Algorithm for selection of optimized EPR distance restraints for *de novo* protein structure determination

Kelli Kazmier [a,b], Nathan S. Alexander [b,c], Jens Meiler [b,c,d,*], Hassane S. Mchaourab [a,b,e,*]

[a] Chemical and Physical Biology Program, 340 Light Hall, Vanderbilt University, Nashville, TN 37232, USA
[b] Center for Structural Biology, Nashville, TN 37232, USA
[c] Department of Chemistry, 7330 Stevenson Center, Station B 351822, Vanderbilt University, Nashville, TN 37235, USA
[d] Institute of Chemical Biology, 896 Preston Building, Nashville, TN 37232, USA
[e] Department of Molecular Physiology and Biophysics, 702 Light Hall, Nashville, TN 37232, USA

ARTICLE INFO

ABSTRACT

A hybrid protein structure determination approach combining sparse Electron Paramagnetic Resonance (EPR) distance restraints and Rosetta *de novo* protein folding has been previously demonstrated to yield high quality models (Alexander et al. (2008)). However, widespread application of this methodology to proteins of unknown structures is hindered by the lack of a general strategy to place spin label pairs in the primary sequence. In this work, we report the development of an algorithm that optimally selects spin labeling positions for the purpose of distance measurements by EPR. For the α-helical subdomain of T4 lysozyme (T4L), simulated restraints that maximize sequence separation between the two spin labels while simultaneously ensuring pairwise connectivity of secondary structure elements yielded vastly improved models by Rosetta folding. 54% of all these models have the correct fold compared to only 21% and 8% correctly folded models when randomly placed restraints or no restraints are used, respectively. Moreover, the improvements in model quality require a limited number of optimized restraints, which is determined by the pairwise connectivities of T4L α-helices. The predicted improvement in Rosetta model quality was verified by experimental determination of distances between spin labels pairs selected by the algorithm. Overall, our results reinforce the rationale for the combined use of sparse EPR distance restraints and *de novo* folding. By alleviating the experimental bottleneck associated with restraint selection, this algorithm sets the stage for extending computational structure determination to larger, traditionally elusive protein topologies of critical structural and biochemical importance.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

Decades into the structural biology revolution, tens of thousands of structures have been deposited in the Protein Data Bank (PDB) cataloging protein folds, defining motifs of catalysis, and revealing architectures of protein complexes. The overarching goal of delineating the biochemical and physiological circuitry that interconnect to form cells and organisms requires further progress on two fronts. The sampling of structure space has been uneven; primarily skewed towards classes of proteins amenable to analysis by the leading structural methods. Undersampled protein structure space includes proteins of high functional and pharmacological significance such as multispan membrane proteins (Tusnady et al., 2004) and large, conformationally heterogeneous soluble proteins (Haley et al., 2000). In addition, protein function often involves the transitions between conformational states or shifts in the equilibrium between such states. Static crystallographic snapshots represent a limited and sometimes biased view of the conformational space of dynamic proteins. Structures trapped in the confines of the crystal lattice may not be defined mechanistically or may be distorted by non-native environments such as detergent solubilization or osmotically active molecules (Fanucci et al., 2003).

These two challenges motivated the development of both theoretical and experimental methods to accelerate the speed of

structure determination and to describe protein dynamic dimensions. Electron Paramagnetic Resonance (EPR) spectroscopy in conjunction with site-directed spin labeling (SDSL) (Hubbell et al., 1996, 1998) has been extensively applied to map conformational changes in soluble (Persson et al., 2001; Claxton et al., 2008) and membrane proteins (Wegener et al., 2000; Dong et al., 2005; Borbat and Freed, 2007; Smirnova et al., 2007; Altenbach et al., 2008; Padmavathi and Steinhoff, 2008; Zou et al., 2009; Zou and Mchaourab, 2009; Hanelt et al., 2010) and to probe the structure of dynamic oligomers (Koteiche and Mchaourab, 2002; Mchaourab et al., 2009) and amyloids (Margittai and Langen, 2006; Jao et al., 2008). Combining residue-specific measures of solvent accessibility and local dynamics with global geometric distance restraints describing packing of secondary structures and domains, this approach provides enough restraints for modeling protein structures and their rearrangements (Jeschke et al., 2004; Hilger et al., 2007, 2009; Bhatnagar et al., 2010). High sensitivity, absence of size limits and restriction on environment and/or solvent enables the evaluation of crystallographic structures and comparative models under native-like, well defined biochemical conditions.

However this approach is intrinsically limited by the need for incorporation of spin labels into protein sequences. Compared to other restraint-based approaches such as Nuclear Magnetic Resonance (NMR) Spectroscopy, this reduces the experimental throughput effectively reducing the practical number of obtainable restraints. Moreover, the linking arm of the spin label tethering it to the protein introduces uncertainty in the interpretation of EPR parameters in terms of backbone structure. In the case of distance measurements, the translation of a precisely measured distance between spin labels to a restraint between corresponding β-carbons (C_β) is model dependent. Models derived from molecular dynamic simulations (Sale et al., 2005; Borovykh et al., 2006; Fajer et al., 2007; Jeschke and Polyhach, 2007), crystallographic rotamer libraries (Altenbach et al., 2008), or based on simple geometric considerations (Alexander et al., 2008) have been used to rationalize the experimental EPR distances.

A general approach for protein structure determination from EPR restraints was developed by Alexander et al. (2008). It capitalizes on the *de novo* protein structure prediction algorithm, Rosetta (Simons et al., 1997, 1999; Bonneau et al., 2001; Rohl et al., 2004; Bradley et al., 2005a,b; Yarov-Yarovoy et al., 2006; Das and Baker, 2008; Raman et al., 2009), to overcome the sparseness of EPR experimental restraints. The premise of this work was that restriction of conformational space by the EPR restraints increases Rosetta's efficiency in finding native folds. That a limited number of distances between pairs of spin labels significantly improved the quality of models put to rest concerns regarding the value of EPR distances as restraints for modeling. Experimental EPR distances were translated into C_β–C_β restraints using a simple cone model with virtually no restriction of spin label rotameric states.

The limited throughput of EPR methods and the ensuing restraint sparseness encourages a rational approach in the selection of spin labeled sites. Alexander et al. (2008) demonstrated the importance of high information content (defined as the ratio between sequence separation and Euclidean distance) as a criterion for restraint quality. The improvement in model quality was attributed to a third of the restraints with the highest information content. However, for proteins of unknown structures where the Euclidian distance is not known, using the numerator (i.e. sequence separation) as a proxy for information content will cluster restraints between the ends of the primary sequence.

This paper reports the development and experimental application of a general algorithm for selection of optimized distance restraint patterns for protein structure determination. Starting from sequence information, an iterative computational approach validated by Rosetta *de novo* folding yielded the best scoring

scheme for restraint selection. Using the α-helical domain of T4 lysozyme (T4L) as a model system (Matthews, 1996), we demonstrate that restraints selected to simultaneously optimize sequence separation and pairwise connectivity of secondary structures led to high quality models. To test the robustness of the algorithm, distances were experimentally measured between pairs of spin labels selected by the algorithm. Rosetta folding using these distances yielded high quality models as predicted.

2. Methods and materials

2.1. Algorithm development

Input parameters of secondary structure and solvent exposure predictions of the C-terminal 107 amino acids of T4L were obtained using psipred (McGuffin et al., 2000) and NetSurfP (Petersen et al., 2009) analyses, respectively. The ideal secondary structure definitions were obtained directly from the crystal structure of T4L (PDB ID: 2LZM). The ideal solvent exposure definitions were generated from the T4L crystal structure (2LZM) using a Rosetta neighbor count protocol. A neighbor count threshold of smaller than or equal to 9 defines solvent exposed residues (Durham et al., 2009).

The Monte Carlo protocol is initiated with a random distribution of spin label pairs that yield a total score for the distribution terms being tested. Each iteration of the Monte Carlo optimization involves random reassignment of label positions for a single pair. New label positions that improve or equal the best previous score are accepted. A typical optimization included 10,000 iteration steps and 10 optimization trajectories after which scores converged. Restraint patterns were generated on local clusters using a perl script.

2.1.1. Sequence separation term

The sequence separation score (S_{SS}) is calculated by taking the natural log of the number of amino acids separating the two spin labels in each restraint pair (d_i), averaging over all restraint pairs (r), and normalizing to the natural log of the sequence length (g) to yield a value between 0 and 1 (Eq. (1)).

$$S_{SS} = (\sum_{i=1}^{r} \ln d_i)/(r \times \ln g) \tag{1}$$

Thus, the sequence separation term effectively applies a penalty function for pairs separated by a small number of amino acids. This penalty logarithmically decreases with increased label separation. The logarithmic scaling is a modification of the original information content measure (Alexander et al., 2008). We found that the improvement in model quality measures becomes less dependent on sequence separation as d_i increases (data not shown).

2.1.2. Secondary structure term

The secondary structure term distributes the spin labels evenly among the secondary structural elements (SSE). First, an ideal number of spin labels per SSE (Q) is calculated by dividing the number of spin labels (l) (twice the number of restraints) by the number of SSEs (s). We define $Q' = \text{div}(l,s)$ and $Q'' = Q' + 1$. Note that the floor Q' and ceiling Q'' are acceptable integer values for Q. Further, we define remainder of l/s as $R = \text{mod}(l,s)$. An optimal spin label distribution will have Q'' labels in R SSEs, and Q' labels in all the others.

The secondary structure score (S_{SSE}) has two equally weighted components, $S_{SSE(L)}$ and $S_{SSE(S)}$. The first component, $S_{SSE(L)}$ is the average percentage of labels positioned in each SSE up to the ideal value, Q''. Thus,

$$S_{SSE(L)} = \frac{1}{l} \sum_{i=1}^{s} \min(l_i, Q'') \tag{2}$$

where l_i = number of labels in the ith SSE. As defined, this component favors placement of labels into SSE during the optimization trajectory.

The second component of the score, $S_{\text{SSE(S)}}$ is derived from the fraction of SSEs that contain exactly the ideal number of spin labels:

$$S_{\text{SSE(S)}} = \frac{1}{s}\{\min[E', (s-R)] + \min[E'', R]\} \qquad (3)$$

where E' is the number of SSEs with Q' labels and E'' is the number of SSEs with Q'' labels. While $S_{\text{SSE(L)}}$ determines progress in achieving an optimal spin label placement during the Monte Carlo optimization, $S_{\text{SSE(S)}}$ is needed to arrive at precisely the correct number of spin labels for every SSE (data not shown). The two scores (Eqs. (2) and (3)) are averaged to yield the total S_{SSE} term with values between 0 and 1.

2.1.3. Element connection term

Element connection (S_{EC}) favors patterns that connect each pair of SSEs with restraints. The ideal number of connections for each SSE pair (C) is defined by the ratio between the number of restraints (r) and the number of SSE pairs (p), $p = (s(s-1))/2$, where s = number of SSEs. We define $C' = \text{div}(r,p)$ and $C'' = C' + 1$. In this term, floor C' and ceiling C'' are acceptable integer values for C. In addition, we define remainder of r/p as $M = \text{mod}(r,p)$. An optimal restraint distribution will have C'' restraints in M SSE pairs, and C' restraints in all the others.

Like the secondary structure term, S_{EC} is a composite of two equally weighted component scores, $S_{\text{EC(R)}}$ and $S_{\text{EC(C)}}$. $S_{\text{EC(R)}}$ is the average percentage of restraints in each SSE pair up to the ideal value, C''. Thus,

$$S_{\text{EC(R)}} = \frac{1}{r}\sum_{i=1}^{p}\min(r_i, C''), \qquad (4)$$

where r_i = number of labels in the ith SSE pair. This component favors placement of restraints into SSE pairs during the optimization trajectory.

The second component of this term, $S_{\text{EC(C)}}$ is derived from the fraction of SSE pairs that contain exactly the ideal number of restraints:

$$S_{\text{EC(C)}} = \frac{1}{p}\{\min[F', (p-M)] + \min[F'', M]\}, \qquad (5)$$

where F' is the number of SSE pairs with C' restraints and F'' is the number of SSEs with C'' restraints. As in the secondary structure term, the composite scores of this term are complementary with $S_{\text{EC(R)}}$ measuring progress toward the optimal restraint placement and $S_{\text{EC(C)}}$ determining the correct number of restraints for every SSE pair. The two scores (Eqs. (4) and (5)) are averaged to yield the total S_{EC} term with values between 0 and 1.

2.1.4. Label density Term

The label density score, S_{LD}, imposes equal distribution of spin labels along the sequence. For this purpose, spin label positions are treated as a vector ($a_0, a_1, \ldots, a_l, a_{l+1}$), where a_0 is the N-terminus, a_{l+1} is the C-terminus, and $a_1, \ldots, a_l$ are the positions of the spin labels and l = number of spin labels. An optimal interval between spin labels (I) is the divisor of the ratio of the sequence length (g) to the number of intervals (n), where $n = l + 1$: $I = \text{div}(g,n)$. The score utilizes a harmonic penalty function. A normalization function, $f(x) = (x+1)^{-1}$, is applied to rescale values between 0 and 1. Thus the term is defined as:

$$S_{\text{LD}} = \frac{1}{n}\sum_{i=1}^{n}\{[(a_{i+1} - a_i) - I]^2 + 1\}^{-1} \qquad (6)$$

2.2. Restraint-assisted Rosetta folding simulations and EPR distance interpretation

Rosetta simulations were performed in Rosetta++ (Simons et al., 1997, 1999; Bonneau et al., 2001; Rohl et al., 2004). Specific standard Rosetta procedures were used that are described in details elsewhere (Bradley et al., 2005a). In these course-grained simulations, residues side chains are regarded as centroid superatoms (Simons et al., 1997). All T4L homologs were excluded from the fragment database prior to modeling in order to simulate structure determination of a novel protein fold as closely as possible. Models were obtained in independent simulations on a cluster in Vanderbilt University's Advanced Computing Center for Research & Education (ACCRE). For each simulation, 1000 models were created using the restraints selected by the algorithm for the α-helical subdomain of T4L (residues 58–164). In the algorithm optimization phase, C_α root mean squared deviation (RMSD) distributions and model quality measures for residues 70–164 were reported for all 1000 models resulting from Rosetta folding. Residues 58–69 were excluded from RMSD analysis as these residues link the α-helical subdomain to the excluded β-strand subdomain and tended to vary in our models due to the absence of the β-strand domain. C_α-RMSDs were used due to the course-grained nature of the modeling. In the experimental implementation phase, models were additionally filtered by lowest energy and restraint violation scores.

EPR distance restraints were implemented in Rosetta in a RosettaNMR (Bowers et al., 2000; Rohl and Baker, 2002) protocol as described previously (Alexander et al., 2008). Briefly, distance restraints are used as an additional penalty in the Rosetta energy function. This penalty is zero if the C_β–C_β distance ($d_{C\beta}$) of the restraint residues fall within the range specified. If this distance falls outside this range, a quadratic penalty function is applied. The restraint violation score is defined as the cumulative value of these penalties over all restraints. The boundary range used was based on the motion-on-a-cone model developed by Alexander et al. (2008). This model yielded a function describing the relationship between the experimentally measured spin label distance (d_{SL}) and the $d_{C\beta}$. The d_{SL} defines the range allowed for $d_{C\beta}$ ($d_{SL} - 12.5$ Å to $d_{SL} + 2.5$ Å) which corresponds to the most probable relative spin label orientations. For simulated restraints, the crystallographic $d_{C\beta}$ is used as the experimental distance (i.e. $d_{SL} - d_{C\beta} = 0$ or a parallel spin label orientation).

2.3. Recombinant expression and purification of T4L mutants

Cysteine residues were systematically introduced into a cysless T4L construct through double point mutations at restraint positions identified by the algorithm using QuikChange™ Site-Directed Mutagenesis Kit (Stratagene) as previously described (Dong et al., 2005). Sample preparation has been described elsewhere (Matthews, 1996; Koteiche and Mchaourab, 1999). Briefly, T4L mutants were sequenced, transformed into K38 cells, and expressed in Luria Broth (LB). All mutants were purified using cation exchange chromatography, labeled with a 5-fold excess of MTSSL (S-(2,2,5, 5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl Methanesulfonothioate spin label, Toronto Research Chemicals) at room temperature for 2 h, desalted and concentrated. A total of 21 double mutants resulted in the restraints used for the current analysis.

2.4. EPR distance measurements

Of the 21 restraints, 19 distances were within the distance range appropriate for double electron–electron resonance (DEER) distance measurement (Borbat et al., 2004, 2002; Jeschke, 2002). DEER measurements were performed on a Bruker 580 pulsed EPR spectrometer operating at X-band (10 GHz) using a standard

four-pulse protocol (Jeschke, 2002). Experiments were performed at 83 K. Sample concentrations were 150 μM in a MOPS/Tris buffer (9 mM MOPS, 6 mM Tris, 50 mM NaCl, 0.02% (w/v) sodium azide, 0.1 mM EDTA) with 20% (v/v) glycerol as a cryoprotectant and a sample volume of 50 μl. Spin echo decays were baseline-corrected and analyzed by Tikhonov regularization (Jeschke et al., 2002; Chiang et al., 2005) to determine average distances and distributions in distance (Appendix A). For all data, the selected α parameter corresponds to the elbow of the L-curve (Chiang et al., 2005).

For the two pairs with distances too short for DEER analysis, distance distributions were determined from the continuous wave (CW) EPR spectra using the CWdipfit program developed by Peter Fajer and colleagues (http://www.sb.fsu.edu/~fajer/Programs/CWdipFit/cwdipfit) (Sen et al., 2007). For each pair, fully labeled and underlabeled samples were prepared. Fully labeled samples were prepared as described above. Preparation of the underlabeled samples included incubation with $0.5\times$ MTSSL for 1 h at room temperature followed by addition of 20-fold excess of a diamagnetic MTSSL analog (1-acetyl-2,2,5,5,-tetramethyl-Δ3-pyrroline-3-methyl), methanethiosulfonate (Toronto Research Chemicals). The fully labeled samples display distance-dependant dipolar coupling, while the underlabeled samples represent the EPR spectrum in its absence. CWdipFit assumes Gaussian-shaped distance distributions between spin labels and utilizes Monte Carlo/SIMPLEX algorithm to fit dipolar coupled spectra using the underlabeled spectra as a proxy for the sum of singles (Sen et al., 2007; Sugata et al., 2009). The dipolar coupled spectra and fits are shown in Appendix A.

3. Results

3.1. Methodology

The overall strategy, illustrated by the flowchart in Fig. 1, uses the primary sequence, secondary structure, and solvent exposure definitions as input parameters. For secondary structure and solvent exposure, predicted and ideal (defined by the crystal structure) definitions were compared to assess the impact on model quality. The algorithm relies on a Monte Carlo search to optimize the restraint distribution terms that place pairs of spin labels along the sequence (Supplementary Fig. 1). Briefly, a sequence separation term, defined as the number of intervening amino acids between two spin labels in a pair, was included as an approximation for information content. To balance its tendency to cluster spin labels at the N- and C-termini, three terms favoring uniform sequence coverage were investigated. A secondary structure element (SSE) connection term (element connection) evenly connects all pairs of secondary structures, in this case 7 α-helices, with restraints effectively introducing a triangulation strategy. Alternatively, a label density term which distributes spin labels along the sequence at equal and regular intervals was included. Finally, we tested the efficacy of a secondary structure term that confines spin labels to segments of secondary structures avoiding loops and termini. Term combinations and weight ratios were evaluated for their effectiveness in selecting informative restraints for Rosetta folding (Supplementary Figs. 2 and 3). The combination of sequence separation and element connection terms at a 1:1 weight ratio consistently yielded restraint patterns that resulted in the highest quality models by Rosetta folding. Fig. 2 illustrates how an initial random distribution of labels is shuffled to maximize the sequence separation and element connections scores.

In the algorithm development phase described above, the term combination and relative weight were determined using simulated EPR distances. For this purpose, the distance between the β-carbon of each pair of residues, $d_{C\beta}$ was obtained from the crystal structure
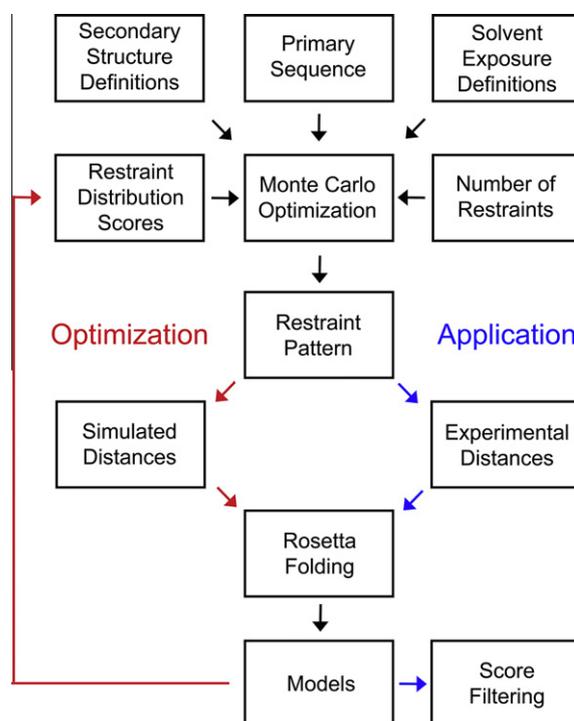

Fig. 1. Methodological flowchart. Using sequence definitions as input, a Monte Carlo approach was applied to iteratively optimize the restraint distribution scores and number of restraints. Simulated restraints were calculated to guide Rosetta folding and modeling outcomes served as indicators of optimal restraint patterns (red arrow). To test the applicability of the algorithm for restraint selection, experimental distances were measured for an optimized restraint pattern and incorporated into the Rosetta folding algorithm (blue arrows). The resulting models were filtered by energy and restraint violation scores to exclusively yield high quality models.

(2LZM) and used as an experimental restraint. To simulate the uncertainty associated with interpretation of distances between spin labels, the corresponding restraint was allowed a range of $d_{C\beta} - 12.5$ Å to $d_{C\beta} + 2.5$ Å based on the motion on a cone model described previously (Alexander et al., 2008) and in Section 2. Models with $d_{C\beta}$ distances outside this range are penalized in the Rosetta energy score.

The output of the restraint-assisted Rosetta folding consisted of 1000 models. Quality measures defined by the models C_α-RMSD to the crystal structure were used as indicators of improvement in the Rosetta sampling of conformational space. To avoid perturbation due to spin label incorporation, the algorithm excluded residues predicted to be buried. This did not affect the quality of models generated by Rosetta (Supplementary Fig. 4). In contrast, the use of predicted secondary structure resulted in a significant decrease in model quality (Supplementary Fig. 4). Therefore, for the purposes of evaluating the effectiveness of the algorithm, secondary structure definitions were based on the crystal structure.

The α-helical subdomain of T4L (residues 58–164) was selected as a model system for this analysis. T4L has been extensively investigated by spin labeling (Mchaourab et al., 1996, 1997) and was the target of a previous study to assess the potential of EPR restraints to increase the efficiency of conformational space sampling by Rosetta (Alexander et al., 2008). The 107 amino acid target region is well within the size limit for efficient structure prediction by Rosetta de novo folding (Bradley et al., 2005b). For the analysis presented here, we excluded structures homologous to T4L from the fragment library to mimic protein structure prediction of novel protein folds. Under these conditions, Rosetta folding in the absence of restraints yields consistently about 8% correctly folded
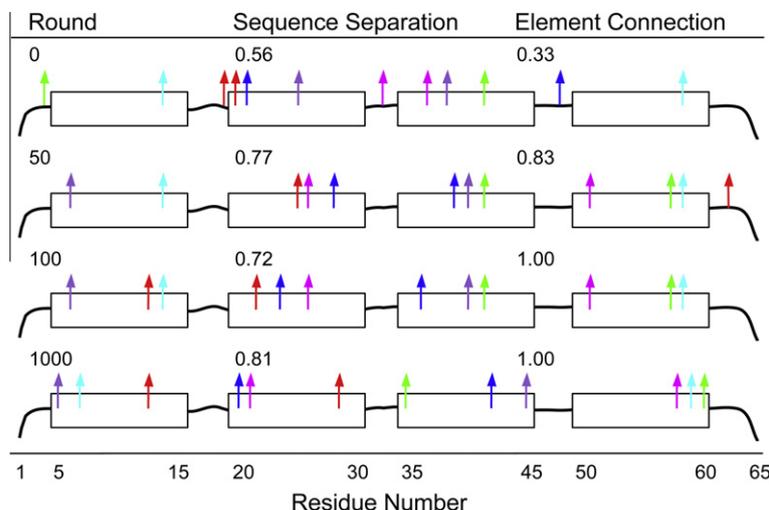
Fig. 2. An example optimization trajectory. In our simplified four helix, 6-restraint system, each arrow represents a spin label and arrows of the same color correspond to spin label pairs. Starting from an initial random distribution, the algorithm repositions the spin labels to maximize sequence separation and element connection scores resulting in an optimized pattern.

models leaving sufficient dynamic range to evaluate the impact of EPR restraints.

3.2. Optimized restraints increase the fraction of correct topology models in Rosetta folding

Following selection of the terms and their relative weights described above, we assessed the degree to which optimized restraint patterns improve the quality of T4L models predicted by Rosetta. For this purpose, 10 sets of 21 restraints were used in conjunction with Rosetta to generate 1000 T4L models. An equivalent number of models was generated by folding without restraints as well as in the presence of 21 randomly selected restraints. Consistently, models obtained using optimized restraints had vastly better quality measures (Fig. 3). A left shift in the RMSD distribution reflects the presence of a major population of models with RMSD below 7.5 Å (Fig. 3A). It is generally accepted that 7.5 Å is the RMSD at which models have the correct overall fold as the native structure (Zemla et al., 1999). Thus using optimized restraints, 54.4% of Rosetta models achieve the general fold compared to 21.0% and

8.0% of models if randomized or no restraints are used, respectively (Fig. 3B).

Optimized restraints also lead to a significant increase in the percentage of models with C_α-RMSDs below 3.5 Å reflecting more effective sampling of conformational space by Rosetta (Fig. 3C). These models being closest to the native structure are ideal candidates for subsequent high resolution refinement (Alexander et al., 2008). Using an RMSD cutoff of 3.5 Å as a criterion, 1.7% of models generated by incorporation of optimized restraints are considered high quality. To achieve 1 Å resolution, a starting set of at least 2000 such models are needed (Bradley et al., 2005b), which is within a computational reasonable time frame. In contrast, only 0.2% of models generated using randomized restraints fulfilled the 3.5 Å RMSD criterion. Thus, to achieve high resolution, one million models are needed which requires substantially more computational resources. If no restraints are used, the computational cost becomes prohibitive, as only 0.04% percent of models have less than 3.5 Å RMSD, therefore requiring tens of millions of models. Furthermore, EPR restraints allow selection of correct topology models for refinement (Alexander et al., 2008).
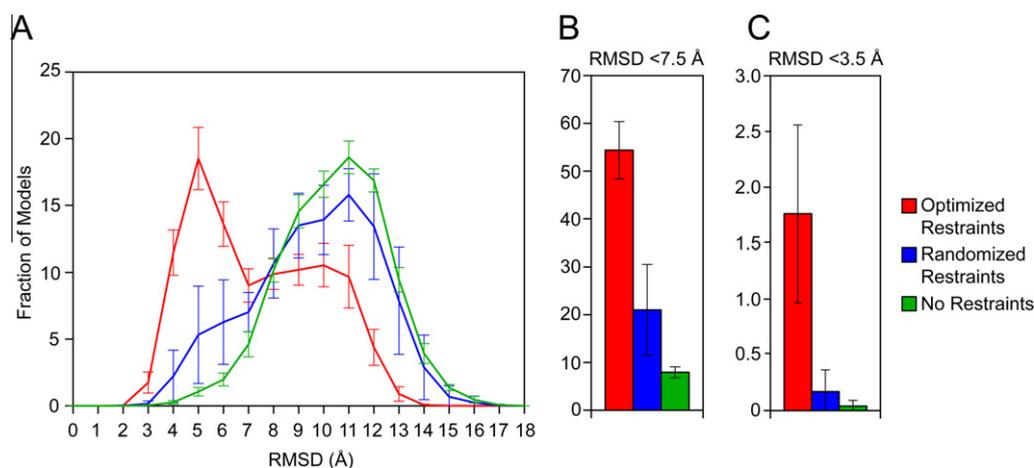

Fig. 3. Effects of incorporation of restraints into Rosetta folding. (A) C_α-RMSD distributions of models generated with optimized restraints, randomized restraints, and no restraints. The distinct left shift in the distribution with optimized restraints is indicative of a more efficient restriction of the conformational search space. Fraction of models fulfilling two measures of model quality, C_α-RMSD < 7.5 Å (B) and C_α-RMSD < 3.5 Å (C) display even more pronounced improvements. Models with RMSD < 7.5 Å to the crystal structure typically have the correct general fold, while models with RMSD < 3.5 Å are considered candidates for high resolution refinement.

3.3. Improvement of model quality requires a limited number of EPR distance restraints

The choice of 21 restraints above was dictated by detailed analysis of the dependence of model quality on restraint number. For this purpose, the Rosetta folding protocol of Fig. 1 was applied successively increasing the number of restraints followed by assessment of model quality. Note that 21 restraints are required to fulfill all pairwise connections between the seven helices of T4L C-terminal domain. Therefore for restraint numbers larger than 21, the algorithm was modified to ensure that the additional restraints duplicating existing secondary structure connectivities are evenly distributed.

Fig. 4A demonstrates that increasing the number of restraints leads to a rapid increase in the percentage of models having the correct fold (C_α-RMSD below 7.5 Å). This effect is pronounced with as few as 5–10 restraints. The trend levels off in the region of 20–22 restraints suggesting that redundant connections between secondary structures add little information (Fig. 4A). In contrast, a more stringent quality measure, the percentage of models with RMSD below 3.5 Å, hardly improves until the number of restraints is well above 10 (Fig. 4B). This lag reflects the significantly lower probability that these models are sampled in the absence of restriction on the search space. Indeed, additional randomized restraints
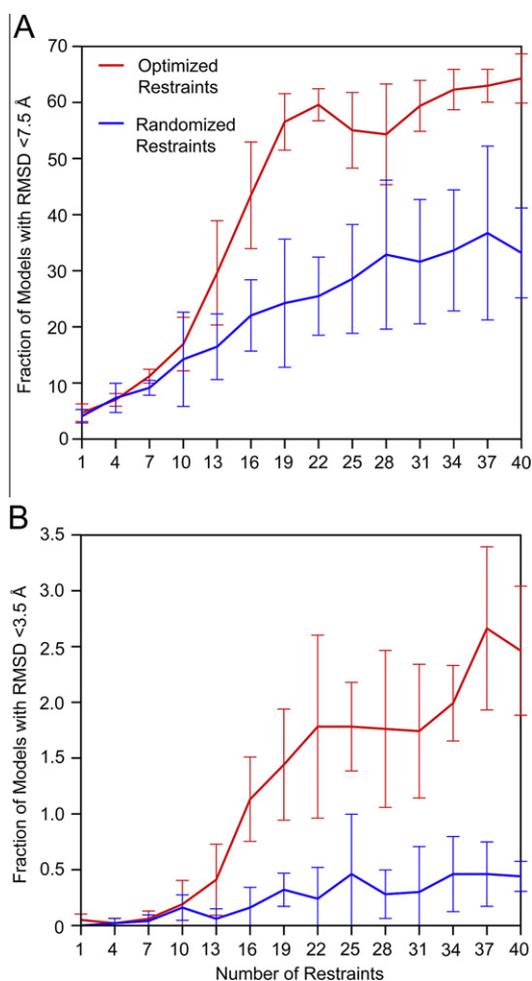
do not appreciably affect the percentage of high quality models for this reason.

The percentage of models with the correct fold plateaus at approximately 60%. The rest typically fulfill the restraints but have incorrect folds. This is not surprising given the soft interpretation of the restraints within a wide error margin (15 Å) by the cone model. It is likely that this limitation also accounts for the relatively limited percentage of high quality models, i.e. with RMSD below 3.5 Å. Interestingly, the incorrectly folded models score worse in Rosetta's knowledge-based potential (see below) allowing for selection of correctly folded models by energy score. In addition, the improved overall quality with few restraints (Fig. 4) provides a plausible explanation for the surprisingly good performance of random restraints in Fig. 3.

3.4. Rosetta Folding of T4L using optimized experimental restraints

The optimization of the algorithm used simulated distances between residue pairs. As described above, this approximation centers the distribution at the $d_{C\beta}$ while the experimental distribution is centered on the distance separating the two spin labels. The offset between these two values is determined by the relative orientation of the labels and represents the major source of uncertainty in interpretation of EPR distance restraints. To assess the consequences of this approximation and validate the optimization strategy, we carried out Rosetta folding of T4L using experimentally determined distances for a set of spin label pairs selected by the algorithm described above. Double cysteine mutants were constructed and the corresponding proteins purified and spin labeled as described in Section 2. Most pairs, except two, were in the distance range suitable for DEER analysis (Jeschke, 2002). Spin echo decays were baseline-corrected and analyzed by Tikhonov regularization (Chiang et al., 2005) to yield distance distributions as described in Section 2 and illustrated in Appendix A. For the short range pairs (86/112 and 127/155), spectral simulation was used to extract a Gaussian distribution of distance from the CW-EPR spectra (Appendix A).

The position of the pairs is mapped onto the T4L crystal structure in Fig. 5. Table 1 reports the average distance between the spin
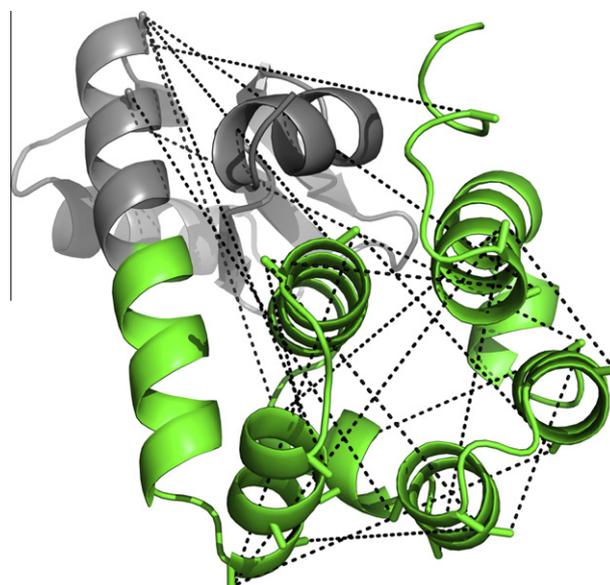

Fig. 4. Impact of restraint number on model quality. Model quality measures, C_α-RMSD < 7.5 Å (A) and C_α-RMSD < 3.5 Å (B) improve with increasing number of restraints. The improvement levels off at a threshold number where the α-helical pairwise connectivities are fulfilled (20–22 restraints for the seven helices of T4L C-terminal domain).

Fig. 5. Location of experimentally measured EPR restraints in the T4L crystal structure. T4L is colored to highlight the subdomains included in (green) and excluded from (silver) RMSD calculations. The dotted lines represent the pairs of spin labels in the restraints.

Table 1
Average distance and distribution width for experimentally measured restraints.

Mutants	Average distance ± distribution width (Å)
59/159	41.9 ± 2.7
60/90	37.8 ± 4.5
60/94	25.5 ± 3.1
60/109	35.2 ± 2.6
60/154	34.1 ± 2.0
62/134	41.1 ± 1.5
64/122	34.1 ± 2.5
82/94	30.7 ± 3.3
82/132	26.3 ± 3.5
82/155	35.8 ± 2.5
83/123	20.5 ± 3.4
93/112	26.1 ± 1.5
93/123	24.8 ± 2.3
93/154	25.1 ± 2.4
94/132	31.7 ± 1.3
108/155	35.2 ± 2.3
109/134	30.6 ± 2.8
115/155	28.2 ± 2.4
116/134	20.2 ± 1.5
86/112[a]	13.0 ± 5.1
127/155[a]	12.1 ± 3.4

[a] Determined by simulation of CW spectra with CWdipFit. All other distances were obtained from Tikhonov analysis of DEER signals.

label pairs as well as the width of the distance distribution. Compared to the $d_{C\beta}$, the deviations show the expected pattern of larger spin label distances (data not shown). The distributions are predominantly narrow despite the surface exposed location of the spin labeled sites. Thus, even though most spin labels are mobile as evidenced by the EPR lineshapes (data not shown), it appears that the sampled rotameric states are restricted.

Fig. 6 demonstrates that Rosetta folding of 10,000 models using the experimental distances leads to improvement in model quality measures that follow the same trends of Fig. 3. These include a left shift in the C_α-RMSD distribution, an increase in the fraction of models with the correct folding topology (RMSD < 7.5 Å), and more importantly of the percentage of high quality models (RMSD < 3.5 Å). However, these improvements underperform those expected from simulated distances. The origin of this underperformance can be rationalized by comparing the upper bound of the simulated and experimental restraints. Experimentally determined distances tend to be larger than the $d_{C\beta}$ thereby increasing the upper bound. Thus, conformational space is less restrained leading to a reduced model quality.

The models generated by incorporation of experimental restraints into Rosetta folding were sorted based on their Rosetta energy and restraint violation scores, as defined in Section 2. While models of vastly different RMSDs have similar Rosetta energy or restraint violation scores, only models with low RMSDs have low scores in both criteria. Fig. 7 demonstrates the improvement in model quality when a Rosetta energy score threshold of below 30 and a restraint violation score threshold of less than 2.5 were applied. This resulted in an enrichment factor of 7.2 for models with RMSDs below 3.5 Å, retaining 44 of the 61 original models. Thus the combination of these two scores can identify the subset of models with topologies closest to the native structure.

4. Discussion

The requirement for incorporation of spin labels into protein sequences shapes the methodology of spin labeling in two fundamental ways. First, the experimental throughput is limited leading to sparse restraints. Second, the arm linking the spin label to the protein backbone introduces an uncertainty in the interpretation of these restraints. The algorithm presented in this paper advances the methodological blueprint of spin labeling and EPR spectroscopy by optimizing the information content of EPR distance restraints and consequently alleviating the experimental bottleneck.

The experimental implementation of this strategy presented here charts a roadmap for future improvements. As expected, using the cone model of Alexander et al. (2008) for interpretation of the EPR distances significantly compromises the quality of the experimental data. Narrow distance distributions at a number of sites imply a tighter limit on the distance range than the 15 Å assumed in the cone model. Furthermore, the shape of the distribution (Appendix A) is in stark contrast to the flat scoring potential implemented in the Rosetta protocol. The consequence of these approximations is that topologies with C_α-RMSDs as large as 12 Å fulfill the EPR restraints. A probability function to describe the offset distribution between the d_{SL} and $d_{C\beta}$ has been developed to address this (Hirst et al., this issue). Furthermore, explicit modeling of the spin label should limit the uncertainty associated with unknown spin label orientation to the backbone. It has been demonstrated that
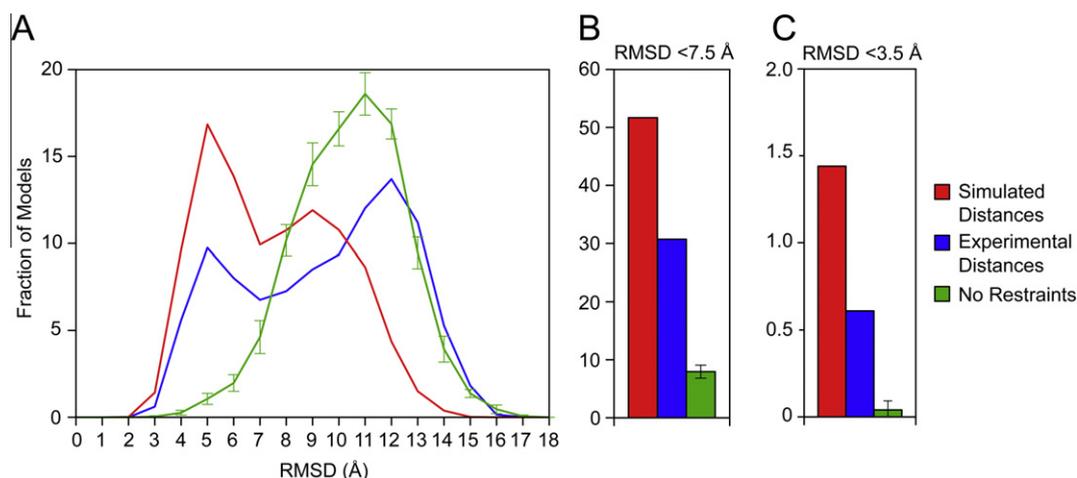

Fig. 6. Comparison of models generated using experimental and simulated restraints. Simulated and experimentally measured distances (Table 1) for the optimized pattern shown in Fig. 5 were used to fold T4L using Rosetta. The resulting C_α-RMSD distributions (A) and quality measures (B and C) were derived from 10,000 models for each set of distances (simulated and experimental). The no restraints curve represents 10 Rosetta trajectories each of 1000 models. The difference between the simulated and experimental results highlights the effect of the typically larger upper bound associated with experimentally determined distances.
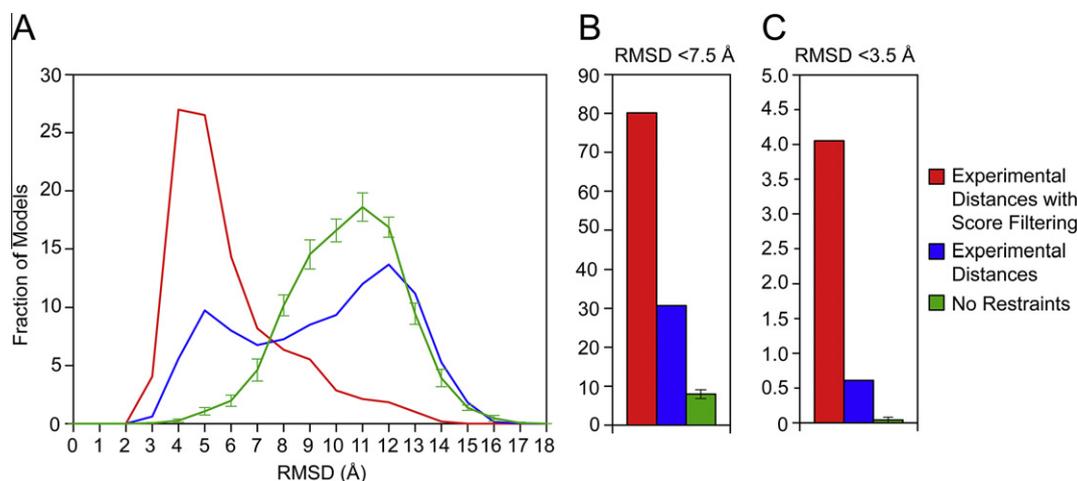
Fig. 7. Effects of filtering by Rosetta energy and restraint violation scores on model quality. (A) RMSD distributions and (B and C) model quality measures shown in Fig. 6 after score filtering. Filtering by Rosetta energy score and restraint violation score excludes almost exclusively low quality models.

molecular dynamics simulations can reproduce average distances between spin labels (Borovykh et al., 2006; Jeschke and Polyhach, 2007). Though more computationally intensive, these approaches will enhance Rosetta models quality specifically increasing the fraction of those below 3.5 Å RMSD.

The performance of the algorithm is also degraded when prediction rather than actual secondary structures are used. The origin of this effect is the inaccurate prediction of the number of secondary structures which for a fixed number of restraints alter the required pairwise connectivities. In the context of the application of this approach to a protein of unknown structure, the location and length of secondary structure can be experimentally determined and/or verified through nitroxide scanning experiments (Berengian et al., 1997; Mansoor et al., 1999; Kaplan et al., 2000).

That the many approximations did not hinder the identification of the correct fold by Rosetta reflects the robustness of its energy function. Similarly, a few EPR restraints lead to a measurable improvement in the quality of the folds highlighting the critical role of these restraints in reducing Rosetta's conformational search space. These findings reinforce the rationale of using *de novo* folding to balance the sparseness of the EPR restraints and their intrinsically lower quality.

Although the algorithm developed in this paper is general, our ultimate goal is to develop a suite of tools to determine structure of membrane proteins. While Rosetta has been successfully used to generate constrained models of membrane proteins (Yarov-Yarovoy et al., 2006; McAllister and Floudas, 2008; Barth et al., 2009), it is likely to be less robust given the limited number of folds and topologies in the Protein Data Bank. Though this may be partially mitigated by the restricted diversity of membrane protein fold imposed by the membrane environment, the number of EPR restraints needed to obtain high quality models is likely to be larger. Furthermore, the rule of one restraint per pair of secondary structures may have to be modified for the longer helices found in these proteins. In this context, redundant restraints may prove important for longer helices common in transmembrane proteins. We expect that additional algorithm terms to optimize the distribution of redundant restraints will be developed. Nevertheless, this algorithm represents a first step in this direction.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jsb.2010.11.003.

References

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

Altenbach, C., Kusnetzow, A.K., Ernst, O.P., Hofmann, K.P., Hubbell, W.L., 2008. High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. Proc. Natl. Acad. Sci. USA 105, 7439–7444.

Barth, P., Wallner, B., Baker, D., 2009. Prediction of membrane protein structures with complex topologies using limited constraints. Proc. Natl. Acad. Sci. USA 106, 1409–1414.

Berengian, A.R., Bova, M.P., Mchaourab, H.S., 1997. Structure and function of the conserved domain in alphaA-crystallin. Site-directed spin labeling identifies a beta-strand located near a subunit interface. Biochemistry 36, 9951–9957.

Bhatnagar, J., Borbat, P.P., Pollard, A.M., Bilwes, A.M., Freed, J.H., Crane, B.R., 2010. Structure of the ternary complex formed by a chemotaxis receptor signaling domain, the CheA histidine kinase, and the coupling protein CheW as determined by pulsed dipolar ESR spectroscopy. Biochemistry 49, 3824–3841.

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

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

Borbat, P.P., Mchaourab, H.S., 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.

Borbat, P.P., Davis, J.H., Butcher, S.E., Freed, J.H., 2004. Measurement of large distances in biomolecules using double-quantum filtered refocused electron spin-echoes. J. Am. Chem. Soc. 126, 7746–7747.

Borovykh, I.V., Ceola, S., Gajula, P., Gast, P., Steinhoff, H.J., Huber, M., 2006. Distance between a native cofactor and a spin label in the reaction centre of *Rhodobacter sphaeroides* by a two-frequency pulsed electron paramagnetic resonance method and molecular dynamics simulations. J. Magn. Reson. 180, 178–185.

Bowers, P., Strauss, C., Baker, D., 2000. De novo protein structure determination using sparse NMR data. J. Biomol. NMR 18, 311–318.

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

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

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

Claxton, D.P., Zou, P., Mchaourab, H.S., 2008. Structure and orientation of T4 lysozyme bound to the small heat shock protein alpha-crystallin. J. Mol. Biol. 375, 1026–1039.

Das, R., Baker, D., 2008. Macromolecular modeling with rosetta. Annu. Rev. Biochem. 77, 363–382.

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

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

Fajer, M.I., Li, H., Yang, W., Fajer, P.G., 2007. Mapping electron paramagnetic resonance spin label conformations by the simulated scaling method. J. Am. Chem. Soc. 129, 13840–13846.

Fanucci, G.E., Coggshall, K.A., Cadieux, N., Kim, M., Kadner, R.J., Cafiso, D.S., 2003. Substrate-induced conformational changes of the periplasmic N-terminus of an outer-membrane transporter by site-directed spin labeling. Biochemistry 42, 1391–1400.

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. J. Mol. Biol. 298, 261–272.

Hanelt, I., Wunnicke, D., Muller-Trimbusch, M., Vor der Bruggen, M., Kraus, I., Bakker, E.P., Steinhoff, H.J., 2010. Membrane region M2C2 in subunit KtrB of the K^+ uptake system KtrAB from Vibrio alginolyticus forms a flexible gate controlling K^+ flux: an electron paramagnetic resonance study. J. Biol. Chem. 285, 28210–28219.

Hilger, D., Polyhach, Y., Padan, E., Jung, H., Jeschke, G., 2007. High-resolution structure of a Na^+/H^+ antiporter dimer obtained by pulsed electron paramagnetic resonance distance measurements. Biophys. J. 93, 3675–3683.

Hilger, D., Polyhach, Y., Jung, H., Jeschke, G., 2009. Backbone structure of transmembrane domain IX of the Na^+/proline transporter PutP of Escherichia coli. Biophys. J. 96, 217–225.

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

Hubbell, W.L., Gross, A., Langen, R., Lietzow, M.A., 1998. Recent advances in site-directed spin labeling of proteins. Curr. Opin. Struct. Biol. 8, 649–656.

Jao, C.C., Hegde, B.G., Chen, J., Haworth, I.S., Langen, R., 2008. Structure of membrane-bound alpha-synuclein from site-directed spin labeling and computational refinement. Proc Natl Acad Sci U S A 105, 19666–19671.

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

Jeschke, G., Polyhach, Y., 2007. Distance measurements on spin-labelled biomacromolecules by pulsed electron paramagnetic resonance. Phys. Chem. Chem. Phys. 9, 1895–1910.

Jeschke, G., Koch, A., Jonas, U., Godt, A., 2002. Direct conversion of EPR dipolar time evolution data to distance distributions. J. Magn. Reson. 155, 72–82.

Jeschke, G., Wegener, C., Nietschke, M., Jung, H., Steinhoff, H.J., 2004. Interresidual distance determination by four-pulse double electron–electron resonance in an integral membrane protein: the Na^+/proline transporter PutP of Escherichia coli. Biophys. J. 86, 2551–2557.

Kaplan, R.S., Mayor, J.A., Kotaria, R., Walters, D.E., Mchaourab, H.S., 2000. The yeast mitochondrial citrate transport protein: determination of secondary structure and solvent accessibility of transmembrane domain IV using site-directed spin labeling. Biochemistry 39, 9157–9163.

Koteiche, H.A., Mchaourab, H.S., 1999. Folding pattern of the alpha-crystallin domain in alphaA-crystallin determined by site-directed spin labeling. J. Mol. Biol. 294, 561–577.

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

Mansoor, S.E., Mchaourab, H.S., Farrens, D.L., 1999. Determination of protein secondary structure and solvent accessibility using site-directed fluorescence labeling. Studies of T4 lysozyme using the fluorescent probe monobromobimane. Biochemistry 38, 16383–16393.

Margittai, M., Langen, R., 2006. Spin labeling analysis of amyloids and other protein aggregates. Methods Enzymol. 413, 122–139.

Matthews, B.W., 1996. Structural and genetic analysis of the folding and function of T4 lysozyme. FASEB J. 10, 35–41.

McAllister, S.R., Floudas, C.A., 2008. Alpha-helical topology prediction and generation of distance restraints in membrane proteins. Biophys. J. 95, 5281–5295.

McGuffin, L.J., Bryson, K., Jones, D.T., 2000. The PSIPRED protein structure prediction server. Bioinformatics 16, 404–405.

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., Oh, K.J., Fang, C.J., Hubbell, W.L., 1997. Conformation of T4 lysozyme in solution. Hinge-bending motion and the substrate-induced conformational transition studied by site-directed spin labeling. Biochemistry 36, 307–316.

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.

Padmavathi, P.V., Steinhoff, H.J., 2008. Conformation of the closed channel state of colicin A in proteoliposomes: an umbrella model. J. Mol. Biol. 378, 204–214.

Persson, M., Harbridge, J.R., Hammarstrom, P., Mitri, R., Martensson, L.G., Carlsson, U., Eaton, G.R., Eaton, S.S., 2001. Comparison of electron paramagnetic resonance methods to determine distances between spin labels on human carbonic anhydrase II. Biophys. J. 80, 2886–2897.

Petersen, B., Petersen, T.N., Andersen, P., Nielsen, M., Lundegaard, C., 2009. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct. Biol. 9, 51.

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.

Rohl, C.A., 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., Baker, D., 2004. Protein structure prediction using Rosetta. Methods Enzymol. 383, 66–93.

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. J. Am. Chem. Soc. 127, 9334–9335.

Sen, K.I., Logan, T.M., Fajer, P.G., 2007. Protein dynamics and monomer-monomer interactions in AntR activation by electron paramagnetic resonance and double electron-electron resonance. Biochemistry 46, 11639–11649.

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. J. Mol. Biol. 268, 209–225.

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

Smirnova, I., Kasho, V., Choe, J.Y., Altenbach, C., Hubbell, W.L., Kaback, H.R., 2007. Sugar binding induces an outward facing conformation of LacY. Proc. Natl. Acad. Sci. USA 104, 16504–16509.

Sugata, K., Song, L., Nakamura, M., Ueki, S., Fajer, P.G., Arata, T., 2009. Nucleotide-induced flexibility change in neck linkers of dimeric kinesin as detected by distance: measurements using spin-labeling EPR. J. Mol. Biol. 386, 626–636.

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

Wegener, C., Tebbe, S., Steinhoff, H.J., Jung, H., 2000. Spin labeling analysis of structure and dynamics of the Na(+)/proline transporter of Escherichia coli. Biochemistry 39, 4831–4837.

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

Zemla, A., Venclovas, C., Moult, J., Fidelis, K., 1999. Processing and analysis of CASP3 protein structure predictions. Proteins Suppl. 3, 22–29.

Zou, P., Mchaourab, H.S., 2009. Alternating access of the putative substrate-binding chamber in the ABC transporter MsbA. J. Mol. Biol. 393, 574–585.

Zou, P., Bortolus, M., Mchaourab, H.S., 2009. Conformational cycle of the ABC transporter MsbA in liposomes: detailed analysis using double electron-electron resonance spectroscopy. J. Mol. Biol. 393, 586–597.