Journal of Structural Biology 173 (2011) 506–514

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

RosettaEPR: An integrated tool for protein structure determination from sparse EPR data

Stephanie J. Hirst [a,b], Nathan Alexander [a,c], Hassane S. Mchaourab [a,d], Jens Meiler [a,c,*]

[a] Center for Structural Biology, Vanderbilt University, Nashville, TN 37212, USA
[b] Chemical and Physical Biology Program, Vanderbilt University, Nashville, TN 37212, USA
[c] Department of Chemistry, Vanderbilt University, Nashville, TN 37212, USA
[d] Department of Molecular Physiology and Biophysics, Vanderbilt University, Nashville, TN 37212, USA

ARTICLE INFO

Article history:
Available online 26 October 2010

Keywords:
De novo protein structure determination
Rosetta
Site-directed spin labeling
Electron paramagnetic resonance
SDSL-EPR

ABSTRACT

Site-directed spin labeling electron paramagnetic resonance (SDSL-EPR) is often used for the structural characterization of proteins that elude other techniques, such as X-ray crystallography and nuclear magnetic resonance (NMR). However, high-resolution structures are difficult to obtain due to uncertainty in the spin label location and sparseness of experimental data. Here, we introduce RosettaEPR, which has been designed to improve *de novo* high-resolution protein structure prediction using sparse SDSL-EPR distance data. The "motion-on-a-cone" spin label model is converted into a knowledge-based potential, which was implemented as a scoring term in Rosetta. RosettaEPR increased the fractions of correctly folded models ($RMSD_{C_\alpha}$ < 7.5 Å) and models accurate at medium resolution ($RMSD_{C_\alpha}$ < 3.5 Å) by 25%. The correlation of score and model quality increased from 0.42 when using no restraints to 0.51 when using bounded restraints and again to 0.62 when using RosettaEPR. This allowed for the selection of accurate models by score. After full-atom refinement, RosettaEPR yielded a 1.7 Å model of T4-lysozyme, thus indicating that atomic detail models can be achieved by combining sparse EPR data with Rosetta. While these results indicate RosettaEPR's potential utility in high-resolution protein structure prediction, they are based on a single example. In order to affirm the method's general performance, it must be tested on a larger and more versatile dataset of proteins.

1. Introduction

1.1. Protein modeling with Rosetta can serve as an alternative means of structure elucidation

The vast majority of proteins in the Protein Data Bank (PDB) have been determined by X-ray crystallography or nuclear magnetic resonance (NMR) (Berman et al., 2002). However, a large number of biomedically relevant proteins continue to evade structural elucidation by these techniques due to membrane environment (Tusnady et al., 2004), high flexibility (Haley et al., 2000), and size (Harrison, 2004). Alternative techniques, such as computational structure prediction methods, can be employed in order to define the structure of such proteins. The usual experimental bottlenecks, such as obtaining highly pure, concentrated samples of protein, are thereby avoided. Rosetta routinely folds soluble proteins of less than 150 amino acids correctly (Bonneau et al., 2002). It is generally among the top performers in the Critical Assessment of protein Structure Prediction, or CASP, experiments (Bonneau et al., 2001; Bradley et al., 2003, 2005; Das et al., 2007; Raman et al., 2009). In addition, Rosetta's ability to obtain the correct fold of membrane proteins of various sizes and topologies has been demonstrated (Yarov-Yarovoy et al., 2006; Barth et al., 2007, 2009). More recently, Das et al., introduced Rosetta FOLD-AND-DOCK, which allows for the *de novo* structure prediction of homomeric proteins (Das et al., 2009).

Rosetta's sampling and scoring capabilities for protein folding have been reviewed extensively elsewhere (Simons et al., 1997; Rohl and Baker, 2002; Bradley et al., 2005; Kaufmann et al., 2010). Briefly, the Rosetta *de novo* protein structure prediction algorithm is divided into two steps: low-resolution protein folding to obtain the overall topology and high-resolution refinement of the backbone and sidechains. Metropolis Monte Carlo peptide

Abbreviations: EPR, electron paramagnetic resonance; NMR, nuclear magnetic resonance; PDB, Protein Data Bank; SL, spin label; SDSL, site-directed spin labeling; MTS, methanethisulfonate; Cα, α-carbon on the protein backbone; Cβ, β-carbon on the amino acid sidechain; d_{C_β}, distance between two β-carbons on the protein in angstroms (Å); d_{SL}, distance between two MTS spin labels on the protein in angstroms (Å); RMSD, root mean square distance in angstroms (Å).

* Corresponding author. Address: Vanderbilt University, 7330 Stevenson Center, Station B 351822, Nashville, TN 37235, USA. Fax: +1 615 936 2211.
E-mail address: jens.meiler@vanderbilt.edu (J. Meiler).

1047-8477/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2010.10.013

fragment insertion is driven by a variety of knowledge-based potentials to rapidly predict protein folds. In high-resolution refinement, the protein backbone φ and ψ angles are perturbed while the overall fold is maintained. Sidechain conformations are predicted via a Metropolis Monte Carlo search of rotamer space, and all torsional degrees of freedom are subjected to gradient-based minimization.

1.2. Sparse NMR restraints can be combined with Rosetta to obtain atomic detail structures

While the algorithm described above performs well in the *de novo* prediction of relatively small, soluble proteins, effectively sampling protein conformational space remains the limiting factor in the accurate prediction of more complex proteins. To this end, distance and orientational restraints, such as those obtained by NMR, have been incorporated into the Rosetta protein folding protocol (Rohl, 2005). Chemical shifts are converted into backbone torsional angle restraints, which are used in the generation of the peptide fragment libraries. Distance restraints from nuclear Overhauser effects (NOEs) are also employed in this process. Additionally, distance and orientational restraints (NOEs and residual dipolar couplings, or RDCs, respectively) have been incorporated into the scoring function and are evaluated during protein folding. Bowers et al., demonstrated that Rosetta, combined with a sparse set of NOEs (approximately one restraint per residue) and backbone chemical shifts, can produce models with atomic detail accuracy (Bowers et al., 2000). Similarly, a combination of sparse RDCs and chemical shifts was used to produce correctly folded models (Rohl and Baker, 2002). Shen et al., have made significant progress in improving the robustness and accuracy of CS-Rosetta with incomplete chemical shift datasets. CS-Rosetta is able to obtain atomic detail models based on data that would otherwise be considered unsuitable for high-resolution structure determination (Shen et al., 2008, 2010, 2009).

1.3. SDSL-EPR offers an advantage over traditional structure determination techniques

Despite such advances, some proteins remain un-amenable to structure determination by these methods. Site-directed spin labeling electron paramagnetic resonance spectroscopy (SDSL-EPR) allows for structural studies of membrane proteins and large macromolecular assemblies in native or native-like environments (Cartes et al., 2001; Hubbell et al., 1996; Hubbell and Altenbach, 1994; Liu et al., 2001; Zou and Mchaourab, 2009; Zou et al., 2009; Mchaourab et al., 2009; Koteiche and Mchaourab, 2002). SDSL involves mutating residues of interest to cysteines, which can be reacted with a paramagnetic spin label, such as methanethiosulfonate (MTS). A sensitive structural probe at a known sequence position is created, forgoing the need to "assign" signals in the spectrum as is necessary in NMR spectroscopy. Additionally, the resolution of SDSL-EPR is not limited by the size of the system. Similar to fluorescence and NMR spectroscopy, however, SDSL-EPR generates information concerning both the local environment of the spin label and the overall global fold of the protein. SDSL-EPR has been used to characterize conformational changes, such as those seen in MsbA (Zou and Mchaourab, 2009; Zou et al., 2009), rhodopsin (Altenbach et al., 1994, 1996, 1999), and KcsA (Liu et al., 2001; Gross et al., 1999; Cordero-Morales et al., 2006). More recently, it has been demonstrated that the fold of a protein can be determined by structural restraints derived from SDSL-EPR data alone (Alexander et al., 2008).

1.4. Atomic detail protein structure determination by SDSL-EPR is difficult and computationally demanding

Challenges in using SDSL-EPR structural data arise from the possible perturbation of the system by introduction of the spin label, sparseness of datasets resulting from the need to construct a dedicated mutant for every data point collected, and uncertainty in the position and dynamics of the spin label relative to the protein backbone. In the past, proteins have displayed a surprising robustness with respect to the introduction of spin labels (Mchaourab et al., 1996; Langen et al., 2000; Vasquez et al., 2006; Perozo et al., 1998; Brown et al., 2002). Molecular dynamics simulations (Sale et al., 2005) and crystallography (Langen et al., 2000; Fleissner et al., 2009) have been employed to explicitly model the spin label in order to help interpret SDSL-EPR structural data. However, these calculations are relatively slow and computationally demanding. In addition, most studies of this nature are designed to examine a specific protein and are not easily expanded to other systems. For the purpose of protein structure determination, a faster, broadly applicable approach to relate the spin label position to the protein backbone is needed. As an exhaustive experimental

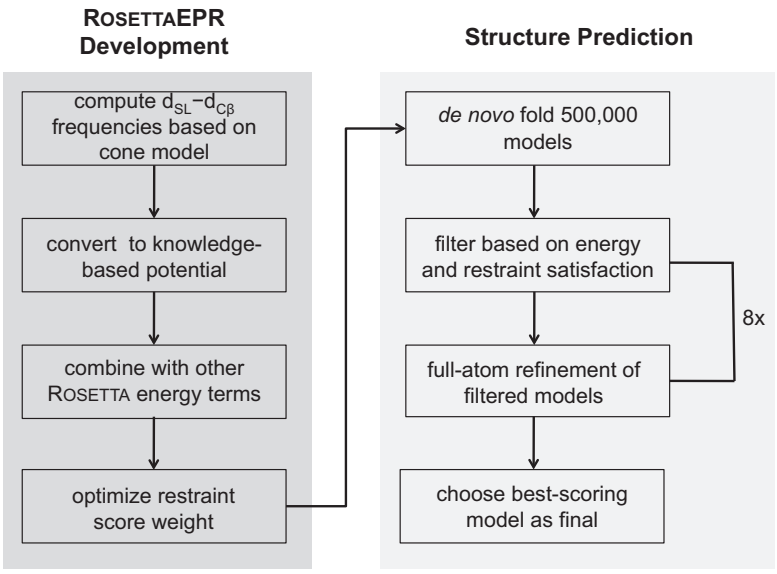

Fig. 1. Flowchart outlining the currently described protocol.

mapping of intra-protein distances is infeasible given the time and labor intensiveness of the SDSL-EPR method, a limited dataset that unambiguously describes the fold of the protein needs to be defined (see Kazmier et al., accompanying article in this issue).

1.5. RosettaEPR is designed specifically to work with sparse SDSL-EPR data

In 2008, Alexander et al. introduced the implicit "motion-on-a-cone" model, or cone model (Fig. 2B), which is based on the structure of the MTS spin label (Fig. 2A) (Alexander et al., 2008). This model was used to convert an observed spin label distance, d_{SL}, into an "allowed" range for the distance of the Cβ atoms, $d_{C_\beta} \in [d_{SL} - 12.5\,\text{Å},\ d_{SL} + 2.5\,\text{Å}]$ (Fig. 2C). The authors demonstrate that these distance restraints are sufficient to determine the structure of T4-lysozyme to atomic detail accuracy from 25 SDSL-EPR restraints. The present study introduces RosettaEPR, which replaces the soft interpretation of the distance constraints used in the previous study with a knowledge-based restraint potential optimized for SDSL-EPR distance data. Alexander et al. utilized RosettaNMR, with the consequence that all d_{C_β} distances falling within the allowed range were considered equally favorable during *de novo* folding. All other distances were disfavored using a quadratic penalty function (Fig. S1). However, while the distance difference, $d_{SL} - d_{C_\beta}$, falls within a wide range, values between 0 and 5 Å are more likely than values outside this range. We used the cone model, in combination with the PDB, to derive a probability function for $d_{SL} - d_{C_\beta}$, which was then converted into a scoring function using the Boltzmann relation. We demonstrate that treatment of SDSL-EPR distance restraints with this scoring function is superior. Following the benchmarking presented in this paper, RosettaEPR will be made available to the scientific community.

2. Materials and methods

The protocol described in the present work is outlined in Fig. 1. It is divided into two subsections corresponding to the implementation and development of RosettaEPR and the prediction of the T4-lysozyme structure to atomic detail.

2.1. Conversion of the motion-on-a-cone model into a knowledge-based potential

The $d_{SL} - d_{C_\beta}$ histogram (Fig. 2D) was generated by placing a cone model-based simulated spin label at every exposed amino acid position in 3584 proteins from a non-redundant protein database (Wang and Dunbrack, 2003). That is, the simulated spin label was placed at residue positions that had a neighbor count (Durham et al., 2009) of less than ten, resulting in over 140 million measured distances. For every pairwise distance within each protein, the protein's d_{C_β} was subtracted from the simulated d_{SL} and stored in 0.5 Å-wide bins. Because the highest frequency of $d_{SL} - d_{C_\beta}$ values was on the order of 10^6, a pseudocount of 10^6 was added to the total counts computed so that less commonly observed values are also considered.

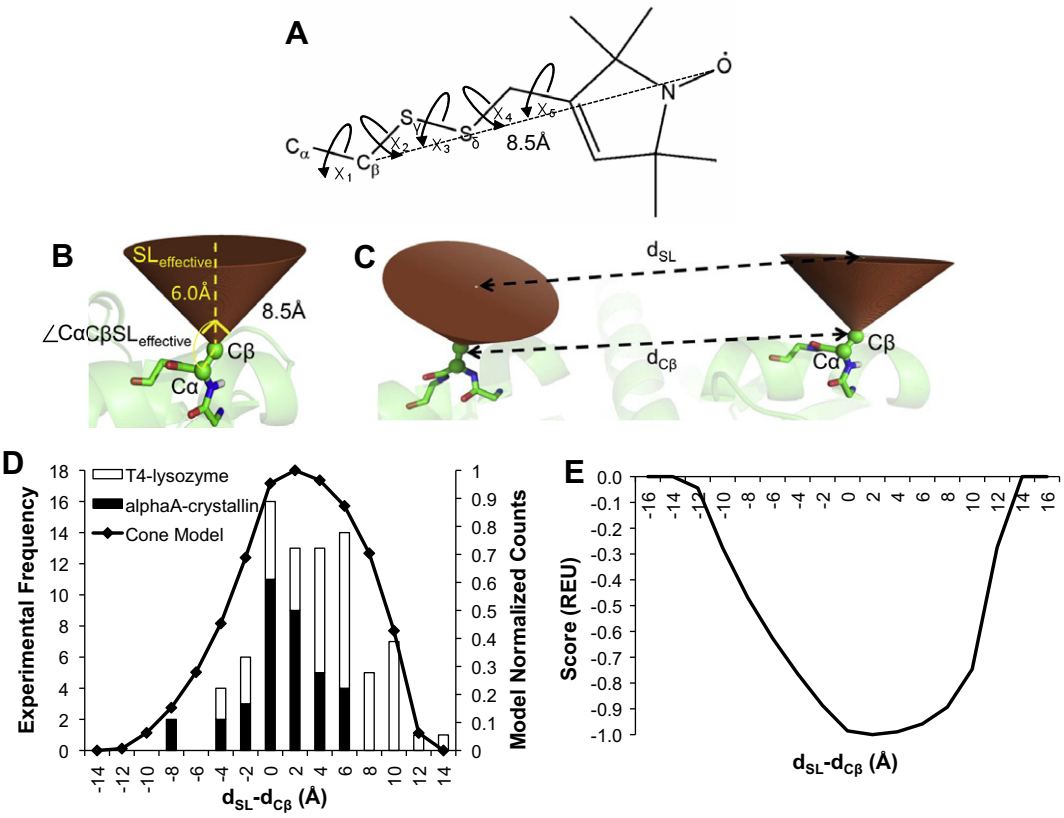

Fig. 2. The "motion-on-a-cone" model. (A) Methanethiosulfonate (MTS) spin label. The Cβ-SL distance is approximately 8.5 Å. (B) In the cone model, the Cβ-SL distance ($SL_{effective}$) is assumed to be 6 Å, and the cone has an opening angle of 90°. The Cα-Cβ-$SL_{effective}$ angle is restrained to angles $135° \leqslant (\angle C\alpha C\beta SL_{effective}) \leqslant 180°$. (C) The cone model is used to calculate $d_{SL} - d_{C_\beta}$ values. (D) The normalized frequency of $d_{SL} - d_{C_\beta}$ values for a database of proteins (black line, right y-axis) compared to experimentally observed values for T4-lysozyme and αA-crystallin (open and filled bars, respectively, left y-axis). (E) The propensity of $d_{SL} - d_{C_\beta}$ values can be converted into a knowledge-based potential according to the Boltzmann relation. The resulting energies were normalized such that the most favored $d_{SL} - d_{C_\beta}$ value correlates with an energy of −1.0 Rosetta Energy Unit (REU), and the least favored $d_{SL} - d_{C_\beta}$ value correlates with a Rosetta energy of 0.0 REU.

The potential (Fig. 2E) was calculated by taking the negative logarithm (−ln) of the propensity of each $d_{SL} - d_{C_\beta}$ value, where the propensity is defined as:

$$propensity = \frac{\left(\frac{frequency + PseudoCounts}{TotalCounts}\right)}{\left(\frac{1}{\#\ bins}\right)}$$

PseudoCount equals 10^6, and *# bins* equals 64. The resulting values were normalized and shifted such that they were all negative. This relationship is based on the Boltzmann relationship, which is used to correlate a population of a species to an associated energy. The potential was re-scaled to give a maximum bonus of −1.0 for $d_{SL} - d_{C_\beta}$ values between −12.0 and 12.0 (observed by the cone model) and a 0.0 penalty for values outside this range.

2.2. Model quality was assessed according to $RMSD_{C_\alpha}$ relative to the 2LZM crystal structure

In order to best assess the ability of RosettaEPR to recover native-like folds, only the α-helical core domain of T4-lysozyme (residues 58–164) was modeled, as experimental restraints for other regions of this protein were not available. The experimentally determined distances used as restraints are reported in Table S1 and are mapped onto the T4-lysozyme crystal structure in Fig. S2. Models of the protein were generated (a) without restraints, (b) with restraints using RosettaEPR's knowledge-based potential, and (c) with restraints defined by the same boundaries as those used by Alexander et al. Model quality was assessed by computing the $RMSD_{C_\alpha}$ relative to the X-ray crystal structure of T4-lysozyme (PDBID: 2LZM, Weaver and Matthews, 1987). Only core residues 70–155, excluding loops, were considered in computing the $RMSD_{C_\alpha}$ (see Table S2).

2.3. Weight optimization for the knowledge-based SDSL-EPR restraint potential

To optimize the factor by which the RosettaEPR scoring function should be applied, 10,000 models of the α-helical region of T4-lysozyme were constructed for a wide variety of weights (Table S3). The fraction of models with $RMSD_{C_\alpha}$ values below 7.5 Å was taken as measure for the correct fold. The fraction of models with $RMSD_{C_\alpha}$ values below 3.5 Å was employed to identify candidate models for successful atomic detail refinement; models generated with this level of accuracy are considered to be "native-like." The knowledge-based potential was implemented as a spline approximation in the Rosetta AtomPairConstraint score. The bounded restraint uses the AtomPairConstraint score as computed according to a bounded quadratic equation (Fig. S1).

2.4. Rosetta was used to de novo fold and refine T4-lysozyme

Secondary structure prediction of the 107 C-terminal residues of T4-lysozyme was performed using Jufo (Meiler and Baker, 2003), Psipred (Jones, 1999), and Sam (Karplus et al., 1997). Peptide fragments to be used in *de novo* structure prediction were generated as previously described, and fragments based on homologous proteins were excluded during folding. Rosetta's low-resolution *de novo* protein folding algorithm was used to generate 10,000 models of T4-lysozyme guided by experimental restraints (Table S1) Alexander et al., 2008 weighted to various extents, resulting in models containing structural information of the protein backbone only. During *de novo* folding, residues are represented as superatoms, or "centroids" (Simons et al., 1997). After determining that the RosettaEPR knowledge-based potential optimally predicts the fold of T4-lysoyzme when multiplied by a factor of 4.0, this weight was used in the generation of 500,000 models of the protein.

The 500,000 models were filtered according to their overall Rosetta energy and the extent to which they satisfied the experimental restraints. Only the top 1% of models by total score that had a restraint score of at least 85% of the optimum value were included in the filtered ensemble. These 1388 models were then refined to atomic detail, in which the centroids were replaced with sidechain rotamers based on a backbone-dependent rotamer library (Dunbrack and Karplus, 1993). During refinement, Rosetta's full-atom scoring potentials are used to guide refinement through an iterative cycle of sidechain repacking and gradient-based minimization (Bradley et al., 2005; Misura and Baker, 2005). Each round of refinement yielded ten times the initial number of models. That is, one round of refinement resulted in 13880 new, refined models. All *de novo* folding and full-atom refinement computations were performed using Rosetta trunk revision 34586.

2.5. Structure determination with RosettaEPR is computationally feasible

All models were generated by independent simulations using Vanderbilt University's Center for Structural Biology computing cluster and the university's Advanced Computing Center for Research and Education (ACCRE). Computations were performed on a combination of AMD Opteron and Intel Nehalem processor nodes. The average time needed to fold one model of the 107 C-terminal residues of T4-lysozyme was approximately 240 s. The same time is required for a single round of high-resolution refinement of one model.

3. Results

3.1. Knowledge-based potential reflects likelihood of model in light of observed SDSL-EPR distance

Cone model-based statistics were collected over a database of non-redundant proteins (see Section 2) and compared to $d_{SL} - d_{C_\beta}$ values determined experimentally for T4-lysozyme and αA-crystallin (Fig. 2D). The set of cone model statistics recovers several features of the experimental data, including the range of $d_{SL} - d_{C_\beta}$ values and a shift towards $d_{SL} - d_{C_\beta}$ values greater than 0 Å. The shift towards positive $d_{SL} - d_{C_\beta}$ values indicates that spin labels are more likely to point away from each other. This is expected for soluble proteins, where mutations of surface residues are not expected to destabilize the protein.

For conversion into a knowledge-based potential, the negative logarithm (−ln) of the propensity of each $d_{SL} - d_{C_\beta}$ value was computed such that less frequently seen $d_{SL} - d_{C_\beta}$ values are considered less favorable than those that are more often observed (Fig. 2E). In result, a restraint that is fulfilled in the most likely area of the distribution improves the total score by one point, and a restraint that is violated is not counted towards the total score. This knowledge-based potential was then incorporated into Rosetta's low-resolution scoring function where it is affiliated with a dedicated weight (see Section 3.2 below). The current model is an improvement upon the original implementation of the cone model, in that a) protein structures, not ellipsoids, were used to generate the statistics, and b) the knowledge-based potential considers the likelihood of $d_{SL} - d_{C_\beta}$ values instead of a simple binary classification.

3.2. Knowledge-based potential achieves up to 55% correctly folded T4-lysozyme models

Ten thousand T4-lysozyme models were folded *de novo* in the presence of the same restraints used previously (Table S1 and Fig. S2) (Alexander et al., 2008). Restraints were incorporated with

various weights, and the results were compared to the bounded potential used by Alexander et al. (Table S3). The usage of restraint scoring functions results in more native-like folds than when folding with no restraints at all (Fig. 3 and Table 1). This reaffirms that experimental data increases sampling of more native-like structures. RosettaEPR recovers the native topology of the T4-lysozyme α-helical region in up to 55% of the models. This compares to 7% if no restraints are used and 42% when using bounded restraints. Furthermore, folding with bounded restraints consistently resulted in approximately 1.0–1.5% of all built models having native-like conformations, compared to 2.1% when using the EPR knowledge-based potential with an optimal weight of 4.0. This improvement is significant, as additional starting structures for high-resolution refinement increase the chance of successfully obtaining atomic detail models (see Section 3.4). Further, conversion to a knowledge-based potential enabled fine-tuning of the weight of the SDSL-EPR potential for optimal performance, while the bounded potential provided constant suboptimal performance over wide ranges of the weight.

3.3. Knowledge-based function improves correlation of score and model quality

The correlation of the scoring function with model quality is key to selection of native-like models when the structure is not known. The correlation coefficient improves from 0.42 in the absence of restraints to 0.51 when using the bounded function and further to 0.62 when using RosettaEPR (Fig. 4). To quantify the value of the score for filtering native-like models, the enrichment for each optimized scenario was also computed (see Table 1). For the knowledge-based potential weighted by a factor of 4.0, the benchmark resulted in an enrichment of 7.0. The same analysis was performed on the models folded with the equally weighted bounded restraint potential, resulting in an enrichment of 5.3. The ensemble of models generated with no restraints contained only three native-like models, all of which were among the 10% best-scoring models, but this method was unable to produce enough native-like models to justify any high-resolution refinement.

3.4. Ten-fold enrichment of low-RMSD models through knowledge-based SDSL-EPR score allows for high-resolution refinement

500,000 models of T4-lysozyme were *de novo* folded in Rosetta guided by 25 EPR distance restraints (weight equals 4.0). From the 1% best-scoring models, models achieving at least 85% of the optimal knowledge-based restraint score were selected for high-resolution refinement. The enrichment of native-like models in the filtered pool was 10.6, while the enrichment of correctly folded models was 2.3, where enrichment was defined as the fraction of native-like or correctly folded models in the filtered

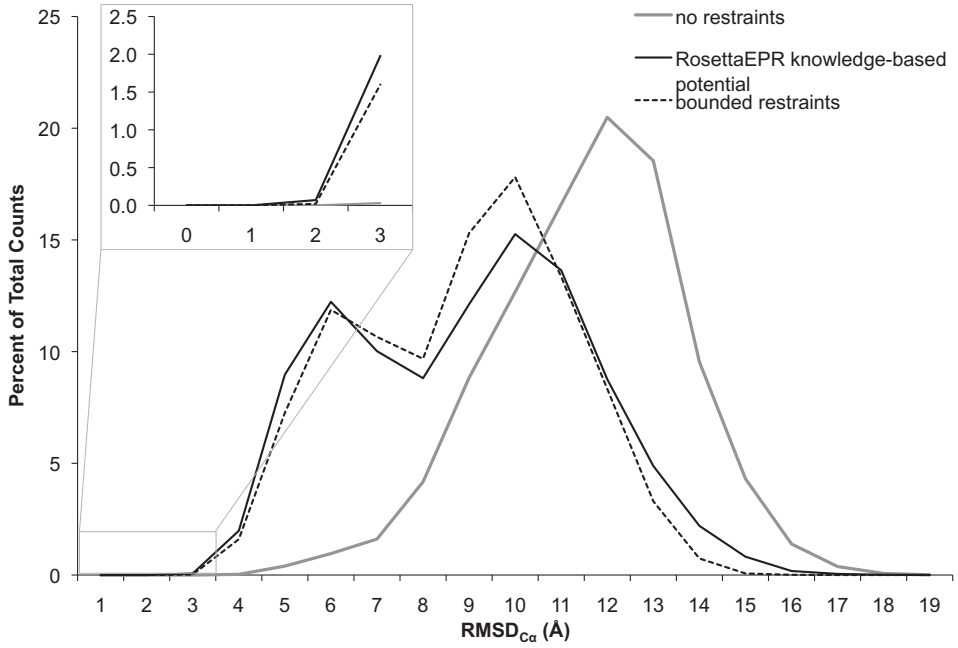

Fig. 3. Comparison of the RosettaEPR knowledge-based potential to the bounded potential. T4-lysozyme was folded *de novo* in Rosetta guided by 25 experimental restraints. Restraint violations were scored according to either a bounded potential or the EPR knowledge-based potential. The RMSD$_{C_\alpha}$ distributions of the resulting models when folded with optimally weighted restraint energies are compared to folding without restraints.

Table 1
Summary of benchmarking results of T4-Lysozyme using no restraints, 25 restraints scored according to the optimally weighted RosettaEPR knowledge-based potential, and 25 bounded restraints with a weight of 4.0[a].

Restraint type	% Models with RMSD$_{C_\alpha}$ < 3.5 Å	% Models with RMSD$_{C_\alpha}$ < 7.5 Å	Enrichment[b]
None	0.03	7.17	–[c]
Knowledge-based potential (weight = 4.0)	2.05	42.08	7.0
Bounded restraints (weight = 4.0)	1.62	41.09	5.3

[a] Results for all tested weights reported in Table S3.
[b] Enrichment = (fraction of low-RMSD models in filtered ensemble) ÷ (fraction of low-RMSD models of all models generated); filtered ensemble = within the top 1% of models by total score, the top 35% of models according to restraint score.
[c] Enrichment could not be computed as with the other data sets due to lack of restraint score.

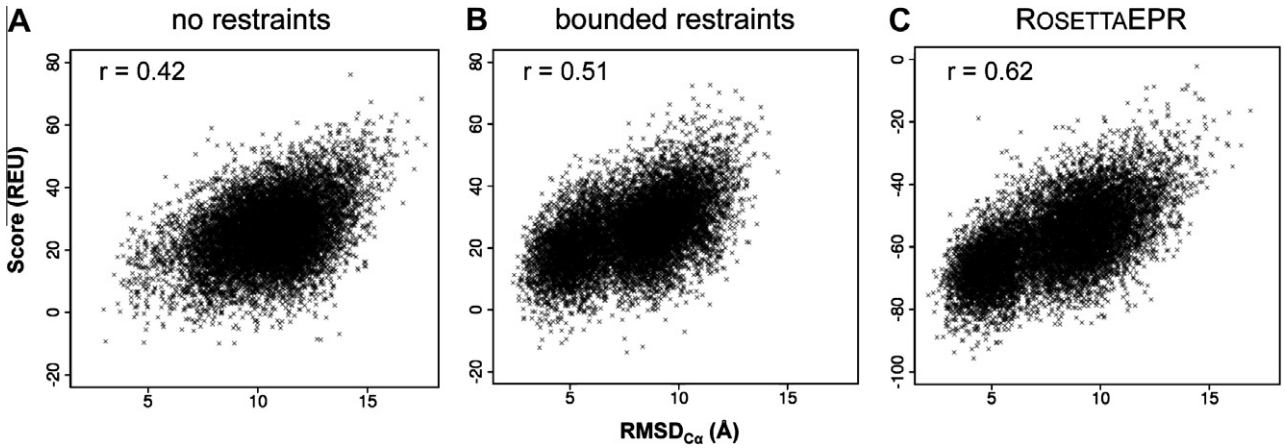

Fig. 4. Correlation between total Rosetta energy and $RMSD_{C_\alpha}$ of *de novo* folded models. Score vs. $RMSD_{C_\alpha}$ for 10,000 models *de novo* folded (A) with no restraints, (B) with 25 bounded restraints, and (C) with 25 restraints guided by the RosettaEPR knowledge-based potential.

pool divided by the fraction of native-like or correctly folded models in the entire ensemble. Filtering decreases the number of models considered for high-resolution refinement to a more manageable ensemble and enriches the fraction of low-RMSD models such that more native-like folds are refined to full-atom detail.

3.5. High-resolution refinement of T4-lysozyme yields structural model that is accurate at atomic detail

The resulting 1388 models of T4-lysozyme were refined to high-resolution using Rosetta's full-atom potentials, which include knowledge-based van der Waals attraction, repulsion, hydrogen bonding, solvation, and electrostatic terms (Bradley et al., 2005). Each input model was refined ten times without experimental restraints, resulting in 13880 models. Ideally, low-RMSD models would be considered energetically favored according to Rosetta's scoring function. Therefore, the models were then filtered such that only the top 10% by total score were carried onto the next round of refinement. This process was repeated through eight iterations, at which point the score of the refined models converged. The total score of each model was plotted against its $RMSD_{C_\alpha}$ (Fig. 5A). The correlation between energetically favorable and low-RMSD models improves after each round of refinement until it converges after the eighth iteration. The lowest energy model produced with this strategy had an $RMSD_{C_\alpha}$ of 1.76 Å relative to the native (Fig. 6), and the lowest $RMSD_{C_\alpha}$ observed was 1.73 Å. The previously reported model was determined to have an $RMSD_{C_\alpha}$ of 1.66 Å.

The ability of Rosetta to recover native-like sidechain conformations was tested by comparing sidechain rotamer agreement of refined models of T4-lysozyme with the X-ray crystal structure. A rotamer of a given amino acid residue is defined by its χ_{1-4} angles. Sidechain conformations are classified by assigning them to the closest rotamer in terms of χ_{1-4} angle deviation (Dunbrack and Karplus, 1993; Dunbrack, 2002). The total Rosetta energy is plotted as a function of the percentage of incorrectly predicted sidechain rotamers (Fig. 5B). In general, the Rosetta energy correlates well with rotamer agreement, with the percent of correct rotamers predicted increasing after each round of refinement.

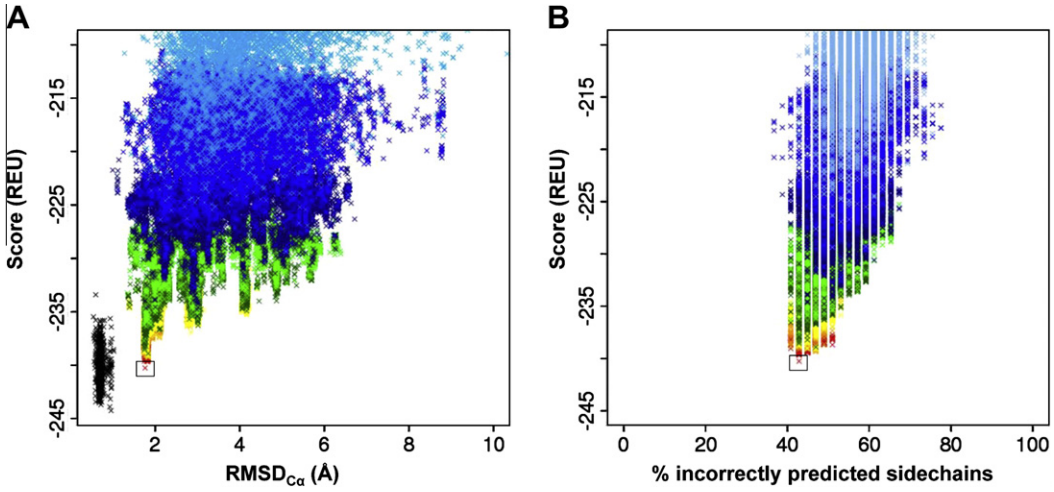

Fig. 5. Correlation between Rosetta energy and $RMSD_{C_\alpha}$ of refined models. (A) Score vs. $RMSD_{C_\alpha}$ plot of T4-lysozyme models for eight cycles of full-atom refinement. Each cycle of refinement resulted in ten times the number of input models. After each cycle, the refined models were filtered by total Rosetta energy, and the top 10% were refined again. Color key: refined crystal structure, black; round 1, sky blue; round 2, bright blue; round 3, dark blue; round 4, light green; round 5, dark green; round 6, yellow; round 7, orange; round 8, red. (B) Percent of incorrectly predicted sidechains of core residues (see Table S2) as a function of total Rosetta score. The same coloring scheme in Fig. 5A was used. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

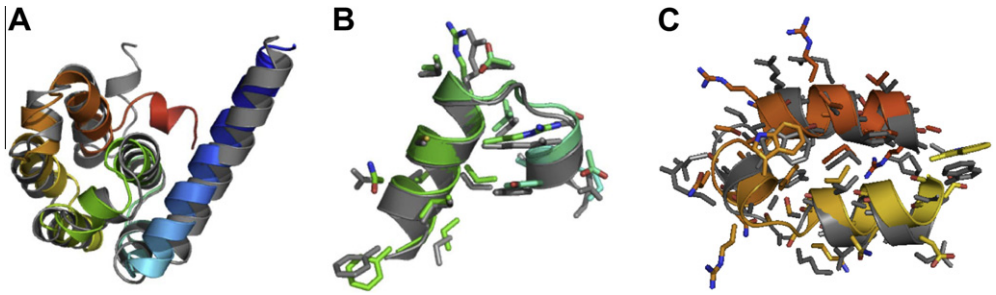

Fig. 6. Atomic detail model of T4-lysozyme *de novo* folded with RosettaEPR. (A) Superimposition of the lowest-scoring model of T4-lysozyme (rainbow) with the 2LZM crystal structure (gray). The $RMSD_{C_\alpha}$ for the lowest-scoring model to the native is 1.76 Å. Sidechains are displayed as sticks. (B) Residues 86–104. (C) Residues 126–154.

4. Discussion

4.1. The RosettaEPR knowledge-based potential proves to be superior to the bounded potential during de novo folding

We have demonstrated the advantages of using a knowledge-based potential to convert EPR distance data into structural restraints. The potential is derived from the cone model (Alexander et al., 2008) and has been shown to perform better than a simple bounded potential. From a conceptual standpoint alone, the energetic bonus correlates with the likelihood of observing $d_{SL} - d_{C_\beta}$ values. As a result, the knowledge-based potential inherently uses the structural information from SDSL-EPR data more completely compared to the bounded scoring function used by Alexander et al. Furthermore, the knowledge-based potential, in combination with Rosetta's low-resolution scoring function and *de novo* folding algorithm, proves more robust in obtaining low-RMSD models of T4-lysozyme, from which atomic detail structures can be generated through full-atom refinement.

4.2. The correlation between score and RMSD improves through multiple rounds of refinements

The Rosetta full-atom scoring function potentially allows the most native-like model to be identified unambiguously by its overall score if model accuracy is better than 2.0 Å. This model should have the lowest overall Rosetta energy and therefore exhibit not only the correct topology, but also native-like sidechain and backbone conformations. Similarly, less favorable conformations should have higher computed energies; these models will also have higher computed RMSDs relative to the native structure. One therefore expects to observe an energy "funnel" after several rounds of full-atom refinement, where both the score and RMSD of the models converge to the native structure. The overall scores of the predicted models of T4-lysozyme are plotted against their $RMSD_{C_\alpha}$ relative to the crystal structure in Fig. 5. The correlation improves after each round of filtering and refinement, resulting in several atomic detail models with Rosetta energies comparable to the 2LZM crystal structure, which was refined using the same potentials as the predicted models.

4.3. RosettaEPR will be developed continuously as more data become available

Although a larger benchmarking set would be ideal, there are a limited number of systems for which both experimentally determined three-dimensional structures and EPR data can be obtained. However, the resulting atomic detail models of T4-lysozyme generally satisfy the experimental EPR data, and benchmarking will be expanded to more diverse systems as more data become available. In the mean time, a larger benchmark on a variety of proteins of known structure using simulated data will be performed to assess the general performance of the method. The current work serves as a proof of principle. It will be interesting to test whether similar results will be obtained for other proteins. It has already been shown that NMR restraints greatly aid Rosetta's ability to recover native-like models (Rohl and Baker, 2002; Rohl, 2005; Bowers et al., 2000; Meiler and Baker, 2003; Raman et al., 2010), a method which is widely applicable to other biological systems, including the fumarate sensor DcuS (Meiler and Baker, 2005) and a chordin-like cysteine-rich (CR) repeat from procollagen IIA (O'Leary et al., 2004). It is believed that the same will be true with RosettaEPR after further testing and refinement.

4.4. Sparse SDSL-EPR distance data alone are not able to yield atomic detail models

SDSL-EPR affords several advantages over other structure determination techniques, such as X-ray crystallography and NMR. No crystallization is required, there are few size constraints, proteins, and membrane proteins in particular, can be studied in a native-like environment, and there is no need to assign resonance signals. Thereby, SDSL-EPR overcomes some experimental limitations in the high-resolution structure determination of proteins that are large, highly flexible, or natively reside in lipid bilayers.

However, while quantitative in nature, the structural information obtained by SDSL-EPR is limited due to the flexibility of the spin label, which adds large uncertainties to the distances determined. Introduction of spin labels into proteins requires removal of native cysteine residues without affecting the protein structure and assumes that the spin label itself does not perturb the structure. Datasets obtained by SDSL-EPR remain sparse due to the requirement to create a dedicated double-mutant for each distance to be measured. Therefore, SDSL-EPR (a) will be applied to systems where crystallography and NMR spectroscopy are not applicable and (b) will be combined with crystallography and other techniques to study structural dynamics of proteins.

The current work and the results presented by Alexander et al. provide the first indication that sparse (approximately 0.25 restraints per residue) SDSL-EPR distance data can be combined with Rosetta for *de novo* protein structure elucidation with atomic detail accuracy. While RosettaEPR can be applied to soluble proteins, it is expected that the need and applicability of RosettaEPR will be highest for the structure determination of membrane proteins, the majority of which continue to evade more traditional techniques. A benchmark of RosettaEPR involving more proteins and membrane proteins in particular will be executed as suitable datasets become available.

4.5. RosettaEPR will be accessible to the scientific community

Other researchers will have access to RosettaEPR via software licenses granted by the RosettaCommons (www.rosettacommons.org). These licenses are free for academic and non-profit

institutions. To encourage usage of RosettaEPR, web tutorials will be made available.

5. Conclusions

RosettaEPR is the first tool designed to generate high-resolution protein structures from sparse EPR data. It can also be used in combination with an optimized restraint-selecting algorithm (see Kazmier et al., accompanying article in this issue) to assist experimentalists in determining protein structures to high-resolution. In the future, RosettaEPR will be modified such that it can be used to effectively determine the structures of membrane proteins, an EPR accessibility knowledge-based potential will be implemented, and high-resolution modeling of the MTS spin label will be included. The ultimate goal of this research is to optimize the structural information that can be achieved through EPR spectroscopy. RosettaEPR will enable the high-resolution structure elucidation of a plethora of proteins for which structures have, until now, not yet been determined.

Acknowledgments

We would like to thank members of the Rosetta community for sharing their knowledge of various aspects of the software. We are specifically grateful to Kristian Kaufmann, Samuel DeLuca, and Kelli Kazmier for their insight and assistance throughout the development of RosettaEPR. This work was funded in part by the NIH R01 GM080403 to Jens Meiler and GM077659 to Hassane Mchaourab. Nathan Alexander is funded by NIH NIMH Award Number F31 MH086222.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jsb.2010.10.013.

References

Alexander, N., Al-Mestarihi, A., Bortolus, M., Mchaourab, H., Meiler, J., 2008. De novo high-resolution protein structure determination from sparse spin-labeling epr data. Structure 16, 181–195.

Altenbach, C., Greenhalgh, D.A., Khorana, H.G., Hubbell, W.L., 1994. A collision gradient method to determine the immersion depth of nitroxides in lipid bilayers: application to spin-labeled mutants of bacteriorhodopsin. Proceedings of the National Academy of Sciences 91, 1667–1671.

Altenbach, C., Yang, K., Farrens, D.L., Farahbakhsh, Z.T., Khorana, H.G., Hubbell, W.L., 1996. Structural features and light-dependent changes in the cytoplasmic interhelical E-F loop region of rhodopsin: a site-directed spin-labeling study. Biochemistry 35, 12470–12478.

Altenbach, C., Klein-Seetharaman, J., Hwa, J., Khorana, H.G., Hubbell, W.L., 1999. Structural features and light-dependent changes in the sequence 59–75 connecting helices I and II in rhodopsin: a site-directed spin-labeling study. Biochemistry 38, 7945–7949.

Barth, P., Schonbrun, J., Baker, D., 2007. Toward high-resolution prediction and design of transmembrane helical protein structures. Proceedings of the National Academy of Sciences 104, 15682–15687.

Barth, P., Wallner, B., Baker, D., 2009. Prediction of membrane protein structures with complex topologies using limited constraints. Proceedings of the National Academy of Sciences 106, 1409–1414.

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., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D., Zardecki, C., 2002. The Protein Data Bank. Acta Cryst. D58, 899–907.

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

Bonneau, R., Strauss, C.E.M., Rohl, C.A., Chivian, D., Bradley, P., Malmström, L., Robertson, T., Baker, D., 2002. *De novo* prediction of three-dimensional structures for major protein families. Journal of Molecular Biology 322, 65–78.

Bowers, P.M., Strauss, C.E.M., Baker, D., 2000. *De novo* protein structure determination using sparse NMR data. Journal of Biomolecular 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., Strauss, C.E., Baker, D., 2003. Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation. Proteins: Structure, Function, and Genetics 53, 457–468.

Bradley, P., Misura, K.M.S., Baker, D., 2005. Toward high-resolution *De novo* structure prediction for small proteins. Science 309, 1868–1871.

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

Brown, L.J., Sale, K.L., Hills, R., Rouviere, C., Song, L., Zhang, X., Fajer, P.G., 2002. Structure of the inhibitory region of troponin by site-directed spin labeling electron paramagnetic resonance. Proceedings of the National Academy of Sciences 99, 12765–12770.

Cartes, D.M., Cuello, L.G., Perozo, E., 2001. Molecular architecture of full-length KcsA role of cytoplasmic domains in ion permeation and activation gating. Journal of General Physiology 117, 165–180.

Cordero-Morales, J.F., Cuello, L.G., Zhao, Y., Jogini, V., Marien Cortes, D., Roux, B., Perozo, E., 2006. Molecular determinants of gating at the potassium-channel selectivity filter. Nature Structural & Molecular Biology 13, 311–318.

Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, M.D., Bhat, D., Chivian, D., Kim, D.E., Sheffler, W.H., Malmström, L., Wollacott, A.M., Wang, C., Andre, I., Baker, D., 2007. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@Home. Proteins 69 (Suppl. 8), 118–128.

Das, R., André, I., Shen, Y., Wu, Y., Lemak, A., Bansal, S., Arrowsmith, C.H., Szyperski, T., Baker, D., 2009. Simultaneous prediction of protein folding and docking at high resolution. Proceedings of the National Academy of Sciences 106, 18978–18983.

Dunbrack, R.L., 2002. Rotamer libraries in the 21st century. Current Opinion in Structural Biology 12, 431–440.

Dunbrack, R.L., Karplus, M., 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. Journal of Molecular Biology 230, 543–574.

Durham, E., Dorr, B., Woetzel, N., Staritzbichler, R., Meiler, J., 2009. Solvent accessible surface area approximations for rapid and accurate protein structure prediction. Journal of Molecular Modeling 15, 1093–1108.

Fleissner, M.R., Cascio, D., Hubbell, W.L., 2009. Structural origin of weakly ordered nitroxide motion in spin-labeled proteins. Protein Science 18, 893–908.

Gross, A., Columbus, L., Hideg, K., Altenbach, C., Hubbell, W.L., 1999. Structure of the KcsA potassium channel from *Streptomyces lividans*: A site-directed spin labeling study of the second transmembrane segment. Biochemistry 38, 10324–10335.

Haley, D.A., Bova, M.P., Huang, Q.L., Mchaourab, H.S., Stewart, P.L., 2000. Small heat-shock protein structures reveal a continuum from symmetric to variable assemblies. Journal of Molecular Biology 298, 261–272.

Harrison, S.C., 2004. Whither structural biology? Nature Structural & Molecular Biology 11, 12–15.

Hubbell, W.L., Altenbach, C., 1994. Investigation of structure and dynamics in membrane proteins using site-directed spinlabeling. Current Opinion in Sturctural Biology. 4, 566–573.

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

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 292, 195–202.

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

Kaufmann, K.W., Lemmon, G.H., Deluca, S.L., Sheehan, J.H., Meiler, J., 2010. Practically useful: what the Rosetta protein modeling suite can do for you. Biochemistry 49, 2987–2998.

Koteiche, H.A., Mchaourab, H.S., 2002. The determinants of the oligomeric structure in Hsp16.5 are encoded in the alpha-crystallin domain. FEBS Letters 519, 16–22.

Langen, R., Oh, K.J., Cascio, D., 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.

Liu, Y.-S., Sompornpisut, P., Perozo, E., 2001. Structure of the KcsA channel intracellular gate in the open state. Nature Structural Biology 8, 883–887.

Mchaourab, H.S., Lietzow, M.A., Hideg, K., 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., Godar, J.A., Stewart, P.L., 2009. Structure and mechanism of protein stability sensors: Chaperone activity of small heat shock proteins. Biochemistry 48, 3828–3837.

Meiler, J., Baker, D., 2003. Coupled prediction of protein secondary and tertiary structure. Proceedings of the National Academy of Sciences 100, 12105–12110.

Meiler, J., Baker, D., 2003. Rapid protein fold determination using unassigned NMR data. Proceedings of the National Academy of Sciences 100, 15404–15409.

Meiler, J., Baker, D., 2005. The fumarate sensor DcuS: Progress in rapid protein fold elucidation by combining protein structure prediction methods with NMR spectroscopy. Journal of Magnetic Resonance 173, 310–316.

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

O'Leary, J.M., Hamilton, J.M., Deane, C.M., Valeyev, N.V., Sandell, L.J., Downing, A.K., 2004. Solution structure and dynamics of a prototypical chordin-like cysteine-rich repeat (von Willebrand factor type C module) from collagen IIa. Journal of Biological Chemistry 279, 53857–53866.

Perozo, E., Marien Cortes, D., Cuello, L.G., 1998. Three-dimensional architecture and gating mechanism of a K+ channel studied by EPR spectroscopy. Nature Structural Biology 6, 459–469.

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., Kellogg, E., DiMaio, F., Lange, O., Kinch, L., Sheffler, W., Kim, B.-H., Das, R., Grishin, N.V., Baker, D., 2009. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins 77 (Suppl. 9), 89–99.

S. Raman, O. Lange, P. Rossi, M. Tyka, X. Wang, J. Aramini, G. Liu, T. Ramelot, A. Eletsky, T. Szyperski, M. Kennedy, J. Prestegard, G. Montelione, D. Baker, NMR structure determination for larger proteins using backbone-only data, Science (2010) science.1183649v1183641.

Rohl, C.A., 2005. Protein structure estimation from minimal restraints using Rosetta. Methods in Enzymology 394, 244–260.

Rohl, C.A., Baker, D., 2002. De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. Journal of the American Chemical Society 124, 2723–2729.

Sale, K., Song, L., Liu, Y.-S., Perozo, E., Fajer, P., 2005. Explicit treatment of spin labels in modeling of distance constraints from dipolar EPR and DEER. Journal of the American Chemical Society 127, 9334–9335.

Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J.M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K.K., Lemak, A., Ignatchenko, A., Arrowsmith, C.H., Szyperski, T., Montelione, G.T., Baker, D., Bax, A., 2008. Consistent blind protein structure generation from NMR chemical shift data. Proceedings of the National Academy of Sciences 105, 4685–4690.

Shen, Y., Vernon, R., Baker, D., Bax, A., 2009. *De novo* protein structure generation from incomplete chemical shift assignments. Journal of Biomolecular NMR 43, 63–78.

Shen, Y., Bryan, P.N., He, Y., Orban, J., Baker, D., Bax, A., 2010. *De novo* structure generation using chemical shifts for proteins with high-sequence identity but different folds. Protein Science 19, 349–356.

Simons, K.T., Kooperberg, C., Huang, E., Baker, D., 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. Journal of Molecular Biology 268, 209–225.

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

Vasquez, V., Sotomayor, M., Marien Cortes, D., Roux, B., Schulten, K., Perozo, E., 2006. Three-dimensional architecture of membrane-embedded MscS in the closed conformation. Journal of Molecular Biology 378, 55–70.

Wang, G., Dunbrack, R.L., 2003. PISCES: A protein sequence culling server. Bioinformatics 19, 1589–1591.

Weaver, L.H., Matthews, B.W., 1987. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. Journal of Molecular Biology 193, 189–199.

Yarov-Yarovoy, V., Schonbrun, J., Baker, D., 2006. Multipass membrane protein structure prediction using Rosetta. Proteins 62, 1010–1025.

Zou, P., Mchaourab, H.S., 2009. Alternating access of the putative substrate-binding chamber in the ABC transporter MsbA. Journal of Molecular Biology 393.

Zou, P., Bortolus, M., Mchaourab, H., 2009. Conformational cycle of the ABC transporter MsbA in liposomes: Detailed analysis using double electron-electron resonance spectroscopy. Journal of Molecular Biology 393, 586–597.