecular orbitals located between two atoms. Therefore the influence decreases here fast with increasing sphere. By far different is the situation in the case of atoms bonded over a conjugated π-electron system. The electron density distribution can be influenced over more than one bond, and larger effects are induced over four, five, or even more bonds.

A further advantage of the NMR chemical shift is the availability of huge amounts of experimental data. Several databases exist today containing hundreds of thousands of chemical shift values in particular for ^{13}C nuclei and the appending information about the chemical environment of these individual carbon. These data are an excellent basis for computer assisted structure determination. Only some examples of electronically stored databases should be mentioned here: SPECINFO,[9] a further development of the ^{13}C spectral database created by Bremser et al.;[10] CSEARCH,[11] a database created by Robien in the beginning of the 1990s; WINDAT,[12] a database created by Trepalin and co-workers; and the CNMR[13] database which was developed few years ago by the ACD company.

Looking closer, ^{13}C NMR spectra databases are statistical tools to establish the relationships between NMR spectral parameters and the chemical environment of individual carbon atoms to propose either chemical structures or spectra. The ^{13}C chemical shift value is extremely suitable for this purpose because of its accuracy, reproducibility, and intelligible structure dependence. Using a large collection of representative molecules for structure determination two approaches are thinkable: starting from an experimentally determined chemical shift value a suitable chemical environment can be determined and starting from a chemical structure the relevant chemical shift can be estimated. Consequently, ^{13}C NMR spectra databases are suitable for three applications: (1) the prediction of NMR parameters for any molecular structure, (2) the verification of existing assignments, and (3) the determination of one or more possible molecular structures corresponding to a ^{13}C NMR spectrum. (This contains the simultaneous assignment of individual NMR signals to the respective carbon of a known structure.)

However, the results can only be offered with a statistical probability, depending on the quantity and quality of the available database entries. In addition to the time required for compiling the data, the greatest problem of database management is the accuracy and reliability of the represented data. In other words, the accuracy of the predicted chemical shifts, chemical environments, or structures cannot be more precise than the stored data. For this reason, careful examinations of shift assignments are just as important as the typographical error checking. Usually, some quality checking procedures are applied. In particular the assignments for quaternary carbons have been twisted in a number of cases, because an unambiguous experimental assignment was often not possible in the past.

Two further possibilities for estimating the assignment between chemical structures and chemical shifts besides database search should be mentioned here: the calculation of the chemical shift values by applications of empirical methods and the computation by quanta chemical procedures. However, quanta chemical computations, e.g. with the IGLO-method (*i*ndividual *g*auge for *l*ocalized *o*rbitals),[14] are relatively extensive. Meanwhile the prediction of ^{13}C NMR spectra is one of the most intensely studied applications of empirical modeling. This method is based on the assumption that the influence of different substituents on the chemical shift of an individual carbon atom can be defined simply by a set of constant values, the "increments". According to the chemical environment of a carbon atom all increments are added up. The chemical environment is defined by the kind and number of the neighboring atoms or atomic groups and by their distances to the considered carbon atom, in this case. The increments themselves were determined by multiple linear regression analysis using data sets of observed chemical shifts from structurally related compounds. The mean advantage of these increments is their simple application and the shortness of its computation. However, the increments are structure class dependent and available only for some substance classes and/or structure groups,[15] e.g. for alkanes,[16] alkenes,[17] substituted benzenes,[18] naphthalenes, and pyridines.[19] Different PC programs were developed in the past which allow the computation of more complex structures, for example the programs SPECTOOL[20] created by Pretzsch et al. and CSPEC2[21] developed by Cheng and Kasehagen. However, the increments employed in these programs were obtained from individual representatives of single substance classes too, and one has to be careful during their application. Possible interactions between several increments are often not considered by these approaches. This was recently shown for the aromatic carbons in substituted benzenes.[22] In such a way, the structure analyst must currently decide between two possibilities: Either for a precise prediction requiring a long time or for the rapid available information afflicted with a larger error. For this reason, several attempts were taken in the past to describe the association between NMR chemical shifts and chemical structure more precisely. Different nonlinear numerical and statistical techniques, such as principal component analysis and artificial neural networks, were used.

Neural Networks. The main advantage of neural networks compared to other methods is their greater capacity to extract general information from a training data set and apply it on presented new data. Neural networks appeared in chemistry at the beginning of the 1990s.[23] The first publications dealing with the determination of ^{13}C NMR chemical shift values using neural networks were also published at this time. Kvasnicka et al. determined the chemical shifts of carbons in monosubstituted benzenes,[24−26] and Doucet and coworkers predicted the shifts of C5 to C9 alkanes.[27] Anker and Jurs[28] trained a three layer neural network with a data set of 391 steroid carbon atoms using the back-propagation learning algorithm. The applied network architecture consisted of 13 input units corresponding to calculated atom-centered descriptors, 40 hidden neurons, and 116 output neurons corresponding to 0.5 ppm chemical shift increments in the range of 8.7−66.7 ppm. The examined results were superior to those achieved with linear regression techniques in every case. Further works followed, all of them with the common characteristic of ^{13}C NMR chemical shift prediction for a group or a class of substances with similar chemical structures, e.g. for alkanes,[29−31] cyclohexanes,[32] alkenes,[33,34] substituted naphthalenes,[35] trisaccharides,[36] dibenzofuranes,[37] ribonucleosides,[38] and substituted benzenes.[22] On the assumption that the influences of substituents on an observed carbon are only similar within an individual substance class, the chemical shift values computed with such a neural network cannot been generalized as desired. Consequently, this is not different from the respective increment systems, certainly more precise but limited in their applicability though. Therefore a fast method for computation of most organic molecular structures is desirable similar as by the use of large databases.

Another approach combining these advantages might be the use of genetic algorithms.[39] However it was not tested yet for this purpose and will not be discussed in this paper.

Molecular Structure Descriptors. In order to calculate the chemical shift values of carbon atoms with neural networks it is necessary to describe the atom types and the chemical environments of all atoms numerically. An optimum must be found for the number of the applied descriptors. On the one hand, the number must be as small as possible for computational reasons; on the other hand, the descriptors must feature clearly the differences in molecular structures. In total 28 descriptors were determined for 28 different atom types and are summarized in Table 1. Additionally, the numbers of representatives of these atom types found in a data base of about 40 000 molecules[9] with up to 100 heavy atoms are indicated. Atom types are derived from element number, hybridization state, and number of bonded hydrogen atoms. As given in Table 1, nine different atom types were defined for the 526 565 carbon atoms which were available for the prediction of ^{13}C NMR chemical shifts in total. In some cases, different atoms were summarized into one type if they only differ in the number of bonded hydrogen atoms. This was necessary since the number of the representatives would have been too small, for example in the case of olefinic carbons with one or two hydrogen atoms (type 6), for sp^2 hybridized nitrogen (type 13), and for sp^3 hybridized phosphorus (type 20) and sulfur (type 22). Two further descriptors were introduced in order to obtain a complete description. One descriptor holds the number of all hydrogen atoms located in the individual sphere (type 29). In order to consider the influence on the ^{13}C NMR chemical shift caused by the formation of rings, a second sum descriptor holds the number of ring closures (type 30). Two different possibilities have to be distinguished: If the ring is closed within one sphere the descriptor increased by two. That is the case in odd-numbered rings (3, 5, ..., $2n + 1$ atoms with $n \epsilon$ N), whereas in even-numbered rings (4, 6, ..., $2n$ atoms with $n \epsilon$ N) the ring is closed over two spheres. In this case the descriptor increased by one for both spheres affected. In the example visualized in Figure 1, the five-membered ring is closed between carbons C-2 and C-3 within the sphere II. In contrast the six-membered ring must be closed between carbons C-4 and C-6, both in sphere II, and carbon C-5 in sphere III.

The chemical environments of the carbons are described by sorting the atoms in spheres as given in Figure 1 and counting the occurrence of every atom type in each sphere. As shown in Figure 1, five spheres (I to V) are formed for every carbon. All atoms in further distances are projected in an additional “sum sphere” (Σ). Consequently, 30 numbers are necessary to describe one sphere, and 180 numbers are necessary for the complete description of the environment of an individual carbon. However, this description is only based on number, kind, and distances of the substituents, which is not sufficient. In addition to this nonspecific enumeration, the descriptors are extended by a so-called “π-contact area”, in order to take the special importance of conjugated π-electronic systems into consideration. Two atoms are in “π-contact” if a conjugated π-electronic system exists from at least one neighbor of the first atom to one neighbor of the other atom. The two atoms themselves are not taken into consideration for this parameter since their belonging to the conjugated system is given by their atom type. Therefore, by description of the environment of carbon C-7a in Figure 1 (focus), the double bonds in the side-chain must not be considered as π-contact, since they do not belong to a conjugated π-electronic system that includes one neighbor of the focused carbon. Otherwise, two sp^3-hybridized carbon atoms have to be considered here. The carbons C-3 and C-6 are neighbors of the conjugated π-electronic system (C3a═C4-C5═O) just as the observed carbon C-7a. In Figure 1 the entire π-contact area of the carbon C-7a is shown as shaded region. The molecular structure descriptors are extended by a second set of 30 numbers for each sphere using the atom types and sum parameters described above in Table 1. But in contrast to the general count in the first set only the atoms being in π-contact with the focused carbon were considered now. Consequently, for each sphere two sets of descriptors result for the encoding of the environment. This is shown in Figure 2. The structure of the descriptors for the calculation of the chemical shift of an individual carbon is represented schematically for the first sphere on the top and for all spheres in the middle of Figure 2.

Table 1: Atom Types (1−28) and "Sum-Types" (29−30) Defined for Describing Atom Environments and Their Frequencies in the Data Set

ID	atom type	frequency of atoms of this type
1	⟩C⟨	19 527
2	⟩CH-	49 556
3	-CH$_2$-	116 175
4	-CH$_3$	73 724
5	=C⟨	50 711
6	=CH-/=CH$_2$	27 556
7	≡C-/≡CH/=C=	3793
8	) ⟩C- (aryl)	68 416
9	) ⟩CH (aryl)	117 107
10	⟩N-	9876
11	-NH-	9115
12	-NH$_2$	3521
13	=N-/=NH	7427
14	≡N	2053
15	-NO$_2$	2688
16	) ⟩N (aryl)	3743
17	-O-	31 641
18	-OH	20 626
19	=O	39 259
20	⟩P-/-PH-/-PH$_2$	383
21	⟩PO-	1053
22	-S-/-SH	4146
23	=S	1214
24	⟩SO$_2$	1789
25	-F	3613
26	-Cl	9585
27	-Br	2718
28	-I	603
29	sum of all hydrogen atoms bond in this sphere	
30	sum of all ring closures in this sphere	

EXPERIMENTAL SECTION

A total of 40 000 molecules were available for the calculations. The 526 565 carbons with well-known ^{13}C NMR chemical shifts and molecular structure assignments[9] were distinguished the nine mentioned atom types (Table 2). All chemical shifts were estimated in deuterated chloroform ($CDCl_3$) or in carbontetrachloride (CCl_4) and refer to tetramethylsilane (TMS) as internal standard. With this, solvent effects were excluded as far as possible. Stereochemical information was not available for the given molecules. The environment of every individual carbon atom was encoded with the descriptors from Table 1. Three hundred sixty descriptors were used as input vector for the neural networks, whereas the individual ^{13}C NMR chemical shift value represents the output. For each of the nine atom types representing a carbon atom an individual neural network was constructed and trained using back-propagation of errors.[40] The amount of available molecules was randomly subdivided into three sets. Ninety percent of data was used for the training of the neural networks. During the training process the percentage of data was increased stepwise up to 90% maximum. A second data set contained 7% of the available data for monitoring. The training and monitoring data sets were used simultaneously. To avoid "overtraining", the iterative training process was stopped if the deviation for the monitoring data set increased again. The hyperbolic tangent (tanh) was found to give best results as a transfer function. The number of hidden neurons was optimized to give the lowest error for the monitoring set of data. Finally, the third data set of 3% randomly selected molecules was used as an independent set for testing the trained networks.

The PC program C_shift[41] which included the trained neural networks and a simply manageable structure-editor for the structure input was written in C++ for Windows 95/98/NT.

RESULTS AND DISCUSSION

In Figure 2 the architecture of neural networks used here is shown schematically. While the number of input and output units was fixed, the number of the hidden neurons had to be determined experimentally. The best results were achieved with a number of 5−20 hidden neurons, depending on the carbon atom type. In Table 2 the descriptions are compiled for the nine different neural networks. The results achieved with these nets for the training and testing data sets are also given in the form of statistical information, respectively. It is shown that a mean deviation of 1.79 ppm and a standard deviation of 2.10 ppm resulted for the more than 15 000 carbons in the test data sets. The accuracy of these calculations is much better in comparison to systems based on fixed increments, especially if complex structures are investigated. In Figure 3 the correlation between the computed and the experimental chemical shift values are shown for all nine carbon atom types obtained from the test data sets. The appropriate correlation coefficients are given in Table 2.

On closer examination of these results it is striking that the largest deviations were obtained for the sp-hybridized carbons (standard deviation of 3.8 ppm for type 7) and for carbons with sp^2 hybridization (standard deviations of 3.7 and 3.6 ppm for types 5 and 6), respectively. For triple bonded carbons the number of examples was relatively small. That causes the higher uncertainty in this calculation.

Furthermore one has to note that all effects influencing the chemical shifts through space were not considered here. Therefore comparable carbons in different stereoisomers do not become distinguishable through their different spatial environments. It is well-known that the different arrangements of substituents either in the *E*- or in the *Z*-direction can influence the chemical shift of the olefinic carbons up to 5 ppm. Influences in this order of magnitude were also observed in different configurational and conformational isomers caused by the dissimilar spatial arrangements of their substituents. Furthermore, one observes an increase of the uncertainty for sp^3 hybridized carbons with an increasing number of non-hydrogen substituents. Here, the standard deviation increases from 1.3 ppm for methyl group carbons up to 3.4 ppm for the quaternary carbons (Table 2). This is expected, since the chemical shift of a methyl group with three fixed substituents in the first sphere is much easier to determine than the chemical shift of a quaternary carbon where four different substituents can interact in the first sphere already.

As already mentioned, the influences of the substituents on the chemical shifts depend on their distances to the observed carbon and on its affiliation to a conjugated π-system. This knowledge can also be obtained analyzing the weights of the trained neural networks. Figure 4 shows the sensitivities of the input units of three different neural networks, subdivided in the six spheres and the types of covalent bonds. These values were determined for every input unit, by variation of their inputs, while all other inputs stay constant at zero. The sensitivity of the selected input is given by the range detected at the output. In Figure 4 the sums of

Table 2: Number of Hidden Neurons, Frequencies of Atom Types (1−9), Correlation Coefficients, Standard Deviations (in ppm), and Mean Deviations (in ppm) for the Training and Test Data Sets of the Nine Different Carbon Atom Types

			training and monitoring data				test data			
ID	atom type	no. hidden neurons	count	corr coeff	std [ppm]	m.d. [ppm]	count	corr coeff	std [ppm]	m.d. [ppm]
1	⟩C⟨	10	18 984	0.995	1.83	1.71	543	0.983	3.42	2.87
2	⟩CH-	10	48 032	0.984	2.37	2.44	1542	0.984	2.39	2.47
3	-CH$_2$-	20	112 639	0.982	1.85	1.51	3536	0.985	2.65	1.39
4	-CH$_3$	20	71 547	0.989	1.47	1.16	2177	0.988	1.30	1.01
5	=C⟨	10	49 139	0.989	2.71	2.71	1518	0.981	3.68	2.72
6	=CH-/=CH$_2$	10	26 691	0.959	3.78	3.28	865	0.966	3.60	3.46
7	≡C-/≡CH/=C=	5	3675	0.995	1.99	2.18	118	0.988	3.80	2.96
8	) ⟩C- (aryl)	20	66 433	0.982	1.88	2.02	1983	0.981	1.72	1.84
9	) ⟩CH (aryl)	20	113 655	0.971	1.57	1.13	3452	0.963	1.81	1.35
all			***510 795***	***0.981***	***1.97***	***1.75***	***15 716***	***0.979***	***2.10***	***1.79***

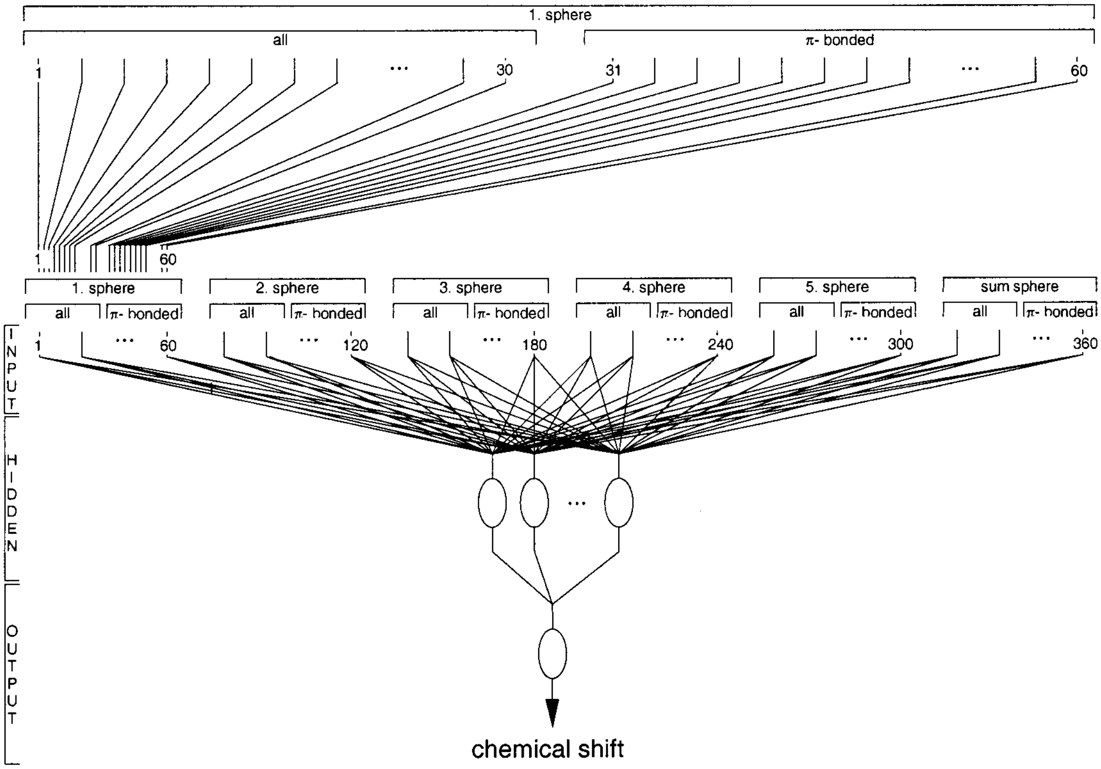


Figure 2. Schema of a three layer neural network for calculating chemical shifts. Sixty numbers divided in two sets result as input for the individual spheres of the environment, respectively. Thirty numbers were used for all substituents (all), and the second set of 30 numbers hold the count of the atoms in the π-contact areas (π). In total 360 input values were required for the five individual spheres and the additional sum sphere. The number of hidden neurons varied between 5 and 20. The single output neuron predicts the chemical shift of the observed carbon atom.

the sensitivities for 30 input neurons for every sphere and every bonding type is shown for three different carbon atom types, respectively. Methyl groups (Figure 4a) shown a fast decrease of the influence of substituents with increasing sphere number. For all atoms beyond the third sphere the influence is low at all. This is in accordance with the knowledge of the substituents induced shifts. In contrast to this the atoms with π-contact show distinctly smaller influences. A significant increase of the sensitivity against atoms with π-contact was observable for the double-substituted sp^2 hybridized carbons (Figure 4b). Since most of the conjugated systems do go beyond the second sphere, a strong decrease was observed after this sphere. Finally the influence of π-conjugated systems can be seen clearly in aromatic systems (Figure 4c). The sensitivity for π-bonded atoms is constantly high over the first three spheres in order to decrease slowly behind the third sphere. Only in the first sphere was observed a smaller influence in comparison to the sum of all substituents.

To test the capability of the method for the determination of ^{13}C NMR chemical shifts, the well-known epothilone A was chosen as an example (Figure 5). Whereas the structure and the NMR spectra of this natural cytotoxic agent is described in detail,[42] this molecule was not included into the database[9] and is therefore used for a comparative determination. Consequently, epothilone A was also not a part of the training or testing data set of the neural networks. In Table 3 all experimentally determined ^{13}C NMR chemical shifts[42] are listed as well as the values calculated by use of the neural networks[41] and the values predicted from a database after spherical coding of all carbons with the HOSE code. Experimental chemical shifts were only available for determination in DMSO. So, the agreements between experimental and computed values were not as precise as

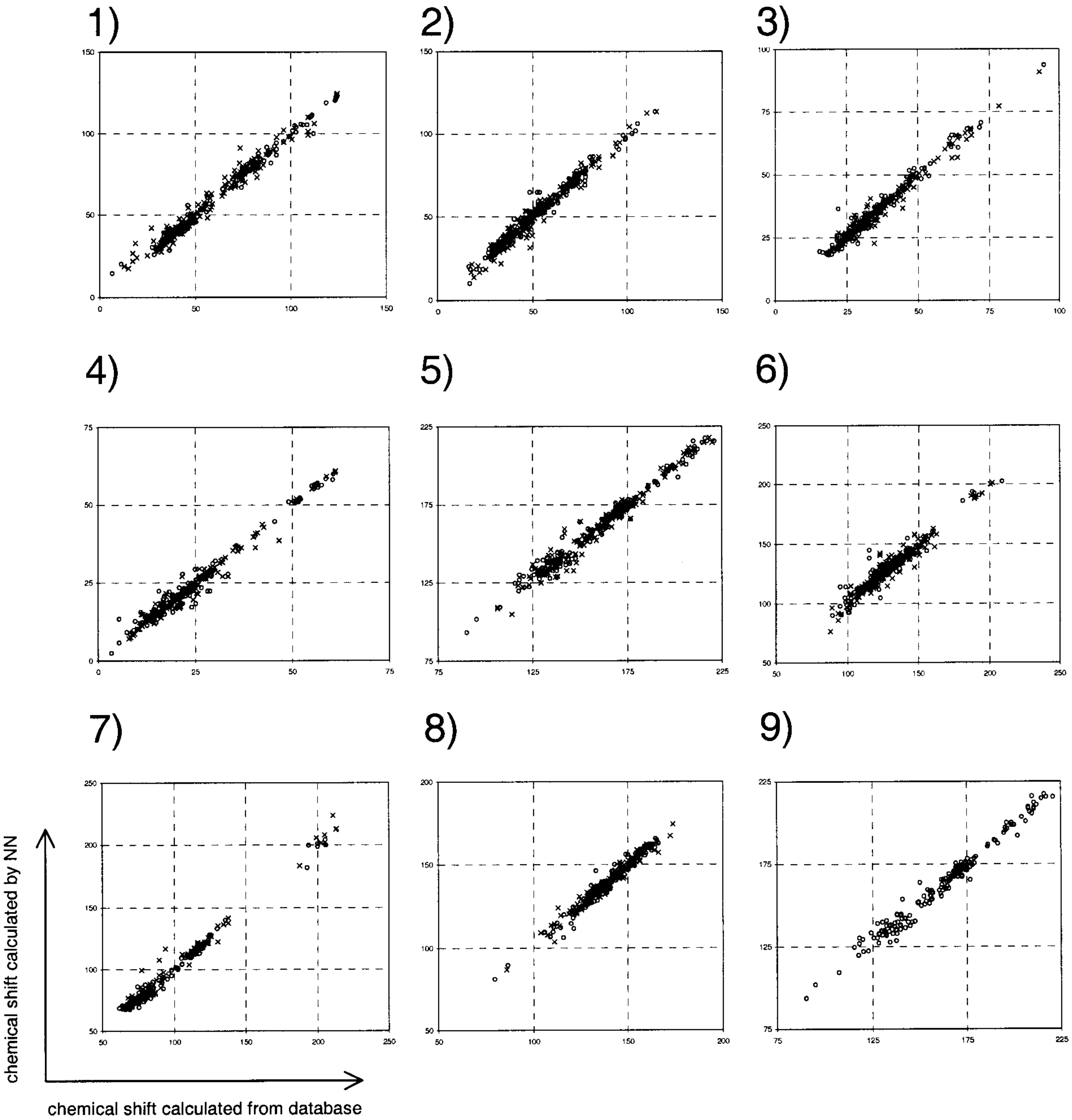


Figure 3. Correlation of chemical shifts of carbons from test data sets determined by use of the data base (x-axis) and by neural networks (y-axis). The results are shown for the nine different types of carbon atoms. The numbering of the diagrams is identical to the ID used in Tables 1 and 2. All data are given in ppm with respect to TMS. For the correlation coefficients and other statistical results look at Table 2.

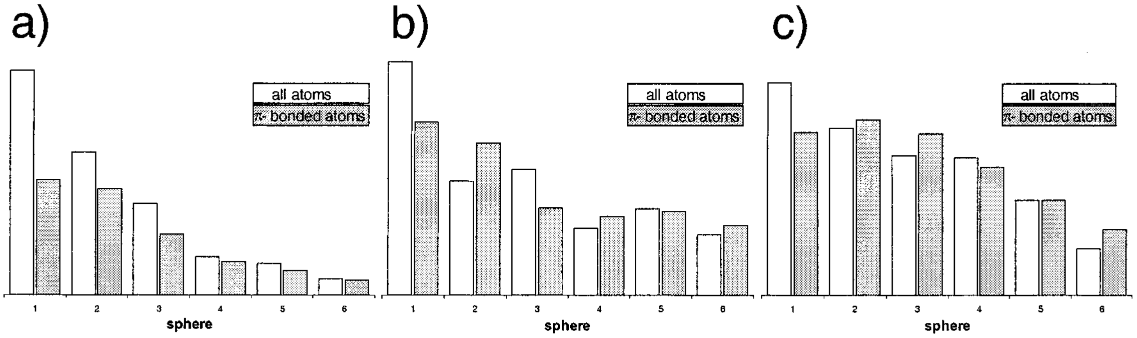


Figure 4. Sum of the sensitivity of the input-units depending on spheres and the types of the covalent bonds for the neural networks trained (a) for methyl carbons (atom type 4), (b) for olefinic carbons (atom type 5), and (c) for aromatic carbons (atom type 8).

expected. However, this restriction concerned all determination procedures uniformly. The standard deviations found here were 2.5 ppm for the neural networks and 2.4 ppm for the determination using the HOSE code description. There-

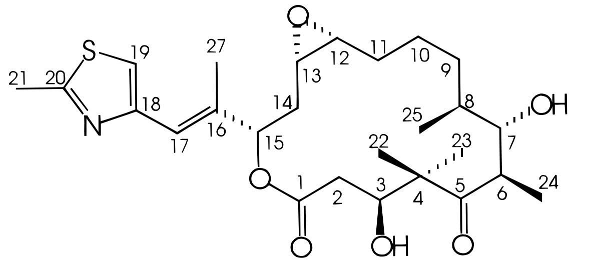


Figure 5. Structure of epothilone A.

Table 3: ^{13}C NMR Chemical Shifts (in ppm) of Epothilone A[a]

	chemical shift [ppm]			
atom ID (acc. Figure 5)	exptl	neural net	specinfo	spectool
1	170.2	173.3	172.2	172.0
2	38.3	36.6	38.9	36.1
3	70.9	73.7	72.8	73.4
4	53.0	52.6	52.8	53.6
5	216.9	216.9	215.9	217.0
6	45.2	45.3	42.3	44.6
7	75.7	75.4	73.5	75.3
8	35.3	33.5	35.0	34.3
9	29.5	32.3	27.7	30.7
10	23.3	24.8	23.7	24.0
11	26.6	30.9	29.9	31.6
12	56.4	61.1	57.9	56.6
13	54.3	58.0	53.6	50.9
14	31.9	35.1	31.0	33.1
15	76.2	76.4	74.7	78.7
16	137.1	137.8	138.7	143.0
17	118.9	122.2	114.2	120.6
18	151.8	159.5	149.2	142.5
19	117.5	116.4	123.1	118.7
20	164.0	169.5	165.7	165.9
21	18.6	18.9	18.9	16.2
22	22.4	21.6	20.6	14.6
23	20.6	21.6	20.6	14.6
24	16.5	15.7	14.6	8.4
25	18.6	15.8	13.2	13.9
27	14.0	16.7	16.6	10.7
mean dev:		2.2	1.9	2.9
std dev:		2.5	2.4	3.8
corr coeff:		0.9991	0.9991	0.9980

[a] The shifts were determined experimentally in DMSO and calculated by the neural network,[41] by the HOSE code based estimation performed with the Specinfo data-base,[9] and by increments (SpecTool).[20]

fore, the results are of a comparable quality. But the advantage of the neural network approach was its drastically shorter computation time. The result was computed 1000 times faster compared to the HOSE code based determination for which the entire database had to be searched. The chemical shifts determined with the increments show a clearly larger deviation (3.8 ppm), as expected. A more detailed look reveals the largest deviations for the adjacent carbons to the epozid (C-11 to C-14), the carbons included in the conjugated π-system of the five-membered ring (C-16 to C-20) and for the methyl groups (C-21 to C-27). Obviously all three methods have difficulties with the conjugated π-system in the heterocyclic ring. While the adjacent carbons to the epoxid were computed quite precisely with the database, relatively large deviations are observed here for the results obtained with the neural networks. This could be caused by stereochemical influences in the high-flexible alicyclic part of the molecule.

CONCLUSIONS

Artificial neural networks offer a fast and accurate possibility to calculate ^{13}C NMR chemical shifts of organic compounds. While the method has an essentially lower standard deviation than increment methods, the computation time is 1000 times faster than using comparable accurate database predictions of chemical shifts. This makes neural networks ideal for screening results of stucture generators or checking the entries of a database. If a large number of ^{13}C NMR spectra has to be predicted or a fast and easy check of a structure is necessary, this approach is a very good opportunity. Moreover the large amount of disk space for saving the database or long time for loading data from external computers are no longer necessary. It would also be possible to perform the training of the network interactively, so that every scientist could create a network specialized in the groups of substances he or she is dealing with.

ACKNOWLEDGMENT

J.M. thanks the “Fond der chemischen Industrie” for a Kekulé stipend.

REFERENCES AND NOTES

(1) Sattler, M.; Schleucher, J.; Griesinger, C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *34*, 93−158.
(2) Munk, E. M. Computer-Based Structure Determination: Then and Now. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 997−1009.
(3) Lindel, T., Junker, J., Köck, M. NMR-Guided Constitutional Analysis of Organic Compounds Employing the Computer Program COCON. *Eur. J. Org. Chem.* **1938**, *3*, 573−577. (A cocon version is available in the Internet under *http://cocon.org.chemie.uni-frankfurt.de*.)
(4) Peng, C.; Yuan, S. G.; Zheng, C. Z.; Hui, Y. Z.; Wu, H. M.; Ma, K.; Han, X. W Application of expert system CISOC-SES to the Structure Elucidation of Complex Natural Products. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 814−819.
(5) Peng, C.; Bodenhausen, G.; Qiu, S. X.; Fong, H. H. S.; Farnsworth, N. R.; Yuan, S. G.; Zheng, C. Z. Computer-assisted structure elucidation: Application of CISOC-SES to the resonance assignment and structure generation of betulinic acid. *Magn. Reson. Chem.* **1998**, *36*, 267−278.
(6) Nuzillard, J. M. Computer-assisted structure determination of Organic Molecules. *J. Chim. Phys.-Chim. Biol.* **1998**, *95*, 169−177.
(7) Pihlaja, K.; Kleinpeter, E. *Carbon-13 NMR Chemical Shifts in Structural and Stereochemical Analysis*; VCH Verlagsgesellschaft: Weinheim, 1994.
(8) Bremser, W. HOSE − a novel substructure code. *Anal. Chim. Acta* **1973**, *103*, 355−365.
(9) *SpecInfo database*; Chemical Concepts, STN: Karlsruhe.
(10) Bremser, W.; Ernst, L.; Franke, B.; Gerhards, R.; Hardt, A. *Carbon-13 NMR Spectral Data*; Verlag Chemie: Weinheim, 1981.
(11) Robien, W. *CSEARCH*; http://felix.orc.univie.ac.at/~wr/csearch_server_info.html.
(12) Trepalin, S. V.; Yarkov, A. V.; Dolmatova, L. M.; Zefirov, N. S.; Finch, S. A. E. WINDAT − an NMR Database Compilation Tool, User Interface, and Spectrum Libraries for Personal Computers. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 405−411.
(13) *CNMR database*; Advanced Chemistry Development Inc.: 133 Richmond Street West, Suite 605 Toronto, Ontario, Canada M5H 2L3.
(14) Schindler, M.; Kutzelnigg Theory of magnetic susceptibilities and NMR chemical shifts in terms of localized quantities. II. Application to some simple molecules. W. *J. Chem. Phys.* **1982**, *76*, 1919−1933.
(15) Pretsch, E.; Clerc, J. T.; Seibel, J.; Simon, W. *Tabellen zur Strukturaufklärung organischer Verbindungen mit spektroskopischen Methoden*; Springer-Verlag: Berlin, 1981.
(16) Fürst, A.; Pretsch, E.; Robien, W. Comprehensive Parameter Set for the Prediction of the ^{13}C NMR Chemical Shifts of sp3-hybridized Carbon Atoms in Organic Compounds. *Anal. Chim. Acta* **1990**, *233*, 213−222.
(17) Pretsch, E.; Fürst, A.; Robien, W. Parameter set for the Prediction of the ^{13}C NMR Chemical Shifts of sp^2- and sp-hybridized Carbon Atoms in Organic Compounds. *Anal. Chim. Acta* **1991**, *248*, 415−428.

(18) Ewing, D. 13C Substituent Effects in Monosubstituted Benzenes. *Org. Magn. Reson.* **1979**, *12*, 499−524.

(19) Thomas, S.; Strohl, D.; Kleinpeter, E. Computer Application of an Incremental System for Calculating the 13C NMR Spectra of Aromatic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 725−729.

(20) Gloor, A., Cadisch, M.; Bürgin-Schaller, R.; Farkas, M.; Kocsis, T.; Clerc, J. T.; Pretsch, E.; Aeschimann, R.; Badertscher, M.; Brodmeier, T.; Fürst, A.; Hediger, H.-J.; Junghans, M.; Kubinyi, H.; Munk, M. E.; Schriber, H.; Wegmann, D. *SpecTool: A Hypermedia Book for Structure Elucidation of Organic Compounds using Spectroscopic Methods*; Chemical Concepts: Weinheim, 1994.

(21) Cheng, H. N.; Kasehagen, L. J. Integrated Approach for C-13 Nuclear Magnetic Resonance Shift Prediction, Spectral Simulation and Library Search. *Anal. Chim. Acta* **1994**, *285*, 223−235.

(22) Meiler, J.; Meusinger, R.; Will, M. Neural Network Prediction of ^{13}C NMR Chemical Shifts of Substituted Benzenes. *Monatsh. Chem./Chem. Monthly* **1999**, *130*, 1089−1095.

(23) Burns J. A.; Whitesides, G. M. Feed-Forward Neural Networks in Chemistry: Mathematical Systems for Classification and Pattern Recognition. *Chem. Rev.* **1993**, *93*, 2583−2601.

(24) Kvasnicka, V.; Sklenak, S.; Pospichal, J. Application of Recurrent Neural Networks in Chemistry. Prediction and Classification of 13C NMR Chemical Shifts in a Series of Monosubstituted Benzenes. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 742−747.

(25) Kvasnicka, V.; Sklenak, S.; Pospichal, J. Application of neural networks with feedback connections in chemistry: prediction of 13C NMR chemical shifts in a series of monosubstituted benzenes. *J. Mol. Struct. (Theochem.)* **1992**, *277*, 87−107.

(26) Sklenak, S.; Kvasnicka, V.; Pospichal, J. Prediction of 13C NMR Chemical Shifts by Neural Networks in a Series of Monosubstituted Benzenes. *Chem. Papers* **1994**, *48*, 135−140.

(27) Doucet, J. P.; Panaye, A.; Feuilleaubois, E.; Ladd, P. Neural networks and 13C NMR shift prediction. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 320−324.

(28) Anker, L. S.; Jurs, P. C. Prediction of Carbon-13 Nuclear Magnetic Resonance Chemical Shifts by Artificial Neural Networks. *Anal. Chem.* **1992**, *64*, 1157−1164.

(29) Svozil, D.; Pospichal, J.; Kvasnicka, V. Neural Network Prediction of Carbon-13 NMR Chemical Shifts of Alkanes. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 924−928.

(30) Svozil, D.; Kvasnicka, V.; Pospichal, J. Introduction to multilayer feed-forward neural networks. *Chemom. Intell. Lab. Syst.* **1997**, *39*, 43−62.

(31) Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D. 13C NMR chemical shift sum prediction for alkanes using neural networks. *Computers Chem.* **1997**, *21*, 437−443.

(32) Panaye, A.; Doucet, J. P.; Fan, B. T.; Feuilleaubois, E.; Azzouzi, S. R. E. Artificial neural network simulation of 13C NMR shifts for methyl-substituted cyclohexanes. *Chemom. Intell. Lab. Syst.* **1994**, *24*, 129−135.

(33) Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J. P. 13C NMR Chemical Shift Prediction of sp2 Carbon Atoms in Acyclic Alkenes Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 644−653.

(34) Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J. P. 13C NMR Chemical Shift Prediction of the sp3 Carbon Atoms in the α Position Relative to the Double Bond in Acyclic Alkenes. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 587−598.

(35) Thomas. S.; Kleinpeter, E. Assignment of the ^{13}C NMR chemical shifts of substituted naphthalenes from charge density with an artificial neural network. *J. Prakt. Chem./Chem.-Ztg.* **1995**, *337*, 504−507.

(36) Clouser, D. L.; Jurs, P. C. Simulation of the 13C nuclear magnetic resonance spectra of trisaccharides using multiple linear regression analysis and neural networks. *Carbohydr. Res.* **1995**, *271*, 65−77.

(37) Clouser, D. L.; Jurs, P. C. The simulation of 13C nuclear magnetic resonance spectra of dibenzofurans using multiple linear regression analysis and neural networks. *Anal. Chim. Acta* **1996**, *321*, 127−135.

(38) Clouser, D. L.; Jurs, P. C. Simulation of the 13C Nuclear Magnetic Resonance Spectra of Ribonucleosides Using Multiple Linear Regression Analysis and Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 168−172.

(39) Meusinger, R.; Moros, R. Determination of Quantitative Structure-Octane Rating Relationships of Hydrocarbons by Genetic Algorithms. *Chemom. Intell. Lab. Syst.* **1999**, *46*, 67−78.

(40) Meiler, J. Smart; http://www.krypton.org.uni-frankfurt.de/∼mj, 1998.

(41) Meiler, J. C_Shift; http://www.krypton.org.uni-frankfurt.de/∼mj, 1999.

(42) Höfle, G.; Bedorf, N.; Steinmetz, H.; Reichenback, H.; Gerth, K. Epothilon A and B - Novel 16-membered macrolides with cytotoxic activity: Isolation, crystal structure, and conformation in solution. *Angew. Chem. Int. Ed. Engl.* **1996**, *35*, 1567−1569.

CI000021C

Validation of Structural Proposals by Substructure Analysis and ^{13}C-NMR Chemical Shift Prediction

Jens Meiler[†] and Matthias Köck[‡,*]

Alfred-Wegener-Institut für Polar- und Meeresforschung, Am Handelshafen 12, D-27570 Bremerhaven, Germany and Institut für Organische Chemie, J. W. Goethe-Universität, Marie-Curie-Strasse 11, D-60439 Frankfurt, Germany

Abstract: Systematical constitutional analyses of four model compounds were carried out using the NMR-based structure generator COCON. The resulting sets of up to 28,000 structural proposals were filtered, analyzed and validated using a substructure analysis and a fast method to calculate $\delta(^{13}C)$ as well as a combination of both approaches. A fast and accurate tool for screening large sets of structural proposals is established yielding useful information about quality and diversity of the structures. Structure elucidation of organic compounds can be performed in a highly automated way by combining COCON with these new approaches.

Keywords: NMR, structure elucidation, COCON, ^{13}C-NMR chemical shift calculation, substructure analysis, artificial neural network

* To whom correspondence should be addressed. Phone: +49-471-4831-1497.
Fax: +49-471-4831-1425. E-mail: mkoeck@awi-bremerhaven.de
[†] *Institut für Organische Chemie, Frankfurt*
[‡] *Alfred-Wegener-Institut für Polar- und Meeresforschung, Bremerhaven*

Introduction

NMR-based structure generators are of special interest for the analysis of under-determined systems in respect to NMR correlation data as such as proton-poor compounds. Because of the small number of protons and therefore small number of NMR correlations a rather large number of constitutions is in accordance with the NMR correlation data. If only ^{1}H,^{1}H-COSY and ^{1}H,^{13}C-HMBC data are available for these compounds (e. g. because the amount of material is not sufficient to obtain a ^{1}H,^{15}N-HMBC or ^{13}C,^{13}C correlation data) the number of theoretically possible structures easily exceeds 10,000. This large number leads to problems in the analysis. Therefore, computer-assisted methods are required to validate these results. In a recent paper we have demonstrated the successful application of calculating the ^{13}C-NMR chemical shifts with SpecEdit[1] for ranking the structural proposals.[2] For large data sets (>10,000 structural proposals) these calculations are rather time consuming because of an approximate calculation time of 1s per structure (SpecEdit). There are two possible methods available to circumvent this problem (see Fig. 1):

- an essential acceleration in the calculation of $\delta(^{13}C)$
- systematic investigation of the generated structures by substructure analysis

Neural networks have become an effective method in chemistry as a flexible tool for data handling and analysis. Several examples of neural networks were published for the analysis[6-18] and the prediction of NMR spectra[19-30]. A neural network approach is used here to ensure a fast and accurate chemical shift prediction.[5] Algorithms for determining common substructures in a set of structural proposals are still subject of development[31-35]. A new implementation of a substructure analysis based on the comparison of atomic environments will be introduced. A substructure analysis allows to investigate the diversity of a set of structural proposals. Common and highly populated substructures can be extracted. In a combination of substructure analysis and chemical shift prediction, ^{13}C-NMR chemical shifts are calculated for the substructures by averaging the chemical shifts of combined structures.

To demonstrate the efficiency of these approaches several sets of structural proposals generated by COCON[2-4] are investigated. COCON uses connectivity information from 2D NMR spectroscopy to generate all possible structures of a molecule which fulfill this information. The theoretical background and several applications of the NMR-based structure generator COCON were already described in the literature.[2-4] Here we want to focus on the analysis of structural proposals generated by COCON. Especially for large sets of structural proposals, as observed for underdetermined systems, an additional analysis becomes necessary. It will be demonstrated that both, the comparison of calculated δ(^{13}C) with experimental data as well as a substructure analysis are efficient tools to perform this validation.

The data sets of four model compounds are used as input for COCON calculations. The generated structure ensembles contained from 33 up to over 28,000 structural proposals. These are investigated with a fast method for the calculation of δ(^{13}C) and a substructure analysis as well as with a combination of both approaches to demonstrate the possibility to reduce the number of probable structure proposals dramatically. In all examples the correct structure is ranked within the first 2‰ in a hit list of all structures.

Insert Scheme

The first example is oroidin (**1**) which already served as model system for other COCON investigations.[4] Oroidin (**1**) was first isolated in 1971 from the marine sponge *Agelas oroides*[36-38] and is a key intermediate of the bromopyrrole alkaloid family. The correlation data used here were described in ref. [4]. Also manzacidin A (**2**) was used as a model compound for COCON calculations before.[3] It was first isolated in 1991 from the marine sponge *Hymeniacidon* sp., the published correlation data of this investigation were used as input for the COCON analysis.[39] 5-deoxyenterocin (**3**) was isolated from a tunicate of the genus *Didemnum* in 1996.[40] The published correlation data served as input for the COCON calculation. Ascididemin (**4**), a pyridoacridine alkaloid, was first isolated in 1988 from the tunicate *Didemnum* sp.. It stands as an example of a proton-poor compound for which only ^{1}H,^{1}H-COSY and ^{1}H,^{13}C-HMBC correlations are available. A theoretical data[A] set was used for this example in order to demonstrate the

[A] Theoretical data set means that all $^3J_{HH}$ and $^2J_{CH}$/$^3J_{CH}$ correlations of the given constitution of ascididemin (**4**) were extracted.

consequences of a small number of protons for a structural analysis. This results in a low number of correlation data (even if all theoretical possible correlations are obtained) leading to a large number of structural proposals.

Calculation of ^{13}C chemical shift

NMR-based structure generators as COCON often produce thousands of possible solutions which can be impossibly validated by hand. Therefore, there is a need for fast and accurate filters. The inclusion of orthogonal (not correlated) information with respect to the experimental data used in the structure generator optimizes the efficiency of such filters. COCON uses connectivity information from NMR spectra for the generation of structural proposals. The orthogonal information used in the following examples is the ^{13}C-NMR chemical shift.[B] This leads to large deviations between the experimental and the predicted chemical shifts for many carbons in the generated structures. This wide distribution of deviations provides an effective filter. A fast and accurate method for determining ^{13}C chemical shifts of organic substances is available using artificial neural networks[5].

So far the ^{13}C chemical shift prediction was carried out using large computer stored databases or incremental methods. While databases like Specinfo[42], SpecEdit[1] or CSearch [43, 44] provide an accurate shift prediction they have rather slow calculation times and are often not available. Incremental methods[45-50] are usually very fast but lead to large deviations for complex structures because interactions between substituents are not considered[30]. Combining the advantages of both methods artificial neural networks are used in this approach. They allow to calculate ^{13}C chemical shift as fast as using incremental methods but are about 10^3 times faster in comparison to a database search with no loss in accuracy. The details of this approach are decribed in ref. [5]. Therefore, only a brief summary of this approach is given.

[B] Chemical shift information is only considered very crudely in COCON. The chemical shift rules of COCON are as follows: a) C=S and C=O bonds are forbidden if δ_C < 150 ppm, b) aliphatic C–O bonds are forbidden if δ_C < 45 ppm, c) olefinic C–O bonds are forbidden if δ_C < 130 ppm, d) olefinic C–N bonds are forbidden if δ_C < 105 ppm and e) Methyl–C bonds are forbidden if δ_{CH3} > 35 ppm.

From the spherical environment of a carbon atom[C] a numerical code is derived containing the number of atoms, the atom type and the hybridization state. The first five spheres and an additional sum sphere (which considers all atoms from the sixth to higher spheres) are taken into consideration. All atoms are subdivided in 28 atom types according to their order number, hybridisation state and the number of attached protons. In each sphere the frequency of atoms for every atom type, the number of protons and the number of ring closures is determined, first for all atoms and in the next step only for atoms belonging to a conjugated π-electronical system. Therefore, the environmental code of a carbon atom consists of (28+2) parameters for each of the six spheres and for σ- and π-bonded atoms which leads to 360 numbers. Nine out of the 28 defined atom types are carbon atoms.[D] For each of them an individual neural network was established which uses a vector with 360 numbers as input and predicts the chemical shift. After training these neural networks with 40,000 compounds from the Specinfo data base the average deviation of the ^{13}C-NMR chemical shift calculation was determined to be 1.8 ppm for an independent data set of 5,000 molecules (depending on the atom type and the hybridization state of the carbon atom).[5]

Substructure analysis

Computer programs like COCON often generate similar structures with equivalent basic structural elements (e.g. closed ring systems) but a different arrangement of substituents. To separate both information, a substructure analysis is of special interest. This allows to investigate a small number of basic common substructures and the different substitution patterns. For a chemist it would be very time consuming to perform this analysis by hand. But it would be an important information to find e.g. 10 common substructures out of 500 generated constitutions.

Furthermore, this analysis can be easily combined with a ^{13}C chemical shift calculation in two ways:

[C] The environment of a carbon atom can be subdivided in spheres. This is carried out by counting the minimal number of bonds between the carbon atom of interest and every other atom, respectively.

[D] The nine carbon atom types are: >C<, >CH-, -CH_2-, -CH_3, =C<, =CH- (=CH_2), -=C- (-=CH, =C=).

a) only the generated structures with the smallest ^{13}C chemical shift deviation to the experimental data could be used. This might become necessary if the number of generated constitutions is too large to perform a full substructure analysis or the resulting set of substructures would become too complex for further investigations.

b) it is possible to calculate an average chemical shift value for every carbon atom in a substructure. This is carried out by averaging the chemical shift values of the corresponding carbon atoms in molecules which contain this particular substructure. As we will show later, this averaging leads to smaller deviations of the chemical shift to the experimental values if the substructure is a part of the correct structure.

The set of substructures is calculated by combining all structural proposals pair wise. For every pair of molecules the largest common substructures is computed. A substructure of two molecules is defined that all superimposed atoms within the substructure are: a) of the same element type (C, N, O ...) and b) are connected by exactly the same bond types (single, double, triple or aromatic). If the investigated ensemble contains n molecules, n (n – 1) / 2 substructures have to be generated. As mentioned, this can become very time consuming for large sets of molecules. If necessary the number of proposals considered for the substructure analysis can be limited to the molecules with the lowest $\delta(^{13}C)$ deviation to the experimental values. The resulting ensemble of substructures has to be further analyzed with respect to two questions:

- Are substructures generated more than once?
- Exist further structures that include a generated substructure?

Every substructure is taken into consideration only once. A newly generated substructure is tested to be a part of every structural proposal. Together with every generated substructure links are saved that point to all molecules and also to all other substructures that contain this particular fragment.

The key function of this analysis is a procedure that generates the largest common substructure from two given structures. The largest common substructure can be found by an algorithm that associates atoms of the first structure to atoms of the second structure. Two atoms can be potentially associated, if they have the same element number and are connected to all other atoms of the new substructure by identical bond types. Due to this definition more than one association can be usually found for two molecules. The association with the maximal number of atoms is the largest common substructure. Hydrogen atoms are not taken into consideration explicitly. Figure 2 gives two molecules with a bold marked largest common substructure as example.

The problem to find the largest possible association is a tree search type analysis in a mathematical sense. Nodes of two trees have to be assigned to each other. Figure 2 illustrates this problem. Similar to the earlier discussed spherical definition of an atomic environment, a recursive function is used for this purpose which starts from one atom and compares its environment sphere by sphere with the environment of another atom (Figure 2). Two atoms of the same element number are selected from both molecules and superimposed to become the first part of the new substructure. Its neighbors are assigned now sphere by sphere. If the element type (C, N, O, ...) and the bond type (single, double, triple or aromatic) is equivalent, the atom is added to the substructure. The substructure increases until no further superimposition is possible.

However, some special problems have to be considered performing this assignment of two structures. The selection of the two starting atoms influences the result of the procedure and has therefore to be changed incrementally over all possible atom-atom combinations in an outer loop. Furthermore, more than one possibility in the recursive sphere by sphere assignment can occur and all possibilities have to be tested in these cases. This procedure is tree based and has therefore to consider the mathematically special case of ring closures in this tree.

During the development and the testing of this procedure it turned out that additional options are necessary which allow the generation of “intelligent” sets of substructures and limiting their number to a reasonable amount. Therefore, several options are introduced:

- Defining a minimum number of atoms in a substructure.
- Defining a minimum number of rings in a substructure, to prefer substructures that include large closed ring systems (acyclic substructures are not very hepful for polycyclic molecules).
- Defining a maximum number of “non ring atoms” in a substructure, to prefer substructures with large closed ring systems without substituents.
- Defining a minimum number of molecules combined in a substructure, to find substructures that are common in many of the generated structures.
- Analyzing only a part of all structures (for example the first 1% of the structural proposals with the lowest deviation from the experimental ^{13}C-NMR spectrum) to reduce the number of the generated substructures.
- Generating reduced sets of substructures by selecting small set of substructures out of all generated substructures. This selection is preformed in order to find the smallest “complete” substructure ensemble which covers every generated molecule with exactly one substructure.

Several options for the visualization of the substructure analysis are introduced and used for the described problems:

- Sorting the results by the deviation from the ^{13}C-NMR spectrum, to rank the substructures according to their probability to occur in the correct structure.
- Sorting the results by the number of atoms in the substructures, to rank the substructures according to their size.
- Sorting the results by the number of molecules that contain a particular substructure, to rank the substructures according to their frequency of occurrence.
- Reorganizing the substructures as a tree. This reorganization is performed by validating the relations between the substructures, by testing if a substructure is part of another substructure. The result is a plot which starts with small substructures in a first generation. All substructures containing this small substructures are given in a second generation and so on until the last generation of substructures is reached and the generated structures that contain these

substructures are given. This tree or a part of it allows to analyze the relations between the substructures (see figures 6, 7, 9 and 12).

The ^{13}C-NMR chemical shift calculation as well as the substructure analysis are combined in the program *"Analyze"*[51].

Results and Discussion

Oroidin (**1**) was already intensively discussed in the literature.[4] It is used here to demonstrate both approaches on a small ensemble. The results of the substructure analysis for **1** can therefore be validated by hand allowing the approach to be tested and optimized. COCON generates 33 structural proposals for the experimental data set of oroidin (**1**) including: 6 ^{1}H,^{1}H-COSY, 23 ^{1}H,^{13}C-HMBC and 8 1,1-ADEQUATE correlations. The substructure analysis was applied to the 33 structures and identified 10 different substructures (Fig. 2). This result is in accordance with a substructure analysis carried out by hand.[4] As mentioned before the substructure analysis can be combined with the carbon chemical shift calculation (Fig. 1). The carbon chemical shifts for all substructure families of **1** are calculated and used for ranking (see Fig. 4). Two substructures (**1**-S1 and **1**-S2) are clearly favored over the others. Substructures with a small deviation of their ^{13}C-NMR spectrum with respect to the experimental spectrum have a high probability to be a part of the correct structure, since statistical errors in the chemical shift deviation are averaged out combining large numbers of structures.[E] The two substructures differ in the connection of the pyrrole with the other part of molecule. In **1**-S1 the pyrrole is connected to the carbonyl carbon of the amide whereas in **1**-S2 it is connected to the imidazole. Both could be distinguished by their different UV absorption. The final substructure family (**1**-S1) consists of four structures (**1**-27, **1**-29, **1**-30, **1**-32, see Fig. 5). Structural proposals **1**-30 and **1**-32 which contain aminopyrrole and bromoimidazole substructures can be neglected. The distinction of the 3,5-dibromopyrrole (**1**-29) versus the 4,5-dibromopyrrole (**1**-27) is possible by comparison of $\delta(^{13}C)$ of C-2, C-3, C-4 and C-5. The correct structure of oroidin (**1**, **1**-27) shows the lowest ^{13}C chemical shift deviation in this family (**1**-S1) and the second lowest of all 33 structural proposals. The absolute ^{13}C chemical shift deviations are rather high for this particular ensemble (from 7.8 to 20.5 ppm). However, only the relative information is of interest for this analysis. The relative large absolute $\delta(^{13}C)$ deviation has only a minor influence on the result.

[E] If a multiple determination of a property value is possible, it is a known fact from statistical analysis that the precision and the accuracy of the prediction increases. In the described approach this fact leads to small deviations in the $\delta(^{13}C)$ prediction for substructures which are part of the correct solution. Since in a substructure several molecular structures are combined, $\delta(^{13}C)$ becomes the average value of $\delta(^{13}C)$ calculated for the individual molecules and tends to approach the experimental value.

For the experimental data set of manzacidin A (**2**) including 6 ^{1}H,^{1}H-COSY and 18 ^{1}H,^{13}C-HMBC correlations COCON generated 190 structural proposals. The results of the ^{13}C chemical shift calculation for the best 10 structures are given in Table 2. A part of the generated substructure tree of the manzacidin A (**2**) data set including all 190 structures is shown in Fig. 6. The requirements for the substructures are: a) the minimum number of atoms per substructure is 2, b) the minimum number of molecules per substructure is 8 and c) the substructures contain not more than 2 atoms that are not part of a ring system. The substructure analysis identified all different ring systems present in the ensemble. The chemical shift deviation for both ring systems (pyrrole and tetrahydropyrimidine) of manzacidin A (**2**) are smaller than for other possibilities. The pyrrole subunit is found to be a part of 52 molecules and the next generation of substructures is given here. The 3-bromopyrrole subunit of manzacidin A (**2**) is clearly preferred by its ^{13}C chemical shift deviation. Note, that the sum of structures combined in two substructures in a subtree can be larger than the number of structures given at the root, since often two or even more substructures of a tree are present in one structure at the same time.

COCON proposed 82 structures for the experimental data set of 5-deoxyenterocin (**3**) which consists of 4 ^{1}H,^{1}H-COSY (plus fixed phenyl ring) and 52 ^{1}H,^{13}C-HMBC correlations. The results of the ^{13}C chemical shift calculation for the best 10 structures are given in Table 2. The substructure analysis of the 5-deoxyenterocin (**3**) data set presented as a substructure tree is shown in Fig. 7, allowing only one "non ring atom" per ring system. The first two generations of substructures are given and the substructures contained in the correct structure are indicated by bold bonds. The phenyl ring is found in all 82 structures, since it was fixed. But two major groups can be found: 54 phenols and 28 carbon substituted benzenes. The second group is clearly favored by the chemical shift deviation (0.9 versus 3.0 ppm) and is also part of the correct proposal.

The bicyclic system is found to be part of 78 out of the 82 structures. Again two major groups were obtained in the next step introducing an additional bridge (tricyclic systems) one with and one without an oxygen. The oxygen bridged substructure (oxymethylene) is favored by the lower chemical shift deviation in comparison to the methylene (0.80 versus 5.02 ppm) and is part of the correct solution.

In contrast to molecules **1** to **3** ascididemin (**4**) is more underdetermined with respect to the NMR correlation data set. To get some idea about the underdetermination of this system a theoretical data set was generated including 14 ^{1}H,^{1}H-COSY and 35 ^{1}H,^{13}C-HMBC correlations. With this data COCON generated 28,672 structural proposals which shows the requirement of C,C correlations or a fast method to analyse all structural proposals. The ^{13}C chemical shifts for all 28,672 structures were calculated (see Fig. 8). The correct structure of ascididemin (**4**) is ranked as 25th, which is within the first 1‰ of all structural proposals! The distribution over the carbon chemical shift deviation shows a Gaussian type behavior (see Fig. 8). The substructure analysis cannot be applied to all generated structures due to computational requirements. Here, it is applied to the 300 structures with the lowest deviation of the calculated versus experimental δ(^{13}C) (about 1%). In contrast to examples **1** to **3**, the ^{13}C chemical shift deviations can not be used as an argument for discrimination of substructures because these values are approximately the same for these structures (see Fig. 8). However, substructure analysis can be used to investigate differing ring systems present in this ensemble. Figure 9 shows the substructure analysis of **4** which results in 10 ring systems containing a) at least 2 rings, b) 10 atoms and c) occur in at least 20 molecules. Again the substructure of the correct solution has a rather small chemical shift deviation, but the differences to the others are not significant as mentioned before. However, the extraction of the basic ring systems in **4** give an overview about the set of structural proposals. Increasing the minimum number of required ring systems from 2 to 3 leads to 97 instead of 10 different ring systems!

For this example the complete way to the final structure will be discussed. Out of the best 60 structural proposals (appr. 2‰ of 28,672) there are only 6 non-strained structures which do not violate Bredt's rule (see Fig. 10). A further distinction is possible by taking δ(^{15}N) into account. Structures **4**-26112 (diazo), **4**-28613 (lactam) and **4**-28672 (nitroso) can be neglected using this argument. ^{15}N-HMBC correlations

would be of help to distinguish between **4**-27927, **4**-28646, and **4**-28656. In structural proposal **4**-27927 there is no nitrogen in β position to the carbonyl. A correlation from the phenyl ring to the nitrogen in β position to the carbonyl is only possible for **4**-28646 which represents the correct constitution of ascididemin.

Conclusions

A general approach is presented which allows a fast and efficient analysis of large sets of structures as generated e.g. by structure generators. The usage of ^{13}C chemical shifts and a substructure analysis is successfully demonstrated. This approach is able to analyze structural proposals calculated by COCON. The short calculation times to obtain δ(^{13}C) by a neural network and the usage of a substructure analysis allow a structure elucidation with less correlation data from 2D NMR spectra. Therefore, the presented method is an alternative approach to obtain an almost complete correlation data set (including ^{1}H,^{15}N-HMBC and ^{13}C,^{13}C correlation data) for an underdetermined structure. The number of structures that have to undergo a further analysis to obtain the correct result can be safely decreased to about 1% of the original number of structures for large ensembles without a reasonable risk of loosing the correct proposal. This approach is independent on the structure generator COCON and can therefore also be used in combination with other structure generators. However, a combination of this approach with COCON is an essential step towards an automatic structure elucidation of organic compounds.

Acknowledgements

This research was supported by the Deutsche Forschungsgemeinschaft (Ko1314/3-3 and 3-4). J. M. is supported by a Kekulé stipend of the Fonds der Chemischen Industrie. We would like to thank Christian Gödde for testing the substructure analysis extensively. We would like to acknowledge Dr. Martin Will (BASF AG) and Dr. Reinhard Meusinger for their work at the ^{13}C chemical shift prediction and Dr. Walter Maier (BASF AG) for analyzing substructure ensembles with SpecEdit.

References

[1] W. Maier, *Comput.-Enhanced Anal. Spectrosc.*, New approaches to computer-aided NMR interpretation and structure prediction, S. 37-55 (1993).

[2] M. Köck, J. Junker, W. Maier, M. Will, T. Lindel, *Eur. J. Org. Chem.* **1999**, 579-586.

[3] T. Lindel, J. Junker, M. Köck, *J. Mol. Model.* **1997**, *3,* 364-368.

[4] T. Lindel, J. Junker, M. Köck, *Eur. J. Org. Chem.* **1999**, 573-577.

[5] J. Meiler, M. Will, R. Meusinger, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1169-1176.

[6] B. Bienfait, *J. Chem. Inf. Comput. Sci.* **1994**, *34,* 890-898.

[7] A. K. Sharma, S. Sheikh, I. Pelczer, G. C. Levy, *J. Chem. Inf. Comput. Sci.* **1994**, *34,* 1130-1139.

[8] J. Meiler, R. Meusinger, J. Gasteiger, *Software - Entwicklung in der Chemie* **1995**, *10,* 259-263.

[9] R. Meusinger, R. Moros, J. Gasteiger, *Software - Entwicklung in der Chemie* **1995**, *10,* 209-216.

[10] L. Montanarella, M. R. Bassani, O. Breas, *Rapid Commun. Mass Spectrom.* **1995**, *9,* 1589-1593.

[11] Y. Isu, U. Nagashima, T. Aoyama, H. Hosoya, *J. Chem. Inf. Comput. Sci.* **1996**, *36,* 286-293.

[12] L. Michon, B. Hanquet, B. Diawara, D. Martin, J.-P. Planche, *Energy & Fuels* **1997**, *11,* 1188-1193.

[13] G. d. V. Rodrigues, I. P. d. A. Campos, V. d. P. Emerenciano, *Spectroscopy* **1997**, *13,* 191-200.

[14] D. E. Zimmerman, C. A. Kulikowski, Y. Huang, W. Feng, R. Powers, *J. Mol. Biol.* **1997**, *269,* 592-610.

[15] S. R. Amendolia, A. Doppiu, M. L. Ganadu, G. Lubinu, *Anal. Chem.* **1998**, *70,* 1249-1254.

[16] J. Kaartinen, S. Mierisova, J. M. E. Oja, J.-P. Usenius, Y. Hiltunen, *J. Magn. Res.* **1998**, *134,* 176-179.

[17] J. Moult, *Curr. Opin. Biotechnology* **1999**, *10,* 583-588.

[18] J. Shockcor, *personal information* **1999**, .

[19] V. Kvasnicka, S. Sklenak, J. Pospichal, *J. Chem. Inf. Comput. Sci.* **1992**, *32,* 742-747.

[20] J.-P. Doucet, A. Panaye, E. Feuilleaubois, P. Ladd, *J. Chem. Inf. Comput. Sci.* **1993**, *33,* 320-324.

[21] S. Sklenak, V. Kvasnicka, J. Pospichal, *Chem. Pap.* **1994**, *48,* 135-140.

[22] D. L. Clouser, P. C. Jurs, *Carbohydr. Res.* **1995**, *271,* 65-77.

[23] A. Panaye, J.-P. Doucet, E. Feuilleaubois, S. R. E. Azzouzi, *AIP Conf. Proc.* **1995**, *330,* 734-739.

[24] D. Svozil, J. Pospichal, V. Kvasnicka, *J. Chem. Inf. Comput. Sci.* **1995**, *35,* 924-928.

[25] S. Thomas, E. Kleinpeter, *J. Prakt. Chem./Chem.-Ztg.* **1995**, *337,* 504-507.

[26] D. L. Clouser, P. C. Jurs, *J. Chem. Inf. Comput. Sci.* **1996**, *36,* 168-172.

[27] D. L. Clouser, P. C. Jurs, *Anal. Chim. Acta* **1996**, *321,* 127-135.

[28] O. Ivanciuc, J. P. Rabine, D. Cabrol-Bass, A. Panaye, J.-P. Doucet, *J. Chem. Inf. Comput. Sci.* **1996**, *36,* 644-653.

[29] S. R. E. Azzouzi, B. T. Fan, A. Panaye, J.-P. Doucet, *Org. React.* **1997**, *31,* 3-20.

[30] J. Meiler, R. Meusinger, M. Will, *Monatshefte für Chemie* **1999**, *130,* 1089-1095.

[31] N. E. Shemtulskis, D. Weininger, C. J. Blankley, J. J. Yang, C. Humblet, *J. Chem. Inf. Comput. Sci.* **1996**, *36,* 862-871.

[32] K. Ozawa, T. Yasuda, S. Fujita, *J. Chem. Inf. Comput. Sci.* **1997**, *37,* 688-695.

[33] D. D. Robinson, T. W. Barlow, W. G. Richards, *J. Chem. Inf. Comput. Sci.* **1997**, *37,* 943-950.

[34] T. Wang, J. Zhou, *J. Chem. Inf. Comput. Sci.* **1997**, *37,* 828-834.

[35] D. J. Klein, I. Gutman, *J. Chem. Inf. Comput. Sci.* **1999**, *39,* 534-536.

[36] S. Forenza, L. Minale, R. Riccio, F. E., *J. Chem. Soc. Chem. Comm.* **1971**, 1129-1130.

[37] E. E. Garcia, L. E. Benjamin, R. I. Fryer, *J. Chem. Soc. Chem. Comm.* **1973**, 78-79.

[38] R. P. Walker, D. J. Faulkner, D. van Engen, J. Clardy, *J. Am. Chem. Soc.* **1981**, *103,* 6772-6773.

[39] J. Kobayashi, F. Kanda, M. Ishibashi, Shigemori, *J. Org. Chem.* **1991**, *56,* 4574-4576.

[40] H. Kang, P. R. Jensen, W. Fenical, *J. Org. Chem.* **1996**, *61,* 1543-1546.

[41] J. F. Cheng, Y. Ohizumi, M. R. Wälchli, H. Nakamura, Y. Hirata, T. Sasaki, J. i. Kobayashi, *J. Org . Chem.* **1988**, *53,* 4621-4624.

[42] C. Concepts, *SpecInfo database*, STN, Karlsruhe.

[43] W. Robien, *Monatshefte für Chemie* **1983**, *114,* 365.

[44] W. Robien, *Nachr. Chem. Tech. Lab.* **1998**, *46,* 74-77.

[45] J.-T. Clerc, H. Sommerauer, *Anal. Chim. Acta* **1977**, *95,* 33-40.

[46] D. F. Ewing, K. Abe, *Org. Magn. Res.* **1979**, *12,* 499-524.

[47] W. Bremser, L. Ernst, B. Franke, R. Gerhards, A. Hardt, *Carbon-13 NMR Spectral Data*, Verlag Chemie, Weinheim, **1981**.

[48] R. A. Hearmon, *Mag. Res. Chem.* **1986**, *24,* 995-998.

[49] A. Fürst, E. Pretsch, *Anal. Chim. Acta* **1990**, *229,* 17-25.

[50] S. Thomas, D. Ströhl, E. Kleinpeter, *J. Chem. Inf. Comput. Sci.* **1994**, *34,* 725-729.

[51] J. Meiler, http://www.krypton.org.chemie.uni-frankfurt.de.

Table 1: Results of the COCON and $\delta(^{13}\text{C})$ calculations for compounds **1** to **4**.

	COCON				$\delta(^{13}\text{C})$		
	Data[a]	Corr.[b]	No.[c]	Calc. time	Range[d]	Best[e]	Calc. time
Oroidin (**1**)	Exp	111000	33	0.3s	7.8 – 20.5	9.1 (2)	< 1s
Manzacidin (**2**)	E+T	111000	190	3.9s	5.2 – 27.1	5.2 (1)	< 1s
Enterocin (**3**)	Exp	110000	82	4.9s	3.7 – 21.7	3.7 (1)	< 1s
Ascididemin (**4**)	Theo	110000	28672	1m43s	4.9 – 29.5	6.7 (25)	7m56s

[a] Origin of the correlation data set: Exp (E) stands for experimental, Theo (T) for theoretical.

[b] Correlation data used for the COCON calculations. The six columns stand for ^{1}H,^{1}H-COSY, ^{1}H,^{13}C-HMBC, 1,1-ADEQUATE, ^{1}H,^{15}N-HMBC, fixed and forbidden (1 indicates that the data is used and 0 that it is not used).

[c] Number of structural proposals generated by COCON under consideration of the correlation data given in column Corr.

[d] Range of the $\delta(^{13}\text{C})$ deviations [ppm] (calculated – experimental) for all structural proposals.

[e] $\delta(^{13}\text{C})$ deviation [ppm] for the best structural proposal. The ranking of the correct strcuture is given in parenthesis.

Table 2: $\delta(^{13}\text{C})$ deviations for the 10 best structural proposals of manzacidin (**2**) and 5-deoxyenterocin (**3**).

Manzacidin (**2**)		5-Deoxyenterocin (**3**)	
No.	$\Delta\delta(^{13}\text{C})$	No.	$\Delta\delta(^{13}\text{C})$
1	5.2 ppm	1	3.7 ppm
2	7.2 ppm	2	8.5 ppm
3	7.4 ppm	3	9.1 ppm
4	10.9 ppm	4	9.4 ppm
5	11.8 ppm	5	10.1 ppm
6	11.8 ppm	6	10.5 ppm
7	12.0 ppm	7	10.7 ppm
8	12.2 ppm	8	11.6 ppm
9	12.3 ppm	9	11.7 ppm
10	12.4 ppm	10	11.8 ppm

Figure Captions

Figure 1: General scheme for a systematical analysis of natural products based on NMR spectroscopical data as presented in this contribution.

Figure 2: Largest common substructure of two model compounds. The highlighted carbon atoms are superimposed first. Subsequently, the atoms in the increasing numbered spheres are compared and added to the substructure until no further superimposition is possible. The largest substructure is bold marked.

Figure 3: Results for the substructure analysis carried out for the 33 structural proposals of oroidin (**1**). A minimum of 15 atoms per substructure as well as a reorganization in a tree was applied.

Figure 4: Results of the $\delta(^{13}C)$ calculation for all substructures generated for oroidin (**1**).

Figure 5: Four structures of the best substructure family of oroidin (**1**).

Figure 6: Substructure analysis of manzacidin A (**2**) allowing only 2 atoms not to be part of a ring.

Figure 7: Substructure analysis of enterocin (**3**) with at least 6 atoms and only 1 "not ring atom" per ring in a substructure. Only 200 structural proposals with lowest chemical shift deviation were analyzed and only substructures combining at least 20 molecules are given.

Figure 8: Results of the $\delta(^{13}C)$ calculation for all structural proposals of ascididemin (**4**) generated by COCON.

Figure 9: Substructure analysis of ascididemin (**4**). Substructures of the 300 structural proposals with lowest chemical shift deviation are given that combine: a) at least 20 molecules, b) 10 atoms per molecule and c) 2 rings within a molecule.

Figure 10: The six non-strained structural proposals for ascididemin (**4**) out of the first 2‰ of 28672 structures.

Fig. 1

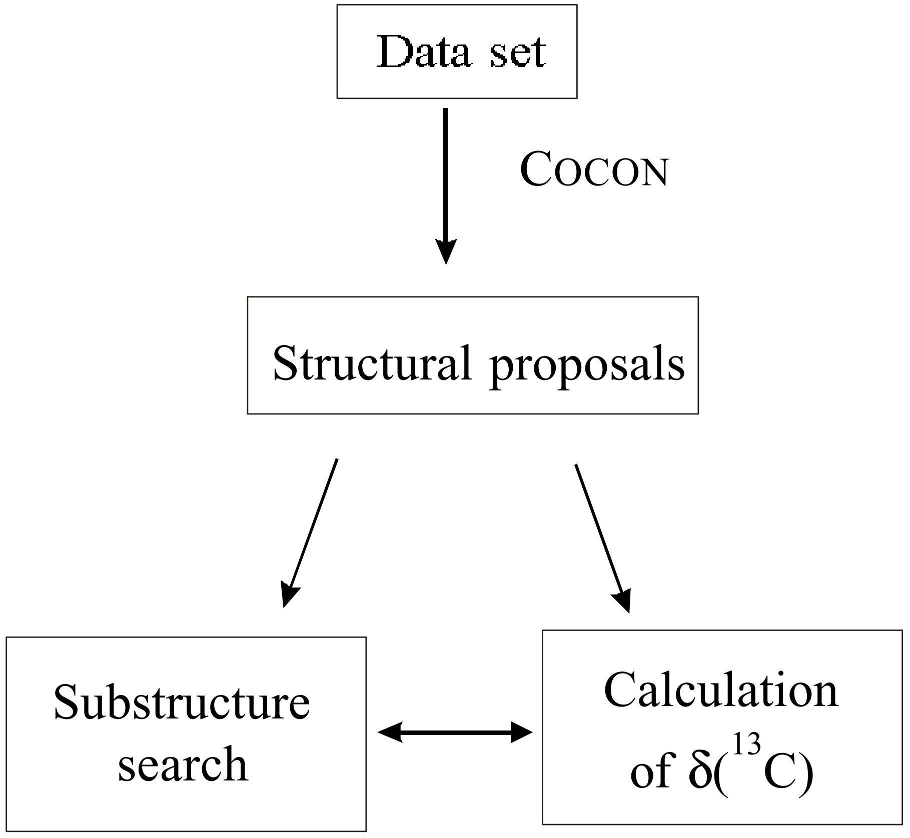
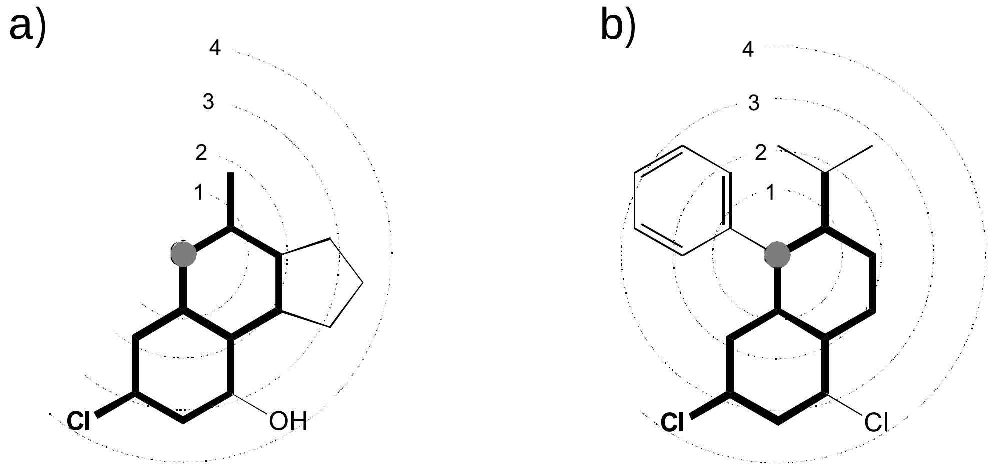
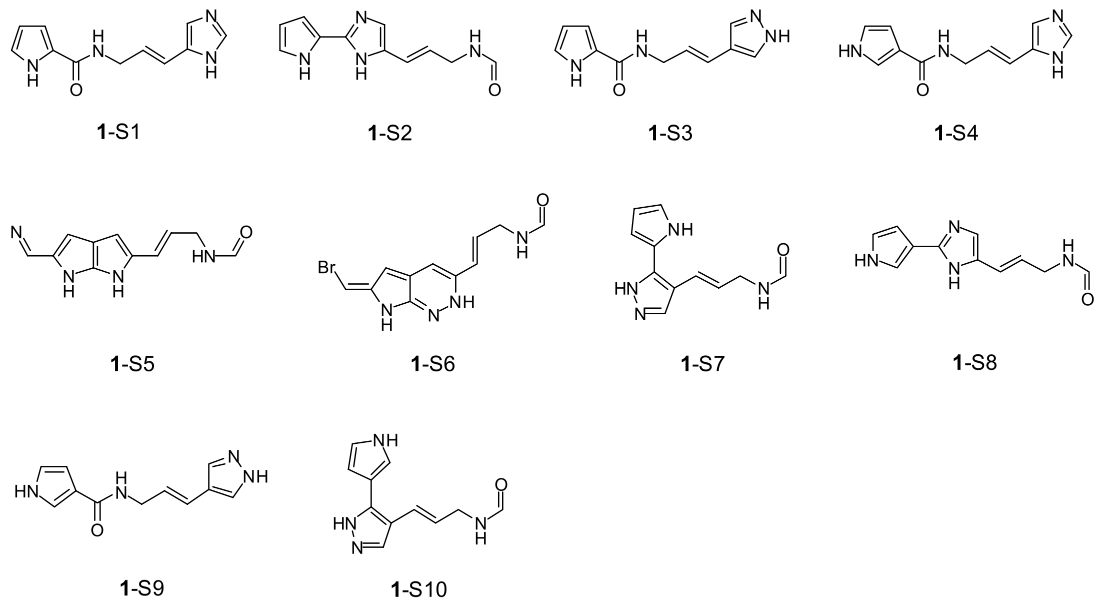


Fig. 2





1-S1

1-S2

1-S3

1-S4

1-S5

1-S6

1-S7

1-S8

1-S9

1-S10

Fig. 3

Fig. 4

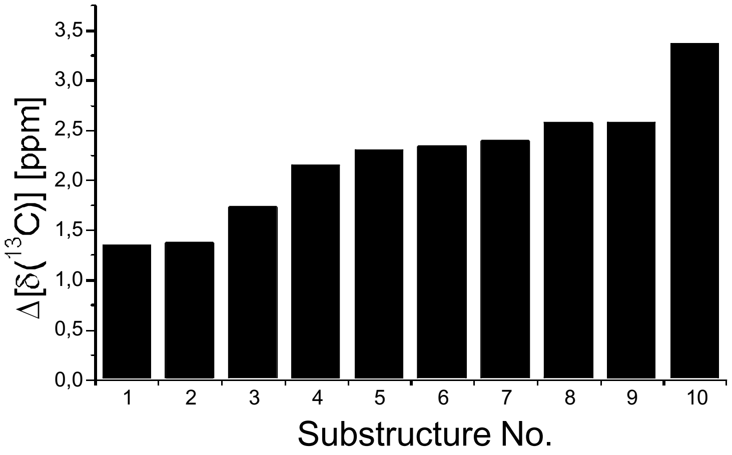
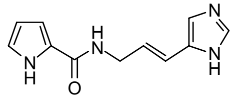
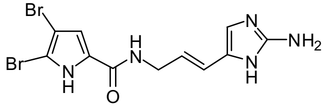


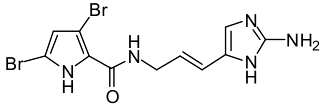
Fig. 5



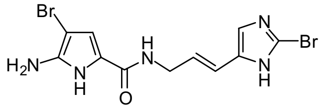
1-S1



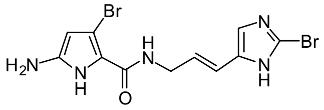
1-27 (Δδ(^{13}C) = 9.05 ppm)



1-29 (Δδ(^{13}C) = 11.85 ppm)



1-30 (Δδ(^{13}C)=18.40 ppm)



1-32 (Δδ(^{13}C) = 13.08 ppm)

Fig. 6

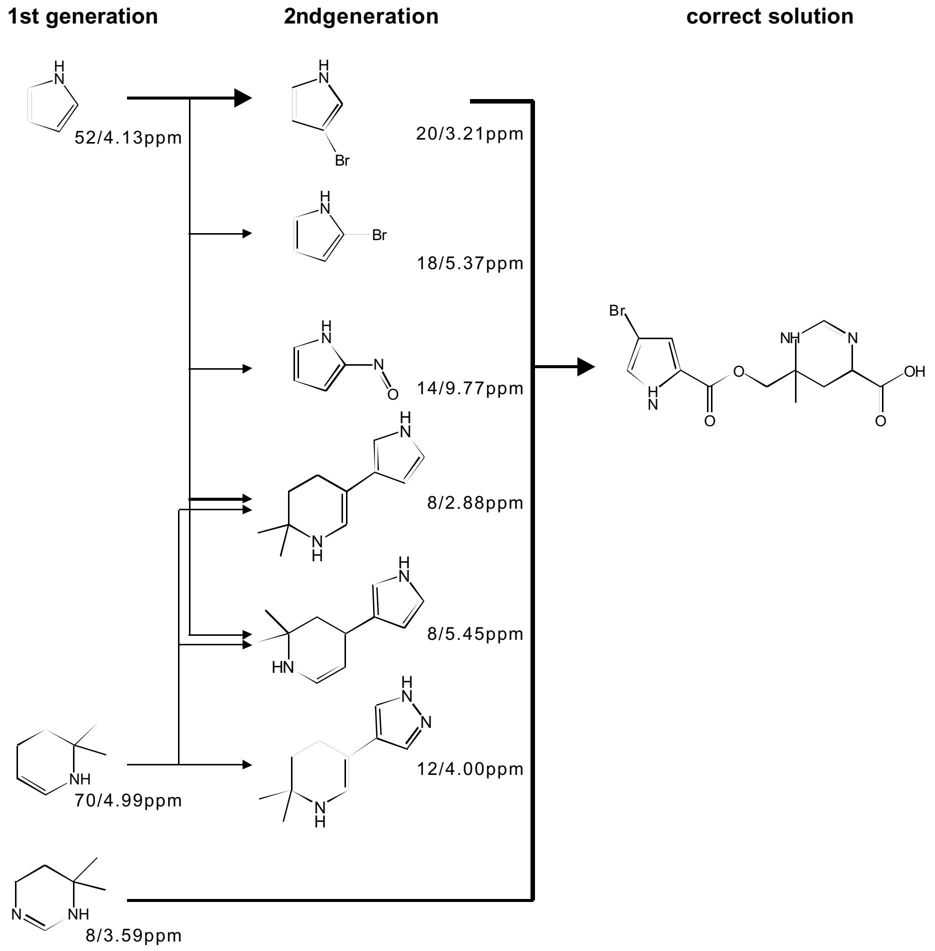

1st generation
2ndgeneration
correct solution
52/4.13ppm
20/3.21ppm
18/5.37ppm
14/9.77ppm
8/2.88ppm
8/5.45ppm
70/4.99ppm
12/4.00ppm
8/3.59ppm

Fig. 7

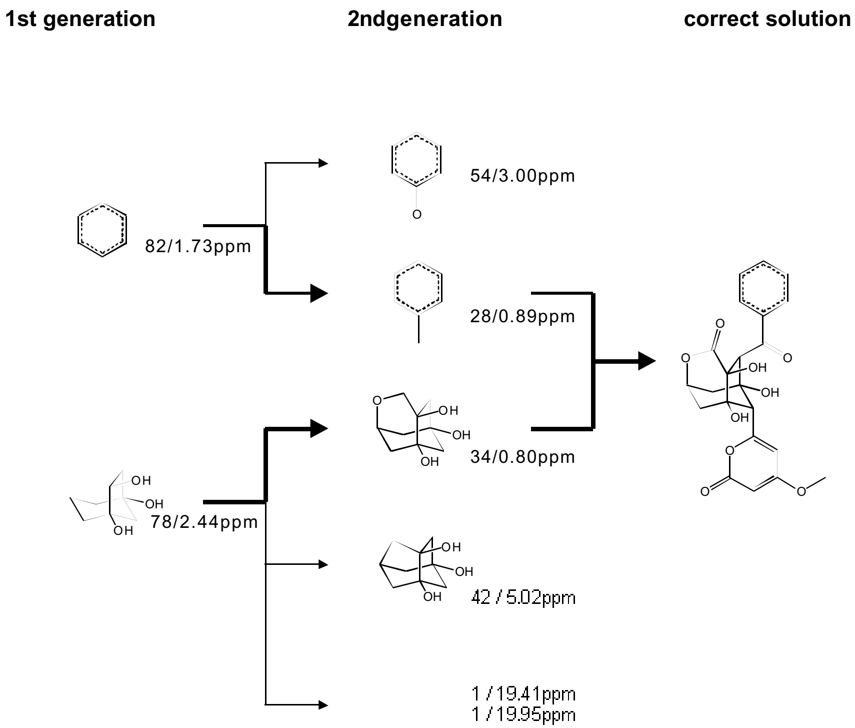
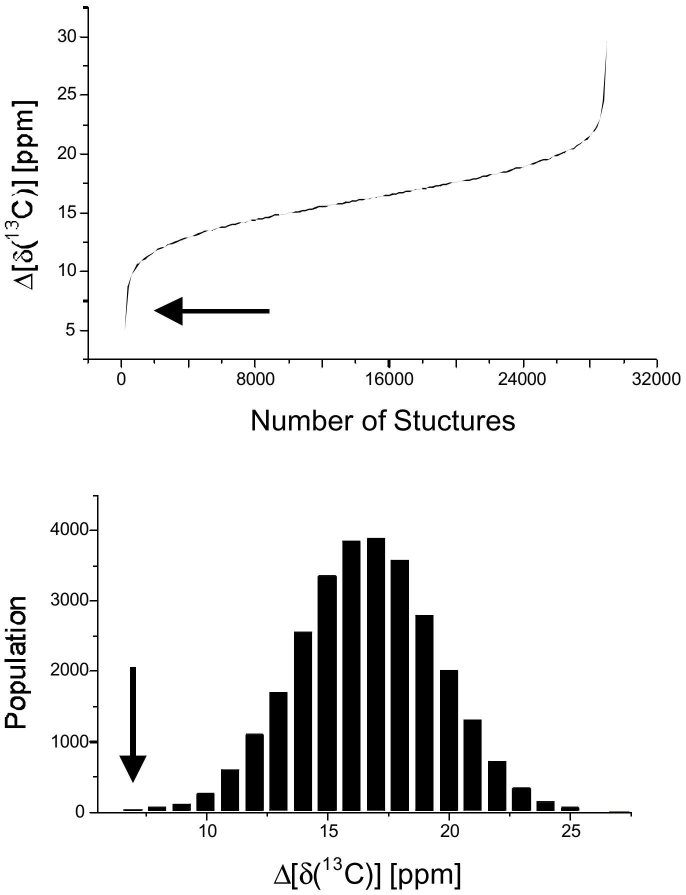
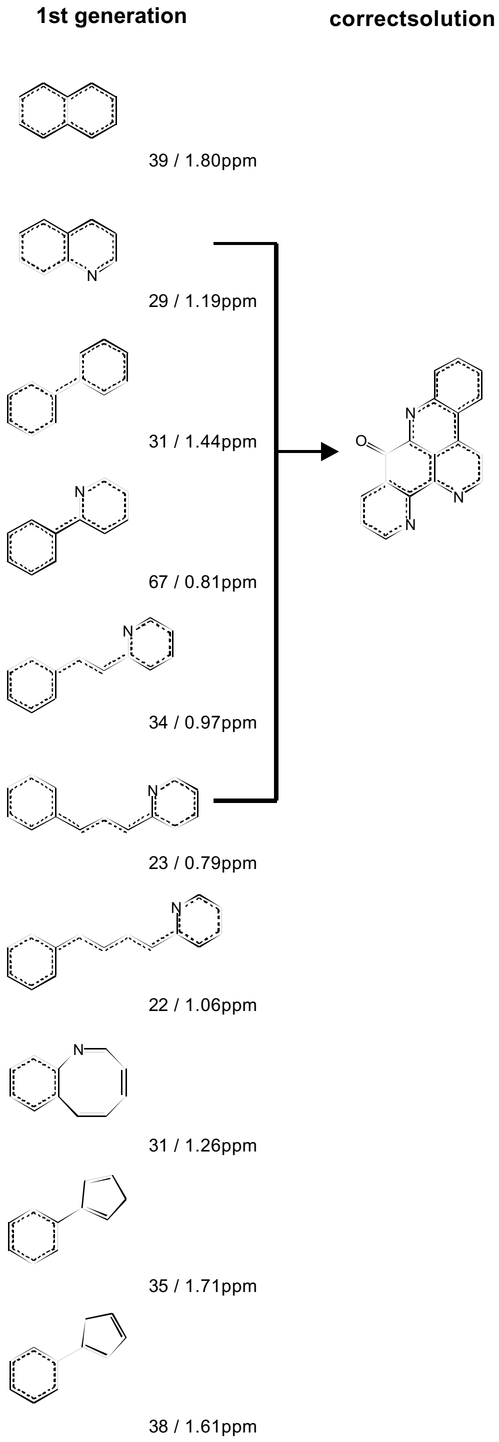

1st generation
2ndgeneration
correct solution
54/3.00ppm
82/1.73ppm
28/0.89ppm
34/0.80ppm
78/2.44ppm
42 / 5.02ppm
1 / 19.41ppm
1 / 19.95ppm

Fig. 8



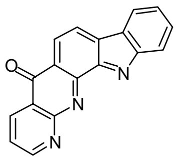
30
25
20
15
10
5
Δ[δ(^{13}C)] [ppm]
0
8000
16000
24000
32000
Number of Stuctures
4000
3000
2000
1000
0
Population
10
15
20
25
Δ[δ(^{13}C)] [ppm]

Fig. 9

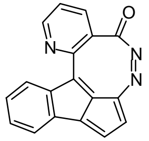


1st generation
correctsolution
39 / 1.80ppm
N
29 / 1.19ppm
31 / 1.44ppm
N
67 / 0.81ppm
N
34 / 0.97ppm
N
23 / 0.79ppm
N
22 / 1.06ppm
N
31 / 1.26ppm
35 / 1.71ppm
38 / 1.61ppm
O
N
N
N

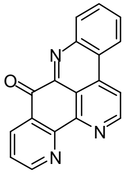
Fig. 10



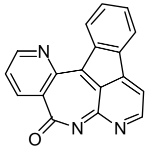
4-27927 ($\Delta\delta(^{13}C)$ = 6.29 ppm)



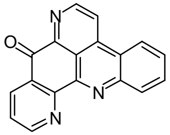
4-26112 ($\Delta\delta(^{13}C)$= 6.50 ppm)



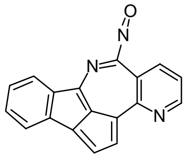
4-28646 ($\Delta\delta(^{13}C)$= 6.70 ppm)



4-28613 ($\Delta\delta(^{13}C)$= 6.71 ppm)



4-28656 ($\Delta\delta(^{13}C)$= 6.80 ppm)



4-28672 ($\Delta\delta(^{13}C)$=6.89 ppm)

Automated Structure Elucidation of Organic Molecules from ^{13}C NMR Spectra using Genetic Algorithms and Neural Networks

JENS MEILER[1]*, and MARTIN WILL[2]

[1]Universität Frankfurt
[2]BASF AG Ludwigshafen

Key words. neural networks; genetic algorithms; quantitative structure property relation; ^{13}C nuclear magnetic resonance spectroscopy; automated structure elucidation; structure generator; database

*To whom correspondence should be addressed

Contact address.

Jens Meiler

Universität Frankfurt (AK GRIESINGER)
Marie-Curie-Str. 11
60439 Frankfurt am Main

Tel.: 069 798 29 798
Fax.: 069 798 29 128
Mail: mj@org.chemie.uni-frankfurt.de

Abstract. The automated structure elucidation of organic molecules from experimentally obtained properties is extended by an entirely new approach: A genetic algorithm is implemented that uses molecular constitution structures as individuals. With this approach the structure of organic molecules can be optimized to meet experimental criteria, if in addition a fast and accurate method for the prediction of the used physical or chemical features is available. This is demonstrated using ^{13}C NMR spectrum as readily obtainable information. ^{13}C NMR chemical shift, intensity and multiplicity information is available from ^{13}C NMR DEPT spectra. By means of artificial neural networks a fast and accurate method for calculating the ^{13}C NMR spectrum of the generated structures exists. The approach is limited by the size of the constitutional space that has to be searched and by the accuracy of the shift prediction for the unknown substance. The method is implemented and tested successfully for organic molecules with up to 20 non hydrogen atoms.

Introduction. Thousands of substances are synthesized every day and their structures need to be elucidated or validated. Consequently the daily routine work of structure elucidation of molecules produced by organic synthesis, especially by combinatorial synthesis, is still one of the most important and demands on chemists. Tools that assist sprectroscopists to elucidate the structure of organic molecules, or are even able to predict the structure of unknowns automatically, are therefore of general importance.

In order to automatically determine from any source of experimental data the structure of an organic molecule, two steps of data processing are performed: A structure generator creates proposals for the unknown molecular structure. A filter validates these proposals usually by comparing easily derivable properties of the generated molecules with the corresponding experimental values and minimizing the deviation. In theory it is possible to determine the chemical structure of any compound from just one experimental quantity, provided that every compound has its own characteristic experimental quantity and this value is obtainable without any experimental uncertainty. Even if these prerequisites are obtained two additional requirements are necessary:

- The experimental value can be calculated from the molecular structure with infinite precision, and
- Infinite computational power is available.

Under these conditions all possible structures can be created, the known experimental value can be computed and a comparison with the experimentally observed value yields an unambiguous answer as to which of the hypothetically proposed structures is the unknown.

However, in practice these requirements are impossibly. Even if the experimental parameter differs for every molecule, it can only be measured within experimental uncertainty. If the number of possible structures is large enough and if the error of the property calculation is taken into consideration, “false” positives will be found with smaller deviation of the calculated and the experimental value than the true solution has. Therefore, it is only possible

to obtain a hit list of structural proposals ranked according to their similarity to the experimental data. In this hit list the correct solution is provided together with false positives. The introduction of additional experimental data helps to overcome this limitation. A more challenging problem is the infinite number of possible structures. It is impossible to compute an infinite number of proposed structures in a finite period of time. Therefore, the key point is the development of “intelligent” structure generators that include already available experimental data during the generation of structures, and create therefore only a finite number of probable structures, each having only a small deviation from the experimentally obtained data. The earlier the comparison of the experimental value with the values calculated for the generated structures is performed and the result is incorporated into the further structure generation process, the more exact the structural space can be defined that has to be searched. This decreases the required computation time.

A frequently used first restriction is the molecular formula. This boundary condition ensures a finite number of possible structures, which then allows a computation of the entire structural space in a finite period of time. The generation of a structural space can be separated into two steps: The generation of all possible constitutions (a constitution formula contains all connectivity but no stereo chemical information) and the subsequent generation of all possible stereo isomers for every constitution formula.

MOLGEN is a powerful structure generator that performs both those steps and creates all possible structures having a given molecular formula[1,2]. A subsequent calculation of a predictable parameter (for example the ^{13}C NMR spectrum) for all these structures and a comparison with the experiment would provide a straightforward approach for automated structure elucidation. However, even for a small number of atoms the computation time increases to an impractically large size.

The COCON approach by Lindel, Köck and Junker uses connectivity information from two dimensional NMR Spectroscopy in addition to the molecular formula and so becomes usable for much larger molecules[3,4]. COCON produces all constitutions that fulfill the introduced connectivity information. However, since COCON uses only connectivity information, it does not differentiate stereoisomers that may occur in the generated constitutions. Thus MOLGEN generates a set of all possible constitutions, whereas COCON reduces this number. In some cases only one constitution fulfills all the connectivity information. However, often COCON presents a large set of possible constitutions, more than can be validated by hand.

The collection of connectivity information from two dimensional NMR spectra is time consuming and difficult to automate. The one dimensional ^{13}C NMR chemical shift is much easier to obtain but also contains diverse constitutional and stereochemical information. Further, artificial neural networks offer a fast and accurate tool for calculating the ^{13}C NMR spectrum of organic compounds[5]. Recently we demonstrated that a combination of COCON with a subsequent comparison of the experimental and calculated ^{13}C NMR chemical shifts is an effective and efficient possibility to decrease the number of possible constitutions presented by COCON alone[6].

A very powerful approach named SPECSOLV uses the ^{13}C NMR spectrum in combination with the Specinfo database[7,8]. The molecular constitution formula can be elucidated only from their ^{13}C NMR chemical shifts by a search for similar substructure spectra in the database and reassembling the substructure fragments found.

In contrast to all these approaches, the implementation introduced here uses an entirely new procedure of intelligent structure generation – a genetic algorithm. This allows one to circumvent certain limits and disadvantages of previous approaches:

- The time consuming determination of connectivity information for COCON by two dimensional NMR spectroscopy is replaced by the much easier and rapidly obtainable chemical shift value.
- The genetic algorithm is able to use the generated structures immediately as a basis for further optimization process. While MOLGEN or COCON generate structural spaces of predefined size and content, the structural space generated by this genetic algorithm is dynamically determined.

- In contrast to "SPECSOLV" the generation of a structural database by reference to thousands experimental spectra no longer necessary and does not limit the searched structural space. The entire space, including all possible structures, can be investigated unaffected by either preferred and neglected regions in a reference database.
- Additional structure information (including connectivity information from two dimensional NMR spectroscopy) can be implemented easily as boundary conditions.
- The generated structures can be ranked by their chemical shift deviation to the target spectrum and not only by a binary quality factor (e. g. in line or not in line with the connectivity information – COCON).

Since not all structures of the structural space are generated with such an implementation there can be no guaranty that the correct solution structure is actually created. We will describe one implementation of such a genetic algorithm and discuss its advantages and its limitations. First we will give a brief summary of existing neural networks and genetic algorithms used in context with NMR spectroscopy. The usefulness of the ^{13}C chemical shift values for structure elucidation is reprised.

Methods of artificial intelligence are wide spread and accepted methods for data analysis in chemistry and biology[9]. Neural networks have been suggested for more than ten years as solutions for a wide range of optimization problems. They are intensively used for the prediction of NMR chemical shifts of organic substances, in particular for Carbon atoms[5,10-19]. Genetic algorithms are of special interest due to their ability to solve complex optimization problems on complex hyper dimensional surfaces with many local minima. In combination with NMR spectroscopy they are implemented for the assignment and analysis of spectra[20-25]. Due to its combination of spectral simplicity and the large content of complex chemical information, the ^{13}C NMR chemical shift value is well suited for the storage in data bases[7,26] and application in intensive numerical analysis. A triple of three values: chemical shift, intensity, and multiplicity contains detailed information about the chemical environment of the most common atom in most organic molecules, carbon. Many approaches for the prediction of the chemical shift value and for its use in further analysis are suggested beside these already listed applications involving neural networks. Only a few are mentioned here[27-30]. Moreover the chemical shift plays an important role in the daily routine work of structure elucidation and validation in organic chemistry. Consequently a fully automated structure elucidation program based exclusively on ^{13}C NMR data is a dream of NMR spectroscopists.

Theory. A genetic algorithm is a method of producing new individual examples from combinations of previous individuals. The algorithm has the same logical structure as inheritance in biological systems and much of the terminology is similar by analogy. So, for example, a genetic algorithm describes the previous examples as "parents" and the combinations produced as "children" or "offspring" or individuals belonging to the next generation. The identity of a particular individual is determined randomly but by a process which is probability weighted. The probability that an individual will be produced and participate as a parent in a succeeding generation must be defined by some standard. For an optimization process, the suitability of an offspring can be assessed using some "fitness" function. This is a direct analogy to Darwin's evolutionary rules of selection, survival of the fittest. One algorithmic process that mimics biological evolution is described as generating mutation. How these relationships work out for a non-biological sequence of events, a synthetic calculation occurring entirely within a computer, will now be described in more detail.

The implementation of every genetic algorithm invokes three data processing steps: selection, recombination (cross over) and mutation. For optimizing molecular structures, a genetic code needs to be defined that describes them. Figure 1a) visualizes how a vector of bond states

between all (atom – atom) pairs can be defined from the connectivity matrix of a molecule. This vector provides a suitable genetic code for the constitution of an organic molecule. Stereochemistry is not considered in this implementation, since it was not possible to distinguish between stereo isomers using the selection procedure as discussed below. A set of randomly generated constitutions is taken as the starting parent population. The members of this population satisfy only the molecular formula, which has to be known in advance. Iteratively, the population undergoes the processes of selection, recombination and mutation to form a child generation which can then be used as the next parent generation (Figure 1b)):

Selection. While recombination and mutation can be implemented independently from boundary conditions, the selection process is affected by using the ^{13}C NMR spectrum as a "fitness function". As mentioned earlier, a fast and exact calculation method for this fitness function is necessary to implement a genetic algorithm. The ^{13}C NMR chemical shift can be determined most efficiently for this purpose using artificial neural networks. Once trained, they are fast and exact. The implementation of neural networks used in the following approach is described in the literature[5] and is therefore only briefly summarized below:

The spectra can be predicted for all organic substances that contain exclusively C, H, N, O, P, S or the four halogens (F, Cl, Br, I). To obtain the spectrum of a molecule the chemical shift of every carbon atom is successively calculated in an individual run. The environment around the carbon atom of interest is subdivided into six spheres. All atoms inside these spheres are again classified as belonging to one of 28 previously defined atom types. The types consider the atomic number, hybridization and number of bound hydrogen atoms. The 28 dimensional vector containing the number of atoms of every atom type in a particular sphere is accomplished by two sum parameters holding the number of hydrogen atoms in the sphere (hydrogen is not considered as one of the 28 atom types) and the number of ring closures. This vector contains now 30 numbers and is collected for atoms belonging to one out of six spheres separately. Moreover the information is collected a second time, but only for atoms that belong to a conjugated together with the carbon atom of interest to consider the special influence of such systems on the chemical shift value. Therefore a vector of 360 (=30·6·2) numbers describes the carbon atom environment and serves as input for the neural networks. Nine of these 28 atom types describe carbon atoms. For each of these nine types an individual neural network is trained using the overall number of about 1,300,000 chemical shifts out of the Specinfo database[7]. The number of hidden neurons in the single hidden layer varies from 5 to 40, depending on the number of training examples for the carbon atom type. One output neuron calculates the chemical shift. The average deviation of this method is as low as 1.6ppm relative to an independent database of about 50,000 chemical shifts. Essential advantages of this method are the fast, exact, and database independent shift prediction for all organic molecules. Since spherical description of the carbon atom environment used in the method does not contain stereochemical information, the predicted chemical shift spectrum is the same for all stereo isomers for any particular constitution formula. Consequently the genetic algorithm implemented here can only optimize the molecular constitution relative to the NMR spectrum. Therefore, the genetic code needs also only to define the constitution. If stereochemistry had been considered in estimating the chemical shift, the introduction of stereochemical descriptors in the genetic code would allow the definition of stereochemistry in the structures elucidated.

The chemical shifts of all Carbon atoms of a constitution are calculated by the artificial neural networks and sorted by their size. The "fitness" of every single Carbon atom i is now the deviation between its experimental and the corresponding calculated chemical shift value: $\left(\delta^i_{calc}\left({}^{13}C\right)-\delta^i_{exp}\left({}^{13}C\right)\right)$. The fitness of the whole molecular constitution formula is given by the root mean square deviation (RMSD) over all N Carbon atoms:

$$\Delta\left({}^{13}C\right)\equiv\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\delta^i_{calc}\left({}^{13}C\right)-\delta^i_{exp}\left({}^{13}C\right)\right)^2}$$

(Figure 2). The multiplicity of a signal can be easily incorporated, if experimentally obtained: The absolute deviation between the experimental and the calculated multiplicity for every carbon atom i $\left|M^i_{calc}-M^i_{exp}\right|$ is multiplied by a factor ("multiplicity deviation factor" = MDF (in *ppm*)), and added to the absolute deviation of the chemical shift. The fitness $\Delta\left({}^{13}C\right)$ becomes now

$$\Delta\left({}^{13}C\right)\equiv\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left|\delta^i_{calc}\left({}^{13}C\right)-\delta^i_{exp}\left({}^{13}C\right)\right|+MDF\cdot\left|M^i_{calc}-M^i_{exp}\right|\right)^2}\ .$$

The lower the $\Delta\left({}^{13}C\right)$ value of a structure the higher is its fitness and the higher is its probability of participation in the recombination step of the genetic algorithm:

Recombination: Two molecules from the parent generation are selected to form the child molecule. The smaller the $\Delta\left({}^{13}C\right)$ value of a molecule the higher is its probability to be considered as parent. This probability for a single molecule j out of a population of m constitutions is given by $p_j=\left[\Delta_j\left({}^{13}C\right)\right]^{-1}/\sum_{i=1}^{m}\left[\Delta_i\left({}^{13}C\right)\right]^{-1}$. After this selection, all possible (atom – atom) pairs in both parents are taken and the bond type between them is analyzed (0 = non bounded, 1 = single bond, 2 = double bond, or 3 = triple bond). Randomly one out of the two possibilities for every (atom – atom) pair is taken for the newly generated child structure (Figure 3).

Since hydrogen atoms are not explicitly taken into consideration but are added to the free valences afterwards, the molecular formula needs to be checked after a new child constitution is generated. If the number of potential hydrogen atoms in the generated constitution is not the same as defined in the target molecular formula, bonds must be added or deleted until this deviation is corrected to zero. For this purpose the same function is used as for mutation (see below). Moreover it is necessary to ensure that a single molecule is formed and not a set of two or three fragments with the correct overall molecular formula but not connected to each other. After both boundary conditions are fulfilled, the new molecule is accepted as a member of the newly formed population.

Mutation. A "mutation" is implemented by modifying simply bonds. Two atoms are randomly selected and a bond is deleted (or the bond type is decreased by one) while, for two other randomly selected atoms, a bond is inserted (or the bond type is increased by one, Figure 4). A deletion is always combined with an insertion, so that the total number of hydrogen atoms remains constant. Also this process has to be controlled so that only one molecule and not a set of fragments is created.

Figure 5 illustrates the formation of a single generation in the form of a simplified flow chart diagram. By repeating the procedure of subsequent recombination and mutation, *m* molecules for the child generation are created out of the *m* parent molecules. Optionally the *l* fittest molecules of the parent generation replace the *l* worst molecules of the child generation to ensure that the fittest constitutions are not lost. To enable the optimal use of multiprocessor computers *n* populations can be calculated in a parallel manner without interactions. This procedure takes advantage of the well known fact, that in a genetic algorithm a set of small independent populations converges faster than one large population.

The recombination probability (RP) and the mutation probability (MP) are parameters systematically varied during the iterative calculation process. RP is the probability that a child is generated by combining two parents (recombination) and not by copying a molecule already in the parent population (no recombination). MP defines the probability that a child generated undergoes a mutation (compare Figure 5). It is well known that a high mutation rate at the beginning of a genetic algorithm ensures a fast convergence but later high mutation rates are rather contra productive and simple recombination achieves a better fitting. This fact is also comparable to the evolution of life on the earth:

- High intensity UV irradiation caused high mutation rates in the beginning of evolution of live but both UV levels and mutation rates are lower now.

- Creatures at a low level of evolution frequently reproduce without recombination, whereas creatures on a high level of evolution exclusively reproduce by recombination.

In keeping with the biological analog, RP and MP are changed during the genetic algorithm. In principle they could be independently defined for every evolutionary step of the optimization procedure. However, it is sufficient to predefine the RP and MP for certain evolutionary steps and change the values linearly between these points to approach the defined values at the fixed points.

The incorporation of additional information is possible by defining a good and a bad list of fragments that either need to be part of the molecule or are forbidden to use. In the first case the fragments are incorporated during the initial creation of random molecules and not changed during the further optimization process. In the second case generated structures that contain forbidden fragments are excluded and not used in the child generation. To avoid a reduction of the genetic pool of a population it is excluded that identical individuals that might be formed during the algorithm are considered for one population more than a single time.

Results and Discussion. Three experiments are performed to evaluate a genetic algorithm implemented as discussed above: In the first experiment the parameters are optimized and both the structural space and the generated populations are analyzed for a relatively small molecule. In a second experiment the previously optimized parameters are used to perform a fully automated structure elucidation for a small database consisting of molecules with 9 to 16 non hydrogen atoms. In a third class of calculations, the limitations of this method are examined by investigating larger molecules with up to 20 non hydrogen atoms. In this case an individually optimized setup and the use of additional boundary conditions become necessary.

For reasons of computation time the introduced parameters need to be optimized for a relatively simple example. For the same reason not all possible interactions between the parameters can be analyzed in detail. It is further assumed that the optimized values for these parameters can be later scaled for larger molecules. Moreover the investigation of a small molecule allows a hand analysis of the generated populations resulting in a deeper insight into the operations performed during the optimization. Isoleucin ($C_6H_{13}NO_2$) is chosen because it contains hetero atoms and a double bond. Since it has only nine non hydrogen atoms and only one double bond the number of possible constitutions is comparably low (23,946). Therefore the algorithm finds the correct solution in a reasonably short time period. This allows the optimization of parameters and an intensive analysis of the algorithm itself. The total deviation between the experimentally obtained and the neural network calculated chemical shifts is $\Delta\left({}^{13}C\right)=1.12\,ppm$ for Isoleucin.

The parameters that need to be optimized are: the size of the population, *m*; the number of fittest individuals conserved for the next generation, *l*; the number of parallel calculated populations, *n*; the multiplicity deviation factor, MDF; the recombination probability, RP; and the mutation probability, MP. The product out of *m* and *n* defines the size of the genetic pool since it defines the overall number of individuals, whereas *l* defines the degree of conservative character. It is responsible for the fraction of replaced individuals in each generation. multiplicity deviation factor MDF weights the influence of the deviation in the chemical shift values with respect to a deviation in the obtained multiplicity. The recombination probability RP and the mutation probability MP select the pathway for generating a new individual and are therefore strongly interacting parameters. One out of the two operations (mutation or recombination) has to be performed in order to generate a individual different from its parent(s). In order to ensure this mutation is forced if no recombination was carried out. RP and MP define which fraction of the newly generated constitutions is obtained by recombination or mutation only and which fraction is obtained by a subsequent application of both operations (Figure 5). Figure 6 summarizes the results of the optimization of all six parameters. Since *m* and *n* as well as RP and MP are not independent with respect to each other, experiments were applied to investigate those dependencies.

In a first experiment the size of the population is chosen to be *m* = 8, 16, 32, 64, 128. All other parameters are set to be constant with *l* = 0.25•*m*, *n* = 1, MDF = 1ppm. The mutation probability is 100% during the first four steps and decreases linearly to become 50% between the fifth and the eighth generation. The recombination probability is set to be 0% during the first four generations and increases to become 100% after the eighth generation. For simplicity such a program of RP and MP values will be given in the following notation (MP: ${}^{0}1.0^{4}$ => ${}^{8}0.5$ | RP: ${}^{0}0.0^{4}$ => ${}^{8}1.0$) from now on. In order to obtain realistic results and avoid the influence of the random start population the average $\Delta\left({}^{13}C\right)$ value of the best individual in 16 independent test runs with varying randomly generated start populations is computed for all experiments. All runs are stopped after the 16th generation. As visualized in Figure 6a) the $\Delta\left({}^{13}C\right)$ is generally smaller if larger populations are used. Since more molecules are generated, the probability of creating a molecule with a smaller $\Delta\left({}^{13}C\right)$ increases. However, the generation of more molecules necessarily requires greater computation time. Obviously, a direct comparison of experiments with variable population sizes is not fair.

A realistic picture is given in Figure 6b). In these experiments the number of calculated generations is increased by a factor of 2, 4, 8, and 16 as the number of individuals is decreased from 128 to 64, 32, 16, and 8, respectively. Similarly, the fix points for the MP and RP values are adjusted by these factors. The result of this systematic adjustment is that the overall calculation time as well as the number of generated structures are the same for all populations and a direct comparison becomes possible. The segment between the second and the ninth generation for the population with 128 individuals is plotted in comparison with the corresponding parts of test runs with a smaller number of individuals. As shown in Figure 6b) an optimum size of the population is achieved between 32 and 64 individuals. Although the overall differences in results between the setups are small, the setup with 32 individuals provides the fastest decrease of $\Delta\left({}^{13}C\right)$ in the first period of the algorithm until the fourth generation. An increased number of subsequent mutations (as mentioned recombination does not take place in this period) has advantages compared to a further increase of the number of parallel calculated individuals. However, if the number of randomly parallel generated structures is too small (below 32 in this case) the starting points for the optimization are badly sampled and the optimization velocity suffers. After the fourth generation the setup using 64 individuals seems to become slightly favored. This is due to the fact that here recombination becomes active and therefore the number of individuals in a generation plays an increasing role. It defines the size of the "genetic pool" which is incorporated into the recombination process.

In the next experiment (Figure 6c)) the number of conserved individuals *l* is optimized for a setup using *m* = 32 individuals. *l* is set to be 1, 8, 16, 24, and 31. The overall influence is small. However, an optimum is obtained for *l* = 0.25•*m*. A remarkable worse convergence is obtained in the case of 31 conserved individuals. This behavior is plausible in this case due to the small number of changes in the population with each new generation. Therefore the constitutional space searched by the genetic algorithm is very small.

The number of populations calculated parallel, *n*, is optimized with the constraint of a constant overall calculation time. A setup with $(n\,|\,m)=(1|128)$ is compared with $(n\,|\,m)=(2|64),(4|32),(8|16),\text{and}(16|8)$ in Figure 6d). The optimum is obtained for four parallel calculated populations with 32 individuals each. Only a slight decrease in convergence is obtained going to $(n\,|\,m)=(2|64)$. The algorithm is more sensitive for a further decrease in the size of the population *m*. The "genetic pool" becomes too small in these cases.

The multiplicity deviation factor MDF is optimized in Figure 6e). The optimal value is MDF = 1*ppm* for this example. This result is a compromise between the additional usable information coded by the multiplicity, which causes a better convergence compared to MDF = 0*ppm*, and the higher complexity of the $\Delta\left({}^{13}C\right)$ hyper surface, which causes a worse convergence in the case of higher values for MDF.

In the last plot of Figure 6 RP and MP are systematically changed. All combinations of RP and MP equal 0.0 or 1.0 for the two periods between 0..4 generations and 8..16 generations are tested except the case where RP and MP values are 0.0 at the same time (which would be of cause meaningless since mutation is forced if no recombination takes place, compare Figure 5). The test runs are sorted along the left axis by the lowest $\Delta\left({}^{13}C\right)$ value for the fittest molecule in the population after the 16th generation. Best convergence is obtained for a high RP during the whole run and especially in the second part of the algorithm. A high mutation probability in the second part of the algorithm seems to be contra productive. The same is true for having no recombination at all. However, the differences between the several test runs are again relatively small.

To get a better impression about the decision making processes during the genetic algorithm Figure 7 visualizes tracks on that Isoleucin is formed for a test run with *n* = 1, *m* = 32, *l* = 8, MDF = 1*ppm* and (MP: $^{0}1.0^{4}$ => $^{8}0.5$ | RP: $^{0}0.0^{4}$ => $^{8}1.0$). In the 10th generation Isoleucin itself occurs for the first time. Up to this point the $\Delta\left({}^{13}C\right)$ values for all 32 molecules in the population are given. The constitutions that participate in the formation of Icoleucin at any point in the algorithm are visualized together with the performed mutation and recombination steps. Consistent with the high mutation and low recombination probabilities during the first part of the calculation only mutations take place in this period. A rapid decrease of the $\Delta\left({}^{13}C\right)$ values is obtained, which can be interpreted as a local minimization of the constitutions on the $\Delta\left({}^{13}C\right)$ hyper surface which can then be used in subsequent iterations as an optimal starting point for the recombination process. With the increase of the RP value the number of successful recombinations increases too. However, the first effectively used recombination for the formation of Isoleucin takes place during the generation of the eighth population in this example. During this eighth step recombination is performed in combination with a subsequent mutation. In the next two steps the Isoleucin constitution structure is formed by recombination steps without an additional mutation. The average decrease of the $\Delta\left({}^{13}C\right)$ is more moderate during these later steps. This second part of the optimization can be interpreted as a search through the low $\Delta\left({}^{13}C\right)$ ranges of the hyper surface for the global minimum. Although constitutions with small $\Delta\left({}^{13}C\right)$ values play statistically a more important role in the eventual formation of the Isoleucin molecular constitution, the influence of some individuals with high $\Delta\left({}^{13}C\right)$ value is also observed in the process. This behavior is typical for a genetic algorithm.

The second experiment addresses two questions: (i) the possibility of automated structure elucidation by this approach (ii) statistical analysis of a database of molecules is performed. Six groups of 20 molecules containing 9 to 16 non hydrogen atoms, respectively are randomly selected from the Specinfo database[7]. The experimental ^{13}C NMR chemical shifts are used as input for the algorithm. The setup is equivalent for all 160 molecules with $m = 8$, $n = 32$, $l = 8$, MDF = 1*ppm*, and (MP: $^{0}1.0^{25}$ => $^{50}0.0$ | RP: $^{0}0.5^{25}$ => $^{50}1.0$). The algorithm is repeated until either the right constitution is formed, another constitution with a $\Delta\left({}^{13}C\right)$value smaller than the $\Delta\left({}^{13}C\right)$value of the correct solution molecule is created (accuracy limit) or a maximum of 500 generations is achieved (time limit).

In the first case the automated structure elucidation is successful, whereas in the second or the third cases the method treated as a failure. If a constitution with a smaller $\Delta\left({}^{13}C\right)$value than

the correct solution exists. This happens because the ^{13}C chemical shift calculation is not exact enough to determine unambiguously the correct constitution of the unknown. The probability that such constitutions can be found increases with the number of possible structures. It is therefore the first limiting factor for the maximal size of a molecular constitution structure solvable by this algorithm. If a maximum of 500 generations is calculated without the formation of the correct solution, the optimization process is stopped and treated as a failure. The second limiting factor is the time necessary to search the structural space.

Table 1 summarizes the results of this experiment. The average $\Delta\left({}^{13}C\right)$ value for the 20 structures is within the statistical deviation constant with about 1.1*ppm*. Of 20 molecules tested in each case, the number of correct solutions found decreases slowly from 20 to 14 as the number non hydrogen atoms increases from 9 to 14 atoms. It decreases rapidly to 5 and 4 correct solutions for 15 and 16 non hydrogen atoms, respectively. For one out of twenty molecules with 14 non hydrogen atoms the calculation is stopped for the first time because of a smaller $\Delta\left({}^{13}C\right)$ value for a generated constitution than the $\Delta\left({}^{13}C\right)$ value of the correct solution. For 15 and 16 non hydrogen atoms 2 and 4 calculations are stopped for this reason (accuracy limit). Both, the average calculation time and the average number of generations performed increase dramatically with an increasing number of non hydrogen atoms. All calculations are performed on a PII 450MHz Processor equipped with 512MB memory. The number of generated structures per minute increases more moderately. In comparing the last value with errors typical for other structure generators (e. g. MOLGEN) it has to be kept in mind that in this method not only the constitutions are generated but also aromatic ring systems must be identified and ^{13}C chemical shift must be calculated. Therefore, the genetic algorithm so implemented is slower compared to MOLGEN, if only the number of generated constitutions per unit time is compared.

The results prove that an automated elucidation of the constitution is possible for up to 14 non hydrogen atoms with this setup. However, some points have to be addressed here: For reasons of comparability all calculations are started with the same parameter setup. The drop in the percentage of correct solutions going from 14 to 15 non hydrogen atoms suggests that a setup optimized for Isoleucine (with only nine non hydrogen atoms) may no longer be optimal for 15 and 16 non hydrogen atoms. Specifically too small values for *n*, *m* and a maximum of only 500 generations avoid a higher percentage of correct solutions for larger molecules. Also, the inaccuracy in predicting of the ^{13}C chemical shift information apparently plays an increasing role for molecules with 15 and more non hydrogen atoms.

While the first problem can be solved by slightly modifying the setup of the algorithm, in the second case additional information beside the ^{13}C chemical shift is necessary to obtain an unambiguous result. This information can be a list of forbidden substructures (a "bad" list) or a list of substructures to use (a "good" list) in the easiest case. For twenty cases randomly selected form the non solved structures the test run is repeated using a modified setup of *l* = 16, *m* = 64 and a bad list of only four fragments (directly bounded hetero atoms: N – O, N – N, O – O and allenyl fragments: C = C = C). For 17 out of these 20 examples the correct solution is found. However, both limits are present and have to be discussed. The algorithm would run in case of a real unknown fully automatic until it is stopped by hand. Afterwards all generated structures with a $\Delta\left({}^{13}C\right)$ smaller than the average error of the neural network prediction plus the experimental deviation have to be considered as the possible solution to avoid loosing the correct solution due to the accuracy limit. Due to the computational time limit, there is no guaranty that the correct solution is within the generated set of constitutions. Therefore this program also does not replace the spectroscopist. It is able to solve a part of routine problems in a fully automated mode. Afterwards, the spectroscopist must validate the solutions and concentrate on the cases not unambiguously solved. As discussed below, the algorithm can however still assist solving those more complex, and interesting cases.

We will now demonstrate using a few examples the ability of a more individual setup to solve the constitution for molecules with up to 20 non hydrogen atoms. Table 2 summarizes the results. Isoleucin is again listed as first example molecule to enable comparison. The second example, Histidine, is much more complex due to the increased number of non hydrogen atoms and double bond equivalents. More than 89.5 million constitutions are possible for this molecular formula. However, this problem is still solved using a relative simple setup. Unlike all examples previously discussed, a bad list is used here for the first time. If not constrained, the genetic algorithm tends to create molecules that contain bonds between hetero atoms: e. g. N – N, O – O or O – N. Because of the absence of chemical shift information for such fragments, it is difficult to recognize such structures as false solutions in cases where their overall $\Delta\left({}^{13}C\right)$ value is small. If these structural fragments can be excluded, an accelerated convergence is obtained. It happens that fragments like C = C = C are also favored by the genetic algorithm. The exclusion of this fragment also accelerates the optimization process. For Tryptophan with 15 heavy atoms the determination of all possible structures is essentially impossible within a practical period of time. About 36 billion structures exist. The genetic algorithm creates only about 20,000 structures out of this huge number before the correct solution is found. Examples 4 and 5 demonstrate the power of the approach on molecules with 20 non hydrogen atoms. While the first example has only three double bonds, in the second case the number of double bond equivalents is already nine. Due to this fact, the number of possible structures is much larger and the problem requires a factor of 15 in computation time. However, computers are becoming faster and time can also be saved by computing parallel populations on parallel processors. A high degree of automation is a significant advantage for the method. By increasing the number of "intelligent" interventions, even molecules with more non hydrogen atoms might become solvable. This is demonstrated in examples 5', 6, 7 and 8, where parts of the molecule are defined in advance. In real application such information might be known from the NMR spectrum (e. g. examples 5' and 6), from the synthesis or even from the genetic algorithm itself. The letter case is emphasized if one fragment is generated with a high frequency in the process of the genetic algorithm and the $\Delta\left({}^{13}C\right)$ values of the corresponding Carbon atoms are low. Such a fragments has a high probability of being part of the solution structure. This fragment could then be defined as fixed. Example 5' differs from example 5 only by the fact that two benzene fragments need to be part of the generated constitutions. The dramatic acceleration of structure elucidation by this small intervention demonstrates how the approach can assist the spectroscopists. All known structural information can be introduced into the initial setup and the remaining constitutional space can be searched quickly and effectively using the genetic algorithm. Similar essential significant decreases in computational time are observed for the other three example structures.

Conclusions. A general implementation of a genetic algorithm that uses molecular constitutions as individuals is described. This algorithm is able to optimize a molecular constitution structure to fulfill experimentally observed properties. The ^{13}C NMR spectrum of organic molecules can be accurately determined by experiments and also rapidly predicted by neural network calculation. Consequently the constitution of organic molecules can be optimized relative to an unknown organic sample by combining the genetic algorithm with the neural network spectral prediction. An automated structure elucidation is possible for molecules with up to 14 non hydrogen atoms. Molecular structures with up to 20 non hydrogen atoms can be determined using only their ^{13}C NMR chemical shifts by introducing a small list of forbidden fragments. Larger molecular structures become solvable if fragments that need to be in the molecule are known and introduced as a good list. The number of overall possible solution structures is drastically reduced by the inclusion of such known fragments.

The maximal size of the solvable molecule is limited by the size of the structural space accessible (time limit) or by the accuracy of the property determination, either by experiment or by calculation (accuracy limit). Since the ^{13}C NMR chemical shift prediction is the most time consuming step of the algorithm and is also responsible for the introduction of the calculation error, it is the critical step for both limits. With a fast and accurate chemical shift prediction by neural networks the implementation of such a genetic algorithm becomes possible for the first time.

The described procedures are combined in the program "GENIUS"[31] that should become a helpful tool for structure elucidation of organic molecules.

Acknowledgement. The authors would like to thank Dr. Reinhard Meusinger, Dr. Matthias Köck and Prof. Christian Griesinger for useful discussions. The authors would like to thank the reviewers for useful comments and suggestions. Reviewer #2 added the first introducing paragraph to the theory section. J.M. is supported by a Kekulé stipend of the Fonds der Chemischen Industrie.

Figure Captions.

Figure 1: The connectivity matrix of a randomly created constitution with the molecular formula $C_6H_{13}NO_2$ is given. From this connectivity matrix the genetic code is obtained by rearranging the a triangular half matrix into a vector (a). This vector contains now the bond state between all (atom – atom) pairs of the molecule. Part (b) of the figure visualizes the general implemented procedure. The molecular formula (obtained e.g. from mass spectroscopy) is used to generate a random set of *m* constitutions that fulfill that molecular formula. This set is now valuated by calculating the ^{13}C NMR spectrum and comparing it with the experimental data. The lower the $\Delta\left({}^{13}C\right)$ value the higher is the probability that a molecule is considered for recombination. A new generation is formed by recombining two parent molecules *m* times. Optionally some of these *m* new constitutions can undergo a mutation or *l* of them can be replaced by the *l* fittest parent constitutions. This cycle of selection, recombination and mutation is repeated until $\Delta\left({}^{13}C\right)$ is minimized.

Figure 2: The ^{13}C NMR spectrum for a new generated constitution (a) is calculated by artificial neural networks (b). By comparing this spectrum with the experimentally obtained spectrum (c) the $\Delta\left({}^{13}C\right)$ value can be computed as the RMSD of all single deviations (d). The $\Delta\left({}^{13}C\right)$ is taken as the fitness function in the selection process of the genetic algorithm.

Figure 3: Recombination of two molecules out of the parent generation (a,b) to form a new molecule (c) that becomes a part of the population in the next generation. The gray shaded areas of the parent molecules are linked to form the new molecule. Below each constitution formula the corresponding genetic code of the molecule is given. The vector representing the newly formed child constitution contains at every position exactly one of the both possible values obtained at the corresponding positions in the parent molecule vectors.

Figure 4: The process of Mutation is illustrated on two example constitution (a) and (b). The gray shaded area is again conserved in the mutated constitution (c) while the white marked bond is changed. Below each constitution formula the corresponding genetic code of the molecule is given. The vector changes consequently exactly at two positions.

Figure 5: Flow chart diagram for one single generation performed during the genetic algorithm. The formation of *n* child populations containing *m* constitutions out of *n* parent populations is illustrated. The selection is performed by the calculation of the $\Delta\left({}^{13}C\right)$ values. Recombination and Mutation are performed according to the probabilities set with the RP and MP value. To ensure that a new constitutions is calculated one of both processes, recombination or mutation, have to take place. Finally for every generated population the *l* constitution with largest $\Delta\left({}^{13}C\right)$ value are replaced by the *l* fittest molecules out of the parent set of structures.

Figure 6: The optimization of the parameters is illustrated for elucidating the Isoleucin structure by the implemented genetic algorithm. The $\Delta\left({}^{13}C\right)$ of the best found solution is always displayed on the z axis. The number of calculated generations is given at the x axis (in the diagram (b) the axis is scaled) and the optimized parameter is displayed at the y axis. Regions on the surfaces coded in the same color are isobars of $\Delta\left({}^{13}C\right)$. Diagram (a) proves that an increase of the population size *m* causes a faster convergence of the algorithm due to a higher number of structures generated. Diagram (b) obtains the comparison of differing population sizes *m* rescaled to an equal number of generated structures for a fair comparison. The results of the optimization of the part of conserved individuals *l* is optimized in diagram (c). The number of parallel calculated populations *n* is again optimized with special care of a comparable number of generated structures in experiment (d). The introduced multiplicity deviation factor MDF is investigated in experiment (e) and the influence of mutation probability MP as well as recombination probability RP are visualized in diagram (f).

Figure 7: Optimization process for the example molecule of Isoleucin. For the 11 performed optimization steps the $\Delta\left({}^{13}C\right)$ values for all generated constitutions are given. All constitutions that participate in the formation of the correct solution Isoleucin in the 11th generation are displayed and the performed operation is marked by different arrows as indicated in the legend. The right column of the figure

illustrates the used recombination and mutation probability values in the single steps.

References and Notes.

(1) Benecke, C.; Grund, R.; Hohberger, R.; Kerber, A.; Laue, R.; Wieland, T. MOLGEN+, a generator of connectivity isomers and stereoisomers for molecular structure elucidation, *Anal. Chim. Acta* **1995**, *314*, 141-147.

(2) Wieland, T.; Kerber, A.; Laue, R. Principles of the Generation of Constitutional and Configurational Isomers, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 413-419.

(3) Lindel, T.; Junker, J.; Köck, M. COCON: From NMR Correlation Data to Molecular Constitution, *J. Mol. Model.* **1997**, *3*, 364-368.

(4) Köck, M.; Junker, J.; Maier, W.; Will, M.; Lindel, T. A COCON Analysis of Proton-Poor Heterocycles - Application of Carbon Chemical Shift Predictions for the Evaluation of Structural Proposels, *Eur. J. Org. Chem.* **1999**, 579-586.

(5) Meiler, J.; Will, M.; Meusinger, R. Fast Determination of 13C-NMR Chemical Shifts Using Artificial Neural Networks, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1169-1176.

(6) Meiler, J.; Köck, M. Structure Elucidation by Automatic Generation and Analysis of Molecule Databases from NMR Connectivity Information Using Substructure Analysis and 13C-NMR Chemical Shift Prediction, *submitted* **2001**.

(7);; Chemical Concepts: Karlsruhe, 2001.

(8) Will, M.; Fachinger, W.; Richert, J. R. Fully Automated Structure Elucidation - A Spectroscopist's Dream Comes True, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 221-227.

(9) Zupan, J.; Gasteiger, J.; VCH Verlagsgesellschaft mbH: Weinheim, 1993.

(10) Kvasnicka, V.; Sklenak, S.; Pospichal, J. Application of Recurrent Neural Network in Chemistry. Prediction and Classification of 13C NMR Chemical Shiftsin a Series of Monosubstituted Benzenes, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 742-747.

(11) Doucet, J.-P.; Panaye, A.; Feuilleaubois, E.; Ladd, P. Neural networks and carbon-13 NMR shift prediction, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 320-324.

(12) Panaye, A.; Doucet, J.-P.; Fan, B. T.; Feuilleaubois, E.; Azzouzi, S. R. E. Artificial neural network simulation of 13C NMR shifts for methyl-substituted cyclohexanes, *Chemom. Intell. Lab. Syst.* **1994**, *24*, 129-135.

(13) Sklenak, S.; Kvasnicka, V.; Pospichal, J. Prediction of 13C NMR chemical shifts by neural networks in a series of monosubstituted benzenes, *Chem. Pap.* **1994**, *48*, 135-140.

(14) Clouser, D. L.; Jurs, P. C. Simulation of the 13C nuclear magnetic resonance spectra of trisaccharides using multiple linear regression analysis and neural networks, *Carbohydr. Res.* **1995**, *271*, 65-77.

(15) Svozil, D.; Pospichal, J.; Kvasnicka, V. Neural Network Prediction of Carbon-13 NMR Chemical Shifts of Alkanes, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 924-928.

(16) Thomas, S.; Kleinpeter, E. Assignment of the 13C NMR chemical shifts of substituted naphthalenes from charge density with an artificial neural network, *J. Prakt. Chem./Chem.-Ztg.* **1995**, *337*, 504-507.

(17) Clouser, D. L.; Jurs, P. C. Simulation of the 13C Nuclear Magnetic Resonance Spectra of Ribonucleosides Using Multiple Linear Regression Analysis and Neural Networks, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 168-172.

(18) Ivanciuc, O.; Rabine, J. P.; Cabrol-Bass, D.; Panaye, A.; Doucet, J.-P. 13C NMR Chemical Shift Prediction of sp2 Carbon Atoms in Acyclic Alkenes Using Neural Networks, *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 644-653.

(19) Meiler, J.; Meusinger, R.; Will, M. Neural Network Prediction of 13C NMR Chemical Shifts of Substituted Benzenes, *Monatshefte für Chemie* **1999**, *130*, 1089-1095.

(20) Pearlman, D. A. Automated detection of problem restraints in NMR data sets using the FINGAR genetic algorithm method, *J. Biomol. NMR* **1999**, *13*, 325-335.

(21) Fisher, B.; Dillon, N.; Carpenter, T.; Hall, L. Design by Genetic Algorithm of a Z Gradient Set for Magnetic Resonance Imagine of the Human Brain, *Measurment Science & Technology* **1995**, *6*, 904-909.

(22) Meusinger, R.; Moros, R., In *Software - Entwicklung in der Chemie*; Gasteiger, J., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1995; Vol. 10, pp 209-216.

(23) Pearlman, D. A. FINGAR: A newgenetic algorithm-based method for fitting NMR data, *J. Biomol. NMR* **1996**, *8*, 49-66.

(24) Li, L.; Darden, T. A.; Freeman, S. J.; Furie, B. C.; Furie, B.; Baleja, J. D.; Smith, H.; Hiskey, R. G.; Pedersen, L. G. Refinement of the NMR Solution Structure of the γ-Carboxyglutamic Acid Domain of Coagulation Factor IX Using Molecular Dynamics Simulation with Initial Ca2+ Positions Determined by a Genetic Algorithm, *Biochemistry* **1997**, *36*, 2132-2138.

(25) Choy, W. Y.; Sanctuary, B. C. Using Genetic Algorithms with a a Priori Knowledge for Quantitative NMR Signal Analysis, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 685-690.

(26) Robien, W. Das CSEARCH-NMR-Datenbanksystem, *Nachr. Chem. Tech. Lab.* **1998**, *46*, 74-77.

(27) Clerc, J.-T.; Sommerauer, H. A Minicomputer Program Based On Additivity Rules For The Estimation Of 13C NMR Chemical Shifts, *Anal. Chim. Acta* **1977**, *95*, 33-40.

(28) Ebraheem, K. A. K.; Webb, G. A. Semi-Empirical Calculations of the Chemical Shifts of Nuclei other than Protons, *Progress in NMR Spectroscopy* **1977**, *11*, 149-181.

(29) Bremser, W.; Ernst, L.; Franke, B.; Gerhards, R.; Hardt, A.; Verlag Chemie: Weinheim, 1981.

(30) Fürst, A.; Pretsch, E. A computer program for the prediction of 13C NMR chemical shifts of organic compounds, *Anal. Chim. Acta* **1990**, *229*, 17-25.

(31) Meiler, J., *www.jens-meiler.de* **2001**.

Table 1: fully automated structure elucidation for a database containing 8•20 molecules with 9 to 16 heavy atoms

number heavy atoms:	9	10	11	12	13	14	15	16
average Δ(^{13}C) (ppm) for correct solutions[a]:	1.23	1.13	1.19	1.00	0.97	1.10	1.18	1.15
number correct solutions[b]:	20	20	18	16	14	14	5	4
algorithm stopped with smaller deviation than target[c]:	0	0	0	0	0	1	2	4
average calculation time (min)[d]:	2	2	13	23	37	51	85	123
average number steps[e]:	18	13	82	154	197	258	375	414
generated structures per minute[f]:	1952	1733	1605	1708	1375	1293	1133	863

[a] average Δ(^{13}C) (ppm) value of the 20 molecules representing the correct solutions

[b] total number of molecules with correctly determined solutions out of the 20 tested molecules

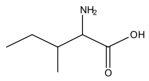
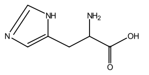
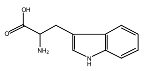
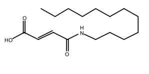
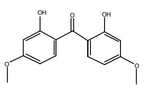
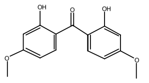
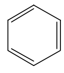
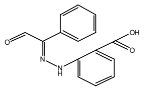
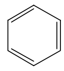
[c] number of test runs stopped because a structure with a smaller Δ(^{13}C) than the correct solution structure was created out of the 20 testes molecules

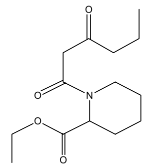
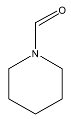
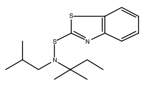
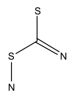
[d] total average calculation time in minutes on a PII processor with 450 MHz and 512 MB RAM

[e] total average number of steps until the algorithm was stopped

[f] average number of generated and tested structures per minute

Table 2: molecular structures, parameters and results obtained for some example molecules solved by the genetic algorithm approach

ID	**molecule properties**						**parameters for genetic algorithm**						**results**		
	name	molecular formula	numb. heavy atoms	structure	Δ(^{13}C) (ppm) [a]	number possible structures [b]	*n* [c]	*m* [d]	*l* [e]	MDF (ppm) [f]	MP[g] | RP[h]	excluded or included substructures [i]	numb. steps [j]	calcu-lation time[k] (min)	struct. / time[l] (min^{-1})
1	Isoleucin	$C_6H_{13}NO_2$	9		1.12	23,946	1	32	8	1.00	$^{0}1.0^{4}$ => $^{8}0.5$ | $^{0}0.0^{4}$ => $^{8}1.0$	---	11	<1	1408
2	Histidin	$C_6H_9N_3O_2$	11		1.63	89,502,542	5	50	25	2.00	$^{0}1.0^{10}$ => $^{15}0.5$ | $^{0}0.0^{10}$ => $^{15}1.0$	For these examples a bad list of fragments was introduced that excludes allens (C=C=C) and any bonds between hetero atoms (X---X)	19	4	1187
3	Trypto-phan	$C_{11}H_{12}N_2O_2$	15		1.44	≈36•10^{9}	1	60	30	2.00	$^{0}1.0^{100}$ => $^{200}0.5$ | $^{0}0.0^{100}$ => $^{200}1.0$		327	20	981
4	-	$C_{16}H_{29}NO_3$	20		0.81	≈66•10^{9}	64	64	16	1.00	$^{0}1.0^{25}$ => $^{50}0.0$ | $^{0}0.5^{25}$ => $^{50}1.0$		71	453	642
5	-	$C_{15}H_{14}O_5$	20		1.77	?	64	64	16	1.00	$^{0}1.0^{25}$ => $^{50}0.0$ | $^{0}0.5^{25}$ => $^{50}1.0$		1113	6704	680
5'	-	$C_{15}H_{14}O_5$	20		1.77	?	1	64	16	1.00	$^{0}1.0^{25}$ => $^{50}0.0$ | $^{0}0.5^{25}$ => $^{50}1.0$	include 2x 	38	3	810
6	-	$C_{15}H_{12}N_2O_3$	20		1.89	?	16	128	32	1.00	$^{0}1.0^{25}$ => $^{50}0.0$ | $^{0}0.5^{25}$ => $^{50}1.0$	include 2x 	4	18	455

7	-	$C_{14}H_{23}NO_5$	20		1.36	?	8	32	8	1.00	${}^{0}1.0^{25}$ => ${}^{50}0.0$ | ${}^{0}0.5^{25}$ => ${}^{50}1.0$	include 	10	3	853
8	-	$C_{16}H_{24}N_2S_2$	20		0.79	?	16	64	16	1.00	${}^{0}1.0^{25}$ => ${}^{50}0.0$ | ${}^{0}0.5^{25}$ => ${}^{50}1.0$	include 	302	540	573

[a] $\Delta(^{13}C)$ (ppm) value

[b] total number of possible constitutions generated by MOLGEN if available

[c] number of parallel calculated populations (**n**)

[d] number of individuals in the populations (**m**)

[e] number of best ranked (small $\Delta(^{13}C)$ value) individuals in the parent generation that are conserved for the new child generation (**l**)

[f] **M**ultiplicity **D**eviation **F**actor defines the penalty added to $\Delta(^{13}C)$ for a wrong multiplicity of a ^{13}C carbon signal in a generated structure

[g] **M**utation **P**robability is the probability that a generated structure undergoes a mutation in a certain part of the algorithm.

[h] **R**ecombination **P**robability is the probability that a structure generated by recombination of two individuals of the parent generation and not by just copying and mutating an individual of the parent generation.

[i] included structural fragments that have to be used in all generated structures and excluded structural fragments that are forbidden in all generated structures

[j] total number of steps until the correct solution was found

[k] total calculation time in minutes on a PII processor with 450 MHz and 512 MB RAM

[l] number of generated and tested structures per minute

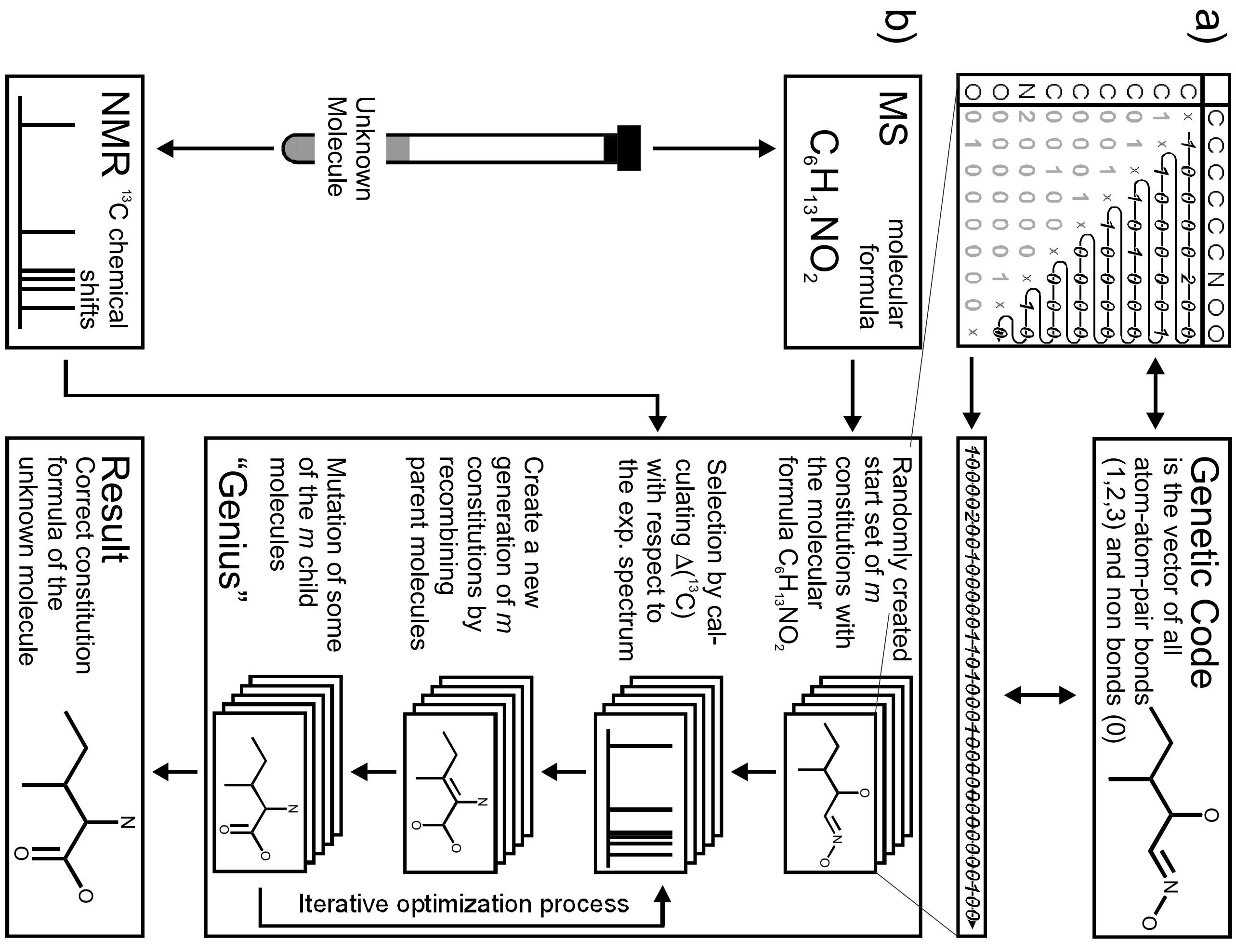

a)
C C C C C C N O O
Genetic Code
is the vector of all atom-atom-pair bonds (1,2,3) and non bonds (0)
100002001000001101000100000000000100
b)
MS molecular formula
$C_6H_{13}NO_2$
Unknown Molecule
NMR ^{13}C chemical shifts
Randomly created start set of *m* constitutions with the molecular formula $C_6H_{13}NO_2$
Selection by calculating $\Delta(^{13}C)$ with respect to the exp. spectrum
Create a new generation of *m* constitutions by recombining parent molecules
Mutation of some of the *m* child molecules
"Genius"
Iterative optimization process
Result
Correct constitution formula of the unknown molecule

Fig 1

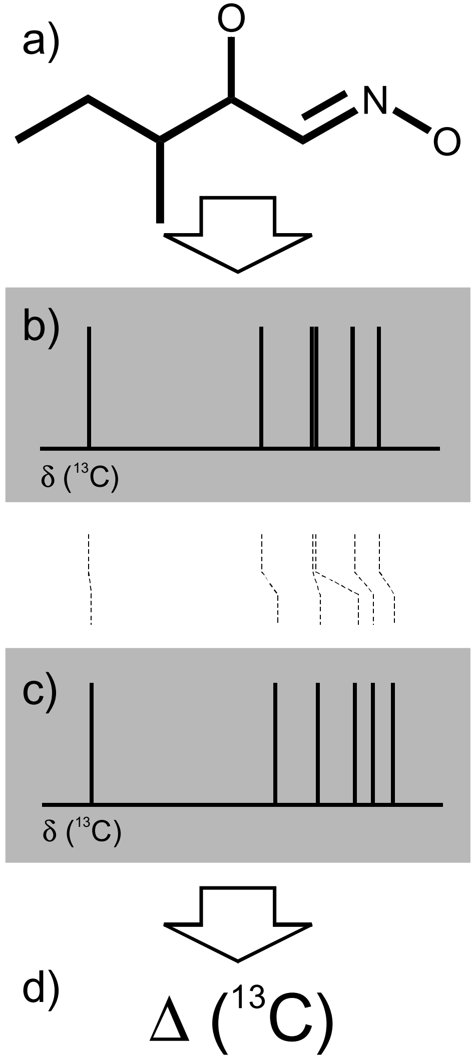

a)
O
N
O
b)
δ (^{13}C)
c)
δ (^{13}C)
d)
Δ (^{13}C)

Fig 2

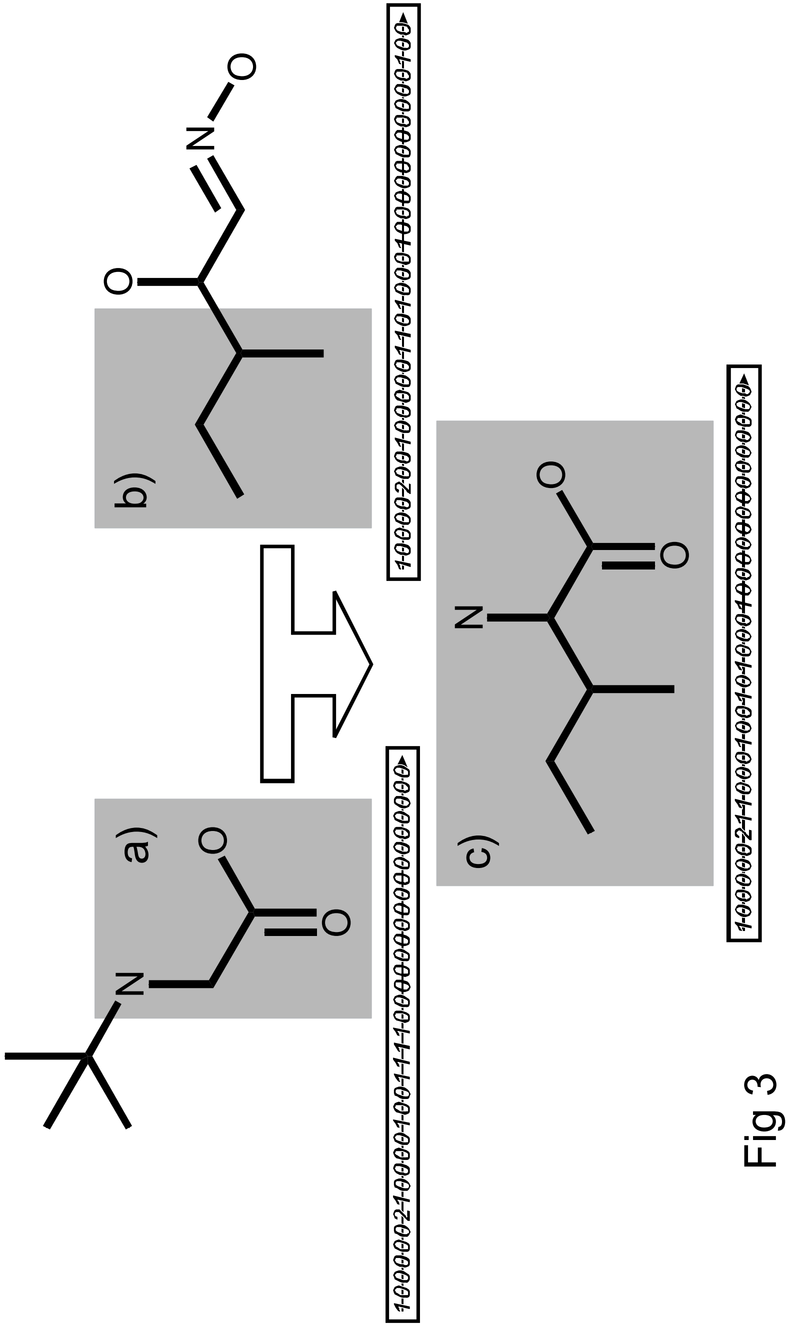

a)
N
O
O
1-0-0-0-0-0-2-1-0-0-0-0-1-0-0-1-1-1-1-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0
b)
O
N
O
1-0-0-0-0-2-0-0-1-0-0-0-0-0-1-1-0-1-0-1-0-0-0-1-0-0-0-0-0-0-0-0-0-0-1-0-0
c)
N
O
O
1-0-0-0-0-0-2-1-1-0-0-0-1-0-0-1-0-1-0-1-0-0-0-1-0-0-0-0-0-0-0-0-0-0-0-0

Fig 3

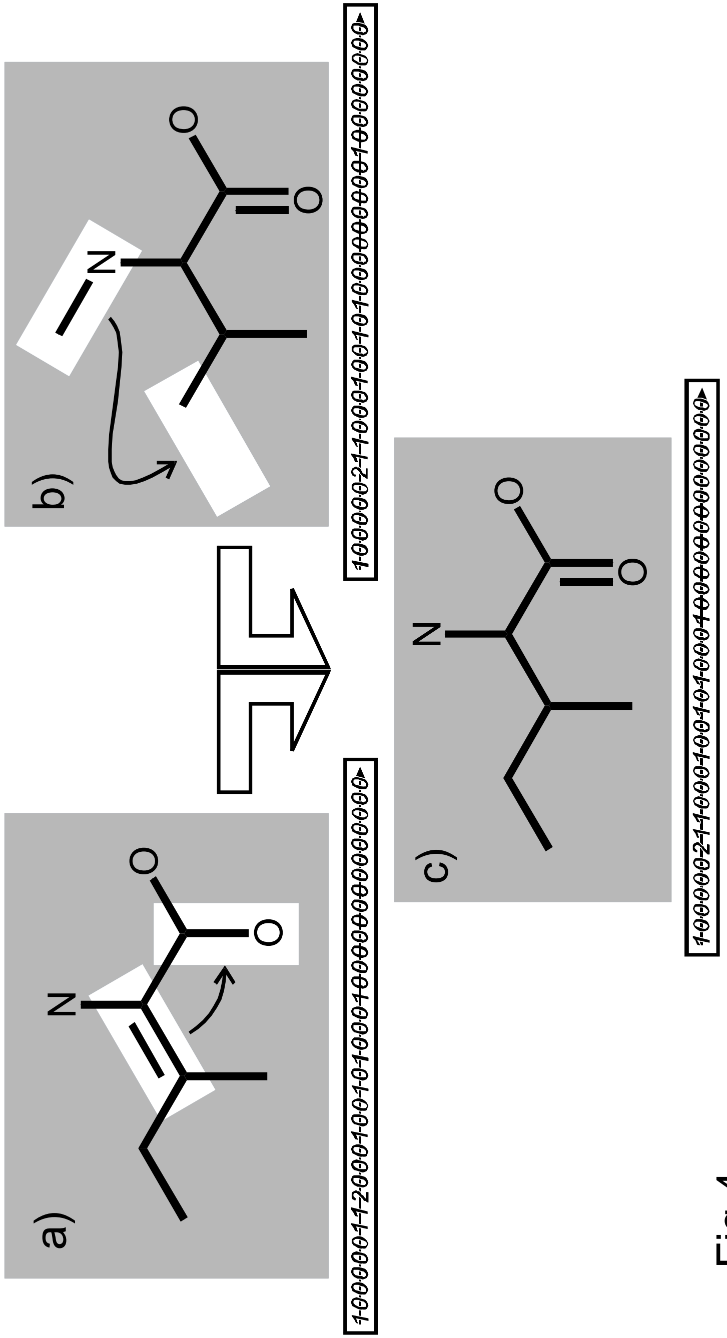

a)
b)
c)
N
O
O
1-0-0-0-0-0-1-1-2-0-0-0-1-0-0-1-0-1-0-0-0-1-0-0-0-0-0-0-0-0-0-0-0-0-0-0
1-0-0-0-0-0-2-1-1-0-0-0-1-0-0-1-0-1-0-0-0-0-0-0-0-0-1-0-0-0-0-0-0-0-0-0
1-0-0-0-0-0-2-1-1-0-0-0-1-0-0-1-0-1-0-0-0-1-0-0-0-0-0-0-0-0-0-0-0-0-0-0

Fig 4

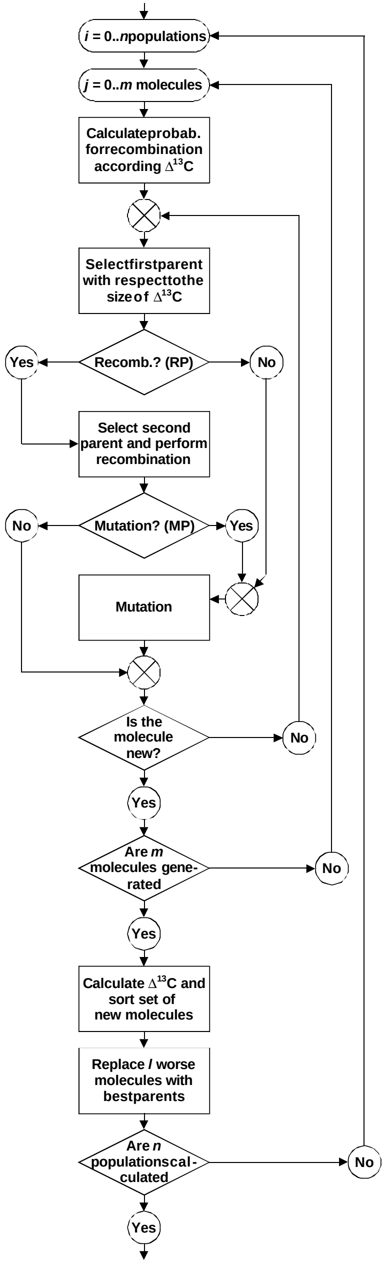

i = 0..*n*populations
j = 0..*m* molecules
Calculateprobab. forrecombination according Δ^{13}C
Selectfirstparent with respecttothe sizeof Δ^{13}C
Recomb.? (RP)
Yes
No
Select second parent and perform recombination
Mutation? (MP)
No
Yes
Mutation
Is the molecule new?
No
Yes
Are *m* molecules gene-rated
No
Yes
Calculate Δ^{13}C and sort set of new molecules
Replace / worse molecules with bestparents
Are *n* populationscal-culated
No
Yes

Fig 5

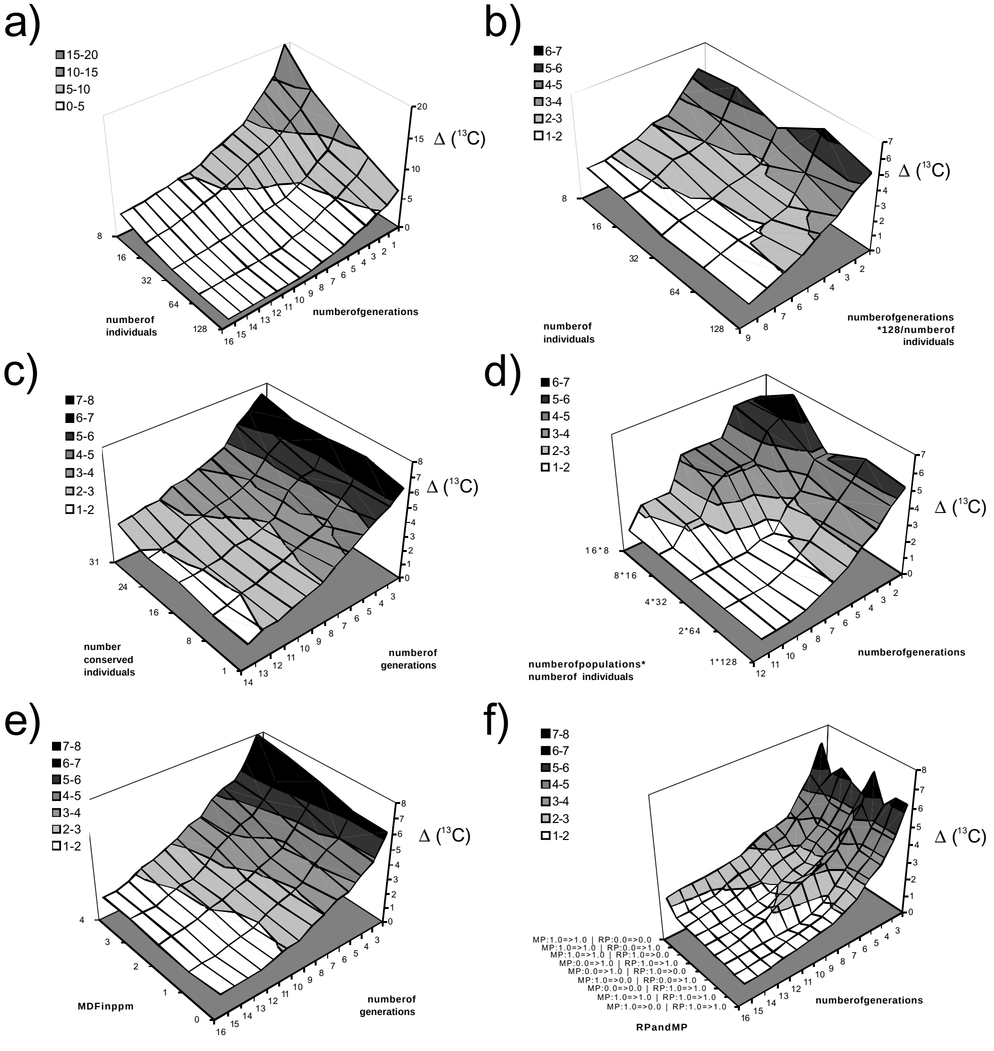

a)
15-20
10-15
5-10
0-5
Δ (¹³C)
numberof individuals
numberofgenerations
b)
6-7
5-6
4-5
3-4
2-3
1-2
Δ (¹³C)
numberof individuals
numberofgenerations *128/numberof individuals
c)
7-8
6-7
5-6
4-5
3-4
2-3
1-2
Δ (¹³C)
number conserved individuals
numberof generations
d)
6-7
5-6
4-5
3-4
2-3
1-2
Δ (¹³C)
numberofpopulations* numberof individuals
16*8
8*16
4*32
2*64
1*128
numberofgenerations
e)
7-8
6-7
5-6
4-5
3-4
2-3
1-2
Δ (¹³C)
MDFinppm
numberof generations
f)
7-8
6-7
5-6
4-5
3-4
2-3
1-2
Δ (¹³C)
MP:1.0=>1.0 | RP:0.0=>0.0
MP:1.0=>1.0 | RP:0.0=>1.0
MP:1.0=>1.0 | RP:1.0=>0.0
MP:0.0=>1.0 | RP:1.0=>1.0
MP:0.0=>1.0 | RP:1.0=>0.0
MP:1.0=>0.0 | RP:0.0=>1.0
MP:0.0=>0.0 | RP:1.0=>1.0
MP:1.0=>1.0 | RP:1.0=>1.0
MP:1.0=>0.0 | RP:1.0=>1.0
RPandMP
numberofgenerations

Fig 6

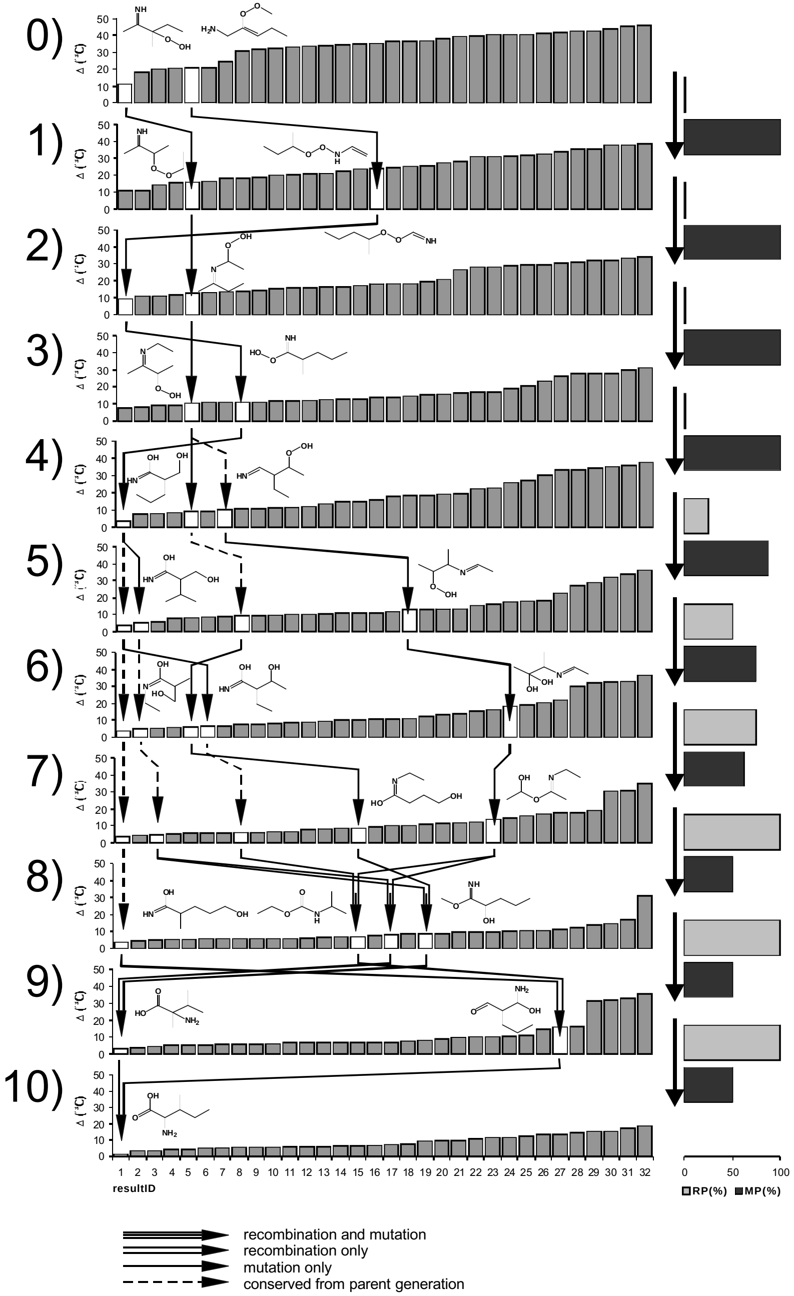

0)
1)
2)
3)
4)
5)
6)
7)
8)
9)
10)
Δ (¹³C)
0
10
20
30
40
50
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
resultID
0
50
100
RP(%)
MP(%)
recombination and mutation
recombination only
mutation only
conserved from parent generation

Fig 7

Generation and Evaluation of Dimension Reduced Amino Acid Parameter Representations by Artificial Neural Networks

JENS MEILER*, MICHAEL MÜLLER, ANITA ZEIDLER, FELIX SCHMÄSCHKE

Key words. Amino acid parameters; neural networks; quantitative structure property relation; secondary structure prediction

*To whom correspondence should be addressed

Contact address:

Jens Meiler
Universität Frankfurt (AK Griesinger)
Marie-Curie-Str. 11
60439 Frankfurt am Main

Tel.: 069 798 29 798
Fax.: 069 798 29 128
Mail: mj@org.chemie.uni-frankfurt.de

Abstract. In order to process data of proteins often a numerical representation for an amino acid is necessary. Many suitable parameters can be derived from experiments or statistical analysis of databases. To ensure a fast and efficient use of these source of information, a reduction and extraction of relevant information out of these parameters is one basic need. In this approach established methods like principal component analysis (PCA) are supplemented by a method based on symmetric neural networks. Two different parameter representations of amino acids are reduced from five and seven dimensions, respectively to one, two, three or four dimensions by using a symmetric neural network approach alternatively with one or three hidden layers. It is possible to create general reduced parameter representations for amino acids. To demonstrate the ability of this approach, these reduced sets of parameters are applied for the ab initio prediction of protein secondary structure from primary structure only. Artificial neural networks are implemented and trained with a diverse representation of 430 proteins out of the PDB. An essentially faster training and also prediction without a decrease in accuracy is obtained for the reduced parameter representations in comparison with the complete set of parameters. The method is transferable to other amino acids or even other molecular building blocks, like nucleic acids and therefore a general approach.

Introduction. In between artificial neural networks are a wide spread and intensively discussed method for analyzing data and describing relationships in chemistry [1]. Neural networks are often the preferred solution if the dependence cannot be expressed by a simple mathematical equation or this equation is unknown and also not important for solving the problem. Moreover neural networks are particularly suitable for the work with blurred information and able to apply learned correlations to unknown examples. A lot of applications of these networks already exist in chemistry [2-19] and in particular also for secondary structure prediction of proteins [20-50].

However, the focus of this paper is the introduction of a so far not intensively discussed possibility of using neural networks: Reducing the number of dimensions for a parameter representation like traditional performed by principal component or cluster analysis. The method is introduced by Livingstone, et. al. [51] and also used by Kocjancic and Zupan [52] as a mapping device. Here it is applied to reduce the dimension of property representations of amino acids. The potential of these reduced parameter representations is demonstrated by comparing them with the complete property representations in its ability to serve as input for another neural network that predicts the secondary structure of proteins from primary structure only.

System and Methods. A dataset of l individuals containing m properties for each of this individuals is called a m-dimensional property representation of these l individuals. These m properties can be projected into n dimensions with $n < m$ by principal component analysis or cluster analysis. This is of special interest for obtaining linear dependencies between these properties and of course, for visualizing relations between the l individuals. However, these purposes can also be gained using artificial neural networks. As given in figure 1 a three layer neural network can be built with m inputs and m output neurons but only n hidden neurons. Due to their symmetric architecture: m inputs, n hidden neurons, m output neurons these

networks are called symmetric networks. Note, that this is somehow a critical name, since the nodes in the input layer act mainly as distributors. Therefore they are also not regarded as neurons in figure 1. They process the non weighted input data without summarizing and with a linear transfer function $y = x$ to pass the information to the hidden neurons. So the three basic features of an artificial neuron: The summation of previously weighted information to process it with a transfer function are hardly recognizable in this case. Therefore the nodes of the first layer have another structure than the neurons in further layers. Thus the phrase “symmetric” just targets at the symmetric distribution of all neurons including the nodes of the input layer relative to the central hidden layer. It does not reflect, that the nodes of the input layer have a significant different structure than the neurons of the output layer. Moreover also the weights are not restricted to be symmetric in their value with respect to the central hidden layer.

However, training this network with the m properties of l individuals to predict again the same m properties provides a network where in the hidden layer all the information is represented by n numbers. If now these n numbers for the l individuals are obtained, a n-dimensional representation of the m properties for each of the l individuals is found. Depending on the transfer function, the representing parameters lay between 0 and 1 (e.g. $\mathrm{sigmoid}(x) = \frac{1}{1+e^{-x}}$) or between –1 and 1 (e.g. $\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$). The completeness of this representation can be obtained by investigating the deviation of the original properties with the properties back calculated by the three layer network.

Using a neural network with only one hidden layer, the dependence of the reduced parameter set $p'_j\left(1 \le j \le n\right)$ to the original one $p_i\left(1 \le i \le m\right)$ is linear accept of the sigmoidal transfer function: $p'_j = \left(1+e^{\left(-\sum_{i=1}^{m} w_{ij} p_i\right)}\right)^{-1}$. Also the back transformation of the reduced parameters to the original uses the same “pseudo linear” equation. So the similarity of this approach to a principal component analysis is obvious. Each of the derived parameters is a linear combination of the original parameters processed with the transfer function.

To analyze the data due to a more complex dependence a five layer network can be created, again with m input and output neurons and n neurons in the third layer as given in figure 1. Now an additional hidden layer allows a modification of the data before the reduced parameter representations are obtained in the central hidden layer and another hidden layer is inserted after the central hidden layer (figure 1). The model is much more complex than linear and therefore able to simulate even complicate polynomial functions due to the very flexible network model.

The number of neurons in these additional hidden layers could obviously be varied between n and m. Less than n neurons in these layers would force the network to reduce the dimension in the second and the fourth layer even further than just to n dimensions, whereas more than m neurons would distribute m input values in even more parameters in the second and the fourth layer. As expected the lowest RMSD value in the optimization of these networks is obtained using m neurons in the second and in the fourth layer. In this case for every input and output neuron one working neuron in the additional layers is provided to process the information. The two latter points justify the use of m neurons in the second and the fourth layer. Since it is the largest possible value between n and m it is consequently the most complex model and therefore capable to achieve the biggest contrast to the three layer case. Again the property information is represented by only n numbers in the third layer.

For a more detailed description of the method as well as a comparison to other methods see literature [51-53].

Algorithm and Implementation. The generation and the testing of the reduced parameter representations is performed in two consecutive steps:

- Training of symmetric neural networks with property representations of all twenty naturally occuring amino acids and obtaining the reduced parameter representations from these networks.

- Testing the reduced parameter representations by comparison with the complete parameter representations. Therefore both parameter representations are used to code protein sequences and their secondary structure is predicted by training another artificial neural network with these numbers.

For the calculation of the reduced parameter sets five properties of amino acids are used: a sterical parameter, hydrophobicity, volume, polarizability [54], and isoelectric point (Table 1).

The sterical parameter is the graph shape index Ξ. This parameter encodes complexity, branching and symmetry of a group and can be directly calculated from the graph structure of the amino acid side chain.

The hydrophobicity π is defined as a side chain parameter as π(side chain) = log P(amino acid) – log P(glycine) in which P is the partition coefficient of the amino acid in octanol/water.

The normalized van der Waals volume υ_v is defined by υ_v(side chain) = {V(side chain) – V(H)} / V(CH_2). The measure is therefore 0 for glycine and 1 for alanine.

The polarizability α is related to the molar refractivity. It is given by:

$$\alpha = \frac{3}{4\pi N}\frac{M}{d}\frac{n^2-1}{n^2+2}$$

(n: index of refraction, M: molecular weight, d: density, N: number of atoms)

These values are used alone (*five parameter set*) as well as in combination with two statistical parameters: helix and sheet probability (*seven parameter set*, Table 1).

The secondary structure probabilities are extracted from a sub space of the PDB [55] containing 430 proteins with over 60,000 residues. So every out of the twenty naturally occurring amino acids is represented by five or seven properties, respectively.

Overall 24 small neural networks are trained with different number of hidden neurons to obtain the reduced parameter representations:

number hidden neurons	*three layer network*	*five layer network*
five parameter set	1 , 2 , 3 , 4 , 5	1 , 2 , 3 , 4 , 5
seven parameter set	1 , 2 , 3 , 4 , 5 , 6 , 7	1 , 2 , 3 , 4 , 5 , 6 , 7

For the mapping procedure no testing set of data is required. All 20 amino acids are therefore a part of the training set of data. The training is continued until the root mean square deviation of the recalculated property values is minimized. After the training process the networks are "cut" after the second (three layer network) or third layer (five layer network) to obtain the reduced parameter sets. The central hidden layer becomes the output layer and the values detected by applying the property values at the inputs for all 20 amino acids provide the parameter representations.

In order to test the information conserved in these reduced parameter representations, neural networks are trained to predict secondary structure of proteins from primary structure only: Basically the secondary structure is predicted for every amino acid in one run. The sequence information is provided as input for a symmetric window around this amino acid of interest. The secondary structure of a protein is calculated by moving this window over the sequence and calculating the secondary structure for every amino acid individually. This concept of a moving window widely used for this purpose (e.g.[42]).

The input values for every amino acid are the m properties of the amino acid or their reduced n-dimensional parameter representations. The number of values necessary to describe one amino acid is therefore not fixed but varies between one and seven. In our example the size of the window is optimized to contain the central amino acid as well as 15 amino acids before and after this central amino acid (figure 2). So this initial window has a size of 31 amino acids. This is a compromise between a window as large as possible to provide a most possible complete sequence information and a preferable small input layer to minimize the number of connections. Since the number of connections is equal to the degrees of freedom in the artificial neural network, this parameter determines the necessary training information (number of sequences with known secondary structure) as well as the time for the training procedure.

Although the next neighbors of the amino acid of interest have a larger impact on the formation of secondary structure fragments, also the part of the sequence not represented by the 31 amino acid window influences the formation of secondary structure. This is due to the fact, that the secondary structure is not only influenced by short primary structure sequences but also by long range interactions that occur in the tertiary structure and can therefore contain interactions between amino acids that are far apart from each other in the primary sequence.

To enable the network to use at least some information about the non considered parts of the sequence, all amino acids before and after this window of 31 amino acids are incorporated into average property and parameter values. These averages somehow represent the character of the remaining part of the protein that surrounds the 31 amino acid window (e. g. mainly aliphatic amino acids, pH character, …). The averages are computed for four groups of amino acids before and after the 31 amino acid window, respectively. The first three groups contain the average values for five amino acids each. They hold therefore information for another 30 amino acids, 15 on both sides of the window. The fourth group on both sides contain the average values of all remaining amino acids before or beyond this window of now 61 amino acids to the start or to the end of the sequence. As visualized in figure 2 this leads to 31 + 8 = 39 groups of input properties (or parameters). For every out of this 39 groups either m properties or n parameters are used as description, which leads to $39 \cdot m$ and $39 \cdot n$ input neurons, respectively. The use of these additional four groups with averaged parameters on both sides leads to an overall improvement of about 2% in the prediction (compare with the later discussed Q_3 values).

All established three layer neural networks contain 39 hidden neurons. This number is optimized for the network trained with the seven property representation and remains constant to ensure comparable conditions for all experiments.

For every amino acid probabilities for being a part of a α-helix, β-sheet or coil (in the range of 0..1) are obtained at the output layer. "Coil" covers in our case all other secondary structure elements, loops and turns accept α-helix and β-sheet. For the training of these networks these probabilities are set to be "1 0 0", "0 1 0" or "0 0 1", respectively.

By optimizing these networks it turned out, that an increase of correct predicted secondary structure is obtained by predicting the secondary structure of more than one amino acid in one run. The optimum is found by calculating probabilities for 5 amino acids before and after the central amino acid, respectively. Therefore eleven amino acid probability sets are predicted parallel in one network run which leads to 33 output neurons for all neural networks (Figure 2). By moving the window over the amino acid sequence, every single amino is part of the output window exactly eleven times. These eleven predicted probabilities for one amino acid are combined by a triangular weighted average. The prediction is weighted with one, if the amino acid is the central and the weight is reduced as the amino acid moves to the edges of this window of eleven amino acids. The vector of the eleven weights is consequently w = (0.166, 0.333, 0.500, 0.666, 0.833, 1.000, 0.833, 0.666, 0.500, 0.333, 0.166). The three probabilities are computed by $p^{H,S,C} = \sum_{i=1}^{11} p_i^{H,S,C} \cdot w_i / \sum_{i=1}^{11} w_i$ where $p_i^{H,S,C}$ are the eleven

predicted probabilities for an amino acid to be part of a helix, a sheet or a coil region. This procedure allows to correct a possible wrong judgment for the central amino acid alone. In about 3% of all cases this correction takes place and the overall accuracy is therefore improved by 3% using this procedure (compare with the later discussed Q_3 values).

All in all 16 neural nets using different property and parameter representations as input are trained. Two networks with the complete five and seven parameter representations as well as fourteen networks using one, two, three and four (only for the seven property representation) dimensional parameter representations obtained from three and five layer networks with five and seven properties used for training. To ensure the use of all known folds in this procedure, the FSSP database (http://www.ebi.ac.uk/dali/fssp/) introduced by Holm and Sander [55] is used.

For the training of these networks 430 peptides derived from this database are separated into two set of data: first with 95% of the peptides for training the networks and second with the remaining 5% for the testing process. The networks are trained until the root mean square deviation of the testing data set is minimized. The probabilities between 0 and 1 are obtained using sigmoidal function for transfer and back propagation of errors as training method. All neural networks are trained and analyzed using the program "Smart"[56].

Discussion. Figure 3 gives the root mean square deviation of the recalculated and normalized property values in dependence of the number of layers, number of hidden neurons and number of properties. As expected a network with m hidden neurons is able to recalculate the m given properties totally. The small deviation obtained for the networks with five or seven hidden neurons, respectively, is due to the use of an optimization instead of a direct calculation of the weights. However, differences become observed for networks with a reduced number of hidden neurons. First of all, networks with five layers are able to reproduce the data with a essentially smaller deviation than nets with only three layers using the same number of hidden neurons. This is in line and expected due to the ability of these nets to simulate more complicate dependencies and proves, that in these cases more information can be projected into a n parameter representation. In this case for both, the five property representation as well as the seven property representation, a nearly complete description of the properties is given by only three numbers. The RMSD is about 0.020 in both cases. Essentially larger is the RMSD for the three layer networks with about 0.053 and 0.080 for the five property and for the seven property fit, respectively. These values are in the order of the according networks with only one hidden neuron but five layers. However, one has to keep in mind, that in the networks with five layers, two layers are used to recalculate the property values from the parameter representation instead of only one layer in the three layer network. So the recalculation uses again a much more complicated model and of course more weights. The improvement of the prediction going from five to seven properties and especially for going from three to five layer networks is clearly observable.

All derived reduced parameter representations are given in tables 2 and 3 and some examples are visualized in figure 4. Groups of amino acids with similar properties are plotted closer together and become better separated by increasing the number of layers from three to seven and also by using seven instead of five properties, already for the one dimensional representations:

For example the aromatic amino acids are all together plotted in a range of 0.20 in the three layer case but in a range of 0.05 in the five layer case. Also the basic amino acids are plotted closer together and separated clearly from the rest of amino acids. Only Histidin is found between the aromatic and the basic amino acids due to its ambiguous character. Glutamin and Asparagin are plotted closer together and become clearly separated from Glutaminic and Asparaginic acid going from three to five layer networks. While also Methionin and Cystein are plotted close together, aliphatic amino acids are relatively wide spread but sorted by the size of the side chains. Also Serin and Threonin are projected with a large difference. This

marks the incomplete representation given in this one dimension and is improved by introducing further parameters. This incompleteness causes also the ambiguous position of Proline in the one dimensional representations (0.34, 0.45, 0.77, 0.00). The network obviously learns the large and easily obtainable differences first. The single neuron in the hidden layer reaches its capacity fast, and the Proline parameter becomes ambiguous. The RMSD values of the back calculated properties (figure 3) prove, that only a part of the information can be saved by the network. However, an increase of the dimensions will overcome this problem and the parameters for Proline will become rather well defined too. The still observable usually relatively small differences between the individual parameter representations are than only based on differences in the applied property sets (five via seven property representation) or on different models (three versus five layer network).

For the two and three dimensional parameter representations again an improvement can be obtained using two additional hidden layers. Moreover the representation changes essentially going from five to seven properties. A better use of space while using the seven property values and a clearer separation of the amino acid groups is obtained increasing the number of layers in the network.

Beside the information obtainable by this projection method about similarities in a data set, these reduced parameter representations given in tables 2 and 3 can be used as general reduced parameter sets for amino acids representation. Using the trained and cut neural networks, the same parameter representation can be calculated also for other amino acids. Of course the parameters represent a combination of the used properties and therefore they cannot be directly interpreted as easily understandable properties. Their advantage is the reduction in number. The visualization becomes possible and allows a graphical analysis of the parameter space. Calculation time can be saved by using for example three instead of seven parameters. For a 200 amino acid protein the representing code can so be reduced from 1,400 using all seven properties to only 600 numbers using a three dimensional parameter representation. In our special case the calculation of all secondary structure probabilities for the used database with 430 proteins can be reduced from 120s for the seven property values to 50s for the three dimensional parameter representation. However, this is only a relative small improvement considering the high speed of computers today. Much more impressive is the gain of time in the training process. The necessary time for stabilizing the weights depends not linearly on the number of weights, but increases much faster with higher number of weights. This is difficult to calculate in general, but for the same example the complete training process lasts about 24h for the seven property values but less than 4h in case of the three dimensional parameter representation. (All CPU times are obtained from a 450MHz Pentium II processor equipped with 512MB RAM.)

The method could also become more effective if a higher number of parameters can be projected in two or three dimensions. Considering the enormous number of amino acid and nucleic acid sequences the potential of this method to provide general parameter representations combining a large number of relevant properties is remarkable and can lead to better mapping pictures as well as substantial gain in processing time.

Moreover the influence of each property on a parameter can be extracted by a “sensitivity analysis” of the neural network. In order to do this, the value for each input is varied within the experimental input range respectively, while all other input are set to be zero. The covered range of the output neurons gives a sensitivity value between 0 and 1. If the influence of a particular input on a particular output value is high, the covered output range is large and the sensitivity becomes 1. However, if the input has no influence at all, the output value will remain constant and the sensitivity stays 0.

It has to be mentioned that this method gives only a qualitative picture for several reasons: The non changed input values are set to be zero which is just one realistic input signal. However, there are usually a lot more realistic input values, which are not used for this analysis. The use of another static realistic input signal might influence the obtained

sensitivities. Moreover, due to the constant chosen signals on all other neurons the method is unable to detect cross correlation effects between different input values. However, the method works reliable in our hands. The sensitivity values change below 10% by setting the non changed input data to other realistic input values. Even better preserved are the relations between the different sensitivities, that change below 5%.

The one dimensional parameter representation is found to be dominated by the volume while for example the isoelectric point has no influence at all. In the two parameter representation the first parameter is still dominated by the volume but the isoelectric point is projected into the second parameter together with the polarizability, which is also only badly represented in the one dimensional parameter. The three parameter representation takes already all the five parameters into consideration. The possibility of projecting a part of the sterical information together with hydrophobicity in parameter 1 and polarizability, volume and again hydrophobicity in parameter 2 are inline with the empirical understanding of these parameters.

In order to test reduced parameter representation, secondary structure prediction of proteins is chosen as an example problem. Methods for predicting secondary structure of proteins are widely discussed and therefore optimal for testing the derived amino acid parameters. However, this prediction is carried out to compare the results for the full and reduced parameter representations only and not to achieve an optimal secondary structure prediction. Therefore the setup is not optimized to give best results in this manner.

The Q_3 values (as introduced in the literature [21]) for the 16 trained neural network obtained for the test data set are given in Table 4. The results for the training data set are slightly better than the results obtained for the testing data as expected, and therefore not reported here. The correct prediction achievable with this straight forward use of the primary sequence of one protein only is 67% for the total seven parameter representation and 63% for the five parameter representation. However, the three and four reduced parameter representations offer results of the same quality. The prediction accuracy stays in all cases in a range of ±2%. In case of the five layer three dimensional representation it becomes even slightly better than the complete parameter representation.

The prediction accuracy increases with every new introduced dimension by about 5% for the first three dimension. No further increase is obtained introducing further dimensions. Since the prediction accuracy reaches the level obtained for the full parameter representations, the optimal prediction using this method of coding is reached by introducing the third parameter. The completeness of the description with these three parameters is already obtained analyzing the RMSD values provided in figure 3 and becomes now also visible in these results.

It is not surprising, that the reduced parameter set obtained from the five layer networks give only slightly better results than the reduced parameter set obtained from the three layer networks. The step of data interpretation not performed in the first case can be completed by the three layer neural network used for deriving secondary structure information. Thus a more significant improvement for parameter sets derived with symmetric five layer networks compared with three layer networks might become obtainable, if linear methods instead of neural network are used for the further data processing.

Figure 5 provides the analysis of the input sensitivities for the calculation of secondary structure with the seven property representation. These analysis leads to comparable results for all networks with a prediction accuracy better than 60%, so that the network using the complete seven parameter representation is chosen as an example. All sensitivities for one of the 39 input data blocks are summarized in order to derive information about the influence of each of this input data blocks. The values are normalized to give 1 for the highest sensitivity. As expected the sensitivity for the central amino acid is the highest. Moreover a comparable high sensitivity is obtained for the four bordering input blocks on both sides containing averaged information for the surrounding amino acids. Of special interest is moreover the

higher influence of amino acids located after the central amino acid in the PDB file. These are the amino acids synthesized after this central amino acid in nature and would therefore suggest that these amino acids have a higher influence on the folding behavior of the central amino acid. A plot of the sensitivities obtained only for the helix probability provides an increased sensitivity for the –8th, –4th, +4th and +8th amino acid. This result shows the influence of the local primary structure on the formation of hydrogen bonds to form an α-helix, which is also the reason for the higher accuracy in the prediction of α-helices.

Figure 6 gives the sensitivity summarized for the seven and five parameter representation, respectively obtained for the both networks using the complete parameter representations. The dominating influence of the helix and sheet probability in the seven parameter representation is easy to obtain. The improvement in the prediction obtainable for going from five to seven properties is caused by the two statistically derived values and therefore also the reason for their high sensitivity. The neural network extracts in this case an essential part of the information not on the basis of amino acid parameters but on the basis of the database knowledge. While the amino acid properties only reflect primary structure information, the database averages already contain information about preferred secondary and tertiary structure of proteins. This additional relevant information has consequently to improve the prediction of secondary structure.

A relative increase of the sensitivity for the volume and the sterical parameter is obtained for the five parameter representation with respect to the seven parameter representation as compensation for not provided secondary structure probabilities in this case. Both parameters have high influence on the formation of secondary structure as well known. However, better results obtained for the seven parameter representation prove that this information can not be replaced in total.

As widely discussed, errors in secondary structure prediction occur often at the begin and at the end of secondary structure elements so that their length or their position becomes ambiguous. However, most of the secondary structure elements are found. The comparison of the three predicted probabilities for helix, sheet and coil allows to decide, whether the network is “sure” about its judgment or whether a second possibility or even all three possibilities are of similar high probability.

Without going into to much detail, the available information is illustrated on one small protein, ubiquitin in figure 7. Beside the probabilities for helix, sheet and other (sum is normalized to be one) the true secondary structure obtained from X-Ray structure as well as the predicted secondary structure are given. The overall prediction of the network is as good as 68% for this protein. The network misses the small β-sheet region 49-51 and makes the β-sheet 65-72 basically to an α-helix. The other secondary structure elements are found at a correct position. The α-helix is one period to short and two of the β-sheets are one amino acid to short. However, the black dashed line offers one minus the quotient of the second highest probability divided by the highest probability in %. This value would be 100%, if one of the three types would have been predicted with 100% probability and the other with zero and can reach 0% if two probabilities are the same and higher than the third. We suggest therefore to use this value as a confidence measure for secondary structure predictions. As obtainable from figure 7, the value is high if changes in secondary structure occur and especially for the wrong predicted α-helix at the end of the sequence, since the β-sheet probability is close to have the same size. If only the predictions with a confidence value smaller 50% are considered, 91% of the predicted secondary structure types would be correct.

Conclusion. Symmetric neural networks are successfully implemented to reduce dimension of amino acid parameter sets. The relevant information of these parameter sets is projected in three numbers, which can be used for further data analysis. The use of reduced parameter representations has the general potential to increases the speed of data analysis. In this special case the number of necessary parameters is decreased by more than 50% representing seven

properties by three parameters without loosing essential information. The computation time for the special problem of using these numbers as input values for a neural network decreases by a factor of six.

The ability of the reduced parameter representations to provide the complete information is proved by predicting secondary structure of proteins from primary structure. The reduced parameter representations with three parameters gave comparable good results to the complete parameter representations. The speed of computing the secondary structure is increased about linearly compared to the complete parameter representations by more than 100%. The time necessary for training the network is decreased even more essentially.

The approach can be adopted to predict reduced parameter representations for other amino acids and of course also for other structural building blocks, as for example nucleic acids. Moreover additional or other parameter sets can be used to create reduced parameter representations for the solution of special problems.

The data processing inside a neural network is illustrated by analyzing input sensitivities of the networks. A confidence value for secondary structure prediction is suggested, that allows the critical analysis of the network suggestion of secondary structure and might be useful to mention false positives in the predicted secondary structure elements.

The derived reduced parameter representations for amino acids, the used part of the protein data base (PDB), the program for predicting secondary structure: “Secondary” as well as the program for training and analyzing artificial neural networks “Smart” are available from www.jens-meiler.de. The software (Windows based) is free for academic use.

Acknowledgement. The authors thank B. Coligaev and W. Peti for downloading the subset of the PDB. The authors are especially thankful to Prof. C. Griesinger and Dr. R. Meusinger for useful discussion. J.M. is supported by a Kekulé stipend of the Fonds der Chemischen Industrie.

Figures.

1: A m-dimensional parameter representation for an amino acids is presented to an artificial neural network with three or five hidden layers. The data are processed thru a central hidden layer with $n \leq m$ neurons and recalculated by the output layer containing again m neurons. The network is trained with parameter representations for all twenty amino acids occur in natural systems. The n values obtained for the amino acids at the hidden neurons are used as n-dimensional parameter representation as illustrated by the three dimensional representation at the left.

2: Three layer network for predicting secondary structure of amino acids is visualized. 39 n-dimensional parameter or m-dimensional property representations of amino acids are presented to the network. The number of input neurons is therefore $39 \cdot n$ or $39 \cdot m$, respectively. The gray shaded central amino acid is surrounded by 15 amino acids represented by individual parameters as well as 15 additional amino acids represented by average parameters computed for three groups of five amino acids, respectively. All other amino acids surrounding this window of 61 amino acids are represented by the average value of their parameters ("edge"). The data pass a hidden layer containing 39 hidden neurons and the network is trained to predict probabilities for α-helix, β-sheet and unknown secondary structure for the central amino acid as well as the next five amino acids on both sides of the central amino acid.

3: Final RMSD values for neural network back calculation of a m-dimensional parameter set drawn on the y-axes. The number of used hidden neurons is given on the x-axes. Number of layers and used parameter set are varied according to:

a) three layer networks trained with five parameter set

b) three layer networks trained with seven parameter set

c) five layer networks trained with five parameter set

d) five layer networks trained with seven parameter set

the assigned letters correspond to the letters used in tables 2 – 4.

4: Obtained reduced parameters for the five layer neural network mapping the seven parameter set in one (a), two (b) or three (c) dimensions.

5: Input sensitivity for the 39 input blocks of the neural network predicting secondary structure probabilities from the seven parameter set of amino acids. The sensitivity is normalized to give 1 for the central amino acid. Summarized is over all output values (gray bars) and only over the helix probability (white bars).

6: Input sensitivity for the five or seven parameters out of the complete parameter set obtained from the neural network predicting secondary structure probabilities. The sensitivity is normalized to give 1 as maximum value.

7: Secondary structure of ubiquitin is given as obtained from the X-Ray structure (bars above 100) and as seen from the neural network using the complete seven parameter representation (bars below 0). Light gray stands for β-sheet and dark gray stands for α-helix. The individual probabilities normalized to give a sum of 100% are plotted. Light gray squares are again for β-sheet and dark gray squares for α-helix. The coil probability is given by a black line. The final network prediction is given by the highest out of these three values. The dashed black line is the one minus the quotient out of the second highest and the highest probability in percent. This value provides a confidence value, monitoring how save the judgment of the network is for each individual residue.

Literature.

(1) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists*; VCH Verlagsgesellschaft mbH: Weinheim, 1993.

(2) Doucet, J.-P.; Panaye, A.; Feuilleaubois, E.; Ladd, P. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 320-324.

(3) Lohninger, H. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 736-744.

(4) Cherqaoi, D.; Villemin, D. *J. Chem. Soc. Faraday Trans.* **1994**, *90*, 97-102.

(5) Panaye, A.; Doucet, J.-P.; Fan, B. T.; Feuilleaubois, E.; Azzouzi, S. R. E. *Chemom. Intell. Lab. Syst.* **1994**, *24*, 129-135.

(6) Sklenak, S.; Kvasnicka, V.; Pospichal, J. *Chem. Pap.* **1994**, *48*, 135-140.

(7) Clouser, D. L.; Jurs, P. C. *Carbohydr. Res.* **1995**, *271*, 65-77.

(8) Meiler, J.; Meusinger, R. In *Software - Entwicklung in der Chemie*; Gasteiger, J., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1995; Vol. 10, p 259-263.

(9) Meusinger, R.; Moros, R. In *Software - Entwicklung in der Chemie*; Gasteiger, J., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1995; Vol. 10, p 209-216.

(10) Panaye, A.; Doucet, J.-P.; Feuilleaubois, E.; Azzouzi, S. R. E. *AIP Conf. Proc.* **1995**, *330*, 734-739.

(11) Svozil, D.; Pospichal, J.; Kvasnicka, V. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 924-928.

(12) Thomas, S.; Kleinpeter, E. *J. Prakt. Chem./Chem.-Ztg.* **1995**, *337*, 504-507.

(13) Clouser, D. L.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 168-172.

(14) Meiler, J.; Meusinger, R.; Will, M. In *Software - Entwicklung in der Chemie*; Fels, G., Schubert, V., Eds.; Gesellschaft Deutscher Chemiker: Frankfurt am Main, 1996; Vol. 11, p 234-238.

(15) Amendolia, S. R.; Doppiu, A.; Ganadu, M. L.; Lubinu, G. *Anal. Chem.* **1998**, *70*, 1249-1254.

(16) Meiler, J.; Will, M.; Meusinger, R. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1169-1176.

(17) Moult, J. *Curr. Opin. Biotechnology* **1999**, *10*, 583-588.

(18) Polanski, J.; Gasteiger, J.; Wagener, M.; Sadowski, J. *Quant. Struct.-Act. Relat.* **1998**, *17*, 27-36.

(19) Kaartinen, J.; Mierisova, S.; Oja, J. M. E.; Usenius, J.-P.; Kauppinen, R. A.; Hiltunen, Y. *J. Magn. Res.* **1998**, *134*, 176-179.

(20) Choy, W. Y.; Sanctuary, B. C.; Zhu, G. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1086-1094.

(21) Rost, B.; Sander, C. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 7558-7562.

(22) Zemla, A.; Venclovas, C.; Fidelis, K.; Rost, B. *Proteins: Structure, Function, and Genetics* **1998**, *34*, 220-223.

(23) Rost, B.; Sander, C.; Schneider, R. *J. Mol. Biol.* **1994**, *235*, 13-26.

(24) Selbig, J.; Mevissen, T.; Lengauer, T. *Bioinformatics* **1999**, *15*, 1039-1046.

(25) Baldi, P.; Brunak, S.; Frasconi, P.; Soda, G.; Pollastri, G. *Bioinformatics* **1999**, *15*, 937-946.

(26) Guermeur, Y.; Geourjon, C.; Gallinari, P.; Deleage, G. *Bioinformatics* **1999**, *15*, 413-421.

(27) Avbelj, F.; Fele, L. *J. Mol. Biol.* **1998**, *279*, 665-684.

(28) Bohr, H.; Bohr, J.; Brunak, S.; Cotterill, R. M. J.; Lautrup, B.; Noerskov, L.; Olsen, O. H.; Petersen, S. B. *FEBS Lett.* **1988**, *241*, 223-8.

(29) Chandonia, J.-M.; Karplus, M. *Protein Sci.* **1995**, *4*, 275-85.

(30) Chandonia, J.-M.; Karplus, M. *Protein Sci.* **1996**, *5*, 768-74.

(31) Chandonia, J.-M.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 293-306.

(32) Hanke, J.; Reich, J. G. *Comput. Appl. Biosci.* **1996**, *12*, 447-454.

(33) Hayward, S.; Collins, J. F. *Proteins: Struct., Funct., Genet.* **1992**, *14*, 372-81.

(34) Jones, D. T. *J. Mol. Biol.* **1999**, *292*, 195-202.

(35) King, R. D.; Sternberg, M. J. E. *Protein Sci.* **1996**, *5*, 2298-2310.

(36) Kneller, D. G.; Cohen, F. E.; Langridge, R. *J. Mol. Biol.* **1990**, *214*, 171-82.

(37) Muskal, S. M.; Kim, S. H. *J. Mol. Biol.* **1992**, *225*, 713-27.

(38) Ouali, M.; King, R. D. *Protein Sci.* **2000**, *9*, 1162-1176.

(39) Petersen, T. N.; Lundegaard, C.; Nielsen, M.; Bohr, H.; Bohr, J.; Brunak, S.; Gippert, G. P.; Lund, O. *Proteins: Struct., Funct., Genet.* **2000**, *41*, 17-20.

(40) Qian, N.; Sejnowski, T. J. *J. Mol. Biol.* **1988**, *202*, 865-84.

(41) Rice, D. W.; Eisenberg, D. *J. Mol. Biol.* **1997**, *267*, 1026-1038.

(42) Rost, B.; Sander, C. *J. Mol. Biol.* **1993**, *232*, 584-99.

(43) Rost, B.; Sander, C. *Proteins: Struct., Funct., Genet.* **1994**, *19*, 55-72.

(44) Rost, B.; Sander, C. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 295-300.

(45) Rost, B. *Methods Enzymol.* **1996**, *266*, 525-539.

(46) Salamov, A. A.; Solovyev, V. V. *J. Mol. Biol.* **1995**, *247*, 11-15.

(47) Salamov, A. A.; Solovyev, V. V. *J. Mol. Biol.* **1997**, *268*, 31-36.

(48) Sasagawa, F.; Tajima, K. *Comput. Appl. Biosci.* **1993**, *9*, 147-52.

(49) Stolorz, P.; Lapedes, A.; Xia, Y. *J. Mol. Biol.* **1992**, *225*, 363-77.

(50) Vivarelli, F.; Giusti, G.; Villani, M.; Campanini, R.; Fariselli, P.; Compiani, M.; Casadio, R. *Comput. Appl. Biosci.* **1995**, *11*, 253-60.

(51) Livingstone, D. J.; Hesketh, G.; Clayworth, D. *J. Mol. Graphics* **1991**, *9*, 115-18.

(52) Kocjancic, R.; Zupan, J. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 985-989.

(53) Devillers, J. E. *Neural networks in QSAR and drug design*; Acad. Press: London, 1996.

(54) Fauchere, J.-L.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. *Int. J. Peptide Protein* **1988**, *32*, 269-278.

(55) Holm, L.; Sander, C. *Science* **1996**, *273*, 595-602.

(56) Meiler, J. *www.jens-meiler.de* **2001**.

table 1: amino acid parameter sets

name	Ξ[a]	α[b]	υ_V[c]	π[d]	I[e]	α[f]	β[g]
ALA	1.28	0.05	1.00	0.31	6.11	0.42	0.23
GLY	0.00	0.00	0.00	0.00	6.07	0.13	0.15
VAL	3.67	0.14	3.00	1.22	6.02	0.27	0.49
LEU	2.59	0.19	4.00	1.70	6.04	0.39	0.31
ILE	4.19	0.19	4.00	1.80	6.04	0.30	0.45
PHE	2.94	0.29	5.89	1.79	5.67	0.30	0.38
TYR	2.94	0.30	6.47	0.96	5.66	0.25	0.41
TRP	3.21	0.41	8.08	2.25	5.94	0.32	0.42
THR	3.03	0.11	2.60	0.26	5.60	0.21	0.36
SER	1.31	0.06	1.60	-0.04	5.70	0.20	0.28
ARG	2.34	0.29	6.13	-1.01	10.74	0.36	0.25
LYS	1.89	0.22	4.77	-0.99	9.99	0.32	0.27
HIS	2.99	0.23	4.66	0.13	7.69	0.27	0.30
ASP	1.60	0.11	2.78	-0.77	2.95	0.25	0.20
GLU	1.56	0.15	3.78	-0.64	3.09	0.42	0.21
ASN	1.60	0.13	2.95	-0.60	6.52	0.21	0.22
GLN	1.56	0.18	3.95	-0.22	5.65	0.36	0.25
MET	2.35	0.22	4.43	1.23	5.71	0.38	0.32
PRO	2.67	0.00	2.72	0.72	6.80	0.13	0.34
CYS	1.77	0.13	2.43	1.54	6.35	0.17	0.41

[a] - sterical parameter (graph shape index)
[b] - polarizability
[c] - volume (normalized van der Waals volume)
[d] - hydrophobicity
[e] - isoelectric point
[f] - helix probability
[g] - sheet probability

Table 1

table 2: reduced parameter representation obtained from the three layer network

name	a)	3 layer networks trained with				b)	3 layer networks trained with									
		5 parameter set					7 parameter set									
	1D	2D		3D			1D	2D		3D			4D			
ALA	0.13	0.15	0.04	0.20	0.84	0.27	0.28	1.00	0.52	0.84	0.50	0.00	1.00	1.00	0.98	0.56
GLY	0.00	0.00	0.03	0.03	1.00	0.32	0.00	0.99	1.00	0.04	0.05	0.20	0.29	0.76	0.18	0.07
VAL	0.55	0.57	0.04	0.69	0.59	0.18	0.76	0.19	0.68	0.68	0.20	0.87	0.38	0.73	0.31	0.93
LEU	0.61	0.62	0.03	0.71	0.53	0.20	0.69	0.45	0.44	1.00	0.43	0.49	0.79	0.82	0.73	0.87
ILE	0.75	0.73	0.03	0.91	0.50	0.16	0.84	0.13	0.60	0.88	0.25	0.88	0.47	0.73	0.38	1.00
PHE	0.81	0.80	0.11	0.83	0.25	0.32	0.83	0.17	0.52	0.88	0.39	0.83	0.43	0.33	0.32	0.78
TYR	0.76	0.75	0.28	0.66	0.09	0.45	0.81	0.18	0.52	0.56	0.44	0.92	0.18	0.03	0.12	0.54
TRP	1.00	1.00	0.20	1.00	0.03	0.32	1.00	0.00	0.36	0.96	0.51	1.00	0.35	0.09	0.31	0.86
THR	0.36	0.37	0.11	0.40	0.56	0.36	0.50	0.47	0.76	0.40	0.22	0.68	0.26	0.42	0.14	0.48
SER	0.15	0.15	0.09	0.14	0.72	0.41	0.26	0.77	0.84	0.24	0.21	0.39	0.32	0.55	0.22	0.26
ARG	0.46	0.40	1.00	0.06	0.00	0.00	0.51	0.92	0.00	0.00	1.00	0.52	0.05	0.21	1.00	0.13
LYS	0.33	0.27	0.82	0.00	0.16	0.05	0.42	0.93	0.20	0.00	0.84	0.47	0.08	0.36	0.87	0.14
HIS	0.52	0.50	0.43	0.37	0.25	0.18	0.57	0.56	0.44	0.32	0.56	0.68	0.18	0.30	0.47	0.38
ASP	0.16	0.20	0.19	0.09	0.44	0.98	0.23	0.87	0.76	0.48	0.40	0.18	0.53	0.00	0.04	0.00
GLU	0.24	0.28	0.24	0.14	0.31	1.00	0.36	0.95	0.40	0.88	0.74	0.00	0.94	0.09	0.46	0.14
ASN	0.21	0.20	0.34	0.09	0.44	0.41	0.25	0.86	0.68	0.12	0.43	0.40	0.19	0.24	0.27	0.06
GLN	0.31	0.32	0.30	0.20	0.31	0.55	0.43	0.84	0.40	0.64	0.67	0.24	0.59	0.30	0.59	0.26
MET	0.58	0.58	0.09	0.60	0.41	0.34	0.67	0.48	0.40	0.96	0.52	0.48	0.74	0.58	0.66	0.73
PRO	0.34	0.33	0.01	0.43	0.72	0.14	0.45	0.44	0.92	0.08	0.00	0.82	0.00	0.58	0.00	0.46
CYS	0.42	0.42	0.00	0.57	0.75	0.16	0.57	0.29	0.88	0.44	0.04	0.81	0.27	0.82	0.20	0.75

Table 2

table 3: reduced parameter representation obtained from the five layer network

name	c)	5 layer networks trained with				d)	5 layer networks trained with									
		5 parameter set					7 parameter set									
	1D	2D		3D			1D	2D		3D			4D			
ALA	0.01	0.13	0.06	0.23	0.19	0.19	0.11	0.55	0.78	1.00	0.19	0.25	0.56	1.00	0.86	0.19
GLY	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.92	1.00	0.90	0.00	0.00	0.05	0.70	0.59	0.17
VAL	0.93	0.26	0.11	0.89	0.29	0.84	0.72	0.26	0.88	0.71	0.18	0.79	0.48	0.77	0.30	0.99
LEU	0.94	0.78	0.23	0.98	0.39	0.48	0.91	0.21	0.57	0.90	0.76	0.79	0.70	0.78	0.89	0.49
ILE	0.94	0.44	0.15	1.00	0.29	1.00	0.74	0.20	0.88	0.71	0.45	0.89	0.94	0.83	0.35	1.00
PHE	0.96	0.92	0.23	0.93	0.59	0.36	0.98	0.14	0.45	0.83	0.77	0.88	0.69	0.50	0.81	0.58
TYR	0.96	0.94	0.30	0.50	0.75	0.16	0.99	0.16	0.27	0.66	0.67	0.89	0.58	0.23	0.70	0.52
TRP	1.00	0.98	0.15	0.95	1.00	0.11	1.00	0.00	0.00	0.95	1.00	1.00	1.00	0.37	0.89	0.66
THR	0.74	0.18	0.13	0.41	0.32	0.52	0.59	0.40	0.92	0.37	0.18	0.73	0.23	0.55	0.38	0.67
SER	0.03	0.12	0.09	0.16	0.30	0.10	0.04	0.56	0.94	0.76	0.11	0.37	0.19	0.55	0.57	0.41
ARG	0.50	1.00	1.00	0.05	0.96	0.00	0.45	1.00	0.02	0.00	0.76	0.66	0.34	0.00	1.00	0.01
LYS	0.47	0.97	1.00	0.02	0.75	0.00	0.51	0.97	0.08	0.15	0.42	0.86	0.27	0.11	0.86	0.08
HIS	0.69	0.90	0.40	0.30	0.65	0.05	0.55	0.31	0.31	0.20	0.34	0.92	0.39	0.25	0.76	0.39
ASP	0.09	0.07	0.40	0.23	0.27	1.00	0.07	0.75	0.88	0.54	0.29	0.22	0.51	0.41	0.11	0.00
GLU	0.12	0.13	0.45	0.27	0.40	0.76	0.21	0.61	0.49	0.68	0.57	0.37	0.90	0.87	0.32	0.01
ASN	0.37	0.59	0.55	0.11	0.49	0.04	0.05	0.61	0.92	0.17	0.27	0.68	0.19	0.18	0.68	0.18
GLN	0.34	0.67	0.45	0.20	0.55	0.09	0.28	0.42	0.45	0.71	0.69	0.65	0.46	0.49	0.84	0.21
MET	0.94	0.84	0.30	0.73	0.52	0.21	0.94	0.23	0.53	0.88	0.78	0.79	0.66	0.67	0.89	0.42
PRO	0.77	0.18	0.09	0.70	0.17	0.71	0.00	0.36	0.96	0.24	0.05	0.82	0.00	0.82	0.00	0.99
CYS	0.92	0.21	0.02	0.93	0.30	0.44	0.61	0.24	0.98	0.88	0.03	0.65	0.09	0.54	0.65	0.74

Table 3

table 4: results for secondary structure prediction of proteins from test data set using the reduced parameter representation in comparison with the complete parameter set

Dimension of parameter representations		**Prediction of artificial neural networks in %**									
		α-helix predicted as			β-sheet predicted as			coil predicted as			Q_3 (Σ–
		α-helix	β-sheet	coil	α-helix	β-sheet	coil	α-helix	β-sheet	coil	correct)
a)	***1***	9.4	3.6	16.8	3.6	8.8	14.7	3.1	4.4	35.5	53.8
	2	10.8	3.4	15.7	2.4	10.9	13.7	3.1	3.7	36.3	58.0
	3	14.3	3.0	12.6	2.2	12.0	12.9	2.8	3.9	36.4	62.7
b)	***1***	9.7	4.0	16.2	3.9	10.0	13.2	3.4	4.8	34.8	54.5
	2	16.1	2.9	10.9	1.9	10.2	15.0	4.6	5.6	32.9	59.1
	3	19.6	2.2	8.1	1.3	11.7	14.0	4.0	4.6	34.4	65.7
	4	19.5	2.8	7.6	2.3	12.2	12.5	4.4	4.7	34.0	65.7
c)	***1***	7.7	3.6	18.6	1.9	10.6	14.5	3.5	4.4	35.2	53.5
	2	13.8	1.9	14.1	1.9	7.3	17.9	3.7	3.1	36.2	57.3
	3	16.4	2.3	11.1	1.8	12.2	13.1	4.5	3.7	34.8	63.4
d)	***1***	11.8	2.7	15.4	3.3	8.7	15.1	3.7	4.0	35.3	55.8
	2	15.4	2.2	12.3	2.8	9.8	14.4	5.2	4.0	33.8	59.0
	3	20.3	1.7	7.9	3.5	9.8	13.8	4.5	3.2	35.3	65.3
	4	20.0	1.8	8.0	1.8	12.0	13.2	4.4	4.9	33.8	65.8
complete 5 parameter set		16.1	2.0	11.8	2.6	11.5	13.0	3.6	4.2	35.3	62.8
complete 7 parameter set		20.1	2.2	7.6	3.3	12.4	11.3	4.1	4.4	34.6	67.1

Table 4

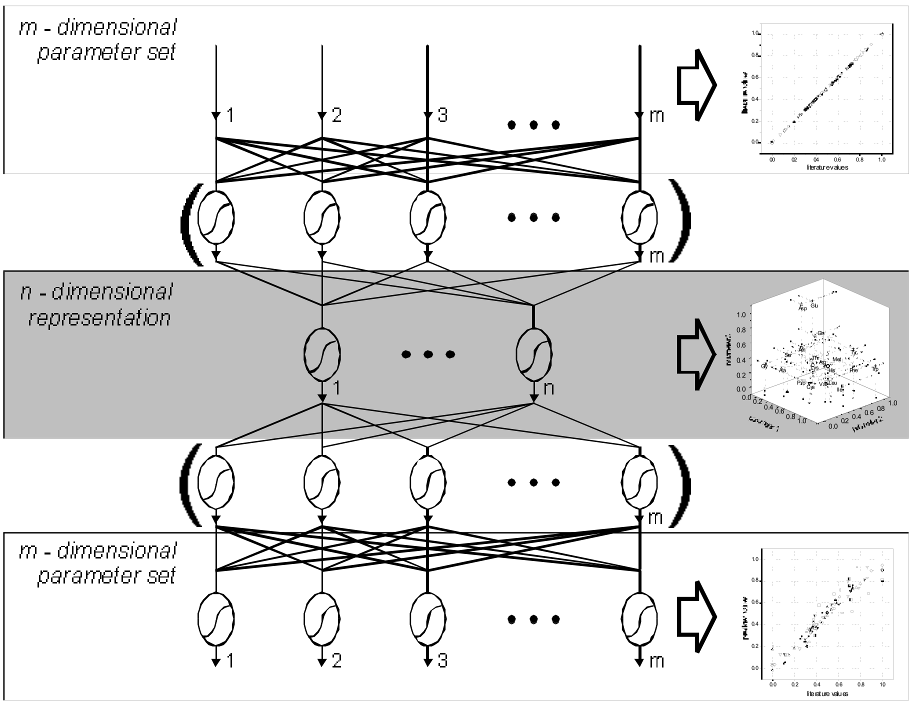

m - dimensional parameter set
1
2
3
m
m
n - dimensional representation
1
n
m
m - dimensional parameter set
1
2
3
m
literature values
literature values

Fig.1

Fig.2

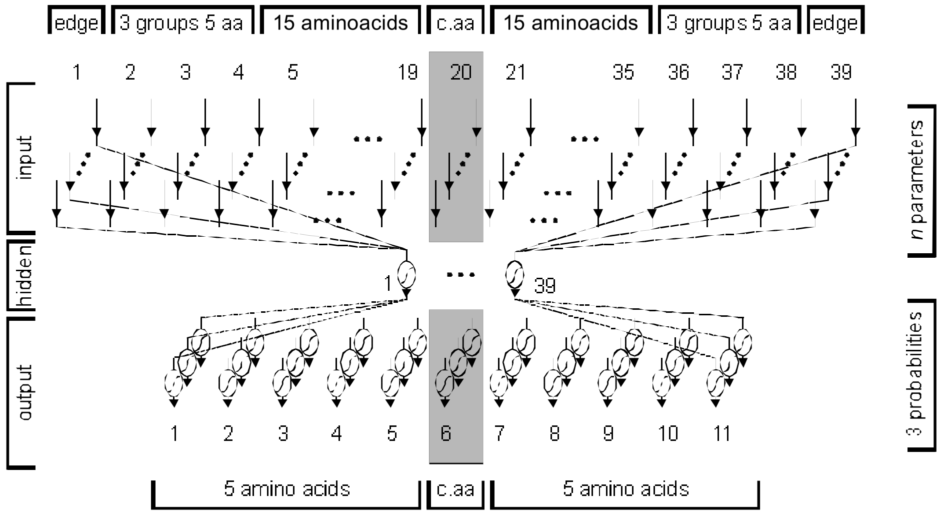

edge
3 groups 5 aa
15 aminoacids
c.aa
15 aminoacids
3 groups 5 aa
edge
1
2
3
4
5
19
20
21
35
36
37
38
39
input
hidden
output
1
39
1
2
3
4
5
6
7
8
9
10
11
5 amino acids
c.aa
5 amino acids
n parameters
3 probabilities

Fig.3

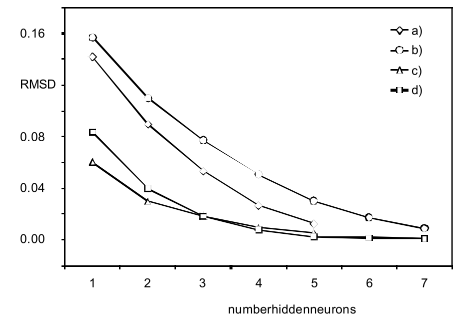

0.16
RMSD
0.08
0.04
0.00
1
2
3
4
5
6
7
numberhiddenneurons
a)
b)
c)
d)

Fig.4

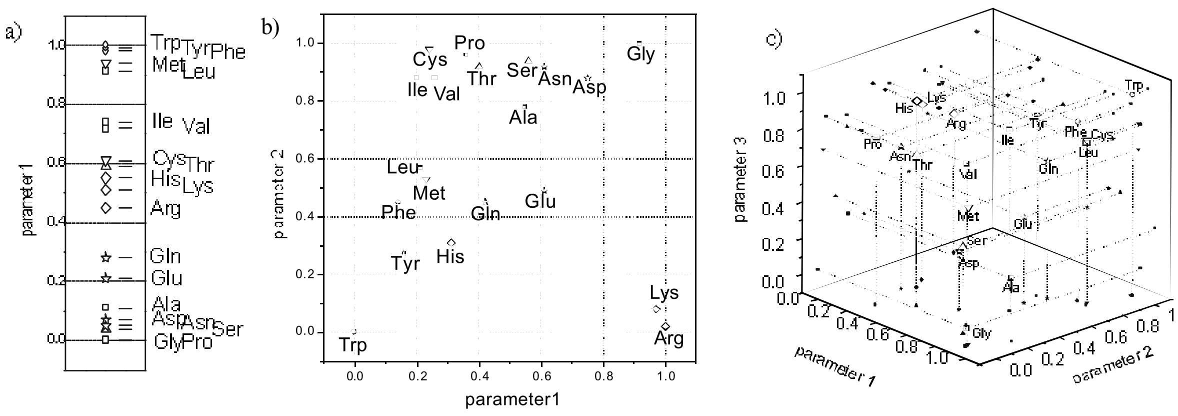

a)
parameter1
1.0
0.8
0.6
0.4
0.2
0.0
Trp Tyr Phe
Met Leu
Ile Val
Cys Thr
His Lys
Arg
Gln
Glu
Ala
Asp Asn Ser
Gly Pro
b)
parameter 2
parameter1
0.0
0.2
0.4
0.6
0.8
1.0
Pro
Cys
Ile Val
Thr
Ser
Asn
Asp
Gly
Ala
Leu
Met
Phe
Gln
Glu
Tyr
His
Lys
Arg
Trp
c)
parameter 3
parameter 1
parameter 2
Trp
His
Lys
Arg
Tyr
Pro
Ile
Phe
Cys
Asn
Thr
Leu
Val
Gln
Met
Glu
Ser
Asp
Ala
Gly

Fig. 5

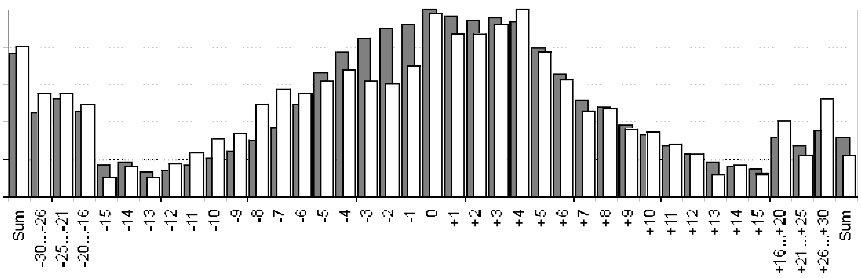

1.0
0.8
0.6
0.4
0.2
0.0
Sum
-30...-26
-25...-21
-20...-16
-15
-14
-13
-12
-11
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
+1
+2
+3
+4
+5
+6
+7
+8
+9
+10
+11
+12
+13
+14
+15
+16...+20
+21...+25
+26...+30
Sum

Fig.6

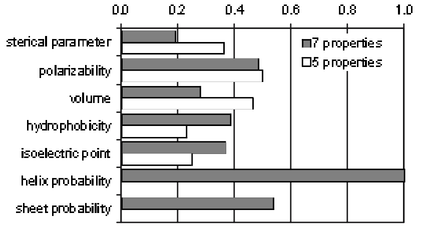
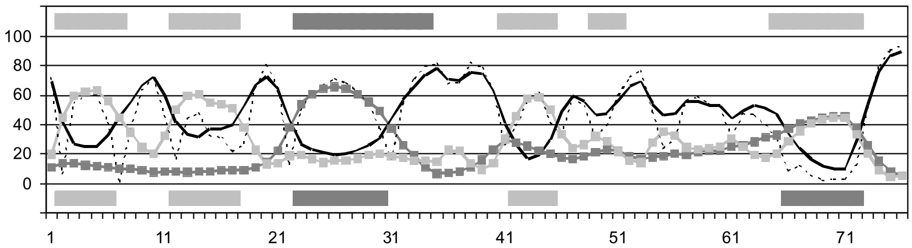

0.0
0.2
0.4
0.6
0.8
1.0
sterical parameter
polarizability
volume
hydrophobicity
isoelectric point
helix probability
sheet probability
▨7 properties
□5 properties

Fig. 7


100
80
60
40
20
0
1
11
21
31
41
51
61
71

Epothilones: Quantitative Structure Activity Relations Studied by Artificial Neural Networks Leading to New Drug Proposals

Jens Meiler* and Annalen Bleckmann

Universität Frankfurt, Marie-Curie-Str. 11, 60439 Frankfurt am Main, Germany

Key words. neural networks; quantitative structure activity relationship; Epothilones; anti-tumor agents; combinatorial chemistry

*To whom correspondence should be addressed

Contact address:

Universität Frankfurt (AK Griesinger)
Marie-Curie-Str. 11
60439 Frankfurt am Main
Germany

Tel.: 069 798 29 798
Fax.: 069 798 29 128
Mail: mj@org.chemie.uni-frankfurt.de

Motivation: Usually a set of molecules with known biological activities is obtained as a result of drug optimization by systematic or combinatorial synthesis of derivatives. These data are shown to be a capable basis to establish structure activity relations. Subsequently the gained quantitative models are successfully combined with structure generators to screen huge structural spaces for biological active molecules.

Results: Artificial neural networks are established to investigate relationships between the structure of Epothilone derivatives and anti-tumor activities. Structural features obtained from the set of molecules are translated into numerical parameters and used for training the network connections. The influence of single structural parameters, interactions as well as the binding site become obtainable. The screening of the structural space spanned by the parameters leads to suggestions for Epothilone derivatives with a possible higher biological activity than all structures known so far.

Availability: The program “Smart” which is used for simulating the artificial neural networks is free for academic use and can be obtained from http://krypton.org.chemie.uni-frankfurt.de/~mj.

Contact: mj@org.chemie.uni-frankfurt.de

Introduction. Epothilones A and B can be isolated from the myxobacterium *Sorangium cellulosum* strain 90 by Höfle et. al. (Höfle, Bedorf et al. 1994; Gerth, Bedorf et al. 1997). The recognition of their cytotoxic action against tumor cells lead to intense research activities in chemistry and biology. Bollag et al.(Bollag, McQueney et al. 1995) discovered the induction of the Tubulin(Nogales, Wolf et al. 1995; Nogales, Wolf et al. 1998) polymerization of these compounds similar to agents like Taxol(Schiff, Fant et al. 1979). The effect of microtubule stabilization even in Taxol-resistant tumor cell lines(Kowalski, Giannakakous et al. 1997) increased their potential in cancer chemotherapy further (Bollag, McQueney et al. 1995; Giannakakou, Gussio et al. 2000; Martello, McDaid et al. 2000).

The complete elucidation of the structure including stereochemistry is published by Höfle(Höfle, Bedorf et al. 1996). Soon after the synthesis of precursors, the natural products themselves and analogues is described in a large number of reports(Schinzer, Limberg et al. 1996; Muhlradt and Sasse 1997; Nicolaou, Sarabia et al. 1997; Nicolaou, Vourloumis et al. 1997; Nicolaou, Vourloumis et al. 1997; Schinzer, Limberg et al. ; Nicolaou, He et al. 1998; Nicolaou, Roschangar et al. 1998; Höfle, Glaser et al. 1999; Höfle, Glaser et al. 1999; Taylor and Zajicek 1999; Altmann, Bold et al. 2000; Johnson, Kim et al. 2000; Lee, Chou et al. 2000; Mulzer 2000; Nicolaou, Scarpelli et al. 2000; Von Angerer 2000). Many of these analogues are biologically investigated and lead to a large source of molecular structures with known biological activities. These data served as basis of qualitative structure activity relations(Winkler and Axelsen 1996; Su, Balog et al. 1997; Nicolaou, Roschangar et al. 1998; He, Jagtap et al. 2000). Moreover, Wang et. al.(Wang, Xia et al. 1999) introduced a unified and quantitative receptor model for microtubule binding of Paclitaxel and Epothilone by

However, to the best of our knowledge no model free quantitative structure activity relation (QSAR) is established so far. This is the first purpose of the work presented here.

Artificial neural networks are known and used for several years in chemistry and biochemistry to describe structure activity relations quantitatively(Zupan and Gasteiger 1993). For this purpose usually no mathematical model is known, many influences interact and the activities are afflicted with comparable large experimental deviations while dealing with biochemical data. Here the special advantages of neural networks play an important role:

- No initial mathematical model is necessary to establish the relation. In contrast to most other methods artificial neural networks are flexible enough to fit a wide range of complex functions on a set of training data.
- The structure of an artificial neuron is copied from a natural brain cell. Although much simpler and smaller than any natural neural network, they can be trained to become an “expert” for a particular problem. Their similarity to the natural system predestinates them to work with blurred information.
- In contrast to simpler methods like multiple linear regression (MLR) the network structure allows to take interactions between several input parameters into consideration. Also dependencies between several output values can be used for establishing the relation.

A lot of applications of these networks already exist in chemistry(Lohninger 1993; Bienfait 1994; Cherqaoi and Villemin 1994; Thomas and Kleinpeter 1995; Devillers 1996; Isu, Nagashima et al. 1996; Anklam, Bassani et al. 1997; Emerenciano, Melo et al. 1997; Rodrigues, Campos et al. 1997; Svozil, Sevcik et al. 1997; Kaartinen, Mierisova et al. 1998; Wang, Dopazo et al. 1998). Neural networks are also applied to biochemical problems like secondary structure prediction of proteins(Rost and Sander 1993; Rost, Sander et al. 1994; Choy, Sanctuary et al. 1997; Zemla, Venclovas et al. 1998; Baldi, Brunak et al. 1999; Guermeur, Geourjon et al. 1999; Selbig, Mevissen et al. 1999), creating reduced amino acid parameter representations[Meiler, submitted #1439] and investigating structure activity relations for pentapeptides at serine proteases(Meiler 1998).

The focus of this paper is the analysis of biological activities for Epothilone analogues with respect to their structure using artificial neural networks. Beyond the investigation of quantitative structure activity relations the method is combined with a structure generator to create and screen a large number of Epothilone analogues with respect to their biological activity, without necessarily synthesizing them before. An extensive acceleration in finding and optimizing drugs can be achieved by this approach. With traditional synthesis and even with modern methods like combinatorial synthesis only a part of the structural space can be screened. This is because already for relatively small molecules like Epothilone, the number of possible derivatives easily exceeds milliards. However, if for a part of the molecules in the structural space the biological activity is known from experiments a neural network can be trained to predict this activity for every member of the structural space. This prediction is very fast and needs especially no further synthesis. The generated derivatives can be ranked with respect to the estimated activities. Depending on the size of the so generated database either candidates or at least more precise ideas for further synthesis can be derived from this hit list. Even an iterative use of this approach alternately with synthesis is thinkable. Therefore and in general this approach does of course not substitute the organic synthesis and biological screening process but it has the potential to accelerate both steps essentially.

Experimental. All biological activities used in the following analysis are taken from Nicolaou et. al.(Nicolaou, Roschangar et al. 1998). The ability to induce the Tubulin assembly is known for over 200 Epothilone derivatives. The values cited in literature are obtained by incubating purified Tubulin for 30 min at 37°C in the presence of the compound. The mixture is filtered, the collected polymerized Tubulin is stained with amido black solution and

quantified by measuring absorbance of the dyed solution. The given %-polymerization is calculated relative to the presumed absorbance of 100% polymerized Tubulin.

The inhibition of carcinoma cell growth for three cell lines is additionally available for 37 out of these derivatives(Nicolaou, Roschangar et al. 1998). The first IC_{50} value is determined for the parental ovarian cell line (1A9), while the second and the third value is obtained from mutated Taxol-resistant cell lines (1A9PTX10: Phe270 Val and 1A9PTX22: Ala364 Thr).

The cell growth is evaluated by measuring the increase in cellular protein. The resistant cell lines are gained by treating the cells with increasing concentrations of Taxol. However, their Taxol resistance does not cause a resistance with respect to the Epothilone analogues, which further increases the potential of Epothilones in cancer chemotherapy.

In order to analyze these data with artificial neural networks, it is necessary to translate both, the structural as well as the biological information in suitable numerical data. Therefore 24 structural descriptors are derived that code 198 of the cited Epothilone derivatives. Not all cited structures could be described with these 24 parameters only. However, for the subsequent neural network analysis it is necessary to have more than only one representative for every possible state (numerical value) of a structural parameter. This is due to the necessity to stabilize the connections outgoing from the corresponding input (usually more than one).

The 24 descriptors are given in table 1 together with their possible states. For coding the 37 structures with published inhibition values for carcinoma cell growth only 8 out of these 24 parameters are necessary.

Biological activities are often a kind of kinetic constants for very complex chemical processes. Therefore they naturally cover a wide range. To decrease this range the natural logarithm of the values is trained and predicted by all neural networks. Moreover the range covered by the activities is linear scaled to lay between 0.25 and 0.75, since the neural network can predict only values between 0 and 1 due to its sigmoid transfer function. The range from 0 to 0.25 and from 0.75 to 1 is excluded to enable the network to predict activities smaller or larger relative to the range covered by the 198 or 37 derivatives, respectively. The numerical parameters are not scaled prior to their submission to the network, since the weights in the first layer can perform this operation. All artificial neural networks are created with the program "Smart"(Meiler 2000). The three layer networks contain one bias in every layer and are trained with the back propagation algorithm. The learning rate η is decreased during the training procedure from 0.01 to 0.0001, the momentum α is constant with 0.5. In all cases the data are separated in a test and a training set of data. The training of the artificial neural network is performed using only the training data. The root mean square deviation (RMSD) of the test set of data is monitored parallel and the training is interrupted if it is minimized. It is assumed that the network understands general relations up to this point in the training procedure. Beyond this point special information is learned that is relevant for the training set of data only. The more diverse examples are used in the test set of data the more exact this break point can be defined. However, every example used in the test set cannot be used for training the network and is therefore lost for establishing the connections. This is one of the major problems while using neural networks together with biochemical data. Usually the amount of data available is relatively small. The main part is necessary to stabilize the weights in the network and has to be a part of the training set of data. A larger training set of data allows to generate larger and therefore more specialized networks, because more weights can be stabilized and hold more information. However, this positive effect can be neutralized if the test set of data becomes too small. The training procedure might be stopped too early because of a small statistically badly chosen test set. We will see this at one example later.

"Input sensitivities" provide insight into the finally achieved structure of the neural network. In order to compute these values only one input is varied within the experimental input range while all other inputs are set to be zero. The signals obtained at the output neurons provide the searched information. The covered range is the sensitivity of this particular output with

respect to the investigated input. The value lays naturally between 0 and 1. If the influence of a particular input on an output value is high, the covered output range is large and the sensitivity becomes 1. However, if the input has no influence at all, the output value will remain constant and the sensitivity stays 0. One should mention that this method does not visualize the cross correlation between different input values.

For the set of Epothilone analogues three different experiments are performed combining different structural parameters with different biological activities:

- The 198 Epothilones analogues where separated into five groups. All substances modified in one of the four regions **A**, **B**, **C**, **D** (figure 2, table 1) are combined in one set of data, respectively. The fifth data set contains all substances modified in more than only one of the four regions. Four different artificial neural networks are trained to predict the induction of Tubulin assembly with one of the first four sets of data, respectively. Only the parameters necessary for covering the particular region out of all 24 parameters are used as input vector. Three neurons in the hidden layer process the data and one output neuron is trained to predict the polymerization constant. Between 15% and 20% of the available substances out of each of the four sets of data are randomly selected as test set of data.

  (It has to be mentioned, that the separation into these five groups is partially ambiguous. Some of the input parameters have to be used, although they are not part of the particular region. For example substances modified in region ***C*** *can have the Epothilone A or the Epothilone E structure in region* ***B****. They are still counted to be modified in region* ***C****, however the parameters necessary for coding the structure at C-12 and C-13 are necessary input data for this network too. Moreover P13 and P21 are not used for any of the four neural networks as input. The number of substances modified at these positions is in all four cases too small to stabilize the connections of*

the network. However, they will be used in the second experiment using a combined data set of all Epothilone analogues.)

- In a second experiment one neural networks that contains 24 inputs and predicts again the induction of Tubulin assembly is established. This network considers the whole structural space described by the 24 parameters. For training and testing all 198 derivatives can be used, in particular also the members of the fifth group. 20 example structures are randomly selected out of all 198 Epothilones to form the test set of data.
- For the last experiment only the 37 structures with known values for carcinoma cell growth inhibition are taken into consideration. Only 8 out of the 24 parameters are necessary to code them. Three neural networks are created predicting one of the three constants, respectively. Since these three biological activities are related, in a last experiment the question is addressed, whether a neural network can use these relations to enhance the model. The fourth net has therefore three output neurons and predicts all biological activities parallel.

If the neural network describes a quantitative relation between the structure and the biological activity, it can now be used to predict the activities not only for the test data as done during the training procedure but also for “unknown” structures. The 24 introduced parameters span a structural space that is defined by all possible permutations of the parameter states. Every molecule in this space corresponds to one permutation. This leads to the idea to “generate” all possible structures that way and test them with the trained neural networks to predict their biological activity. The 198 used derivatives are part of the so generated set of structures together with a large number of new derivatives not synthesized and tested so far. Within the error of the method the neural network will calculate a biological activity for all structures. A hit list of substances is gained by ranking all structures with respect to the predicted activity. This leads to suggestions for highly active substances, that are not necessarily synthesized so

far. If so, they can be serve as target in the further drug optimization process. This procedure allows a fast and effective way to screen the enormous number of members in a structural space in a very short time. Even the test of more than milliards of substances becomes possible. A number, that cannot be synthesized even by combinatorial synthesis in a reasonable time. Moreover the subsequent synthesis of only the most promising derivatives should be straight forward. All necessary reaction are established and well known since used already for the synthesis of the existing structures. Only the combination of educts and subsequently performed reactions change. This approach is carried out for two out of all trained networks:

- First the network out of experiment 2 that uses all 24 parameters to predict the induction of Tubulin assembly is taken. With this network the structural space spanned by all 24 parameters can be analyzed. 2.6 milliards (!) structures are possible in this space. They are generated by permutation, checked with the neural network and ranked with respect to the predicted activity.
- 624 possible structures are members in the structural space spanned by the eight parameters necessary for coding the 37 Epothilone analogues. They are also generated and screened with the neural net, that predicts all three inhibition constants for the carcinoma cell growth parallel. The substances are ranked according to the average of all three IC_{50} constants. This procedure ensures a higher quality of the prediction for two reasons: The model itself is better as discussed below and moreover permutations with a high activity for all three cell lines are ranked best. This minimizes the possibility of predicting false positives.

Results and Discussion. The results for all trained neural networks are given in Table 2 together with the net sizes and the number of trained and tested data. Figure 3 shows the correlation diagrams and the input sensitivities for the five neural networks trained in the first two experiments. Comparing the individual networks generated for the four regions of modification in Epothilone one notices the differences in the prediction quality. While especially network 1a) (region **A)** and also network 1c) (region **C**) are able to establish a relation between the structural parameters and the activity, network 1b) (region **B**) is essentially poorer. Network 1d) (region **D**) is unable to suggest any correlation. It predicts the same induction of Tubulin polymerization for all presented substances. This quality of prediction is obviously correlated with the input sensitivity, which is essentially larger in regions **A** and **C** than in region **B** or even **D**.

For region **A** one derives a high sensitivity for changes in the substituents on C-8 (P02, P03) as well as for the number of CH_2 groups (P04). In the region **B** the stereochemistry of C-12, C-13 and C-15 (P09, P11, P12) have a high impact as well as the substituent at C-12 (P05 - P07) and the bond type between C-12 and C-13 (P08). However, all sensitivities are smaller than 0.2. For the region **C** especially the Nitrogen in position 22 seems to be essential for biological activity and moreover the substituent at C-12 has impact. As mentioned network 1d) does not find any relation. Consequently all input sensitivities are zero.

As mentioned earlier the behavior of network 1d) is often observable for a small set of data with complex relations. Of course it would be possible to train the neural network even in region **D** to predict the training set of data. However the test set will not show any meaningful correlation. In other words: the network is not able to establish a general relation between the presented structures and the corresponding activities. Three reasons can cause this behavior: The relation is not present in the data, the training set of data is too small to establish the relation, or the test set of data is too small (or badly chosen) and the break point to stop the training process is therefore not good enough defined. In this case the small set of data is responsible for the bad result obtained. The experiment is not reproduced with another test set

of data, because the general problem of too small set of data cannot be solved that way. Rather the data basis is enlarged by combining the small set of data:

Overcoming the problem of too little information the substances of the experiments 1a) to d) are combined in one network using now all 24 parameters as input data. The overall number of substances used in the analysis increased by 30. These additional structures are either modified in more than one of the regions **A**, **B**, **C** or **D** (24 substances) or they are modified at P13 or P21 (6 substances). Their number is too small to use them in the individual networks for regions **A**, **B**, **C** and **D** but together with the 24 substances out of the fifth set of data they could be used in the second experiment.

The network uses 178 substances for training and 20 randomly selected substances for testing the network. The RMSD values for the training and also for the test set of data are 0.76, which correspond to a factor of about 2 on an exponential scale. The neural network shows an increase in the input sensitivity values compared with the networks obtained in experiment 1. Especially the atom type in position 22 (P15) and the number of CH_2-groups reflect a very high impact. Moreover the structure at C-12 and C-13 (P08), the stereochemistry at C-15 (P12) and the two substituents at C-12 (P05..P07) and C-21 (P19..P21) are important. However, nearly all input sensitivities differ from zero. The neural network uses therefore all presented parameters to establish the relation which indicates a complex model with extensive interactions between the input parameters.

The overall correlation is anyway not very good. However, it is shown that the network is able to extract at least a part of the necessary information out of these readily obtainable structural parameters. Figure 4 plots the input sensitivities as circles on a structure of Epothilone A. The area of the circles is proportional to the corresponding sensitivity. The special importance of the region around C-9..C-11, C-12 and C-13 and of region **C** becomes clearly observable. A smaller but still obtainable influence has the region from C-3 to C-8. This site of the Epothilone molecule has therefore a high probability for binding at Tubulin.

This result is inline with qualitative suggestions by Winkler and Axelsen(Winkler and Axelsen 1996). They overlay the discussed regions of the molecule with Taxol and report a reasonable structural similarity. Su, et. al.(Su, Balog et al. 1997) discussed already the high influence of the region **C**. However, the influence of the several structural elements could be quantified for the first time and a more detailed picture is achieved that way.

Figure 5 shows the results evaluated with the neural networks trained in the third experiment. The correlation coefficients as well as the RMSD values are well for the inhibition of carcinoma cell growth of the parental cell line 3a). Worse are the results for the cell lines with mutated Tubulins 3b,c). The RMSD values for the established relation lay in the range of 0.5 and 1.5 for the inhibition of carcinoma cell growth which is equivalent to a factor between 1.5 ($e^{0.5}$) and 4.5($e^{1.5}$) for the IC_{50} values.

The introduction of a combined network that predicts all three IC_{50} values parallel enhances the model immediately. The correlation coefficients increase and the RMSD values decrease by a factor of up to 2. This example demonstrates impressively the ability of the neural network to use relations between several output values to increase the accuracy of the prediction. The input values have to pass the same hidden layer and are processed up to this point identical for all three biological activities. Therefore in this part of the network the three fold information can be used for establishing the connections compared to the examples discussed before. Only in the last layer every output has to stabilize its own five weights. The relevant information is increased by a factor of three. This allows to stabilize more degrees of freedom (weights). The two additional weights in the hidden layer allow to establish the more complex and precise model.

The increase of the input sensitivities for 3b) and 3c) proves the enhancement of the model. Moreover the sensitivity pattern becomes similar for the prediction of the three constants for the inhibition of carcinoma cell growth, which again mirrors the relation between these values. The substituent at C-12, the structure between C-12 and C-13 as well as the substituent at C-21 have the major impact on this biological activities.

In a second step we use the established neural networks to test computer generated Epothilone analogues for their potential biological activity. This experiment is carried out for the neural net established in the second experiment as well as for the combined network from the third experiment 3d). Figure 6 gives the generated structures with the highest biological activities for both experiments. In case of network 2 the ranking can be directly derived from the calculated activities. In experiment 3 the hit list is gained by sorting the structures by the average of all three predicted biological activities.

2.6 milliards of Epothilone analogues are generated and tested in less than 24h of CPU time (450 MHz PII processor). No other approach is known to be able to perform such an analysis in this time. However, it has to be considered that the deviation in the predicted activities is comparable large due to the small set of data compared with the huge structural space. Moreover the induction of Tubulin polymerization is theoretically covered by 100% in this experimental setup. The neural network does not know this boundary condition and predicts values up to 493% for this activity. However, this is not senseless since substances can of course become more active than the synthesized derivatives. The experimental setup yields 96% for the most active synthesized compounds and is therefore not capable to detect much more active compounds with a significant enhancement. Nevertheless, the comparable bad correlation of 0.71 suggests to handle the values with care although the direction of further synthesis can be derived from these results.

Even more convincing results are obtained for the network from experiment 3. Figure 5 visualizes the distribution of the 37 known structures with respect to their biological activity. The activity distribution of all 624 generated structures is shown for the individual networks in comparison with the combined approach. Most of the structures have a smaller activity (higher value!) than the best synthesized structure. However, structural proposals exist that promise to have a higher activity (lower value) than all known substances. This number increases going from the individual networks to the combined approach. An additional increase in the prediction quality is reached by selecting only structures that promise to have a high activity for all three biological activities, as done in figure 6. The increase in activity by 2 to 3 on this scale corresponds to a factor of 10 to 20 in the IC_{50} value.

Especially low and high activities might be affected by larger deviations in the prediction due to the fact the network has to extrapolate these values outside the training range. However, for the purpose of drug optimization already this “half quantitative” information is essential. It is possible to generate a large number of structures on the computer and rank them according to their activity without the need of synthesis. Out of the results either new target structures for synthesis can be selected or more general rules for active compounds can be derived. Both information are able to accelerate the optimization process of a biological active compound essentially.

After a subsequent synthesis and testing of some out of the new structural proposals, the model could be adjusted with the new data. The prediction can be repeated with a better model and therefore a smaller deviation which would again lead to more precise definition of structural proposals for synthesis and so on. An optimization process using synthesis and QSAR modeling by neural networks alternatively could be established that way.

Conclusion. Neural networks are able to establish relations between the structure of substances and their biological activities if a reasonable large set of data is available for testing and training the neural network. Moreover artificial neural networks take relations between several input parameters (structural information) and also between several output values (biological activities) into consideration. A subsequent analysis of the established network connections allows to gain information about structure activity relations, strong and weak interactions and binding cites.

These quantitative models are able to predict the biological activities for all members of a structural space defined by the introduced parameters. The generated structures can be ranked with respect to their biological activities. This approach is therefore capable to screen a structural space in a short time without necessarily synthesizing all structures. The structure optimization of biological active drugs can be accelerated. In an iterative procedure a subsequent synthesis of the structures predicted to be highly active leads to an enhanced model that can be used to repeat the prediction, and so on.

The described procedure is explained and demonstrated successfully on a set of the Epothilone analogues that induce the polymerization of Tubulin and are therefore able to inhibit carcinoma cell growth. Neural networks are trained to predict the induction of Tubulin assembly as well as the inhibition of carcinoma cell growth. Relevant points of interaction between Epothilones and Tubulin could be identified. Up to 2.6 milliards structures are generated out of a conformational space described by only 198 Epothilone analogues and ranked with respect to their ability to induce the Tubulin polymerization. In a second approach 624 structure are generated out of a conformational space described by 37 Epothilone analogues. These structures are analyzed due to high activity to inhibit carcinoma cell growth in three cell lines. Both analysis suggest structures that promise to have a equal or even higher biological activity than all synthesized Epothilone analogues.

Acknowledgement. The authors would like to thank Professor Christian Griesinger for useful discussion, Stefan Bartoschek for careful reading the manuscript and Stefanie Lamberth for providing the Tubulin graphics. J. M. is supported by a Kekulé stipend of the Fonds der Chemischen Industrie.

Figure captions.

figure 1: Scheme of data processing. A chemical structure is translated into a numerical code. This code is applied to an artificial neural network. The network is trained to predict a certain biological activity.

figure 2: Structure of Epothilone with the four regions of modification.

figure 3: Results of experiment 1 and 2. The correlation diagrams for all five neural networks with the experimental induction of Tubulin polymerization data on the y axis and the network predicted value on the x axis. Training data are indicated by an open circle and test data by a filled square. In the second column the input sensitivities are given. In experiment 1 not all of the input parameters are necessary. Therefore the used parameters are marked with a (#).

figure 4: Plot of the input sensitivities on a scheme of the Epothilone structure. The area of the circles is proportional to the sensitivity obtained for the individual parameter. If more than one parameter for a substituent are used, their sensitivities are added to gain a realistic picture.

figure 5: Results of experiment 3. The correlation diagrams for all three neural networks are visualized with the experimental data on the y axis and the network predicted value on the x axis. Training data are indicated by an open circle and test data by a filled square. In the second column correlation diagrams provide the results for the network that predicts all three biological activities parallel. The diagram in the third column displays the input sensitivity values. Further the distribution of the biological activities for the generated 624 possible structures is given for both networks in comparison with the distribution of the 37 synthesized derivatives. In both diagrams the data for the single networks are indicated by black bars while the data for the combined network approach are indicated by gray bars. White bars visualize the data for the 37 experimentally determined biological activities.

figure 6: Computer generated structures, that promise to have a higher biological activity for induction of Tubulin polymerization only (2) and also for inhibition of carcinoma cell growth (3). Carbon atoms marked with * undergo a change in stereochemistry with respect to the drawn conformation. The structure is marked in the table with ‘x’. The predicted IC_{50} –values for the parental cell line [a] and for the two mutated cell lines [b] and [c] are given.

Literature.

Altmann, K.-H., G. Bold, et al. (2000). “Epothilones and their analogs - potential new weapons in the fight against cancer.” *Chimia* **54**(11): 612-621.

Anklam, E., M. R. Bassani, et al. (1997). “Characterization of Cocoa Butters and Other Vegetable Fazs by Pyrol Mass Spectroscopy.” *Fresenius Journal of Analytical Chemistry* **357**(7)): 981-984.

Baldi, P., S. Brunak, et al. (1999). “Exploiting the past and future in protein secondary *Bioinformatics* **15**(11): 937-946.

Bienfait, B. (1994). “Applications of High-Resolution Self-Organizing Maps to Retrosynthetic and QSAR Analysis.” *J. Chem. Inf. Comput. Sci.* **34**(4)): 890-898.

Bollag, D. M., P. A. McQueney, et al. (1995). “Epothilones, a New Class of Microtubulue-stabilizing Agents with a Taxol-like Mechanism of Action.” *Cancer Research* **55**: 2325-2333.

Cherqaoi, D. and D. Villemin (1994). “Use of a Neural Network to determine the Boiling Point of Alkanes.” *J. Chem. Soc. Faraday Trans.* **90**(1)): 97-102.

Choy, W. Y., B. C. Sanctuary, et al. (1997). “Using neural network predicted secondary structure information in automatic protein NMR assignment.” *J. Chem. Inf. Comput. Sci.* **37**(6): 1086-1094.

Devillers, J. (1996). “Designing Molecules with Specific Properties from Intercommunicating *J. Chem. Inf. Comput. Sci.* **36**: 1061-1066.

Emerenciano, V. d. P., L. D. Melo, et al. (1997). “Application of artificial intelligence in organic chemistry. Part XIX#. Pattern recognition and structural dtermination of flavonoids using 13C-NMR spectra.” *Spectroscopy* **13**: 181-190.

Gerth, K., N. Bedorf, et al. (1997). “Epothilons A and B: Antifungal and Cytotoxic Compounds from Sorangium cellulosum (Myxobacteria) - Production, Physico-chemical and Biological Properties.” *J. Antibiothics* **49**(6)): 560-563.

Giannakakou, P., R. Gussio, et al. (2000). “A common pharmacophore for epothilone and taxanes: molecular basis for drug resistance conferred by tubulin mutations in human cancer cells.” *Proc. Natl. Acad. Sci. U. S. A.* **97**(6): 2904-2909.

Guermeur, Y., C. Geourjon, et al. (1999). “Improved performance in protein secondary structure prediction by inhomogeneous score combination.” *Bioinformatics* **15**(5): 413-421.

He, L., P. G. Jagtap, et al. (2000). “A Common Pharmacophore for Taxol and the Epothilones Based on the Biological Activity of a Taxane Molecule Lacking a C-13 Side Chain.” *Biochemistry* **39**(14): 3972-3978.

Höfle, G., N. Bedorf, et al. (1994). “Epothilone derivatives.” *Chemical Abstracts* **120**: 836.

Höfle, G., N. Bedorf, et al. (1996). “Epthilon A und B - neuartige, 16gliedrige Makrolide mit cytotoxischer Wirkung: Isolierung, Struktur im Kristall und Konformation in Lösung.” *Angewandte Chemie* **108**(13/14)): 1671-1673.

Höfle, G., N. Glaser, et al. (1999). “N-oxidation of epothilone A-C and O-acyl rearrangement to C-19- and C-21-substituted epothilones.” *Angew. Chem., Int. Ed.* **38**(13/14): 1971-1974.

Höfle, G., N. Glaser, et al. (1999). “Epothilone A-D and their thiazole-modified analogs as novel anticancer agents.” *Pure Appl. Chem.* **71**(11): 2019-2024.

Isu, Y., U. Nagashima, et al. (1996). “Development of Neural Network Simulator for Structure-Activity Correlation of Molecules (NECO). Prediction of Endo/Exo Substitution of Norbornane Derivatives and of Carcinogenic Activity of PAHs from 13C- NMR Shifts.” *J. Chem. Inf. Comput. Sci.* **36**(2)): 286-293.

Johnson, J., S.-H. Kim, et al. (2000). “Synthesis, Structure Proof, and Biological Activity of Epothilone Cyclopropanes.” *Org. Lett.* **2**(11): 1537-1540.

Kaartinen, J., S. Mierisova, et al. (1998). “Automated Quantification of Human Brain Metabolites by Artificial Neural Network Analysis from *in Vivo* Single-Voxel ^{1}H NMR Spectra.” *J. Magn. Res.* **134**: 176-179.

Kowalski, R. J., P. Giannakakous, et al. (1997). “Activities of the Microtubule-stabilizing Agents Epothilone A and B with Purified and in Cells Resistant to Paclitaxel (Taxol).” *J. Biol. Chem.* **272**(4): 2534-2541.

Lee, C. B., T.-C. Chou, et al. (2000). “Total Synthesis and Antitumor Activity of 12,13-Desoxyepothilone F: An Unexpected Solvolysis Problem at C15, Mediated by Remote Substitution at C21.” *J. Org. Chem.* **65**(20): 6525-6533.

Lohninger, H. (1993). “Evaluation of Neural Network Based on Radial Basis Functions and Their Application to the Prediction of Boiling Points from Structural Parameters.” *J. Chem. Inf. Comput. Sci.* **33**: 736-744.

Martello, L. A., H. M. McDaid, et al. (2000). “Taxol and discodermolide represent a synergistic drug combination in human carcinoma cell lines.” *Clin. Cancer Res.* **6**(5): 1978-1987.

Meiler, J. (1998). “Untersuchung von Struktur-Eigenschafts-Beziehungen für die Spezifität von Serin-Proteasen gegenüber Polypeptiden mittels NMR-Spektroskopie und künstlicher neuronaler Netze” *Universität Leipzig* **diploma thesis**.

Meiler, J. (2000). *http://krypton.org.chemie.uni-frankfurt.de/~mj*.

Muhlradt, P. F. and F. Sasse (1997). “Epothilone B stabilizes microtubili of macrophages like taxol without showing taxol-like endotoxin activity.” *Cancer Res.* **57**(16): 3344-3346.

Mulzer, J. (2000). “Epothilone B and its derivatives as novel antitumor drugs: total and partial *Monatsh. Chem.* **131**(3): 205-238.

Nicolaou, K. C., Y. He, et al. (1998). “Total Synthesis of Epothilone E and Analogues with Modified Side Chains through the Stille Coupling Reaction.” *Angewandte Chemie* **27**(1/2): 84-87.

Nicolaou, K. C., F. Roschangar, et al. (1998). “Chemie und Biologie der Epothilone.” *Angewandte Chemie* **110**: 2120-2153.

Nicolaou, K. C., F. Sarabia, et al. (1997). “Total Synthesis of Epothilone A: The *Angewandte Chemie* **36**(5): 525-527.

Nicolaou, K. C., R. Scarpelli, et al. (2000). “Chemical synthesis and biological properties of pyridine epothilones.” *Chem. Biol.* **7**(8): 593-599.

Nicolaou, K. C., D. Vourloumis, et al. (1997). “Designed Epothilones: Combinatorial Synthesis, Tubulin Assembly Properties, and Cytotoxic Action against Taxol-Resistant Tumor Cells.” *Angewandte Chemie* **36**(19): 2097-2103.

Nicolaou, K. C., D. Vourloumis, et al. (1997). “Gezielt entworfene Epothilone: kombinatorische Synthese, Induktion der Tubulin-Polymerisation und cytotoxische Wirkung gegen taxolresistente Tumorzellen.” *Angewandte Chemie* **109**: 2181-2187.

Nogales, E., S. G. Wolf, et al. (1998). “Structure of the ab tubulin dimer by electron *Nature* **39**: 199-203.

Nogales, E., S. G. Wolf, et al. (1995). “Structure of tubulin at 6.5A ad location of the taxol-binding site.” *Nature* **375**: 424-427.

Rodrigues, G. d. V., I. P. d. A. Campos, et al. (1997). “Application of artificial intelligence in organic chemistry. Part XX#. Determination of groups attached to the skeleton of natural products using 13C nuclear magnetic resonance spectroscopy.” *Spectroscopy* **13**: 191-200.

Rost, B. and C. Sander (1993). “Improved prediction of protein secondary structure by use of sequence profiles and neural networks.” *Proc. Natl. Acad. Sci. USA* **90**: 7558-7562.

Rost, B., C. Sander, et al. (1994). “Redefining the Goals of Protein Secondary Structure Prediction.” *J. Mol. Biol.* **235**: 13-26.

Schiff, P. B., J. Fant, et al. (1979). “Promotion of microtubule assembly in vitro by taxol.” *Nature* **277**: 665-668.

Schinzer, D., A. Limberg, et al. (1997). “Total Synthesis of (-)-Epothilone A.” *Angew. Chem., Int. Ed.* **36**(5): 523-524.

Schinzer, D., A. Limberg, et al. (1996). “Studies Towards the Total Synthesis of Epothilones: Asymmetric Synthesis of the Key Fragments.” *Chem. Eur. J.* **2**(11)): 1477-1482.

Selbig, J., T. Mevissen, et al. (1999). “Decision tree-based formation of consensus protein secondary structure prediction.” *Bioinformatics* **15**(2): 1039-1046.

Su, D.-S., A. Balog, et al. (1997). “Structure-activity relationships of the epothilones and the first in vivo comparison with paclitaxel.” *Angew. Chem., Int. Ed. Engl.* **36**(19): 2093-2096.

Svozil, D., J. G. Sevcik, et al. (1997). “Neural Network Prediction of the Solvatochromic Polarity/Polarizability Parameter Pi2H.” *J. Chem. Inf. Comput. Sci.* **37**: 338-342.

Taylor, R. E. and J. Zajicek (1999). “Conformational Properties of Epothilone.” *J. Org. Chem.* **64**(19): 7224-7228.

Thomas, S. and E. Kleinpeter (1995). “Zur Zuordnung der 13C-chemischen Verschiebung substituierter Naphtaline aus Ladungsdichtenmit Hilfe Neuronaler Netze.” *J. Prakt. Chem./Chem.-Ztg.* **337**: 504-507.

Von Angerer, E. (2000). “Tubulin as a target for anticancer drugs.” *Curr. Opin. Drug Discovery Dev.* **3**(5): 575-584.

Wang, H.-c., J. Dopazo, et al. (1998). “Self-organizing tree growing network for classifying amino acids.” *Bioinformatics* **14**(4): 376-377.

Wang, M., X. Xia, et al. (1999). “A Unified and Quantitative Receptor Model for the Microtubule Binding of Paclitaxel and Epothilone.” *Org. Lett.* **1**(1): 43-46.

Winkler, J. D. and P. H. Axelsen (1996). “A Model for the Taxol (Paclitaxel)/Epothilone Pharmacophore.” *Bioorganic & Medicinal Chemistry Letters* **6**(24): 2963-2966.

Zemla, A., C. Venclovas, et al. (1998). “A Modified Definition of Sov, a Segment-Based Measure for Protein Secondary Structure Prediction Assessment.” *Proteins: Structure, Function, and Genetics* **34**: 220-223.

Zupan, J. and J. Gasteiger (1993). “Neural Networks for Chemists.” *VCH Verlagsgesellschaft mbH, Weinheim* **3-527-28603-9**.

Table 1: Structural parameters introduced to code the Epothilone derivatives

Par. [a]	**Region** [b]	**Modification**[c]	**Possible values**[d]
P01	A	C-7 stereochemistry	**7S**, 7R
P02	A	C-8 substituent up	**H**, Me
P03	A	C-8 substituent down	**Me**, H
P04	A	C-9..C-11 number of CH_2-groups	1, 2, **3**, 4, 5
P05	B	C-12 volume	H, **Me**, CH_2OH, CH_2OAc, $CH_2OC(O)tBu$, $CH_2OC(O)Ph$, CH_2OMe, CH_2OBn, CH_2Cl, CH_2I, CH_2CH_3, CH_2NHAc, CH=CH_2, C&CH, CHO, CO_2H, =CH_2[e], =CH-CH_3, OH
P06	B	C-12 mass	
P07	B	C-12 max. electro negativity	
P08	B	C-12 / C-13 structure	**epoxid**, double bond cis, double bond trans, single bond
P09	B	C-12 stereochemistry	**12R**, 12S or cis, trans
P10	B	C-13 substituent	**H**, F, OH
P11	B	C-13 stereochemistry	**13S**, 13R or cis, trans
P12	B	C-15 stereochemistry	**15S**, 15R
P13	C	C-16 substituent	**Me**, CH_2CH_3
P14	C	C-17 substituent	**fife membered ring**, six membered ring
P15	C	22 atom type	**N**, C
P16	C	21 atom type	**C**, N
P17	C	19 atom type	**C**, S, O, CCl
P18	C	20 atom type	**S**, O, C, S=O
P19	C	C-21 volume	**Me**, CH_2OH, CH_2OAc, CH_2F, $(CH_2)_5OAc$, piperidyl, SMe, Ph, OEt, OH, OMe, H
P20	C	C-21 mass	
P21	C	C-21 max. electro negativity	
P22	D	C-3 stereochemistry	**3S**, 3R
P23	D	C-4 substituent	**Me_2**[f], three membered ring[g]
P24	D	C-6 stereochemistry	**6R**, 6S

[a]Defined structural parameters for coding the structure of all Epothilone analogues.

[b]The region (see figure 2) in which the modification is performed´.

[c]The kind of modification performed.

[d]All possible states for the particular parameter. The states occurring in **Epothilone A** are indicated by bold letters.

[e]A substituent labeled with "=X" is bond to the respective atom (C-12) with a double bond.

[f/g]At the atom C-4 either two methyl groups[f] or a cyclo propyl ring[g] is bond. This three membered ring includes C-14.

Table 2: Results of experiments 1, 2, and 3.

Exp.	**Biological activity**	**Net size**				**Number of data**		**Correlation coeff.**		**RMSD**	
		input	**hidden**	**output**	**weights**	**train**	**test**	**train**	**test**	**train**	**test**
1a)	TP[c] mod. in region **A**	8	2	1	21	21	5	0.89	0.86	0.44	0.31
b)	TP mod. in region **B**	10	3	1	37	41	7	0.53	0.57	0.84	0.68
c)	TP mod. in region **C**	14	3	1	49	55	9	0.68	0.78	0.92	0.74
d)	TP mod. in region **D**	8	2	1	21	23	5	0.39	-0.15	0.91	0.44
2)	TP	24	6	1	157	178	20	0.73	0.72	0.76	0.76
3a)	CG[d] (parental)	8	2	1	21	32	5	0.77	1.00	1.56	0.45
	combined network	8	4	4	56	31	4	0.87	0.99	0.98	1.07
b)	CG (1A9PTX10)	8	2	1	21	32	5	0.69	0.79	1.08	1.38
	combined network	8	4	4	56	31	4	0.94	0.96	0.47	1.04
c)	CG (1A9PTX22)	8	2	1	21	31	4	0.64	0.94	1.57	1.51
	combined network	8	4	4	56	31	4	0.85	1.00	1.03	0.96

[c]TP is the induction of Tubulin polymerization

[d]CG is the inhibition of carcinoma cell growth.

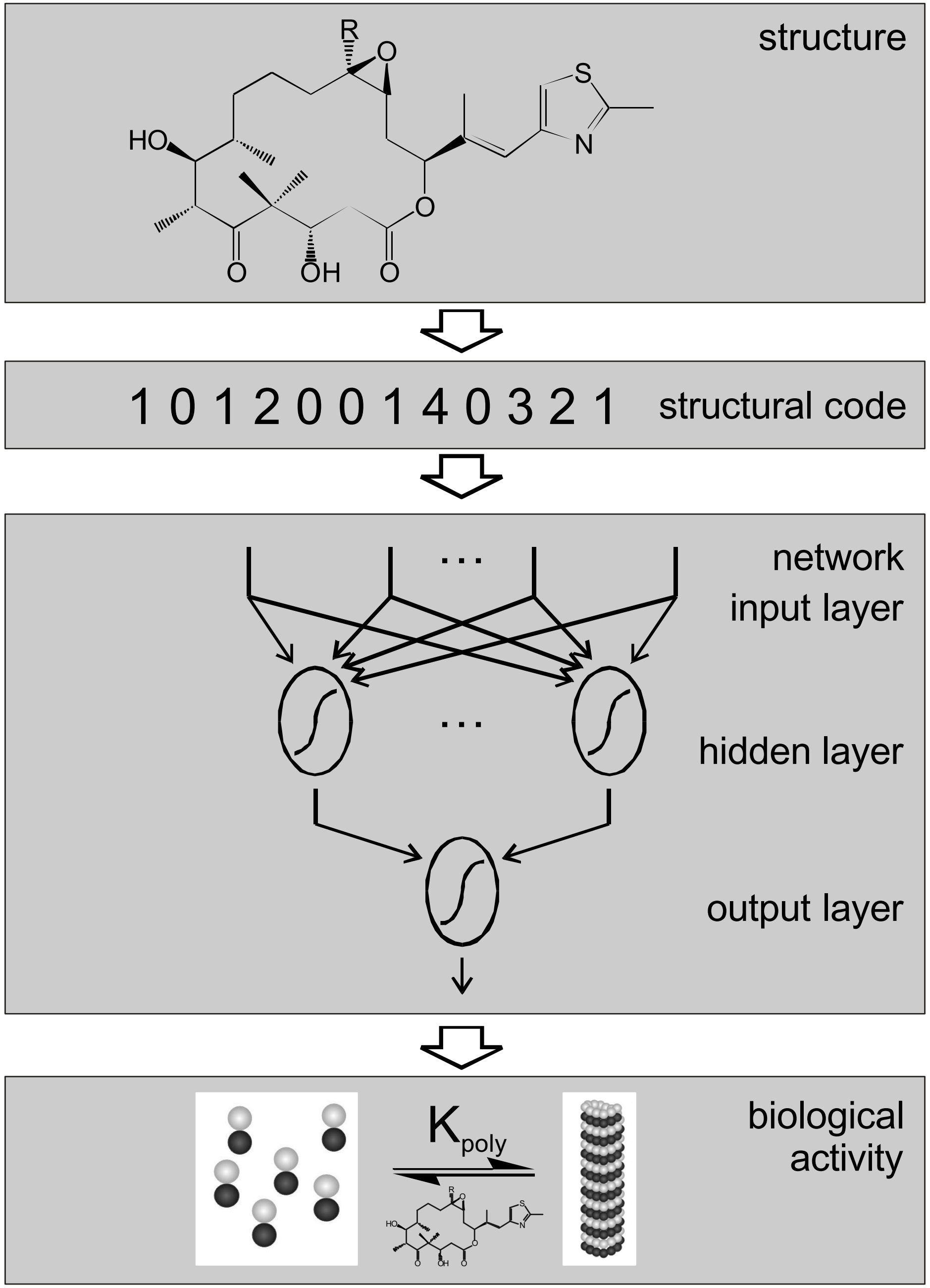


figure 1

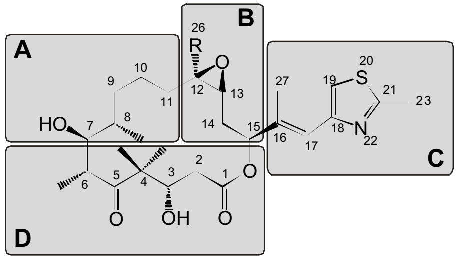


figure 2

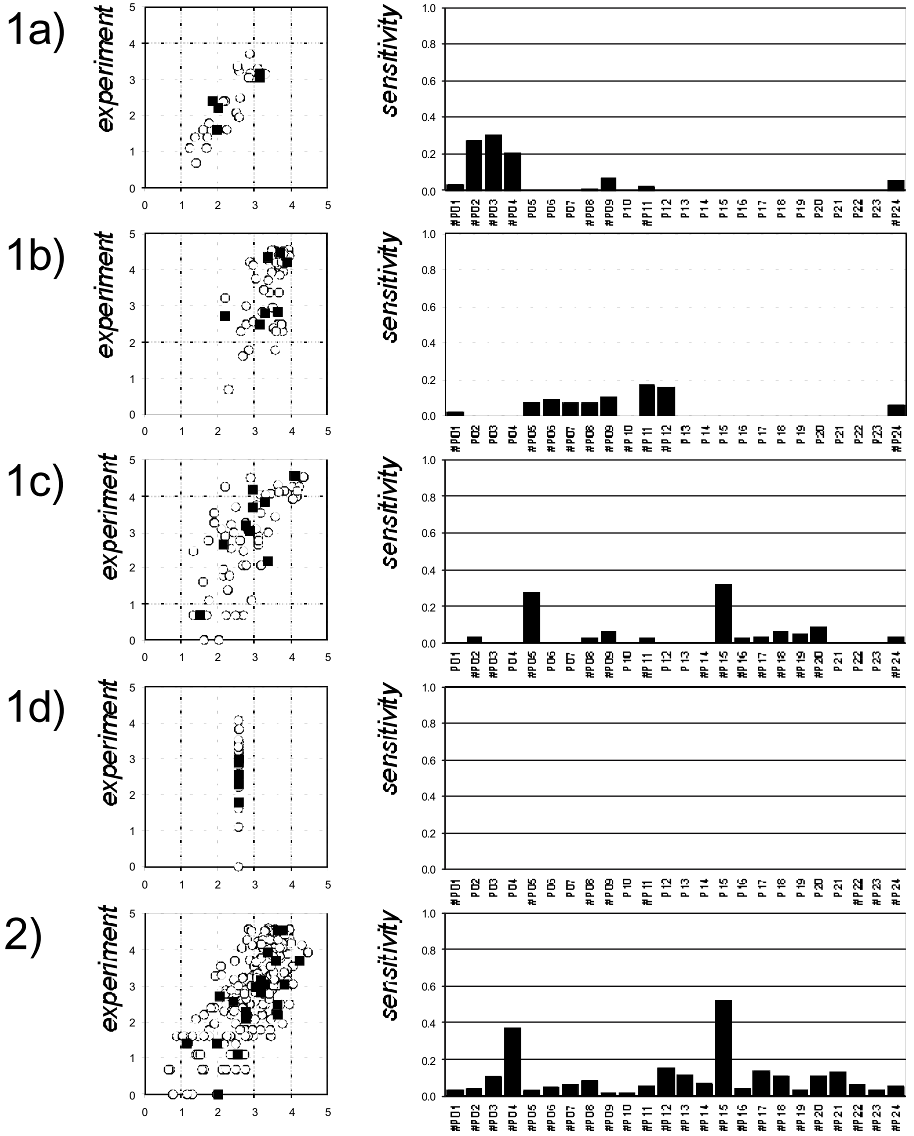


figure 3

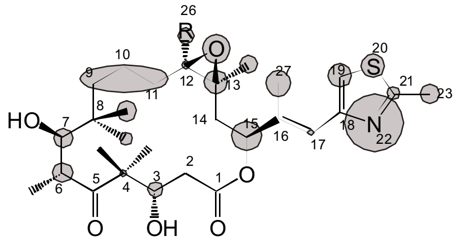


figure 4

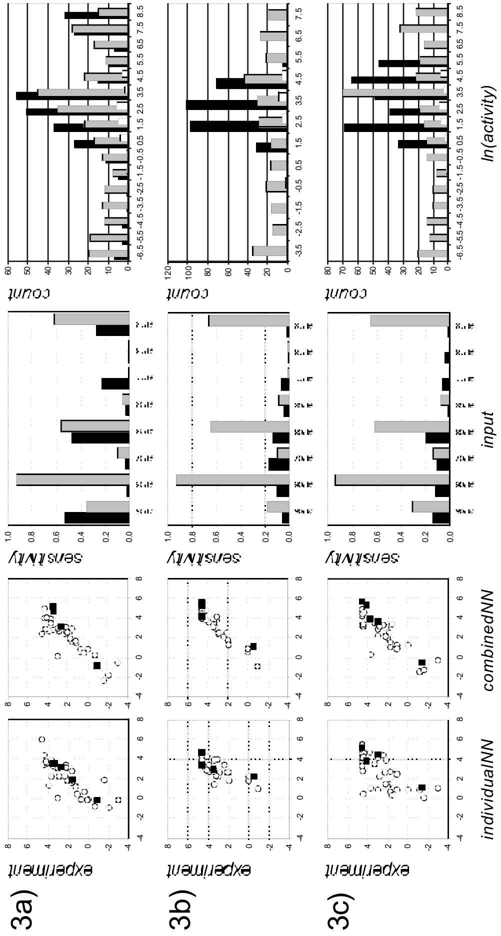
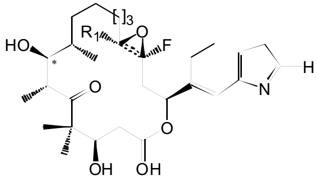


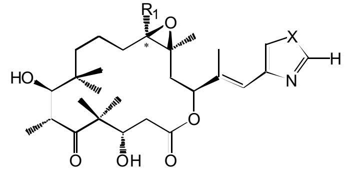
figure 5

2)



R_1	*	= = =	[d]
Epothilone A			76
CH_2OCH_3		epoxid	493
CH_2OH		epoxid	493
CH_2F	x	epoxid	488
OH		= trans	483
CH_2F		= cis	478

3)



R_1	*	X	[a]	[b]	[c]
EpothiloneA			$2.2 * 10^0$	$20.0 * 10^0$	$5.9 * 10^0$
$CH=CH_2$		S	$1.4 * 10^{-3}$	$30.5 * 10^{-3}$	$1.4 * 10^{-3}$
$CH=CH_2$	x	S	$1.4 * 10^{-3}$	$29.6 * 10^{-3}$	$1.3 * 10^{-3}$
$CH=CH_2$		O	$1.4 * 10^{-3}$	$30.8 * 10^{-3}$	$1.4 * 10^{-3}$
CH_3		O	$1.5 * 10^{-3}$	$32.4 * 10^{-3}$	$1.5 * 10^{-3}$

figure 6