Structure
Article

De Novo High-Resolution Protein Structure Determination from Sparse Spin-Labeling EPR Data

Nathan Alexander,[1,3] Marco Bortolus,[2] Ahmad Al-Mestarihi,[1] Hassane Mchaourab,[2,3] and Jens Meiler[1,3,*]
[1]Department of Chemistry
[2]Department of Molecular Physiology and Biophysics
[3]Center for Structural Biology
Vanderbilt University, Nashville, TN 37212, USA
*Correspondence: jens.meiler@vanderbilt.edu
DOI 10.1016/j.str.2007.11.015

SUMMARY

As many key proteins evade crystallization and remain too large for nuclear magnetic resonance spectroscopy, electron paramagnetic resonance (EPR) spectroscopy combined with site-directed spin labeling offers an alternative approach for obtaining structural information. Such information must be translated into geometric restraints to be used in computer simulations. Here, distances between spin labels are converted into distance ranges between β carbons by using a "motion-on-a-cone" model, and a linear-correlation model links spin-label accessibility to the number of neighboring residues. This approach was tested on T4-lysozyme and αA-crystallin with the de novo structure prediction algorithm Rosetta. The results demonstrate the feasibility of obtaining highly accurate, atomic-detail models from EPR data by yielding 1.0 Å and 2.6 Å full-atom models, respectively. Distance restraints between amino acids far apart in sequence but close in space are most valuable for structure determination. The approach can be extended to other experimental techniques such as fluorescence spectroscopy, substituted cysteine accessibility method, or mutational studies.

INTRODUCTION

The accelerated pace of genome sequencing has sparked the development of rapid structure determination methods and ambitious proposals for genome-scale structure determination utilizing primarily X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy (Stevens et al., 2001; Berman et al., 2002; Lesley et al., 2002; Westbrook et al., 2003). However, it has become clear that static and dynamic structural information for a significant subspace of the protein universe continues to evade these tools. Important examples include the static structure of membrane proteins (Tusnady et al., 2004), conformationally heterogeneous water-soluble proteins (Haley et al., 2000), and large protein complexes involved in major cellular processes (Harrison, 2004). Insight into conformational motions that mediate function is restricted to proteins amenable to NMR spectroscopy or to crystallization in multiple intermediate states. Furthermore, the absence of representative structures of entire protein families, whose members often share difficulties in structure determination, reduces the efficiency and accuracy of comparative modeling (Sali, 1998).

A complement of methods with intrinsically lower resolution can provide insight into these problems. Among them are probe-based approaches such as electron paramagnetic resonance (EPR) spectroscopy in combination with site-directed spin labeling (SDSL) (Hubbell et al., 1996; Mchaourab et al., 1997; Koteiche et al., 1998; Perozo et al., 1999; Liu et al., 2001; Brown et al., 2002; Dong et al., 2005). EPR analysis of spin-labeled proteins results in a set of structural restraints that describes, in a native-like setting, local environments as well as aspects of the global fold of the protein. Spin label accessibility and mobility can be used to determine secondary structure location and topology (Farahbakhsh et al., 1992; Altenbach et al., 2005). Distance measurements between pairs of spin labels in the range from 5–60 Å (Rabenstein and Shin, 1995; Borbat et al., 2002) reflect the relative packing of domains and secondary structures. In cases where these parameters are obtained in various conformational intermediates of the protein, they allow for a detailed mapping of structural changes involved in function (Dong et al., 2005). There are relatively few limits on the size and environment of the protein, particularly when compared to X-ray crystallography and NMR spectroscopy.

Despite the widespread application of SDSL (Fanucci and Cafiso, 2006), the use of EPR restraints for structure determination has not been systematically explored. A central question is the number and nature of EPR restraints necessary to obtain a structural model at a biologically relevant resolution. The most extensive use of spectroscopic data along with computational methods for structure determination is in NMR spectroscopy (Wüthrich, 1986). Typically consisting of distances not greater than 5–6 Å with upper and lower bounds, the geometric information is derived from NOE-based experiments. The number of such restraints required for the determination of a structure depends on the range and quality of such restraints, but is generally assumed to be above 15 restraints per residue (Nederveen et al., 2005).

Although EPR distance restraints have a longer range than their NMR counterparts, they are fundamentally less accurate since they report distances between probes introduced into the protein sequence. The significant length of the spin label linking arm implies that the EPR distances will have a rather

large uncertainty when translated into distances between α or β carbons unless the conformation of the spin label is known at every site. Therefore, previous efforts have focused on either using molecular dynamics simulations to define their trajectories (Sale et al., 2005) or on determining a library of rotamers from crystal structures of spin-labeled T4-lysozyme (Langen et al., 2000). These studies are critical since spin-label conformations are likely stabilized by weak specific interactions with neighboring amino acid side chain or backbone atoms. However, such calculations are time and resource intensive and not practical without a high-resolution structural model of the protein.

Sparse experimental data, such as EPR restraints, aid computational protein structure prediction algorithms by restricting the conformational space that must be considered in order to obtain the correct structure. For instance, the Rosetta de novo protein structure prediction algorithm (Simons et al., 1997; Bonneau et al., 2001a, 2001b; Bradley et al., 2003, 2005a, 2005b; Rohl et al., 2004) predicts high-resolution (better than 1.5 Å) structures of proteins with less than 80 amino acids in the absence of experimental restraints (Bradley et al., 2005b). In combination with sparse (less than one restraint per amino acid) NMR NOE distances and/or residual dipolar couplings (Bowers et al., 2000; Rohl and Baker, 2002; Meiler and Baker, 2005), the structure of proteins with up to 200 amino acids can be determined to medium-high resolution (1.5–3.0 Å).

In the present work, Rosetta is combined with sparse EPR distance restraints and solvent accessibility measures for high-resolution structure determination of the mostly helical T4-lysozyme (Weaver and Matthews, 1987) and the all β sheet protein αA-crystallin (Horwitz, 1992, 1993). We address the question of whether the EPR restraints can restrict the protein conformational space without assuming an atomic-detail model for the spin label's dynamics and without accounting for its context-dependent specific interactions. Also addressed are the questions of how many restraints are needed to obtain a high-resolution structure and what type of EPR restraint is most efficient.

The results demonstrate that sparse EPR restraints derived from a nonatomic model of the spin label lead Rosetta to high-resolution structures for both proteins. Also, distance restraints are more efficient in restricting conformational space than spin label accessibilities. Further analysis reveals that those between two amino acids far apart in sequence but close in Euclidian space are the most valuable.

RESULTS

EPR distance and accessibility data were transformed into structural restraints as described in Experimental Procedures. Briefly, distances between spin labels were translated into distances between β carbons by using a motion-on-a-cone model of the spin label location relative to the α carbon. The accessibilities of spin labels were computationally interpreted in terms of the exposed surface area. The effectiveness of the restraints to aid Rosetta in the folding process was then evaluated. Because distance restraints proved vastly more efficient in preliminary experiments, accessibility data was not used during modeling. De novo models were compared to the crystal structure of T4-lysozyme and a comparative model of aA-crystallin.

Evaluation of the "Motion-on-a-Cone" Model for Interpretation of Distance Restraints

The "motion-on-a-cone" model (see Figure 1) yields a predicted distribution for the difference between the distance separating the spin labels (d_{SL}) and that separating the two corresponding β carbons (Cβs; $d_{C\beta}$). Comparison of the predicted d_{SL}–$d_{C\beta}$ distribution with the d_{SL}–$d_{C\beta}$ obtained from the T4-lysozyme and αA-crystallin structures (Figure 1D) demonstrate that they essentially encompass the same range of d_{SL}–$d_{C\beta}$ and reveals a common bias in experiment and model for $d_{SL} > d_{C\beta}$. The comparison also reveals that the model over predicts the frequency with which large (>4 Å) d_{SL}–$d_{C\beta}$ values occur and underestimates the frequency with which low (<4 Å) d_{SL}–$d_{C\beta}$ values occur. However, the present application depends only on the ability of the model to predict the appropriate range of values for d_{SL}–$d_{C\beta}$; the frequency with which these values occur is not a part of the utility of this simple model.

Given a d_{SL}, the "motion-on-a-cone" model provides a restraint in the form of a predicted range for $d_{C\beta}$. The accuracy of the range can be evaluated by comparing it to the $d_{C\beta}$ calculated from the T4-lysozyme structure and the αA-crystallin comparative model. Practically all calculated $d_{C\beta}$ lie within the range predicted by the model (Figures 2C and 2G).

A Linear Regression Relation for Interpretation of Accessibility Restraints

Analogous to an experimental distance measurement, the experimental accessibility of a spin label (e_{SL}) needs to be translated into accessible surface areas of the protein structure for use as a restraint. For this purpose, a consensus linear regression relation between e_{SL} and the number of Cβ atoms within 8 Å of the Cβ of the corresponding amino acid ($e_{C\beta}$) was determined from T4-lysozyme and αA-crystallin structures. The linear relation is given by $e_{C\beta} = (0.76 - e_{SL}) \bullet 20.87$ and has a correlation coefficient of −0.83 to experimental T4-lysozyme and αA-crystallin data (Figures 2D and 2H). The strong correlation suggests the simple method of linearly relating e_{SL} to $e_{C\beta}$ is a sufficient means for obtaining a structural restraint from EPR accessibility data.

Influence of EPR Data on De Novo Fold Determination

To avoid the introduction of noise through unconstrained regions and focus on evaluating the contribution of EPR restraints in structure prediction, regions in both proteins that were not probed with spin labels were excluded from the calculations. For T4-lysozyme, the C-terminal 107 residue helical domain (amino acids 58–164) was modeled (Figure 2A). For αA-crystallin, the C-terminal 88 residue β sandwich domain (amino acids 60–147) was modeled (Figure 2E).

The influence of the experimental EPR restraints on de novo protein folding with Rosetta was evaluated by building 10,000 models for each protein (1) without the use of experimental data, (2) with only the use of distance restraints, (3) with only the use of solvent accessibility restraints, and (4) using both sets of restraints. The average model quality was monitored by the root-mean-square deviation (rmsd). The results for both T4-lysozyme and αA-crystallin follow the same trends: there is an improvement in the quality of models created with distance restraints compared to models created without distance restraints (Figures 3A and 3B, T4-lysozyme; Figures 3I and 3J,

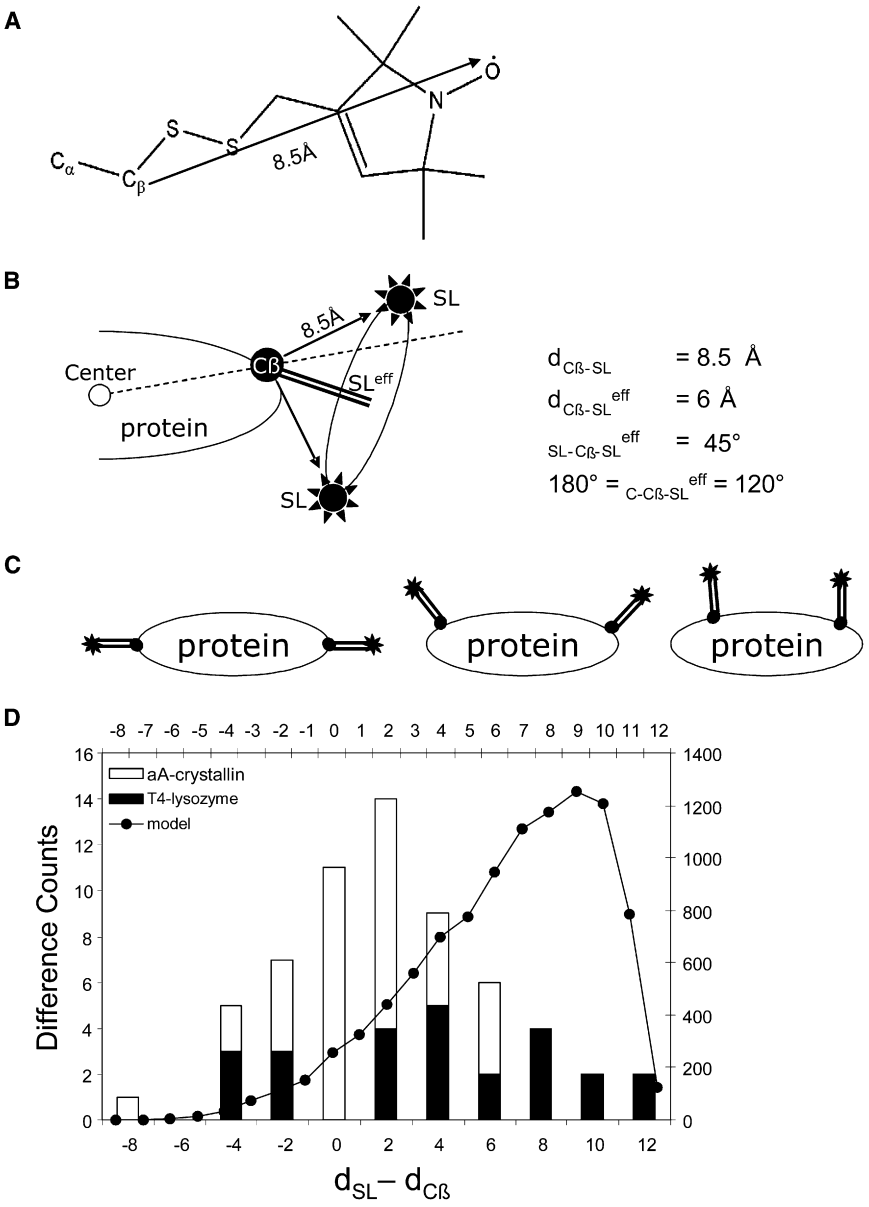


Figure 1. Rational for Translating d_{SL} into $d_{C\beta}$ for Use as a Restraint

(A) Chemical structure of a nitroxide spin-label side chain with the distance from the Cβ atom to the spin label indicated (Borbat et al., 2002).

(B) Illustration of how the maximum distance from Cβ to spin label, SL, is reduced to an effective distance, SL^{eff} (depicted by a double line).

(C) d_{SL} is a starting point for the upper estimate of $d_{C\beta}$, and subtracting the effective distance of 6 Å twice from d_{SL} gives a starting point for the lower estimate of $d_{C\beta}$.

(D) A histogram compares T4-lysozyme crystal structure (black bars, left y axis, bottom x axis) and αA-crystallin comparative model (white bars, left y axis, bottom x axis) d_{SL}–$d_{C\beta}$ values with those obtained from the simple cone model (circles and line, right y axis, top x axis).

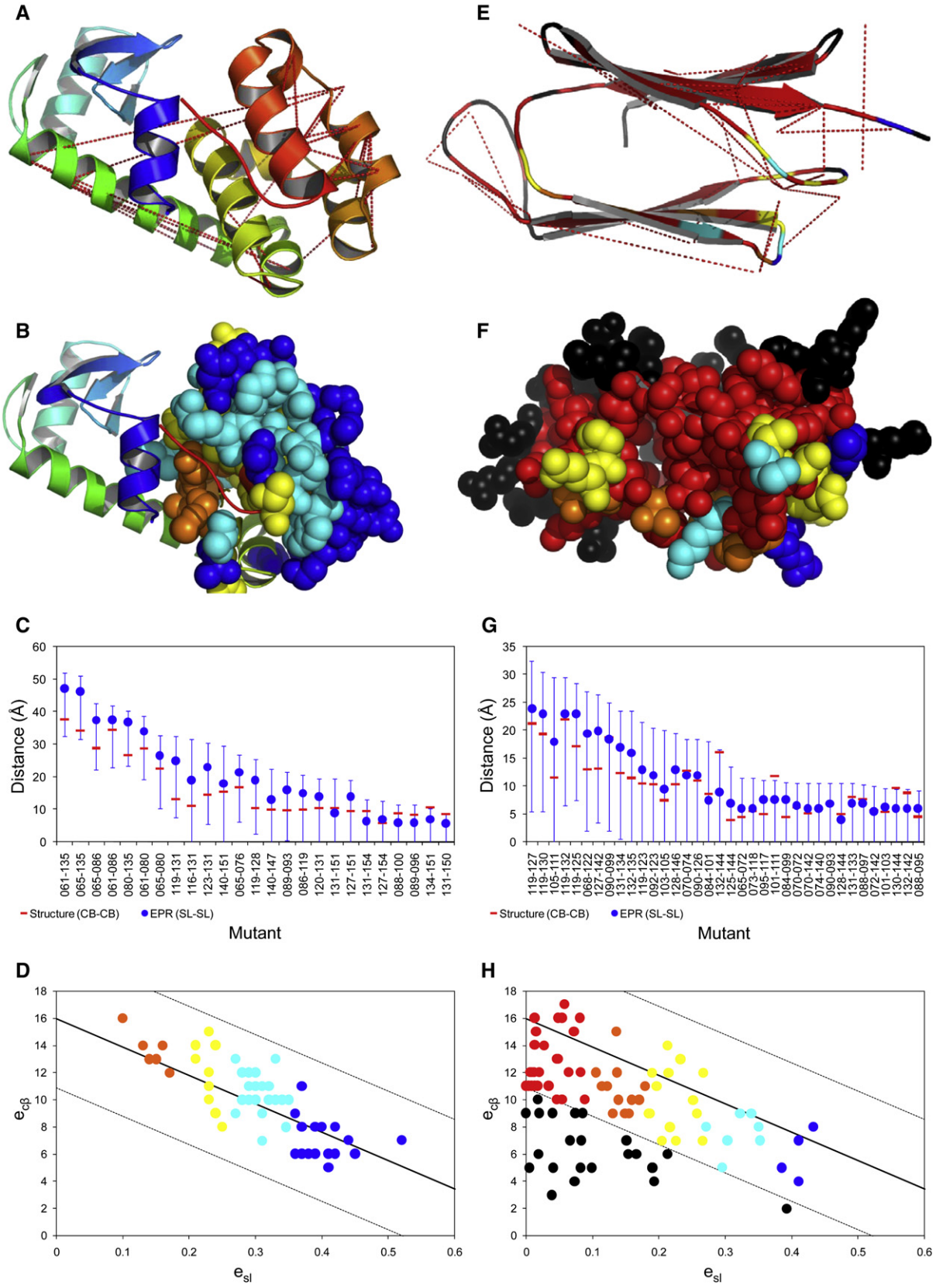
αA-crystallin); there is very little to no improvement when accessibility restraints are included (Figures 3A and 3C, T4-lysozyme; Figures 3I and 3K, αA-crystallin); using accessibility restraints in conjunction with distance restraints provides little to no improvement over using distance restraints alone (Figures 3B and 3D, T4-lysozyme; Figures 3J and 3L, αA-crystallin). It is clear from these analyses that the distance restraints are critical for improving the rmsd distribution of models, while solvent accessibility data only marginally improve the quality of water soluble protein models.

Influence of Spin-Label Placement on De Novo Fold Determination

Spatial contacts of amino acids that are distant in sequence define the protein fold best (Baker, 2000; Bonneau et al., 2002). However, whereas the sequence separation ($s_{C\beta}$) can be chosen when designing an EPR experiment, the Euclidean distance ($d_{C\beta}$) is generally unknown. EPR experiments do not provide contact data but, instead, distances of up to 50 Å. Thus, the information content ($I_{C\beta}$) of an EPR distance restraint can be defined as directly proportional to $s_{C\beta}$ but inversely proportional to $d_{C\beta}$: $I_{C\beta} \sim s_{C\beta}/d_{C\beta}$.

To investigate the influence of spin label location and resulting $I_{C\beta}$ on de novo structure determination, two experiments were designed with subgroups of all available restraints. First, all restraints are ranked by $I_{C\beta}$ to assess their power in an idealized experiment. Second, all restraints are ranked by $s_{C\beta}$ in order to simulate the choices the experimentalist can make when selecting sites for labeling. Ten thousand models for T4-lysozyme and αA-crystallin were built that used (1) the one-third restraints with highest $I_{C\beta}$, (2) the one-third restraints with lowest $I_{C\beta}$, (3) the two-third restraints with highest $I_{C\beta}$, and (4) the two-third restraints with lowest $I_{C\beta}$. The experiment was then repeated using $s_{C\beta}$ instead of $I_{C\beta}$.

Once again, the trends of the results are the same for both proteins and for both $I_{C\beta}$ and $s_{C\beta}$ experiments. (1) Using the one-third restraints with highest $I_{C\beta}$ shifts the rmsd distribution into the same range that is obtained when using all of the available distance restraints (Figures 3E and 3B, T4-lysozyme; Figures 3M and 3J, αA-crystallin). (2) Using the one-third restraints with lowest $I_{C\beta}$ only slightly shifts the rmsd distribution toward lower rmsds and is similar to that in the absence of distance restraints (Figures 3F and 3A, T4-lysozyme; Figures 3N and 3I, αA-crystallin). (3) There is little shift in the rmsd distribution when the two-third restraints with highest $I_{C\beta}$ are used compared to only one-third (Figures 3G and 3E, T4-lysozyme; Figures 3O and 3M, αA-crystallin); the extra restraints increase the number



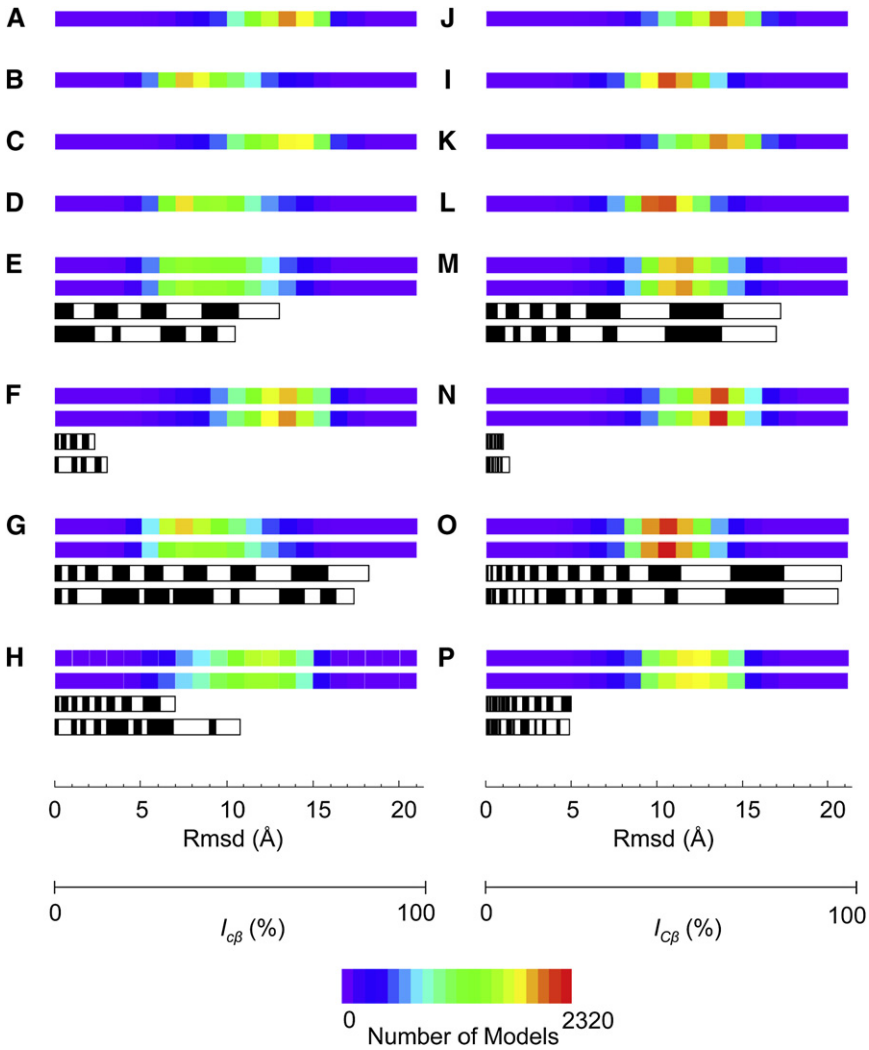


Figure 3. Illustration of the Value of the Experimental Restraints in De Novo Protein Folding for T4-Lysozyme and αA-Crystallin

(A)–(H) corresponds to T4-lysozyme, and (I)–(P) corresponds to αA -crystallin. The backbone rmsd distribution of 10,000 T4-lysozyme de novo models created (A) without the use of EPR restraints, (B) with only the use of EPR distance restraints, (C) with only the use of EPR accessibility restraints, (D) with the use of EPR distance and accessibility restraints. (E) The backbone rmsd distribution of 10,000 T4-lysozyme de novo models created with the use of 1/3 of the EPR distance restraints: top bar, those with the largest information content; second bar, those between amino acids furthest apart in sequence. The third and fourth black and white bars denote the sum percent of information content of the restraints used for the top and second bars, respectively. The width of the blocks comprising the black and white bars denotes the information content of individual restraints. (F) Same as for (E) but using the distance restraints with the lowest information content (top bar) and nearest in sequence (second bar). (G) Same as for (E) but using 2/3 of the total distance restraints. (H) Same as for (F) but using 2/3 of the total distance restraints. (I–P) Same as (A)–(H) but for αA-crystallin.

of lower rmsd models that are created. (4) When the two-third restraints with lowest $I_{C\beta}$ are used, there is a small shift in the rmsd distribution compared to when no distance restraints are used (Figures 3H and 3A, T4-lysozyme; Figures 3P and 3I, αA-crystallin). However, this shift is not nearly as drastic as the shift obtained when the one-third most informative distant restraints are used.

Using $s_{C\beta}$ to select restraints instead of $I_{C\beta}$ results in only a slight reduction in model quality and slight variation in total $I_{C\beta}$ of the selected restraints compared to that of the restraints employed in the first experiment (Figures 3E–3H, T4-lysozyme; Figures 3M–3P, αA-crystallin). This indicates maximal sequence separation can be used to effectively define spin-label placement and select for restraints with large $I_{C\beta}$. However, it should be noted that some of the sites for spin labeling T4-lysozyme were chosen with the crystal structures at hand that might bias the restraint sets for increased information content. Furthermore, this experiment does not test how spin labels should be distributed along the sequence. Additional experiments indicate that, besides maximizing $s_{C\beta}$, a uniform distribution of spin labels over the sequence is optimal (data not shown).

Rosetta Folding of T4-Lysozyme and αA-Crystallin

No accessibility restraints were used in the large-scale folding simulations, due to their minimal influence on structure determination. Of the 500,000 models built for T4-lysozyme, the lowest rmsd obtained was 2.39 Å with a total of 117 models having an rmsd value smaller than 3.5 Å. Of the 500,000 models built for αA-crystallin, the lowest rmsd obtained was 3.36 Å with a total of 46 models having an rmsd value smaller than 4.0 Å.

Filtering the 500,000 models of T4-lysozyme and αA-crystallin reduces the number considered for high-resolution refinement to a manageable number and enriches the high-resolution

Figure 2. Map of the EPR Restraints on the T4-Lysozyme Crystal Structure and on the αA-Crystallin Comparative Model

(A and E) Red dotted lines show $d_{C\beta}$ distances, which are restrained by respective d_{SL}.

(B and F) Residues for which accessibilities e_{SL} were measured are depicted as space-filling models.

(C and G) Diagram shows d_{SL} (blue circle), the range of the derived distance restraints (blue), and the corresponding crystal/comparative model $d_{C\beta}$ (red bar).

(D and H) Diagram illustrating the correlation of e_{SL} with $e_{C\beta}$. The lines indicate the consensus model fit $\pm 3 \bullet \sigma_{C\beta}$, where $\sigma_{C\beta}$ was recalculated based on the consensus fit to be 1.70. In (B), (F), (D), and (H), the residues are color coded with decreasing e_{SL} from blue–cyan–yellow–orange–red; black indicates amino acids in αA-crystallin that show reduced experimental accessibility due to intermolecular contacts with other αA-crystallin units in the oligomeric protein. (A)–(D) corresponds to T4-lysozyme, and (E)–(H) corresponds to αA-crystallin.

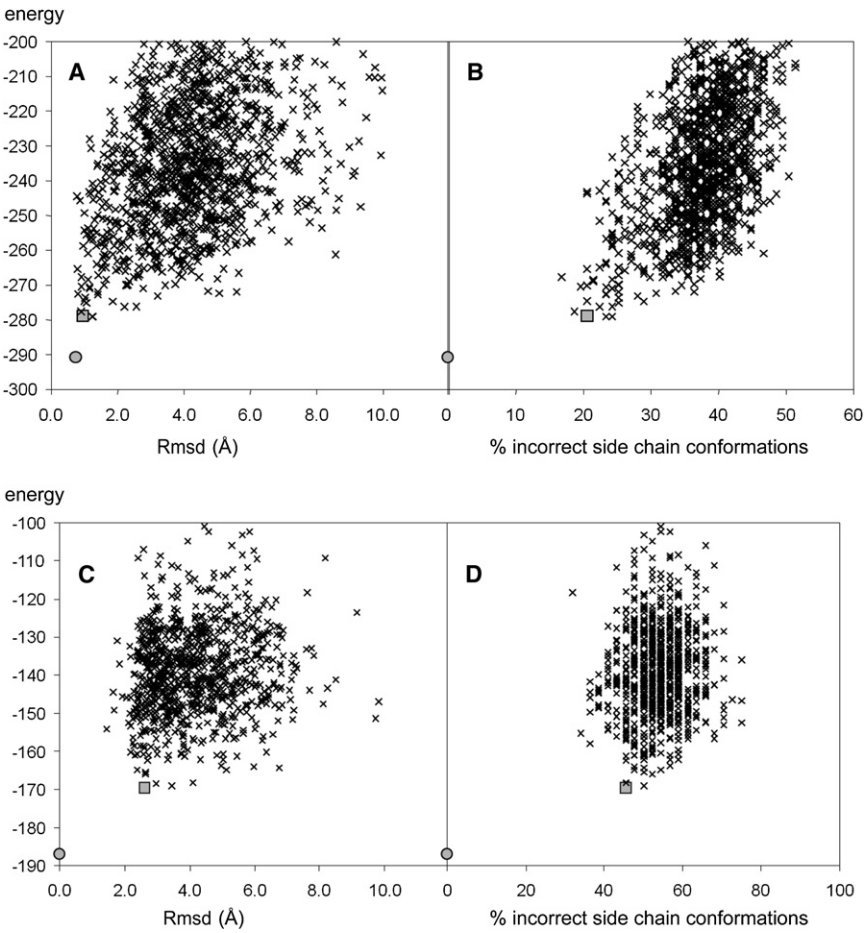


Figure 4. Correlation of De Novo Models' Accuracy with the Energy of the De Novo Models

(A and C) The nonloop rmsd versus Rosetta energy for T4-lysozyme and αA-crystallin models, respectively.

(B and D) The percentage of incorrectly built side chain conformations versus Rosetta energy for T4-lysozyme and αA-crystallin models, respectively. In all diagrams, the minimized crystal structure or comparative model is depicted as a circle; the lowest energy model is shown as a square.

refinement pool for low rmsd models. Enrichment is measured as the fraction of low rmsd models in the filtered ensemble divided by the fraction of low rmsd models in the original ensemble. For T4-lysozyme, requiring full agreement with all distance restraints and an overall Rosetta score better than −35 points prunes down the number of candidate structures to 10,906, keeping 27 models with rmsd values smaller than 3.5 Å. This enriches low rmsd (≤3.5 Å) models in the dataset by a factor of 27/10,906 ÷ 117/500,000 = 10.6.

For αA-crystallin, in order to keep approximately 10,000 structures for high-resolution refinement, models were required to have an overall Rosetta score better than or equal to −75, β strand pairing score better than −31, and total sum of all distance violations smaller than 3.0 Å. These criteria limit the number of structures to 9,796. Of the 9,796 models, 26 models have an rmsd of less than 4.0 Å, which is an enrichment of low rmsd (≤4.0 Å) models of 26/9,796 ÷ 46/500,000 = 28.8.

Structure Determination of T4-Lysozyme and αA-Crystallin

After high-resolution refinement, models were filtered by agreement with distance restraints. T4-lysozyme models were again required to be in full agreement with all distance restraints. αA-crystallin models were required to have sum total distance restraint violations of less than 1 Å. The remaining models were sorted by Rosetta full-atom energy, and the lowest energy model for each protein was compared to the crystal structure of T4-lysozyme and the comparative model of αA-crystallin. For rmsd analysis, loop regions of the α-helical and β sandwich domains were disregarded.

In addition to rmsd analysis, the agreement of side chain conformations can be captured by comparing the dihedral angles $\chi_{1\ldots4}$. A specific set of such angles $\chi_{1\ldots4}$ is called a "rotamer" (Dunbrack and Karplus, 1993; Dunbrack, 2002). If all angles $\chi_{1\ldots4}$ of an amino acid side chain deviate less than 60°, the rotamer is the same, and the conformation is closely recovered. The number reported for side chain conformation comparison is the percentage of nonagreeing rotamers (Figures 4B and 4D).

The Rosetta energy of T4-lysozyme de novo models decreases as their rmsd and side chain rotamer disagreement to the native structure diminish (Figures 4A and 4B). This allows the selection of high-resolution models based on energy alone. The lowest energy de novo model achieves an rmsd to the crystal structure of 1.0 Å in the α-helical domain and 2.0 Å over all modeled residues (Figure 5A). Eighty percent of all rotamers are in agreement with the crystal structure.

The Rosetta energy of αA-crystallin de novo models decreases as their structure approaches that of the comparative model (Figure 4C). However, side chain rotamer agreement does not correlate with Rosetta energy (Figure 4D). The lowest energy de novo model achieves an rmsd of 2.6 Å for the β sandwich and 4.0 Å over the whole protein (Figure 5B). The rotamer agreement in the β sandwich is 54.5%. Note that similar rmsds and side chain agreements are also found between the comparative model and two additional comparative models based on alternative templates (data not shown).

DISCUSSION

Structure Determination from Sparse EPR Restraints

The major conclusion of this paper is that structural restraints obtained from EPR analysis of spin-labeled proteins can be used in combination with de novo prediction methods to

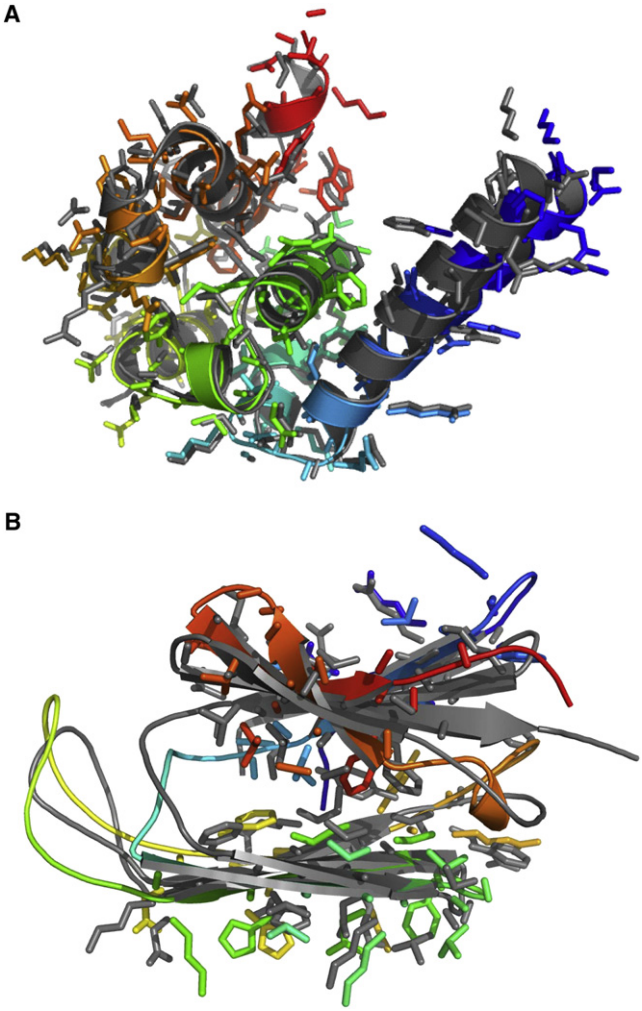


Figure 5. Overlay of Lowest Energy De Novo Models on Crystal Structure or Comparative Model

(A and B) For T4-lysozyme and αA-crystallin, respectively, superimposition of the lowest energy model (rainbow colored) with the crystal structure or comparative model (gray). The backbone is given as a ribbon diagram. Side chains of T4-lysozyme and of the β sandwich of αA-crystallin are shown as stick models without hydrogen atoms.

determine atomic-detail structures of proteins. Furthermore, the structural interpretation of the EPR data for the purpose of de novo modeling does not require a detailed understanding of the position or conformation of the spin label relative to the backbone. A "motion-on-a-cone" model can provide inter-Cβ distance restraints, and the spin-label accessibility can restrain the number of close neighbors when building a structural model.

Only a few specific distance restraints are needed in order to add substantial information and restrict the accessible fold space significantly. Using all 25 distance restraints, 50% T4-lysozyme models have an rmsd smaller than 8 Å compared to less than 5% in their absence. This means that less than one distance restraint per four amino acids was sufficient to focus the de novo structure determination method on sampling the correct fold in more than half of all runs of T4-lysozyme. Only 0.074 distance restraints per residue are needed in order to obtain an equivalent rmsd distribution if the eight restraints between amino acids farthest apart in sequence are used. A significant increase in the quality of models created with distance restraints was also noted for αA-crystallin. By using the distance restraints between amino acids farthest apart in sequence, only 0.136 distance restraints per residue are needed in order to obtain an rmsd distribution similar to that seen when using all restraints. The reduced frequency at which the correct fold is obtained compared with T4-lysozyme can be attributed to the more challenging folding pathway for a β sandwich.

Relative Importance of Accessibility and Spatial Restraints

Distance restraints—even with the comparably large uncertainties resulting from EPR measurements—are more valuable for de novo protein structure determination than solvent accessibility data. Whereas distances reflect specific geometric relationships, the solvent accessibility reflects a convolution of local interactions and rather unspecific interactions with the solvent. The Rosetta energy function already contains knowledge-based terms for these two types of interactions: the amino acid pair potential and the environment potential. The pair potential describes the likelihood of two amino acid types to be spatially close. The environment potential describes the likelihood of an amino acid to be exposed to the solvent or buried in the core. Thus, the EPR accessibility measurements, which reflect the expected buried/exposed distribution, add little information beyond the empirical potentials used in Rosetta.

It should be noted, however, that the importance of the spin label accessibility restraints may be understated by the use of water-soluble proteins as a test case. In general, de novo prediction methods for secondary structure and for sequence-specific residue environment are quite accurate for water-soluble proteins (Rost and Sander, 1993; Jones, 1999; Rost, 2001; Meiler, 2003). In contrast, the apolar character of the membrane core makes a computational distinction between membrane and protein core more difficult. Thus, experimental accessibilities are expected to be critical in defining the secondary structure and topology of initial models for membrane proteins.

Structural Interpretation of EPR Parameters

The fundamental assumption in the "motion-on-a-cone" model is that the spin label cannot point toward the interior of the protein. Thus, possible specific interactions of spin label and protein are disregarded. This is manifest as a bias toward over predicting the frequency with which large (>4Å) d_{SL}–$d_{C\beta}$ values occur. Additionally, the model underestimates the frequency with which low (<4Å) d_{SL}–$d_{C\beta}$ values occur. This is because the spin labels cannot adopt conformations that closely mimic nonspherical arrangements on the surface of proteins such as β strands and α helices. A more precise estimation of the distribution of d_{SL}–$d_{C\beta}$ values might be possible as a comprehensive understanding of spin label rotamers emerges (Langen et al., 2000).

While there is a robust linear relation between the experimental spin-label accessibility, e_{SL}, and the predicted accessibility, $e_{C\beta}$, the applied consensus fit procedure used to obtain the relation has two disadvantages. First, because a comparative model was used in the development of the consensus linear regression

Table 1. T4-Lysozyme EPR Distance Restraints in Comparison with Crystal Structure Distances

AA1–AA2[a]	$d_{C\beta}$ (Å)[b]	d_{SL} (Å)[c]	σ_{SL} (Å)[d]	$d_{SL} + \sigma_{SL} + 2.5$ (Å)[e]	$d_{SL} - \sigma_{SL} - 12.5$ (Å)[f]	Reference
061–135	37.7	47.2	2.2	51.9	32.5	(Borbat et al., 2002)
065–135	34.3	46.3	2.2	51.0	31.6	(Borbat et al., 2002)
061–086	34.5	37.5	2.0	42.0	23.0	(Borbat et al., 2002)
065–086	28.9	37.4	2.7	42.6	22.2	(Borbat et al., 2002)
080–135	26.7	36.8	1.0	40.3	23.3	(Borbat et al., 2002)
061–080	28.7	34.0	2.2	38.7	19.3	(Borbat et al., 2002)
065–080	22.6	26.5	3.8	32.8	10.2	(Borbat et al., 2002)
119–131	13.2	25.0	5.0	32.5	7.5	new data
123–131	14.6	23.0	5.0	30.5	5.5	new data
065–076	16.8	21.4	2.8	26.7	6.1	(Borbat et al., 2002)
116–131	11.1	19.0	10.0	31.5	0.0	new data
119–128	10.4	19.0	4.0	25.5	2.5	new data
140–151	15.5	18.0	9.0	29.5	0.0	new data
089–093	9.8	16.0	3.0	21.5	0.5	new data
086–119	10.0	15.0	3.0	20.5	0.0	new data
120–131	10.5	14.0	3.0	19.5	0.0	new data
127–151	9.6	14.0	2.4	18.9	0.0	new data
140–147	10.1	13.0	7.0	22.5	0.0	new data
131–150	8.7	5.7	0.4	8.6	0.0	new data
127–154	5.9	7.0	3.0	12.5	0.0	new data
131–154	9.5	6.5	4.0	13.0	0.0	new data
134–151	10.7	7.0	0.8	10.3	0.0	new data
131–151	10.4	9.0	8.0	19.5	0.0	new data
088–100	8.9	<6.0	3.0	11.5	0.0	new data
089–096	8.4	<6.0	3.0	11.5	0.0	new data

[a] Indices of spin-labeled amino acids with respect to the crystal structure.
[b] Cβ distance as reported in the crystal structure.
[c] Spin label distance as observed by EPR.
[d] Standard deviation as observed by EPR.
[e] Maximum Cβ atom distance predicted by cone model.
[f] Minimum Cβ atom distance predicted by cone model.

model, using it for de novo folding simulations is somewhat circular. Second, part of the accessibility data for αA-crystallin are influenced by oligomerization and had to be excluded because only a model of the monomeric state of αA-crystallin is computationally feasible. Nevertheless, using this relation in simulations is arguably an acceptable test of the usefulness of such data for de novo protein folding.

Conclusion

De novo protein structure prediction samples as much of the conformational space as possible in order to find the native structure; the number of models reflects the extent of sampling. The increase in the number of high-quality models indicates that the conformational search space has been reduced by the restraints, which allows the remaining space to be sampled more densely. It is remarkable that sparse distance restraints with as large an uncertainty as those obtained from the "motion-on-a-cone" model provide such a drastic improvement in the rmsd distribution of models. The most efficient reduction in conformational search space results from restraints between residues far apart in the primary sequence but close in Euclidean space. A uniform distribution of restraints throughout the protein sequence should be taken into consideration in order to maximize efficiency. Experimental accessibility data add little for structure elucidation of soluble proteins. Less than 0.25 restraints per amino acid are sufficient for protein structure elucidation at atomic detail.

The tendency for side chains to achieve their native rotamer as the protein model backbone approaches its native conformation has been termed "backbone memory" in protein design (Kuhlman and Baker, 2000) and was also observed in very accurate high-resolution de novo protein structure prediction (Bradley et al., 2005b). As a protein model backbone approaches its native conformation, backbone memory allows Rosetta to accurately place side chains into their native rotamer (Kuhlman and Baker, 2000). This is demonstrated with T4-lysozyme; 80% of all rotomers are correct in the lowest energy de novo model, although no side chain conformational restraints are used. For αA-crystallin, definite placement of side chain atoms cannot be conclusively analyzed, since no high-resolution crystal structure is available. The energies between comparative and de novo models are similar; however, no convergence in side chain

Table 2. T4-Lysozyme EPR Solvent Accessibility in Comparison with Crystal Structure

AA[a]	$e_{C\beta}$[b]	e_{SL}[c]	$e_{C\beta}$[d]
086	9	0.36	7.7
093	6	0.41	5.8
094	10	0.28	10.8
096	10	0.23	12.7
097	12	0.17	15.0
100	13	0.15	15.8
101	14	0.13	16.5
102	13	0.14	16.2
103	14	0.16	15.4
104	10	0.28	10.8
105	10	0.33	8.9
106	7	0.31	9.7
108	10	0.32	9.3
109	6	0.45	4.3
111	12	0.23	12.7
113	8	0.40	6.2
114	11	0.30	10.0
115	6	0.36	7.7
116	7	0.52	1.6
117	10	0.29	10.4
118	13	0.21	13.5
119	8	0.35	8.3
120	11	0.23	12.7
121	16	0.10	17.7
122	9	0.36	7.7
123	6	0.39	6.6
124	6	0.42	5.4
125	8	0.40	6.2
126	11	0.32	9.3
127	7	0.44	4.7
128	8	0.42	5.4
129	13	0.14	16.2
130	12	0.30	10.0
131	8	0.37	7.4
132	10	0.35	8.1
133	15	0.23	12.7
134	10	0.30	10.0
135	6	0.38	7.0
136	9	0.31	9.7
137	5	0.41	5.8
138	11	0.29	10.4
139	13	0.33	8.9
140	6	0.37	7.4
141	8	0.25	11.9
142	9	0.27	11.2
143	8	0.39	6.6
144	6	0.36	7.7
145	9	0.24	12.3
146	13	0.27	11.2
147	11	0.31	9.7
148	12	0.28	10.8
149	14	0.21	13.5
150	12	0.29	10.4
151	11	0.37	7.4
153	14	0.24	12.3
154	10	0.34	8.5
155	8	0.25	11.9

[a] Indices of spin-labeled amino acids with respect to the crystal structure.
[b] Number of Cβ atom neighbors in the crystal structure.
[c] Spin label accessibility as observed by EPR (Sompornpisut et al., 2002).
[d] Number of Cβ atom neighbors predicted by the consensus linear regression relation.

conformation was achieved. Therefore, it remains unclear whether the rotamers predicted by the comparative or the de novo model are more accurate.

Overall, this benchmark study sets the stage for application of EPR restraints to protein targets where no structural model is yet available. Advancements in protein-folding algorithms and incorporation of other experimental techniques will further improve the efficiency and accuracy of de novo protein structure determination from sparse experimental data.

EXPERIMENTAL PROCEDURES

Introduction of Site-Directed Spin Labels and EPR Conditions

For the introduction of spin labels, cysteine residues were systematically introduced into the (cysteine-free) T4-lysozyme and αA-crystallin amino acid sequences through single or double point mutations (Koteiche et al., 1998; Borbat et al., 2002; Altenbach et al., 2005). After recombinant protein expression and purification, the mutant was reacted with methanethiosulfonate nitroxide reagent. A total of 25 double mutants and 57 single mutants of T4-lysozyme (Tables 1 and 2) and 36 double mutants and 87 single mutants of αA-crystallin (Tables 3 and 4) resulted in the restraints used for the current analysis. Sample preparation and EPR measurement have been described elsewhere (Mchaourab et al., 1996; Koteiche and Mchaourab, 1999).

EPR Distance Measurements

For T4-lysozyme, 25 distances were measured (Figure 2A and Table 1). Distances derived from double electron-electron resonance (DEER) or DQC experiments (Borbat et al., 2002; Jeschke, 2002; Borbat and Freed, 2007) were distributed in different areas of the molecule with predicted distances larger than 25 Å. They provide geometric restraints on the global fold. CW-EPR was used to measure distances between neighboring helices. For each pair of interacting helices, doubly labeled mutant sets were created by designating a reference spin label in one helix and moving another spin label along the exposed surface of the second helix.

The αA-crystallin EPR data (CW-EPR) (Table 3) consists of 36 distances, including β strand to β strand and β strand to loop distances covering most of the overall topology of the molecule (Koteiche et al., 1998; Koteiche and Mchaourab, 1999) (Figure 2E). For both T4-lysozyme and αA-crystallin, when measurements provided multiple distances, the most contributing distance was used.

For the CW-EPR experiments, dipolar coupling between spin labels was analyzed both in the liquid state and in frozen solutions by using a modification of the deconvolution method (Rabenstein and Shin, 1995). This approach requires two EPR spectra of the double mutant: one in the absence and one in

Table 3. αA-Crystallin EPR Distance Restraints in Comparison with Comparative Model

AA1–AA2[a]	$d_{C\beta}$ (Å)[b]	d_{SL} (Å)[c]	σ_{SL} (Å)[d]	$d_{SL} + 2.5 + \sigma_{SL}$ (Å)[e]	$d_{SL} - \sigma_{SL} - 12.5$ (Å)[f]	Reference
065–072	4.5	6.0	3.0	11.5	0.0	(Koteiche and Mchaourab, 1999)
068–122	13.0	19.5	5.0	27.0	2.0	new data
070–072	6.3	6.6	1.5	10.6	0.0	(Koteiche and Mchaourab, 1999)
070–074	12.8	12.0	4.0	18.5	0.0	(Koteiche and Mchaourab, 1999)
070–142	5.2	6.0	2.0	10.5	0.0	(Koteiche and Mchaourab, 1999)
072–142	5.7	5.5	2.0	10.0	0.0	(Koteiche and Mchaourab, 1999)
073–118	5.6	6.0	3.0	11.5	0.0	new data
074–140	6.0	6.0	2.0	10.5	0.0	(Koteiche and Mchaourab, 1999)
084–099	4.5	7.6	0.6	10.7	0.0	(Koteiche et al., 1998)
084–101	8.6	7.5	8.0	18.0	0.0	(Koteiche et al., 1998)
088–095	4.6	6.0	0.7	9.2	0.0	(Koteiche et al., 1998)
088–097	7.7	7.0	0.8	10.3	0.0	(Koteiche et al., 1998)
090–093	6.8	6.9	1.1	10.5	0.0	(Koteiche et al., 1998)
090–099	18.7	18.5	4.0	25.0	2.0	(Koteiche et al., 1998)
090–126	11.1	12.0	4.0	18.5	0.0	(Koteiche and Mchaourab, 1999)
092–123	10.4	12.0	6.0	20.5	0.0	(Koteiche and Mchaourab, 1999)
095–117	5.0	7.6	1.0	11.1	0.0	(Koteiche et al., 1998)
101–103	5.4	6.3	0.8	9.6	0.0	(Koteiche and Mchaourab, 1999)
101–111	11.9	7.6	1.0	11.1	0.0	(Koteiche et al., 1998)
103–105	7.5	9.5	8.0	20.0	0.0	(Koteiche and Mchaourab, 1999)
105–111	11.6	18.0	9.0	29.5	0.0	(Koteiche and Mchaourab, 1999)
119–123	10.5	13.0	6.0	21.5	0.0	(Koteiche and Mchaourab, 1999)
119–125	17.1	23.0	3.0	28.5	7.5	(Koteiche and Mchaourab, 1999)
119–127	21.2	24.0	6.0	32.5	5.5	(Koteiche and Mchaourab, 1999)
119–130	19.3	23.0	5.0	30.5	5.5	(Koteiche and Mchaourab, 1999)
119–132	21.9	23.0	4.0	29.5	6.5	(Koteiche and Mchaourab, 1999)
125–144	4.0	7.0	4.0	13.5	0.0	(Koteiche and Mchaourab, 1999)
127–142	13.1	20.0	4.0	26.5	3.5	new data
128–144	5.0	4.0	4.0	10.5	0.0	(Koteiche and Mchaourab, 1999)
128–146	10.4	13.0	4.0	19.5	0.0	(Koteiche and Mchaourab, 1999)
130–144	9.7	6.0	1.0	9.5	0.0	(Koteiche and Mchaourab, 1999)
131–133	8.0	7.0	1.0	10.5	0.0	(Koteiche and Mchaourab, 1999)
131–134	12.3	17.0	4.0	23.5	0.5	(Koteiche and Mchaourab, 1999)
132–135	11.5	16.0	5.0	23.5	0.0	(Koteiche and Mchaourab, 1999)
132–142	8.8	6.0	1.0	9.5	0.0	(Koteiche and Mchaourab, 1999)
132–144	16.1	9.0	5.0	16.5	0.0	(Koteiche and Mchaourab, 1999)

[a] Indices of spin-labeled amino acids with respect to the protein sequence.
[b] Cβ atom distance in comparative model.
[c] Spin-label distance as observed by EPR.
[d] Standard deviation as observed by EPR.
[e] Maximum Cβ atom distance predicted by cone model.
[f] Minimum Cβ atom distance predicted by cone model.

the presence of the dipolar interaction (Figures 6A and 6B, respectively). The former is obtained from the digital sum of the spectra of each single mutant. A Levenberg-Marquardt algorithm was used to minimize the difference between the experimental EPR spectrum of the double mutant and the spectrum obtained from the convolution of a broadening function with the EPR spectrum of the corresponding sum of single mutants. The broadening function consisted of either one or two Gaussian distributions for the distance between spin labels (Figure 6C). The relatively wide distance distributions obtained is consistent with a highly dynamic motional state of the spin label obtained at the predominantly exposed sites. The results obtained in the solid and liquid states are in agreement both in terms of the average distance and the overall distribution as previously reported (Altenbach et al., 2001).

DEER measurements were performed on a Bruker 580 pulsed EPR spectrometer, by using a standard four-pulse protocol (Jeschke, 2002). Experiments were performed at 80 K by using Ficoll as cryoprotectant. Sample concentration was 200 μM and sample volume 20 μl. DEER signals were analyzed by the Tikhonov regularization (Chiang et al., 2005) to determine average distances and distributions in distance, P(r), as illustrated in Figures 6D–6F.

Table 4. αA-Crystallin EPR Solvent Accessibility in Comparison Comparative Model

AA[a]	$e_{C\beta}$[b]	e_{SL}[c]	$e_{C\beta}$[d]	AA[a]	$e_{C\beta}$[b]	e_{SL}[c]	$e_{C\beta}$[d]	AA[a]	$e_{C\beta}$[b]	e_{SL}[c]	$e_{C\beta}$[d]
060	5	0.01	10.9	089	11	0.18	9.0	118	12	0.01	10.9
061	5	0.04	10.5	090	9	0.26	8.2	119	7	0.26	8.1
062	11	0.00	11.0	091	4	0.41	6.5	120	11	0.01	10.9
063	6	0.02	10.8	092	7	0.35	7.1	121	5	0.19	8.9
064	10	0.05	10.5	093	12	0.27	8.1	122	8	0.43	6.2
065	9	0.07	10.2	094	17	0.06	10.4	123	7	0.30	7.7
066	10	0.05	10.4	095	13	0.23	8.4	124	16	0.08	10.1
067	9	0.04	10.6	096	15	0.14	9.5	125	9	0.34	7.3
068	5	0.19	8.9	097	12	0.21	8.6	126	7	0.20	8.8
069	6	0.21	8.7	098	12	0.01	10.9	127	5	0.38	6.8
070	10	0.16	9.3	099	9	0.13	9.6	128	9	0.32	7.5
071	16	0.05	10.5	100	16	0.01	10.9	129	14	0.21	8.7
072	12	0.06	10.3	101	9	0.19	9.0	130	9	0.18	9.0
073	15	0.07	10.2	102	10	0.14	9.5	131	9	0.16	9.3
074	13	0.05	10.5	103	6	0.17	9.2	132	8	0.22	8.6
075	12	0.02	10.8	104	6	0.15	9.3	133	11	0.12	9.7
076	11	0.02	10.8	105	12	0.19	8.9	134	7	0.07	10.3
077	12	0.01	10.9	106	5	0.10	9.9	135	4	0.19	8.9
078	9	0.02	10.8	107	7	0.08	10.1	136	5	0.08	10.1
079	11	0.03	10.6	108	4	0.07	10.2	137	12	0.08	10.1
080	14	0.01	10.9	109	10	0.17	9.1	138	10	0.02	10.8
081	10	0.09	10.0	110	3	0.04	10.6	140	13	0.05	10.5
082	8	0.27	8.0	111	7	0.15	9.3	141	15	0.01	10.8
083	5	0.29	7.8	112	9	0.00	11.0	142	11	0.11	9.7
084	10	0.25	8.2	113	9	0.08	10.1	143	14	0.03	10.7
085	16	0.05	10.4	114	11	0.01	10.9	144	11	0.20	8.8
086	7	0.22	8.5	115	9	0.15	9.4	145	12	0.10	9.8
087	12	0.14	9.5	116	11	0.00	11.0	146	7	0.41	6.5
088	8	0.35	7.2	117	10	0.16	9.3	147	2	0.39	6.7

[a] Indices of spin-labeled amino acids with respect to the protein sequence.
[b] Number of Cβ atom neighbors in the crystal structure.
[c] Spin-label accessibility as observed by EPR (Koteiche et al., 1998; Koteiche and Mchaourab, 1999).
[d] Number of Cβ atom neighbors predicted from the consensus linear regression relation.

EPR Accessibility Measurements

For T4-lysozyme, e_{SL} of 57 spin labels was measured (Sompornpisut et al., 2002) (Table 2). For αA-crystallin, e_{SL} of 87 spin labels was measured (Table 4). Accessibility is assessed by measuring the Heisenberg exchange rate between the nitroxide spin label and either molecular oxygen, in the case of T4-lysozyme, or NiEDDA, in the case of αA-crystallin. For the latter, power saturation measurements were carried out under nitrogen and in the presence of 3 mM NiEDDA (Farahbakhsh et al., 1992). e_{SL} was calculated as previously described (Farahbakhsh et al., 1992; Altenbach et al., 2005).

αA-Crystallin Comparative Model Preparation

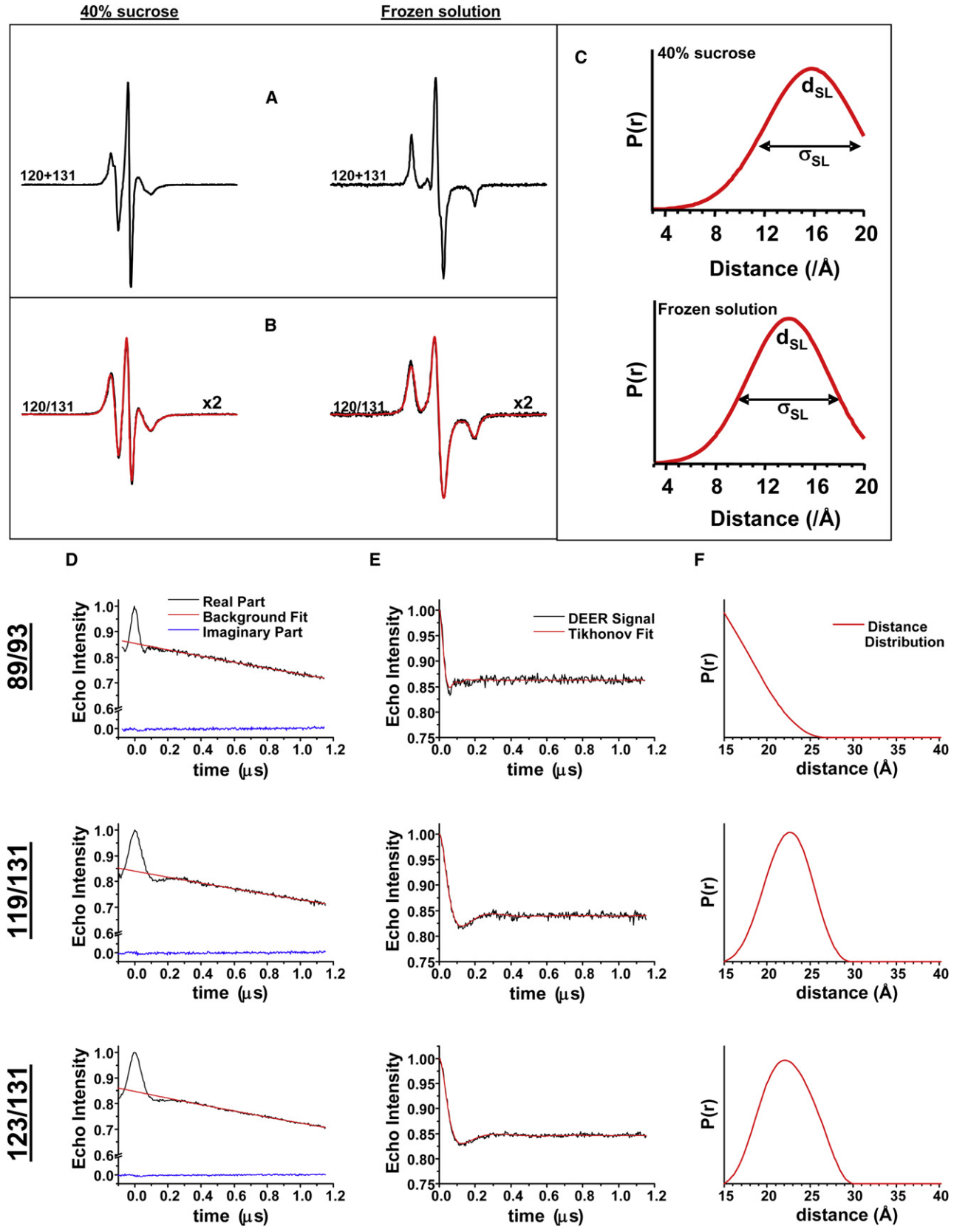
The 88 amino acid C-terminal domain of αA-crystallin was submitted to the BioInfo metaserver (Fischer, 2000; Kelley et al., 2000; Shi et al., 2001; Ginalski et al., 2003, 2004; Karplus et al., 2003; McGuffin and Jones, 2003; Rost et al., 2004; Bryson et al., 2005; Soding et al., 2005; Finn et al., 2006). The server identified three heat-shock proteins (PDB codes: 1gmeA [van Montfort et al., 2001], 1shsA [Kim et al., 1998], and 2bolA [Stamler et al., 2005]) as possible templates with a 3D-Jury score of over 60, where a score over 40 indicates a ~90% chance that the identified proteins have the same fold as the submitted amino acid sequence (Ginalski et al., 2003). Obtaining the correct fold is the most important and difficult aspect of de novo protein folding, so having such a large likelihood that the identified proteins have the same fold as αA-crystallin is essential to ensuring the comparative model provides an adequate benchmark with which the fold of de novo models can be compared. There was approximately 20% sequence homology between the αA-crystallin amino acid sequence and the three template sequences.

A multiple sequence alignment was performed for the αA-crystallin amino acids with the template proteins. The aligned αA-crystallin amino acids were then mapped onto the template proteins' atomic coordinates, and Rosetta was used to reconstruct the loop regions of αA-crystallin while holding the β sandwich region fixed. Afterwards, Rosetta was used to perform a high-resolution refinement of the αA-crystallin comparative models.

The lowest energy comparative model was used to compare against the de novo Rosetta models. This model was based on the PDB structure 1gmeA (van Montfort et al., 2001). The model is 2.3 Å rmsd to a previously published (Koteiche and Mchaourab, 1999) comparative model based on a different template protein (Hsp16.5) but achieves a lower Rosetta energy after both models are refined at high resolution.

The Spin Label "Motion-on-a-Cone" Model

A simple cone model for the relative position of the spin label with respect to the Cβ of an amino acid was developed by using three assumptions. First,


40% sucrose
Frozen solution
A
120+131
120+131
B
120/131
x2
120/131
x2
C
40% sucrose
P(r)
d_{SL}
σ_{SL}
4 8 12 16 20
Distance (/Å)
Frozen solution
P(r)
d_{SL}
σ_{SL}
4 8 12 16 20
Distance (/Å)
D
E
F
89/93
119/131
123/131
Real Part
Background Fit
Imaginary Part
DEER Signal
Tikhonov Fit
Distance Distribution
Echo Intensity
time (µs)
P(r)
distance (Å)

the spin label's motion follows the perimeter of the base of a right circular cone with an opening angle of 90° whose vertex is the Cβ. The average position of this motion is $\sqrt{2}/2$ of length of the extended chain (8.5 Å) (Borbat et al., 2002), which gives a maximal effective distance between spin label and Cβ of 6Å (Figures 1A and 1B). Second, the protein is globular. Third, the angle defined by the center of the protein, the Cβ, and the spin label is between 120° and 180°, and, therefore, spin labels point away from the interior of the protein (Figures 1B and 1C). The "motion-on-a-cone" model can also be adapted for spin labels of different linking arms or ring substituents that restrict the amplitude of its motion.

Using the "Motion-on-a-Cone" Model to Translate EPR Spin-Label Distances into Structural Restraints

The difference between d_{SL} and $d_{C\beta}$ given by the model described above was analyzed with the software package Mathematica (Wolfram Research, Inc., 2005): (1) an ellipsoid with the main radii 10 Å $\leq r_x \leq r_y \leq r_z \leq$ 20 Å was created with otherwise randomly chosen r_x, r_y, and r_z. Its center is C. (2) Two points, $C\beta_i$ and $C\beta_j$, on the surface of this ellipsoid were selected by randomly choosing the polar coordinates $\varphi_{i,j}$ and $\psi_{i,j}$. From these points, $d_{C\beta}$ is computed as the Euclidean distance. (3) Two numbers between 120° and 180° for the angle C–$C\beta_i$–SL_i and C–$C\beta_j$–SL_j are chosen randomly, and the position of the spin labels, SL_i and SL_j, is computed. From these points, d_{SL} is computed as the Euclidean distance. (4) The difference d_{SL}–$d_{C\beta}$ is computed. (5) Steps 1–4 are repeated 10,000 times, and the values d_{SL}–$d_{C\beta}$ are plotted as a histogram (Figure 1D).

This analysis of the difference between d_{SL} and $d_{C\beta}$ showed that (d_{SL} + 2.5 Å) $\geq d_{C\beta} \geq$ (d_{SL} − 12.5 Å) (Figure 1D). σ_{SL}, which is the experimentally determined standard deviation in d_{SL}, is a measure of the magnitude of the spin label's motion or its static distribution relative to the Cβ. Since an increased magnitude of motion increases the ambiguity of the derived $d_{C\beta}$, σ_{SL} is added as an additional allowance to the restraint, which gives ($d_{SL} + \sigma_{SL}$ + 2.5 Å) $\geq d_{C\beta} \geq$ ($d_{SL} - \sigma_{SL}$ − 12.5 Å).

Development of a Model to Translate EPR Spin-Label Solvent Accessibility into Structural Restraints

Obtaining a structural restraint from EPR spin-label solvent accessibility is accomplished by building a consensus linear regression relation of e_{SL} to $e_{C\beta}$ in a three-step procedure: (1) using the crystal structure of T4-lysozyme, $e_{C\beta}$ was computed for all residues with known e_{SL} (Table 2). A linear regression was fit to a plot of $e_{C\beta}$ of a residue versus the corresponding e_{SL}, yielding the relation $e_{C\beta} = (0.71 - e_{SL}) \bullet 24.23$ with a correlation coefficient of −0.80. Using this relation to calculate the number of neighbors for a residue gives $e_{C\beta}^{fit}$ for that residue. The standard deviation ($\sigma_{C\beta}$) of $e_{C\beta}$ from $e_{C\beta}^{fit}$ was calculated to be 1.65. (2) For those residues in αA-crystallin that have an experimentally determined accessibility (Table 4), $e_{C\beta}$ was determined by using the comparative model of αA-crystallin. In addition, the equation from (1) above was used to calculate the number of neighbors, $e_{C\beta}^{fit}$, for each residue. Amino acids were excluded from a linear fitting of $e_{C\beta}$ versus e_{SL} when $|e_{C\beta} - e_{C\beta}^{fit}| > 2 \bullet \sigma_{C\beta}$. This procedure was necessary in order to exclude amino acids in αA-crystallin that show reduced experimental accessibility due to intermolecular contacts with other αA-crystallin units in the oligomeric protein (Figures 2F and 2H). Fitting a linear regression to the remaining data gives a relation similar to that seen for T4-lysozyme: $e_{C\beta} = (0.72 - e_{SL}) \bullet 21.63$ with a correlation coefficient of −0.87. (3) Combining the data for both proteins in a single consensus linear regression model yields $e_{C\beta} = (0.76 - e_{SL}) \bullet 20.87$ with a correlation coefficient of −0.83 (Figures 2D and 2H).

Implementation of Structural Restraints for De Novo Structure Determination

The distance restraints are used as an additional penalty in the energy function of Rosetta. This penalty is zero if $d_{C\beta}$ lay within the range predicted from d_{SL}. As $d_{C\beta}$ ventures outside this range, a quadratic penalty function is applied. The detailed implementation of this penalty function and its use to guide the folding simulation is described in detail in the respective RosettaNMR publications (Bowers et al., 2000; Rohl and Baker, 2002).

Similarly, $(e_{C\beta} - e_{C\beta}^{\text{consensus model}})^2$ is used as a quadratic penalty function for the accessibility data. The relative weight of this penalty function was optimized by a series of experiments varying its weight in a wide range of two orders of magnitude.

Rosetta Folding Simulations

De novo model generation using Rosetta was performed in four steps: step 1, the protein is folded by using RosettaNMR with EPR distance restraints to guide the simulation (Bowers et al., 2000; Rohl and Baker, 2002). In this step, amino acid side chains are embraced in a single superatom—a "centroid" (Simons et al., 1997). Step 2, choosing the models with lowest energy and best agreement with experimental restraints prunes the large number of ~500,000 models from step 1 to ~10,000. Step 3, models obtained from step 2 are refined to high resolution. After replacing side chain centroids with full-atom side chain representations from a backbone-dependent rotamer library (Dunbrack and Karplus, 1993), an iterative protocol of all-atom gradient minimization and side chain repacking is repeated eight times. The details of the protocol are published elsewhere (Bradley et al., 2005b; Misura and Baker, 2005). No restraints were used during these refinement simulations in order to fully leverage the discriminative power of the Rosetta energy function (Kuhlman and Baker, 2000; Bradley et al., 2005a, 2005b; Misura and Baker, 2005; Misura et al., 2006). Step 4, models from step 3 are again filtered for good agreement with the experimental restraints.

Specific standard Rosetta procedures were used that are described in detail elsewhere (Simons et al., 1997, 1999; Bowers et al., 2000; Bonneau et al., 2001a; Rohl and Baker, 2002; Meiler et al., 2003; Rohl et al., 2004; Bradley et al., 2005a). Secondary structure predictions were obtained from the primary sequence of the C-terminal 107 amino acids of T4-lysozyme and the C-terminal 88 amino acid primary sequence of αA-crystallin by using Jufo (Meiler et al., 2001; Meiler and Baker, 2003), PsiPred (Jones, 1999), and Sam (Karplus et al., 1997). All T4-lysozyme and αA-crystallin homologs were excluded from the protein database prior to modeling in order to simulate structure elucidation of an unknown protein fold as closely as possible.

Models were obtained in 500,000 independent simulations on a cluster in Vanderbilt University's Advanced Computing Center for Research & Education (ACCRE) by using up to 300 parallel 2.2 GHz JS20 IBM PowerPC processors. The average time to complete a model was approximately 100 s for T4-lysozyme and 180 s for αA-crystallin. The high-resolution refinement protocol requires about 500 s of computation time per model.

ACKNOWLEDGMENTS

The authors would like to acknowledge Eduardo Perozo for collection of the T4-lysozyme accessibility data. This work was conducted in part by using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. J.M. is supported by grant R01-GM080403 from the National Institute of General Medical Sciences. N.A. is supported by grant National Institutes of Health T32 GM08320, the Molecular Biophysics Training Grant at Vanderbilt University. M.B. and H.S.M. were supported by grant R01-EY12683.

Received: July 16, 2007
Revised: October 22, 2007
Accepted: November 25, 2007
Published: February 12, 2008

Figure 6. Distance Measurements at Room Temperature and in the Solid State between Spin Labels Using EPR

(A) Representative reference EPR in the absence of dipolar coupling obtained from the digital sum of the corresponding single mutant spectra.
(B) Spectra of double mutants along with the nonlinear least-squares fit obtained by the convolution method as described in the Experimental Procedures section.
(C) Distance distributions obtained from CW-EPR spectra.
(D) Distance measurements by DEER for representative double mutants.
(E and F) Raw DEER signals were background corrected and then fit by using Tikhonov regularization to obtain (F) average distances and distance distributions.

REFERENCES

Altenbach, C., Oh, K.J., Trabanino, R.J., Hideg, K., and Hubbell, W.L. (2001). Estimation of inter-residue distances in spin labeled proteins at physiological temperatures: experimental strategies and practical limitations. Biochemistry *40*, 15471–15482.

Altenbach, C., Froncisz, W., Hemker, R., McHaourab, H., and Hubbell, W.L. (2005). Accessibility of nitroxide side chains: absolute Heisenberg exchange rates from power saturation EPR. Biophys. J. *89*, 2103–2112.

Baker, D. (2000). A surprising simplicity to protein folding. Nature *405*, 39–42.

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., et al. (2002). The Protein Data Bank. Acta Crystallogr. D Biol. Crystallogr. *58*, 899–907.

Bonneau, R., Strauss, C.E., and Baker. (2001a). Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins *43*, 1–11.

Bonneau, R., Tsai, J., Ruczinski, I., Chivian, D., Rohl, C., Strauss, C.E., and Baker, D. (2001b). Rosetta in CASP4: progress in ab initio protein structure prediction. Proteins (*Suppl 5*), 119–126.

Bonneau, R., Ruczinski, I., Tsai, J., and Baker, D. (2002). Contact order and ab initio protein structure prediction. Protein Sci. *11*, 1937–1944.

Borbat, P.P., and Freed, J. (2007). Measuring distances by pulsed dipolar ESR spectroscopy: spin-labeled histidine kinases. Methods Enzymol. *423*, 52–116.

Borbat, P.P., McHaourab, H.S., and Freed, J.H. (2002). Protein structure determination using long-distance constraints from double-quantum coherence ESR: study of T4 lysozyme. J. Am. Chem. Soc. *124*, 5304–5314.

Bowers, P.M., Strauss, C.E.M., and Baker, D. (2000). Denovo protein structure determination using sparse NMR data. J. Biomol. NMR *18*, 311–318.

Bradley, P., Chivian, D., Meiler, J., Misura, K.M., Rohl, C.A., Schief, W.R., Wedemeyer, W.J., Schueler-Furman, O., Murphy, P., Schonbrun, J., et al. (2003). Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. Proteins *53* (*Suppl 6*), 457–468.

Bradley, P., Malmstrom, L., Qian, B., Schonbrun, J., Chvian, D., Kim, D.E., Meiler, J., Misura, K.M., and Baker, D. (2005a). Free modeling with Rosetta in CASP6. Proteins *61* (*Suppl 7*), 128–134.

Bradley, P., Misura, K.M., and Baker, D. (2005b). Toward high-resolution de novo structure prediction for small proteins. Science *309*, 1868–1871.

Brown, L.J., Sale, K.L., Hills, R., Rouviere, C., Song, L., Zhang, X., and Fajer, P.G. (2002). Structure of the inhibitory region of troponin by site directed spin labeling electron paramagnetic resonance. Proc. Natl. Acad. Sci. USA *99*, 12765–12770.

Bryson, K., McGuffin, L.J., Marsden, R.L., Ward, J.J., Sodhi, J.S., and Jones, D.T. (2005). Protein structure prediction servers at University College London. Nucleic Acids Res. *33*, W36–W38.

Chiang, Y.W., Borbat, P.P., and Freed, J.H. (2005). The determination of pair distance distributions by pulsed ESR using Tikhonov regularization. J. Magn. Reson. *172*, 279–295.

Dong, J., Yang, G., and McHaourab, H.S. (2005). Structural basis of energy transduction in the transport cycle of MsbA. Science *308*, 1023–1028.

Dunbrack, R.L., Jr. (2002). Rotamer libraries in the 21st century. Curr. Opin. Struct. Biol. *12*, 431–440.

Dunbrack, R.L., Jr., and Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. J. Mol. Biol. *230*, 543–574.

Fanucci, G.E., and Cafiso, D.S. (2006). Recent advances and applications of site-directed spin labeling. Curr. Opin. Struct. Biol. *16*, 644–653.

Farahbakhsh, Z.T., Altenbach, C., and Hubbell, W.L. (1992). Spin labeled cysteines as sensors for protein-lipid interaction and conformation in rhodopsin. Photochem. Photobiol. *56*, 1019–1033.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., et al. (2006). Pfam: clans, web tools and services. Nucleic Acids Res. *34*(Database issue), D247–D251.

Fischer, D. (2000). Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac. Symp. Biocomput. *2000*, 119–130.

Ginalski, K., Elofsson, A., Fischer, D., and Rychlewski, L. (2003). 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics *19*, 1015–1018.

Ginalski, K., von Grotthuss, M., Grishin, N.V., and Rychlewski, L. (2004). Detecting distant homology with Meta-BASIC. Nucleic Acids Res *32*, W576–W581.

Haley, D.A., Bova, M.P., Huang, Q.L., Mchaourab, H.S., and Stewart, P.L. (2000). Small heat-shock protein structures reveal a continuum from symmetric to variable assemblies. J. Mol. Biol. *298*, 261–272.

Harrison, S.C. (2004). Whither structural biology? Nat. Struct. Mol. Biol. *11*, 12–15.

Horwitz, J. (1992). Alpha-crystallin can function as a molecular chaperone. Proc. Natl. Acad. Sci. USA *89*, 10449–10453.

Horwitz, J. (1993). Proctor Lecture. The function of alpha-crystallin. Invest. Ophthalmol. Vis. Sci. *34*, 10–22.

Hubbell, W.L., Mchaourab, H.S., Altenbach, C., and Lietzow, M.A. (1996). Watching proteins move using site-directed spin labeling. Structure *4*, 779–783.

Jeschke, G. (2002). Distance measurements in the nanometer range by pulse EPR. ChemPhysChem *3*, 927–932.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. *292*, 195–202.

Karplus, K., Sjolander, K., Barrett, C., Cline, M., Haussler, D., Hughey, R., Holm, L., and Sander, C. (1997). Predicting protein structure using hidden Markov models. Proteins (*Suppl 1*), 134–139.

Karplus, K., Karchin, R., Draper, J., Casper, J., Mandel-Gutfreund, Y., Diekhans, M., and Hughey, R. (2003). Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins *53* (*Suppl 6*), 491–496.

Kelley, L.A., MacCallum, R.M., and Sternberg, M.J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. J. Mol. Biol. *299*, 499–520.

Kim, K.K., Kim, R., and Kim, S.H. (1998). Crystal structure of a small heat-shock protein. Nature *394*, 595–599.

Koteiche, H.A., and Mchaourab, H.S. (1999). Folding pattern of the α-crystallin domain in αA-crystallin determined by site-directed spin labeling. J. Mol. Biol. *294*, 561–577.

Koteiche, H.A., Berengian, A.R., and Mchaourab, H.S. (1998). Identification of protein folding patterns using site-directed spin labeling. Structural characterization of a β-sheet and putative substrate binding regions in the conserved domain of α A-crystallin. Biochemistry *37*, 12681–12688.

Kuhlman, B., and Baker, D. (2000). Native protein sequences are close to optimal for their structures. Proc. Natl. Acad. Sci. USA *97*, 10383–10388.

Langen, R., Oh, K.J., Cascio, D., and Hubbell, W.L. (2000). Crystal structures of spin labeled T4 lysozyme mutants: implications for the interpretation of EPR spectra in terms of structure. Biochemistry *39*, 8396–8405.

Lesley, S.A., Kuhn, P., Godzik, A., Deacon, A.M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H.E., McMullan, D., Shin, T., et al. (2002). Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. Proc. Natl. Acad. Sci. USA *99*, 11664–11669.

Liu, Y.S., Sompornpisut, P., and Perozo, E. (2001). Structure of the KcsA channel intracellular gate in the open state. Nat. Struct. Biol. *8*, 883–887.

McGuffin, L.J., and Jones, D.T. (2003). Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics *19*, 874–881.

Mchaourab, H.S., Lietzow, M.A., Hideg, K., and Hubbell, W.L. (1996). Motion of spin-labeled side chains in T4 lysozyme. Correlation with protein structure and dynamics. Biochemistry *35*, 7692–7704.

Mchaourab, H.S., Berengian, A.R., and Koteiche, H.A. (1997). Site-directed spin-labeling study of the structure and subunit interactions along a conserved sequence in the α-crystallin domain of heat-shock protein 27. Evidence of a conserved subunit interface. Biochemistry *36*, 14627–14634.

Meiler, J. (2003). JUFO3D: coupled prediction of protein secondary and tertiary structure (server). (http://www.meilerlab.org/).

Meiler, J., and Baker, D. (2003). Coupled prediction of protein secondary and tertiary structure. Proc. Natl. Acad. Sci. USA *100*, 12105–12110.

Meiler, J., and Baker, D. (2005). The fumarate sensor DcuS: progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy. J. Magn. Reson. *173*, 310–316.

Meiler, J., Müller, M., Zeidler, A., and Schmaschke, F. (2001). Generation and evaluation of dimension reduced amino acid parameter representations by artificial neural networks. J. Mol. Model. *7*, 360–369.

Meiler, J., Bradley, P., Chivian, D., Misura, K.M., Rohl, C., Schief, B., Wedemeyer, W.J., and Baker, D. (2003). ROSETTA in CASP5: Progress in De Novo Protein Structure Prediction (poster). *Molecular Modeling Workshop* (Bavaria, Germany: Erlangen).

Misura, K.M., and Baker, D. (2005). Progress and challenges in high-resolution refinement of protein structure models. Proteins *59*, 15–29.

Misura, K.M., Chivian, D., Rohl, C.A., Kim, D.E., and Baker, D. (2006). Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc. Natl. Acad. Sci. USA *103*, 5361–5366.

Nederveen, A.J., Doreleijers, J.F., Vranken, W., Miller, Z., Spronk, C.A., Nabuurs, S.B., Guntert, P., Livny, M., Markley, J.L., Nilges, M., et al. (2005). RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. Proteins *59*, 662–672.

Perozo, E., Cortes, D.M., and Cuello, L.G. (1999). Structural rearrangements underlying K+-channel activation gating. Science *285*, 73–78.

Rabenstein, M.D., and Shin, Y.K. (1995). Determination of the distance between two spin labels attached to a macromolecule. Proc. Natl. Acad. Sci. USA *92*, 8239–8243.

Rohl, C.A., and Baker, D. (2002). De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. J. Am. Chem. Soc. *124*, 2723–2729.

Rohl, C.A., Strauss, C.E., Misura, K.M., and Baker, D. (2004). Protein structure prediction using Rosetta. Methods Enzymol. *383*, 66–93.

Rost, B. (2001). Review: protein secondary structure prediction continues to rise. J. Struct. Biol. *134*, 204–218.

Rost, B., and Sander, C. (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., Yachdav, G., and Liu, J. (2004). The PredictProtein server. Nucleic Acids Res. *32*, W321–W326.

Sale, K., Song, L., Liu, Y.S., Perozo, E., and Fajer, P. (2005). Explicit treatment of spin labels in modeling of distance constraints from dipolar EPR and DEER. J. Am. Chem. Soc. *127*, 9334–9335.

Sali, A. (1998). 100,000 protein structures for the biologist. Nat. Struct. Biol. *5*, 1029–1032.

Shi, J., Blundell, T.L., and Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol. *310*, 243–257.

Simons, K.T., Kooperberg, C., Huang, E., and Baker, D. (1997). Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J. Mol. Biol. *268*, 209–225.

Simons, K.T., Ruczinski, I., Kooperberg, C., Fox, B.A., Bystroff, C., and Baker, D. (1999). Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins Struct. Funct. Genet. *34*, 82–95.

Soding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. *33*, W244–W248.

Sompornpisut, P., Mchaourab, H., and Perozo, E. (2002). http://meeting.biophysj.org/cgi/reprint/82/1/474/a.pdf.

Stamler, R., Kappe, G., Boelens, W., and Slingsby, C. (2005). Wrapping the α-crystallin domain fold in a chaperone assembly. J. Mol. Biol. *353*, 68–79.

Stevens, R.C., Yokoyama, S., and Wilson, I.A. (2001). Global efforts in structural genomics. Science *294*, 89–92.

Tusnady, G.E., Dosztanyi, Z., and Simon, I. (2004). Transmembrane proteins in the Protein Data Bank: identification and classification. Bioinformatics *20*, 2964–2972.

van Montfort, R.L., Basha, E., Friedrich, K.L., Slingsby, C., and Vierling, E. (2001). Crystal structure and assembly of a eukaryotic small heat shock protein. Nat. Struct. Biol. *8*, 1025–1030.

Weaver, L.H., and Matthews, B.W. (1987). Structure of bacteriophage T4 lysozyme refined at 1.7 A resolution. J. Mol. Biol. *193*, 189–199.

Westbrook, J., Feng, Z., Chen, L., Yang, H., and Berman, H.M. (2003). The Protein Data Bank and structural genomics. Nucleic Acids Res. *31*, 489–491.

Wolfram Research, Inc. (2005). Mathematica (computer program). Champaign, IL.

Wüthrich, K. (1986). NMR of Proteins and Nucleic Acids (1H-NMR Shifts of Amino Acids) (New York: John Wiley & Sons).