PROTEIN STRUCTURE ELUCIDATION FROM COMPUTATIONAL TECHNIQUES AND SPARSE EPR DATA

By

Nathan Scott Alexander

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry

August, 2012

Nashville, Tennessee

Approved:

Professor Jens Meiler

Professor Heidi E. Hamm

Professor Terry P. Lybrand

Professor Michael P. Stone

ACKNOWLEDGEMENTS

This research was made possible by funding from the National Institute of Mental Health through a National Research Service Award Individual Fellowship. Funding was also provided by the National Institute of Health Molecular Biophysics Training Grant at Vanderbilt University. Computational resources were expertly provided by the Advanced Computing Center for Research and Education and the Center for Structural Biology at Vanderbilt University.

Thank you to my advisor, Jens Meiler, for the mentoring provided during my time in the lab. Jens allowed the freedom to explore the scientific process but his guidance was readily available to teach the questions that should be asked and foster the ability to answer them.

I would also like to thank other members of the faculty who aided in my scientific progress while at Vanderbilt University including Dr. Hassane Mchaourab, Dr. Tina Iverson, and my dissertation committee Dr. Heidi Hamm, Dr. Terry Lybrand, and Dr. Michael Stone.

Many hours were spent working with friends in the Meiler lab. Thank you for the help you provided to me and the enjoyable activities.

Thank you to my family and, especially, my ever-patient wife. I appreciate it.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The overall focus of this dissertation was to develop computational methods that allow application of electron paramagnetic resonance (EPR) spectroscopy data for protein structure prediction. Also, a main goal of this dissertation was to use these methods to study a protein system of biological significance. Chapter I provides an introduction to protein structure prediction and the relationship of EPR to structural biology. It also provides background on the target protein system of interest for applying the developed methods. Chapter I was written for this dissertation.

Chapter II details the development and rational of a model that allows EPR distance information to be incorporated into structure prediction methods. Chapter II is based on the publication entitled “De novo high-resolution protein structure determination from sparse spin-labeling EPR data” by Nathan Alexander, Marco Bortolus, Ahmed Al-Mestarihi, Hassane Mchaourab, and Jens Meiler.

Chapter III describes the development of a rotamer library and its incorporation with Rosetta allowing EPR distance data to be utilized to atomic detail. It is based on a manuscript in preparation entitled “RosettaEPR: Rotamer library for spin label structure and dynamics” by Nathan Alexander, Richard Stein, Kristian Kaufmann, Hassane Mchaourab, and Jens Meiler.

Chapter IV details the application of the methods developed in Chapter II to investigating the overall structure of a GPCR in complex with a G-protein. It is based on the manuscript entitled “Interaction of a G protein with an activated receptor opens the interdomain interface in the alpha subunit” by Ned Van Eps, Anita Preininger, Nathan Alexander, Ali Kaya, Scott Meier, Jens Meiler, Heidi Hamm, and Wayne Hubbell. The first three authors contributed equally to the overall body of work, but this dissertation contributed to the computational aspects.

Chapter V describes the application of developed computational methods to refine the structure of the GPCR bound to a G-protein. It is based on the manuscript in preparation entitled “A rotation of the C-terminal helix connects binding of the Gi protein and activated receptor to disassociation of helical domain and GDAP release” by Nathan Alexander, Anita Preininger, Ali Kaya, Heidi Hamm, and Jens Meiler. Current contributions to the text from this thesis include methodological descriptions.

Chapter VI describes the application of EPR distance and accessibility data for use in membrane protein structure prediction. It is based on a manuscript in preparation entitled “EPR restraint guided membrane protein structure prediction with bcl::Fold” by Nathan Alexander, Nils Woetzel, Mert Karakas, and Jens Meiler.

Chapter VII was written for this dissertation. It provides major conclusions of the work and the relation of the findings to other work in the field.

The Appendix provides details for protocols not provided in the chapters. The protocol for clustering analysis is given first. This clustering method was used in Chapter IV. Second, the protocol for sampling and finding ensembles of models that fulfill an experimental EPR distance distribution is given. This protocol was used in Chapter V. A guide to membrane protein folding with EPR restraints in the BCL is given, relating to chapter VI. Lastly, the methods used for modeling the receptor-bound conformation of visual arrestin are described. The text is based on the methods and supplemental information sections of the manuscript entitled “The conformation of receptor-bound visual arrestin.”

CHAPTER I

INTRODUCTION

Protein structure prediction methods

The goal of protein structure prediction is to predict the three-dimensional conformation from that protein's amino acid sequence. The conformational flexibility in a protein arises from the rotatable bonds designated as φ and ψ within the amino acids that comprise the protein. Predicting the structure of an amino acid sequence is non-trivial, as demonstrated by Levinthal's paradox. Consider a protein made up of 100 amino acids, which would be considered a small protein, as the median number of residues in the eukaryotic proteome is 361 (Brocchieri and Karlin). Each of the 100 residues in the protein has two rotatable bonds, with an assumed rotational sampling ability of 10°, giving 36 possible rotational conformations for each bond. Now in order to exhaustively sample all possible conformations for the protein would give 72^{100} combinations. Such a number to search through is intractable, but proteins fold biologically in the order of seconds. By introducing energetic considerations into the folding process, the number of conformations and time for folding are reduced to a biological magnitude (Zwanzig, Szabo et al. 1992). Appropriately, protein structure prediction methods can be broken down into two primary components. The first component is the method for evaluating the physical correctness of a given protein conformation. The second component is the strategy for sampling possible conformations of a protein.

Two strategies for evaluating protein conformations are typically used in structure prediction. Evaluation consists of calculating the free energy of a given protein conformation. Energetic scores derived directly from physical properties of a protein

provide the most accurate methods for calculating the energy of a protein (Karplus and McCammon 2002). The downside to such potentials is that they are computationally intensive to compute. An alternative method for protein structure evaluation is to leverage empirically observed information about protein structures (Sippl 1995). More than 80,000 protein structures are currently deposited in the Protein Data Bank (PDB) (Berman, Westbrook et al. 2000). Statistics can be conducted over these proteins in order to extract probabilities of observing specific structural properties within a protein. Some of these properties include preferred amino acid exposures, amino acid pair interactions, and protein compactness, or radius of gyration (Simons, Kooperberg et al. 1997). These knowledge-based potentials are not as accurate as physically derived potentials. However, the knowledge-based potentials are computationally quick to calculate.

Conformational sampling of protein structures is the second component to protein structure prediction. In order to predict the structure of a protein, the amino acids must transition from a starting configuration into the conformation encoded in the amino acid sequence (Kuhlman and Baker 2000). Molecular dynamics is one type of methodology for sampling protein conformations. Conformations are determined by solving Newton's equations of motion to determine the direction and velocity of atoms across a time period (Karplus and McCammon 2002). In order to simulate protein folding, molecular dynamics simulations need to sample for timeframes of millisecond to second, while molecular dynamics simulations typically are limited to up to microseconds (Fenwick, Esteban-Martin et al. 2011). However, success in folding small peptides of more than 20 residues has been demonstrated (Daura 2006). In contrast to molecular dynamics sampling, Monte Carlo (Metropolis 1953) based sampling of protein structures does not require any physically realistic relationship between one conformation and the next. In the stochastic sampling of Monte Carlo methods, a given protein structure is perturbed in

some manner in order test if the new conformation is more energetically favorable than the previous. Defining appropriate perturbations is essential to maximize the efficiency of sampling (Dinner 2000; Ulmschneider and Jorgensen 2003) (Simons, Kooperberg et al. 1997; Karakas, Woetzel et al. 2012). The advantage of stochastic sampling approaches over molecular dynamics is that the protein can rapidly converge into a biologically relevant conformation. The most efficient type of protein structure sampling is template-based modeling. Template based protein structure prediction uses a protein of known structure with a sequence similar to the target protein sequence (Zhang 2008). The structure of the known protein is then imposed on the target sequence. Template based modeling takes advantage of the the hypothesis that similar sequences will give similar structures (Kuhlman and Baker 2000). The limitation of template based modeling is highlighted when templates for a target are not easily identified.

Protein structure prediction coupled with experimental data

In spite of advances in protein structure prediction methods, the successful prediction of a protein structure from sequence remains limited to proteins of approximately 150 residues or less (Yarov-Yarovoy, Schonbrun et al. 2006) (Zhang 2008). The Rosetta protein structure prediction method has had continued success at *de novo* prediction of proteins, and is frequently one of the most accurate methods during the Critical Assessment of Techniques for Protein Structure Prediction (CASP) (Raman, Vernon et al. 2009). Rosetta uses a fragment-based assembly method in conjunction with simulated annealing to predict protein structures from sequence (Rohl and Baker 2002). Peptide fragments of lengths 9 and 5 amino acids are selected from proteins of known structure. The peptide fragments are similar in sequence and secondary structure content to that predicted for the target sequence. The proteins the fragments are taken from are not necessarily homologous to the target protein. This means that protein

structures can be predicted for proteins where templates are not available (Raman, Vernon et al. 2009). During the Rosetta protein structure prediction protocol, the sequence starts out as an extended chain in space. All portions of the sequence are then replaced by fragments of varying conformations in order to produce a native-like protein conformation. To evaluate conformations of the protein, Rosetta used knowledge-based potentials derived from the PDB (Rohl and Baker 2002).

In order to improve the ability of protein structure methods, experimental data can be coupled with the algorithms. This has been demonstrated using medium-resolution cryo-electron microscopy density maps in conjunction with secondary structure element based sampling of possible protein topologies (Lindert, Staritzbichler et al. 2009; Lindert, Alexander et al. 2012). The method can consistently predict protein structures to under 3 Å root mean square deviation (RMSD) from the experimental structure. The combination of nuclear magnetic resonance (NMR) and electron paramagnetic resonance (EPR) spectroscopic data was used for the prediction of the soluble homo-dimer protein Dsy0195 (Yang, Ramelot et al. 2010). NMR chemical shifts have been used in conjunction with molecular mechanics force fields to determine the structures of 11 proteins of up to 123 amino acids to an accuracy of 2 Å RMSD (Cavalli, Salvatella et al. 2007). This method utilized fragment replacement sampling similar to Rosetta. The incorporation of unassigned NMR data such as chemical shifts, NOEs and residual dipolar coupling information allowed models to be predicted to an accuracy of up to 3 Å RMSD (Meiler and Baker 2003). Although these studies did not push the boundaries of protein size during structure prediction they demonstrated the utility of combining experimental data with protein structure prediction. In 2010, the Rosetta protein structure prediction method was coupled with backbone-only NMR data to accurately predict the structures for proteins of up to 200 residues. Such experimental dataset is considered

insufficient for structure determination by classic NMR methodology (Raman, Lange et al.).

Alternative methods for probing membrane protein structure

X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the standard techniques for determining protein structures. Membrane proteins are particularly challenging to these two techniques. Membrane proteins are frequently difficult to purify in large enough quantities to allow characterization by X-ray crystallography or NMR (Bill, Henderson et al. 2011). Membrane protein stability presents another bottle-neck for producing crystals for use with X-ray crystallography. Further, membrane proteins typically large size frequently inhibits widespread use of NMR for structure determination (Kim, Howell et al. 2009; Kang and Li 2011).

Techniques other than X-ray crystallography and NMR are therefore frequently used to investigate aspects of membrane protein structure. Introducing two fluorescent probes into specific residue sites in a protein can provide distance measurements within a (Zou, Surendhran et al. 2007). This can provide distances between 20 Å and 50 Å. Additionally, single fluorescent probes can be used to report on the exposure of positions in membrane proteins (Shinitzky and Rivnay 1977). This provides information on the local environment around the fluorescent probe. Fluorescence measurements are made at room temperature, so dynamic fluctuations of the protein are captured within the measurement protein (Borbat, Surendhran et al. 2007). Small angle X-ray scattering (SAX) can also be performed on proteins in native-like environments. SAX can provide envelopes describing the overall shape of a protein (Francis, Różycki et al. 2011). The SAX data can be used to find ensembles of protein structures which reproduce the SAX curve. Cryo-electron microscopy (CryoEM) can be used to obtain electron density maps at medium resolutions (Lindert, Stewart et al. 2009). At medium resolutions of 5 – 10 Å,

helices within proteins are resolved as cylindrical “density rods”. Such density maps can show the placement of helices relative to one another and therefore the overall architecture of the protein. However, the sequential connectivity of the helices is not easily determined, as the connecting loop residues will not be resolved (Fleishman, Harrington et al. 2006).

Electron paramagnetic resonance spectroscopy is another technique that can be used to probe membrane protein structure. It can provide distances within a protein of up to 60 Å (Borbat, McHaourab et al. 2002) and measurements have been demonstrated of up to 80 Å (Jeschke and Polyhach 2007). The collision frequency of single unpaired electron probes (spin labels) with either molecular oxygen or NiEDDA can be used to measure the exposure of a sequence position to the surrounding environment (Altenbach, Marti et al. 1990; Koteiche, Berengian et al. 1998; Koteiche and McHaourab 1999; Zou and McHaourab 2009). Pairing EPR with site directed spin labeling allows the targeted investigation of specific residues within the protein (Hubbell and Altenbach 1994). For membrane proteins, the combination of measured oxygen and NiEDDA accessibilities allows the determination of residue depth within the membrane (Altenbach, Greenhalgh et al. 1994).

These techniques all have the advantage that they can readily be used to make structural measurements on membrane proteins. The limitations of the approaches arise from the fact that the data gleaned from them cannot be used to unambiguously define the structure of the target protein.

Methods for membrane protein structure prediction

Membrane protein structure prediction techniques follow the same principles as soluble protein structure prediction (Frishman and Barth 2010). However, specific considerations for membrane proteins must be taken into account. Favorably for

structure prediction, the membrane does exhibit some constraints on the orientation of segments that must pass through it. Unfortunately, membrane proteins are typically larger than soluble a protein, which dramatically increases the conformational space. In addition, the membrane introduces a more complex set of environment for residues to reside within. Therefore, membrane specific potentials are needed (Yarov-Yarovoy, Schonbrun et al. 2006). These potentials can include altered residue specific environment scores, radius of gyration scores specific for membrane proteins, and scores favoring placement of transmembrane helices orthogonally to the membrane.

In order to be successful in predicting membrane protein structures, it is important to properly determine the segments which span the membrane. Methods are available for predicting transmembrane spanning segments from the protein sequence. One method is TMMOD (Kahsay, Gao et al. 2005). TMMOD uses hidden markov models to predict helical regions of the protein sequence that span the membrane. TMMOD consistently achieved an accuracy of over 80%. Another tool is Octopus (Viklund and Elofsson 2008), which uses artificial neural networks to predict the helical transmembrane spanning topology of membrane proteins. Octopus achieves an accuracy of 94% across 124 sequences with known structure. Membrane protein structure prediction methods can take advantage of these highly accurate topology predictions. The Rosetta structure prediction algorithm was adapted for membrane proteins with specific potentials and sampling strategies and uses Octopus predictions to help identify transmembrane segments (Yarov-Yarovoy, Schonbrun et al. 2006). Rosetta achieved predictions with accuracy of up to 4 Å RMSD to the native structure for proteins of up to four transmembrane helices and 120 residues.

The TASSER protein structure prediction method was tested on a benchmark set of 38 membrane proteins (Zhang, DeVries et al. 2006). TASSER uses sequence threading to identify template regions of from proteins of known structure. Threaded portions of the

template structures are then combined to produce a complete model of the target protein. This has the advantage over fragment based methods such as Rosetta in that threaded regions of template can span large portions of the target protein if there a highly accurate match can be identified. Seventeen of the thirty-eight proteins were modeled with an accuracy better than 6 Å RMSD over all residues.

Predictions in addition to membrane spanning topologies can be incorporated into membrane protein structure techniques. This was demonstrated using Rosetta (Barth, Wallner et al. 2009). Predicted or experimentally identified contacts between sequentially distance transmembrane segments were used to constraint the topology of membrane proteins during folding. The contact point was held fixed while folding the rest of the protein around the contact points occurred. The contact between transmembrane segments was created without regard for maintaining a continuously intact protein backbone. The method was benchmarked on twelve membrane proteins with up to 300 residues. This study demonstrated two important ideas. First, it showed the utility of being able to efficiently and rapidly sample contacts between sequentially distance segments. This was previously hypothesized as one of the limitations in Rosetta membrane protein structure prediction (Yarov-Yarovoy, Schonbrun et al. 2006). Second, the study showed the large amount of information that can be contributed by a small number of restraints. Just a single restraint was able to focus the sampling and obtain highly accurate models for membrane proteins with up to 6 transmembrane helices.

Methods for structurally interpreting EPR measurements

EPR allows studying membrane protein structure not amenable to classic structural techniques. Information is acquired from EPR measurements using spin label probes which contain unpaired electrons (Jeschke and Polyhach 2007). These spin label probes are covalently attached to cysteine residues within the protein sequence (Hubbell and

Altenbach 1994). Methane thiosulfonate spin label (MTSSL) (Millhauser, Fiori et al. 1995) is commonly used for EPR measurements. The unpaired electron of MTSSL is tethered to the protein backbone at the end of the side chain. There are five rotatable bonds in MTSSL, providing great flexibility for the position of the unpaired electron relative to the protein backbone. MTSSL has been preferred over other spin labels because it has been shown to be sensitive to the secondary structure on which it resides (Isas, Langen et al. 2004). This is advantageous for determining the local structure round a spin label. However, the flexibility of the MTSSL spin label side chain presents challenges when EPR is used to measure distances within a protein. The distance is measured between the unpaired electrons of the two spin labels that have been introduced into the protein. Each of the spin labels project from the backbone with some unknown orientation, making the relation between the distance between the spin labels and the corresponding backbone distance unknown. The difference between the two distances can be up to 12 Å (Borbat, McHaourab et al. 2002).

EPR distance measurements provide a distribution of distances observed in the protein system (Chiang, Borbat et al. 2005). This distribution can be the result of protein backbone fluctuations and conformational sampling of the spin label (Fanucci and Cafiso 2006). Using alternative spin labels with reduced flexibility can remove contributions due to the conformational sampling of the spin label (Columbus, Kalai et al. 2001). However, MTSSL has been shown to be well tolerated when introduced at a variety of positions with a protein, including buried sites in the core of the protein (Guo, Cascio et al. 2007). Even assuming that the spin label is restrained, challenge remains to determine the conformation of the spin label relative to the protein backbone (Sale, Sar et al. 2002). Therefore, multiple experimental and computational efforts have attempted to characterize and predict conformations of the MTSSL spin label. This would allow precisely relating measured distances between spin labels into backbone distances.

Molecular dynamics simulations have been used in order to study MTSSL conformations and relate spin label distances into backbone distances. The conformational dynamics of MTSSL were simulated using MTSSL attached to an α-helical peptide (Tombolato, Ferrarini et al. 2006). The simulations demonstrated distinct energetically preferred conformations for each of the rotatable bonds in MTSSL, including correlated preferences between some of the bonds. When applied to a site within the T4-lysozyme protein, the single mutant continuous wave EPR spectra were able to be accurately reproduced. This provided insights into the contributions of the spin label dynamics to the cw-EPR spectra (Tombolato, Ferrarini et al. 2006). In another approach, a hybrid of both molecular dynamics and Monte Carlo based sampling was used to predict spin label conformations which would accurately reproduces experimentally measured EPR distance measurements (Sale, Song et al. 2005). Monte Carlo sampling provided a coarse grain search of possible spin label conformations. Following this course grained search, the most energetically favorable were then subjected to 1 nanosecond molecular dynamics trajectories. Using proteins of known structure for which EPR distance measurements were available this method was able to recover spin label distances to a mean error of 3 Å.

Experimental structural studies provide an alternative to computational methods for determining conformations sampled by the MTSSL spin label. The structure of sixteen single mutants of T4-lysozyme with MTSSL have been crystallized and determined to high resolution by X-ray crystallography (Langen, Oh et al. 2000; Guo, Cascio et al. 2007; Guo, Cascio et al. 2008; Fleissner, Cascio et al. 2009). In addition, two structures of the membrane protein LeuT have been crystallized each with one site within a transmembrane helix mutated to MTSSL (Kroncke, Horanyi et al. 2010). These structures provide insights into the preferred combinations of angles for the rotatable

bonds of MTSSL. This simplifies the search for MTSSL conformations to interpret EPR data by reducing the number of combinations that must be considered.

Investigations of the structure of the rhodopsin GPRC in complex with the G-protein transducin

G-protein coupled receptor (GPCR) proteins are integral membrane proteins of great biological significance. More than a quarter of all pharmaceutical therapies target GPCR proteins (Overington, Al-Lazikani et al. 2006). GPCR proteins consist of seven transmembrane spanning helices, and the first crystal structure of a GPCR was determined by X-crystallography the bovine rhodopsin protein (Palczewski, Kumasaka et al. 2000). Since then, multiple crystal structures of rhodopsin have been determined. Opsin is the activated form of rhodopsin. Rhodopsin becomes activated by the isomerization caused by light of a bound retinal molecule (Hofmann, Scheerer et al. 2009). A crystal structure of opsin (Park, Scheerer et al. 2008) revealed the structural differences between a rhodopsin structure (Li, Edwards et al. 2004) and opsin. The findings of the structural differences determined by X-ray crystallography were further verified through EPR (Altenbach, Kusnetzow et al. 2008). Sixteen distances were measured on exposed sites of helices on the cytoplasmic side of rhodopsin. The measurements were made before and after photoactivation and the distances changes were compared. The results mimicked those observed crystallographically that the cytoplasmic ends of transmembrane helices five and six undergo conformational changes (Millar and Newton).

Once rhodopsin is activated it can interact with its conjugate G-protein in order to continue the signal cascade into the cell, (Hofmann, Scheerer et al. 2009). Insight into the binding interaction of opsin with its G-protein, transducin, was provided when the crystal structure of opsin was solved with the eleven residue c-terminal peptide of

transducin (Scheerer, Park et al. 2008). This provided hypotheses for the mode of binding of the transducin to opsin. Transducin is comprised of three subunits, αβγ (Lambright, Sondek et al. 1996). The α subunit is the portion of transducin which interacts with opsin, and consists of two domains: a GTPase domain and a helical domain (Hamm 2001). When transducin binds to opsin, the GDP molecule bound to the α subunit is released and a molecule of GTP then binds to the α subunit (Tesmer 2010). Binding of GTP reduces the affinity of βγ for α and they dissociate. The heterotrimer subunits are then free to interact with other proteins in the signaling cascade until GTP is hydrolyzed to GDP and the cycle begins anew (Oldham and Hamm 2007).

The structure of the β_2 adrenergic receptor GPCR in complex with the Gs was determined by X-ray crystallography (Rasmussen, DeVree et al. 2011). Gs is the G-protein that activates adenylyl cyclase. In order to crystallize the complex, the T4-lysozyme protein was interjected into the N-terminus of the receptor. In addition, a nanobody protein is bound to the βγ subunits of Gs. This study provided the first high resolution experimentally determined structure of a GPCR in complex with a G-protein. Gs is in the nucleotide free state, providing a snapshot of the mechanism by which GDP is released from the α subunit. The helical domain of the α subunit in the crystal structure demonstrates remarkable flexibility of movement relative to the nucleotide binding domain. Further studies are needed to fully characterize the conformational ensemble accessible to the helical domain upon GPCR binding and GDP release.

Molecular dynamics simulations investigating the atomic mechanisms of activation and GPCR signaling use determined crystal structures as starting points (Grossfield 2011). Given the large size of GPRC proteins, the computational costs are large, but simulations of GPCRs have reached into the microsecond range. Prior to the publication of an experimental structure, molecular dynamics simulations were conducted on the complex of opsin and transducin (Sgourakis and Garcia 2010). A docked structure was

used as the starting point for the simulation of the complex, and 1.045 μs of simulation time was conducted on the 400000 atom system. The simulation used 2048 processors for 10 months to achieve these calculations.

CHAPTER II

DE NOVO HIGH-RESOLUTION PROTEIN STRUCTURE DETERMINATION FROM SPARSE SPIN LABELING EPR DATA

This work is based on publication (Alexander, Bortolus et al. 2008).

Summary

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

Introduction

The accelerated pace of genome sequencing has sparked the development of rapid structure determination methods and ambitious proposals for genome-scale structure

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

A complement of methods with intrinsically lower resolution can provide insight into these problems. Among them are probe-based approaches such as electron paramagnetic resonance (EPR) spectroscopy in combination with site-directed spin labeling (SDSL) (Hubbell, McHaourab et al. 1996; McHaourab, Berengian et al. 1997; Koteiche, Berengian et al. 1998; Perozo, Cortes et al. 1999; Liu, Sompornpisut et al. 2001; Brown, Sale et al. 2002; Dong, Yang et al. 2005). EPR analysis of spin labeled proteins results in a set of structural restraints that describes, in a native-like setting, local environments as well as aspects of the global fold of the protein. Spin label accessibility and mobility can be used to determine secondary structure location and topology (Farahbakhsh, Altenbach et al. 1992; Altenbach, Froncisz et al. 2005). Distance measurements between pairs of spin labels in the range from 5-60Å (Rabenstein and Shin 1995; Borbat, McHaourab et al. 2002) reflect the relative packing of domains and secondary structures. In cases where these parameters are obtained in various

conformational intermediates of the protein, they allow for a detailed mapping of structural changes involved in function (Dong, Yang et al. 2005). There are relatively few limits on the size and environment of the protein, particularly when compared to X-ray crystallography and NMR spectroscopy.

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

Although EPR distance restraints have a longer range than their NMR counterparts, they are fundamentally less accurate since they report distances between probes introduced into the protein sequence. The significant length of the spin label linking arm implies that the EPR distances will have a rather large uncertainty when translated into distances between α- or β-carbons unless the conformation of the spin label is known at every site. Therefore, previous efforts have focused on either using molecular dynamics simulations to define their trajectories (Sale, Song et al. 2005) or on determining a library of rotamers from crystal structures of spin labeled T4-lysozyme (Langen, Oh et al. 2000). These studies are critical since spin label conformations are likely stabilized by weak specific interactions with neighboring amino acid side chain or backbone atoms. However, such calculations are time and resource intensive and not practical without a high-resolution structural model of the protein.

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

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

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

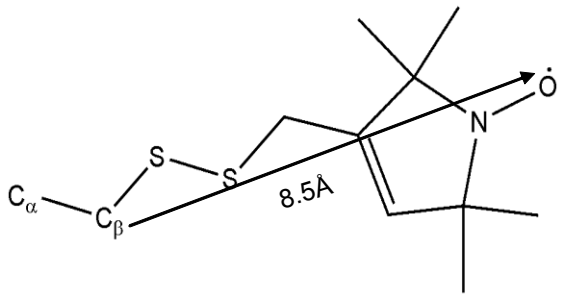
Results

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

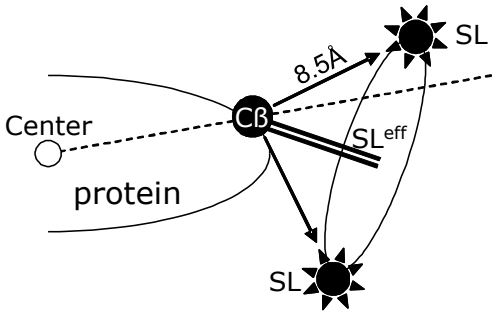
Evaluation of the "Motion-on-a-cone" model for interpretation of distance restraints

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

A)



B)



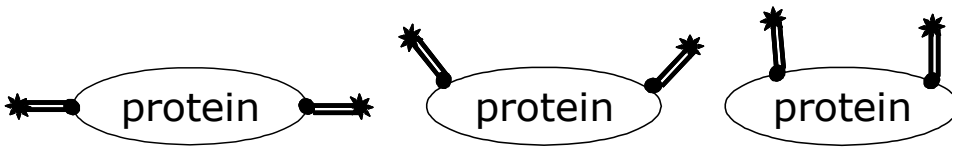
$d_{Cß\text{-}SL}$ = 8.5 Å

$d_{Cß\text{-}SL}^{eff}$ = 6 Å

$_{SL\text{-}Cß\text{-}SL}^{eff}$ = 45°

180° = $_{C\text{-}Cß\text{-}SL}^{eff}$ = 120°

C)



D)

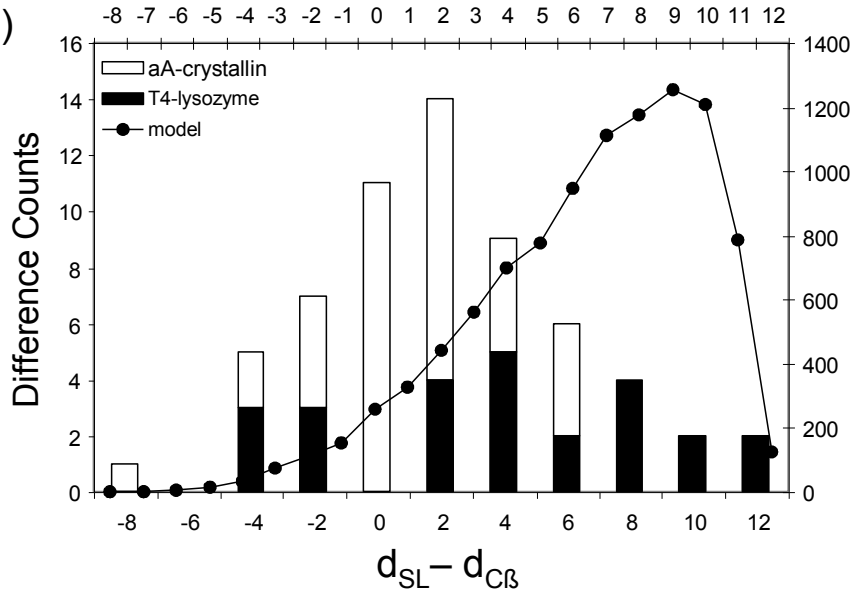


Figure 1 Rational for translating dSL into dCβ for use as a restraint.
A) Chemical structure of a nitroxide spin label side chain with the distance from the Cβ atom to the spin label indicated (Borbat, McHaourab et al. 2002). B) Illustration of how the maximum distance from Cβ to spin label, SL, is reduced to an effective distance, SLeff (depicted by a double line). C) dSL is a starting point for the upper estimate of dCβ, and subtracting the effective distance of 6Å twice from dSL gives a starting point for the lower estimate of dCβ. D) A histogram compares T4-lysozyme crystal structure (black bars, left y-axis, bottom x-axis) and αA-crystallin comparative model (white bars, left y-axis, bottom x-axis) dSL–dCβ values with those obtained from the simple cone model (circles and line, right y-axis, top x-axis).

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

Agreement of the consensus linear regression relation with T4-lysozyme and αA-crystallin

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

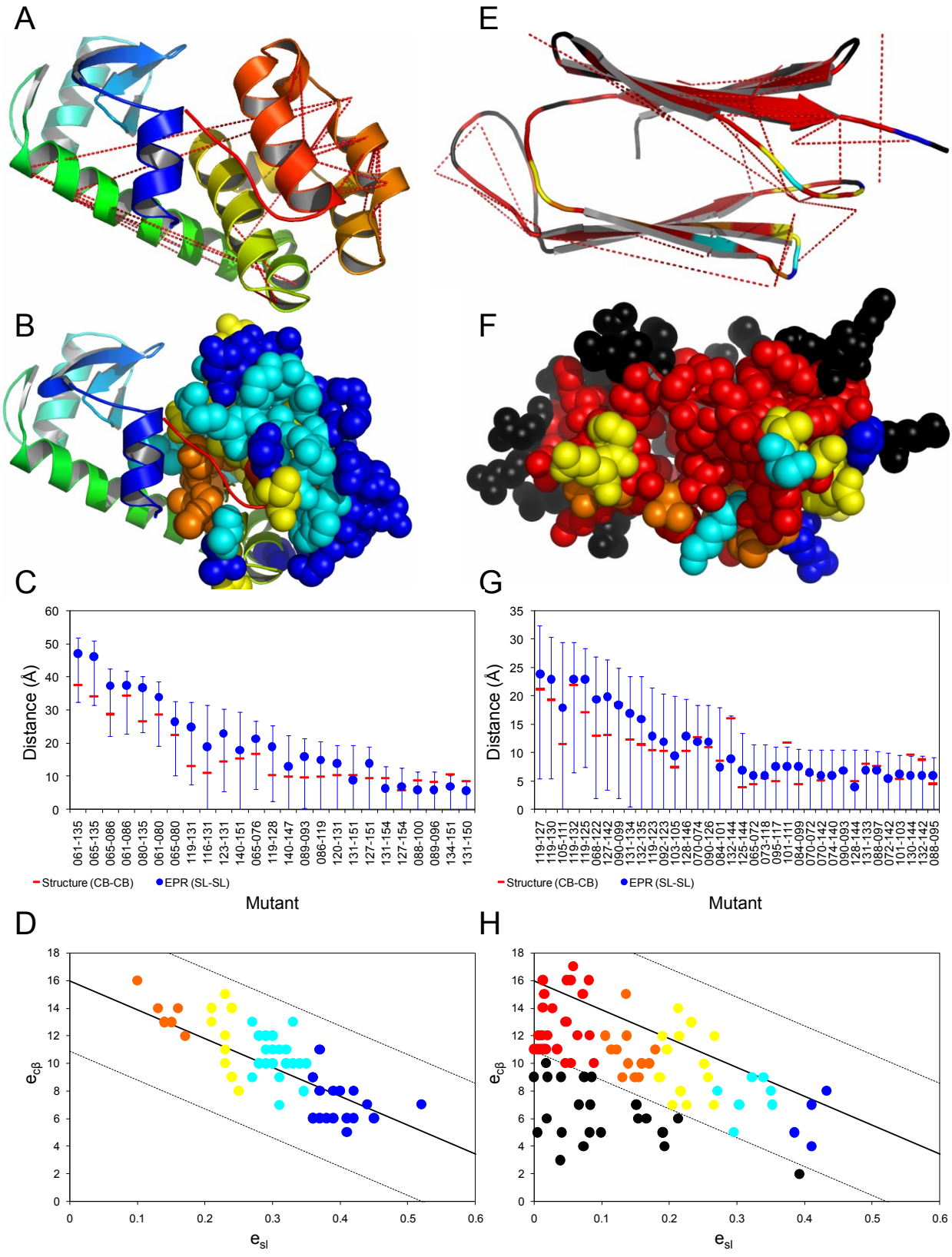

A
E
B
F
C
G
Distance (Å)
Structure (CB-CB)
EPR (SL-SL)
Mutant
D
H
$e_{cβ}$
e_{sl}

Figure 2 Map of the EPR restraints on the T4-lysozyme crystal structure (A-D) and on the αA-crystallin comparative model (E-H).
A and E) dCβ, shown as red dotted lines, which are restrained by respective dSL. B and F) Residues depicted as space-filling models. C and G) Diagram shows dSL (blue circle), the range of the derived distance restraints (blue), and the corresponding crystal/comparative model dCβ (red bar). D and H) Diagram illustrating the correlation of eSL with eCβ. The lines indicate the consensus model fit $\pm 3 \bullet \sigma C\beta$, where σCβ, was recalculated based on the consensus fit to be 1.70Å. In B, F, D, and H the residues are color-coded with decreasing eSL from blue – cyan – yellow – orange – red; black indicates amino acids in αA-crystallin that show reduced experimental accessibility due to intermolecular contacts with other αA-crystallin units in the oligomeric protein.

Influence of EPR data on de novo fold determination

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

The influence of the experimental EPR restraints on *de novo* protein folding with Rosetta was evaluated by building 10,000 models for each protein (a) without the use of experimental data, (b) with only the use of distance restraints, (c) with only the use of solvent accessibility restraints, and (d) using both sets of restraints. The average model quality was monitored by the root-mean-square-deviation (RMSD). The results for both T4-lysozyme and αA-crystallin follow the same trends: there is an improvement in the quality of models created using distance restraints compared to models created without distance restraints (Figure 3A and B, T4-lysozyme; Figure 3I and J, αA-crystallin); there is very little to no improvement when accessibility restraints are included (Figure 3A and C, T4-lysozyme; Figure 3I and K, αA-crystallin); using accessibility restraints in conjunction with distance restraints provides little to no improvement over using distance restraints alone (Figure 3B and D, T4-lysozyme; Figure 3J and L, αA-crystallin). It is

clear from these analyses that the distance restraints are critical for improving the RMSD distribution of models, while solvent accessibility data only marginally improve the quality of water soluble protein models.

Influence of spin label placement on de novo fold determination

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

To investigate the influence of spin label location and resulting $I_{C\beta}$ on *de novo* structure determination, two experiments were designed with subgroups of all available restraints. First, all restraints are ranked by $I_{C\beta}$ to assess their power in an idealized experiment. Second, all restraints are ranked by $s_{C\beta}$ in order to simulate the choices the experimentalist can make when selecting sites for labeling. 10,000 models for T4-lysozyme and αA-crystallin were built which used a) the one-third restraints with highest $I_{C\beta}$, b) the one-third restraints with lowest $I_{C\beta}$, c) the two-third restraints with highest $I_{C\beta}$, and d) the two-third restraints with lowest $I_{C\beta}$. The experiment was then repeated using $s_{C\beta}$ instead of $I_{C\beta}$,.

Once again the trends of the results are the same for both proteins and for both $I_{C\beta}$ and $s_{C\beta}$ experiments. a) Using the one-third restraints with highest $I_{C\beta}$ shifts the RMSD distribution into the range obtained when using all of the available distance restraints (Figure 3E and B, T4-lysozyme; Figure 3M and J, αA-crystalllin b) Using the one-third restraints with lowest $I_{C\beta}$ only slightly shifts the RMSD distribution towards lower RMSDs and is similar to that in the absence of distance restraints (Figure 3F and A, T4-

lysozyme; Figure 3N and I, αA-crystallin). c) There is little shift in the RMSD distribution when the two-third restraints with highest $I_{C\beta}$ are used compared to only one-third (Figure 3G and E, T4-lysozyme; Figure 3O and M, αA-crystallin); the extra restraints increase the number of lower RMSD models that are created. d) When the two-third restraints with lowest $I_{C\beta}$ are used, there is a small shift in the RMSD distribution compared to when no distance restraints are used (Figure 3H and A, T4-lysozyme; Figure 3P and I, αA-crystallin). However, this shift is not nearly as drastic as the shift obtained when the one-third most informative distant restraints are used.

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

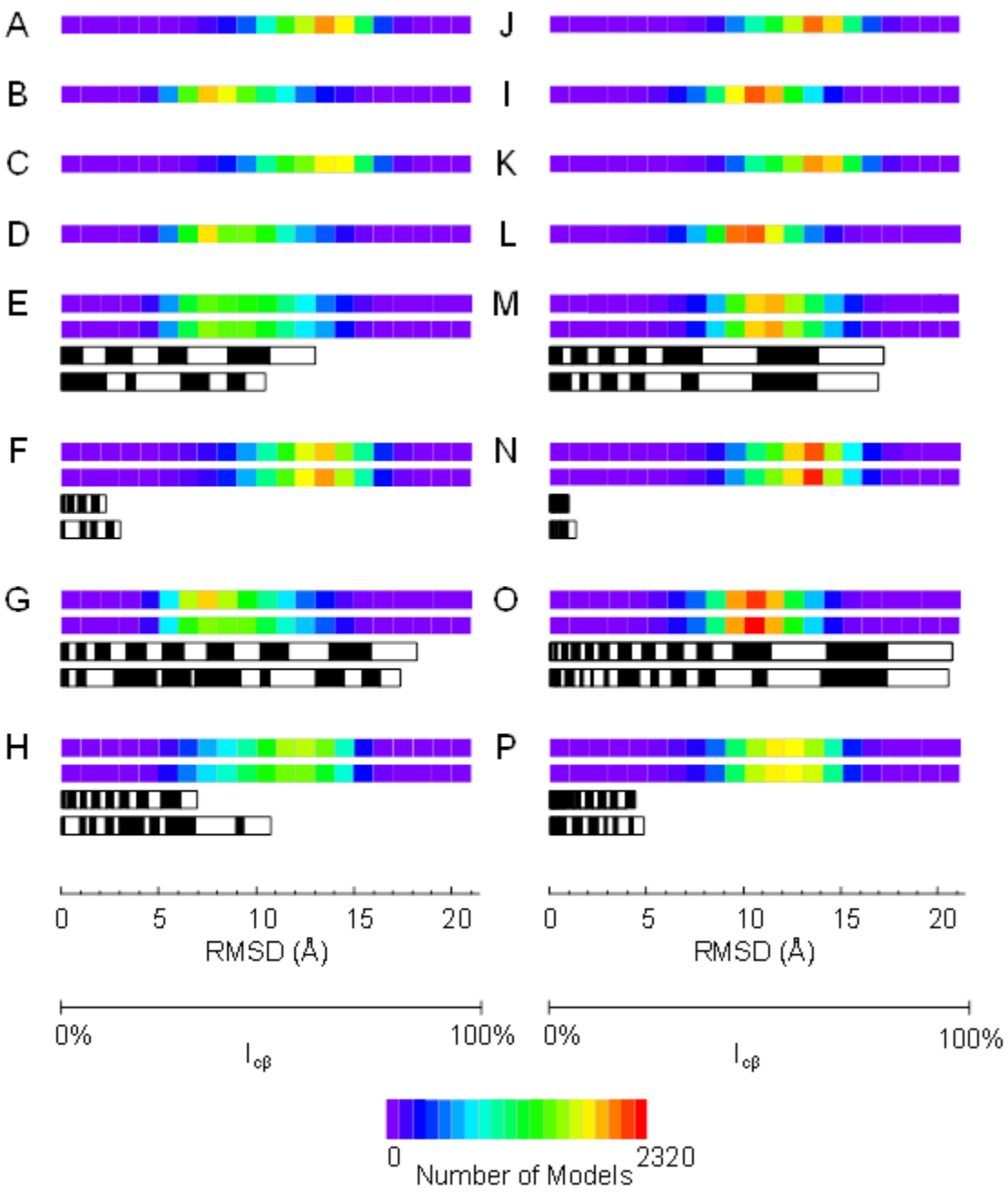


Figure 3 Illustration of the value of the experimental restraints in de novo protein folding for T4-lysozyme (A - H) and αA-crystallin (I - P).
The backbone RMSD distribution of 10,000 T4-lysozyme de novo models created A) without the use of EPR restraints, B) with only the use of EPR distance restraints,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

information content (top bar) and nearest in sequence (second bar). G) Same as for E) but using 2/3 of the total distance restraints. H) Same as for F) but using 2/3 of the total distance restraints. I – P) Same as A - H but for αA-crystallin.

Rosetta folding of T4-lysozyme and αA-crystallin

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

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

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

Structure Determination of T4-lysozyme and αA-crystallin

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

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

The Rosetta energy of T4-lysozyme *de novo* models decreases as their RMSD and side chain rotamer disagreement to the native structure diminish (Figure 4A, Figure 4B). This allows the selection of high-resolution models based on energy alone. The lowest energy *de novo* model achieves an RMSD to the crystal structure of 1.0Å in the α-helical domain and 2.0Å over all modeled residues (Figure 5A). 80% of all rotamers are in agreement with the crystal structure.

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

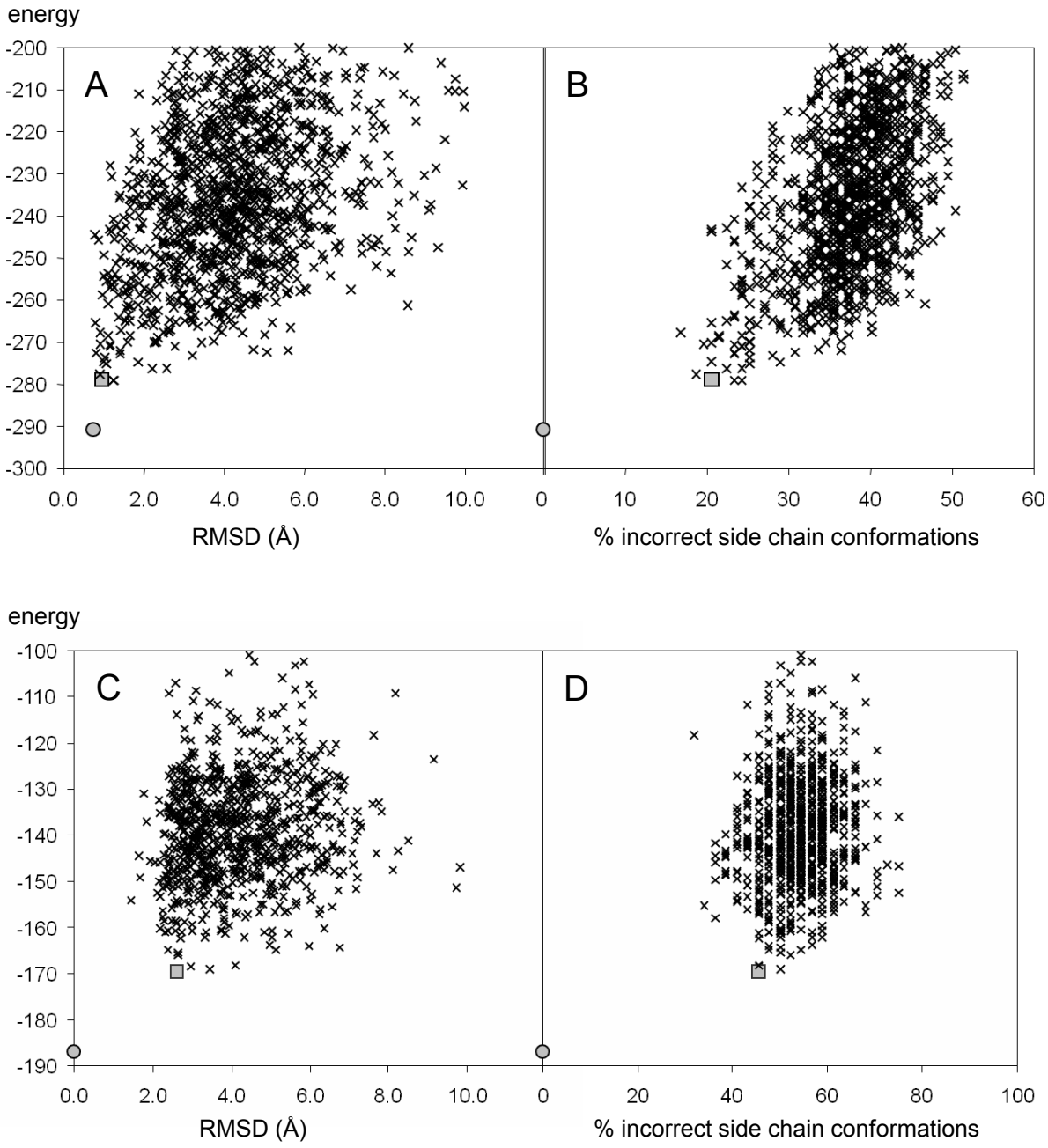
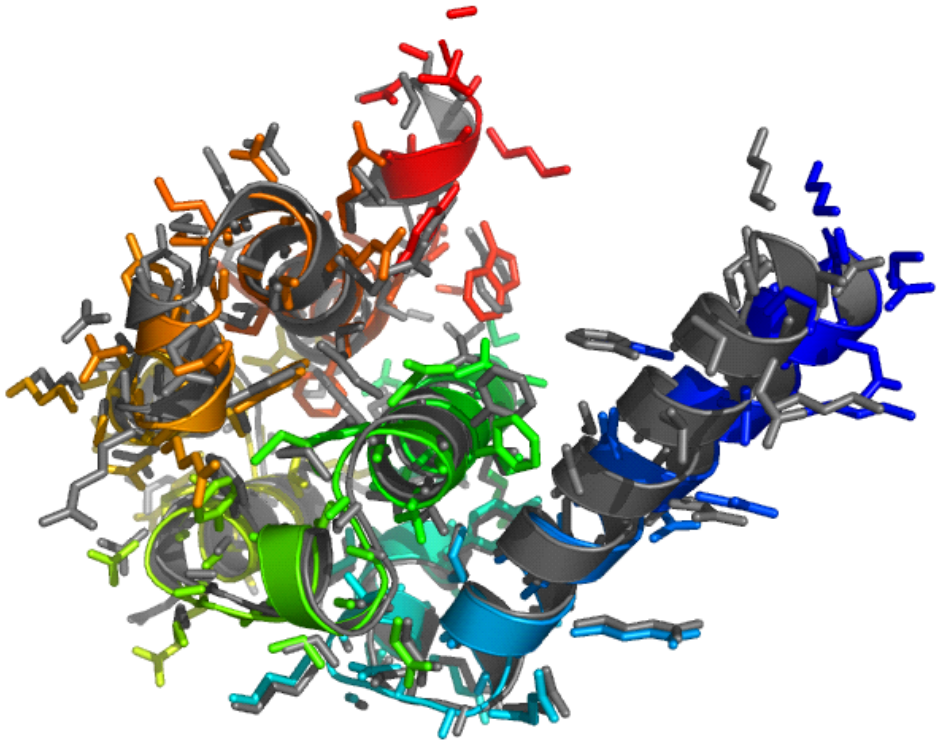


Figure 4 Correlation of de novo models' accuracy with the energy of the de novo models. A) and C) The non-loop RMSD versus Rosetta energy for T4-lysozyme and αA-crystallin models, respectively. B) and D) The percentage of incorrectly built side chain conformations versus Rosetta energy for T4-lysozyme and αA-crystallin models, respectively. In all diagrams the minimized crystal structure or comparative model is depicted as a circle; the lowest energy model is shown as a square.

A



B

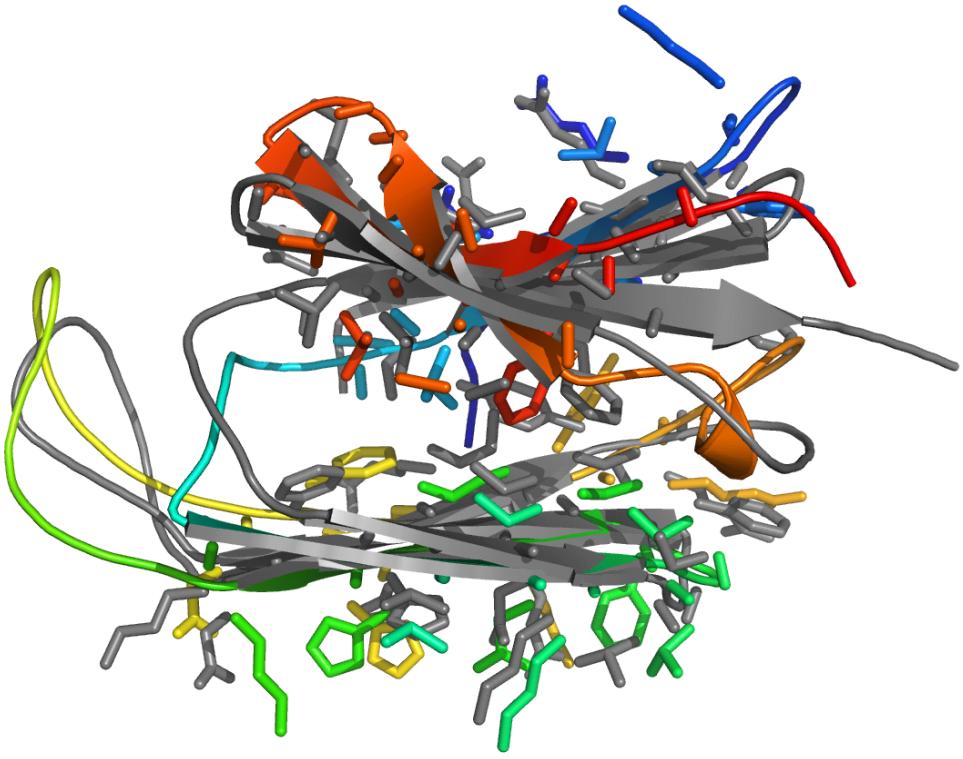


Figure 5 Overlay of lowest energy de novo models on crystal structure or comparative model. A) and B) For T4-lysozyme and αA-crystallin, respectively, superimposition of the lowest energy model (rainbow colored) with the crystal structure or comparative model (gray). The backbone is given as a ribbon diagram. Side chains of T4-lysozyme and of the β-sandwich of αA-crystallin are shown as stick models without hydrogen atoms.

Discussion

Structure determination from sparse EPR restraints

The major conclusion of this paper is that structural restraints obtained from EPR analysis of spin labeled proteins can be used in combination with *de novo* prediction methods to determine atomic-detail structures of proteins. Furthermore, the structural interpretation of the EPR data for the purpose of *de novo* modeling does not require a detailed understanding of the position or conformation of the spin label relative to the backbone. A “motion-on-a-cone” model can provide inter-Cβ distance restraints and the spin label accessibility can restrain the number of close neighbors when building a structural model.

Only a few specific distance restraints are needed in order to add substantial information and restrict the accessible fold space significantly. Using all twenty-five distance restraints, 50% T4-lysozyme models have an RMSD smaller than 8Å compared to less than 5% in their absence. This means that less than one distance restraint per four amino acids was sufficient to focus the *de novo* structure determination method on sampling the correct fold in more than half of all runs of T4-lysozyme. Only 0.074 distance restraints per residue are needed in order to obtain an equivalent RMSD distribution if the eight restraints between amino acids farthest apart in sequence are used. A significant increase in the quality of models created with distance restraints was also noted for αA-crystallin. By using the distance restraints between amino acids farthest apart in sequence, only 0.136 distance restraints per residue are needed in

order to obtain an RMSD distribution similar to that seen when using all restraints. The reduced frequency at which the correct fold is obtained compared with T4-lysozyme can be attributed to the more challenging folding pathway for a β-sandwich.

Relative importance of accessibility and spatial restraints

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

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

Structural interpretation of EPR parameters

The fundamental assumption in the “motion-on-a-cone” model is that the spin label cannot point towards the interior of the protein. Thus, possible specific interactions of

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

While there is a robust linear relation between the experimental spin label accessibility, e_{SL}, and the predicted accessibility, $e_{C\beta}$, the applied consensus fit procedure used to obtain the relation has two disadvantages. First, because a comparative model was used in the development of the consensus linear regression model, using it for *de novo* folding simulations is somewhat circular. Second, part of the accessibility data for αA-crystallin are influenced by oligomerization and had to be excluded because only a model of the monomeric state of αA-crystallin is computationally feasible. Nevertheless, using this relation in simulations is arguably an acceptable test of the usefulness of such data for *de novo* protein folding.

Conclusion

De novo structure prediction samples as much of the conformational space as possible in order to find the native structure; the number of models reflects the extent of sampling. The increase in the number of high quality models indicates that the conformational search space has been reduced by the restraints, which allows the remaining space to be sampled more densely. It is remarkable that sparse distance restraints with as large an uncertainty as those obtained from the “motion-on-a-cone” model provide such a drastic improvement in the RMSD distribution of models. The most efficient reduction in conformational search space results from the longest range restraints between residues far apart in the primary sequence. The uniform distribution

of restraints throughout the protein should also be taken into consideration in order to maximize efficiency.

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

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

Experimental Procedures

Introduction of site-directed spin labels and EPR conditions

For the introduction of spin labels, cysteine residues were systematically introduced into the (cysteine-free) T4-lysozyme and αA-crystallin amino acid sequences through

single or double point mutations (Koteiche, Berengian et al. 1998; Borbat, McHaourab et al. 2002; Altenbach, Froncisz et al. 2005). After recombinant protein expression and purification, the mutant was reacted with methanethiosulfonate nitroxide reagent. A total of 25 double mutants and 57 single mutants of T4-lysozyme (Table 1, and Table 2) and 36 double mutants and 87 single mutants of αA-crystallin (Table 3, and Table 4) resulted in the restraints used for the current analysis. Sample preparation and EPR measurement have been described elsewhere (McHaourab, Lietzow et al. 1996; Koteiche and McHaourab 1999).

EPR distance measurements

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

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

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

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

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

EPR accessibility measurements

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

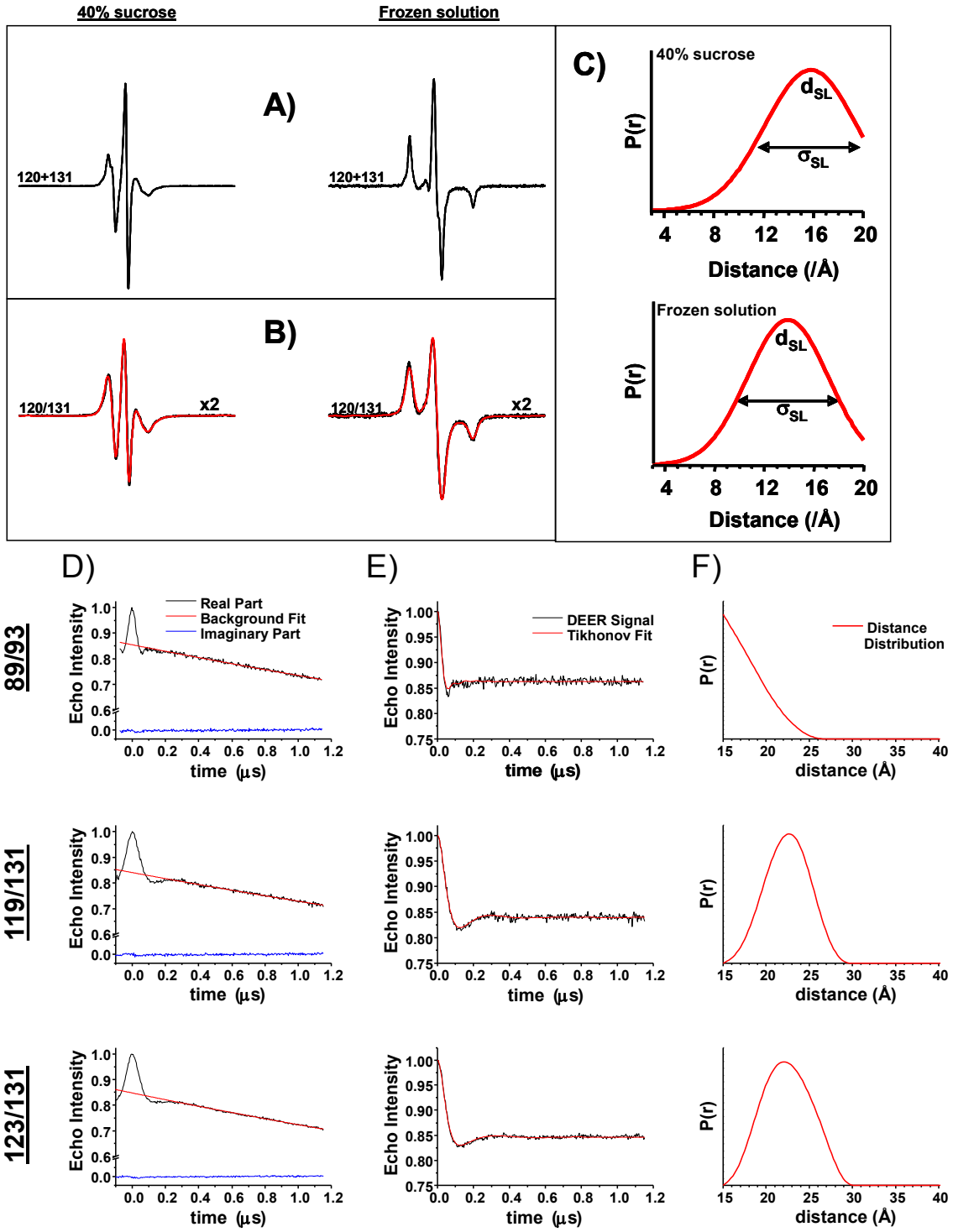


Figure 6 Distance measurements at room temperature and in the solid state between spin labels using CW-EPR.
A) Representative reference EPR in the absence of dipolar coupling obtained from the digital sum of the corresponding single mutant spectra. B) Spectra of double mutants along with the non-linear least squares fit obtained by the convolution method as described in the experimental methods section. C) Distance distributions obtained from CW-EPR spectra. D) Distance measurements by DEER for representative double mutants. E) Raw DEER signals were background corrected and then fit using Tikhonov regularization to obtain F) average distances and distance distributions.

αA-crystallin comparative model preparation

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

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

The lowest energy comparative model was used to compare against the *de novo* Rosetta models. This model was based on the PDB structure 1gmeA (van Montfort, Basha et al. 2001). The model is 2.3Å RMSD to a previously published (Koteiche and

McHaourab 1999) comparative model based on a different template protein (Hsp16.5) but achieves a lower Rosetta energy after both models are refined at high-resolution.

The spin label "motion-on-a-cone" model

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

Using the "motion-on-a-cone" model to translate EPR spin label distances into structural restraints

The difference between d_{SL} and $d_{C\beta}$ given by the model described above was analyzed using the software package Mathematica (2005):

1) An ellipsoid with the main radii 10Å ≤ r_x ≤ r_y ≤ r_z ≤ 20Å was created with otherwise randomly chosen r_x, r_y, and r_z. Its center is *C*.

2) Two points, $C\beta_i$ and $C\beta_j$, on the surface of this ellipsoid were selected by randomly choosing the polar coordinates $\varphi_{i,j}$, and $\psi_{i,j}$. From these points, $d_{C\beta}$ is computed as the Euclidean distance.

3) Two numbers between 120° and 180° for the angle $C–C\beta_i–SL_i$ and $C–C\beta_j–SL_j$ are chosen randomly and the position of the spin labels, SL_i and SL_j, is computed. From these points, d_{SL} is computed as the Euclidean distance.

4) The difference $d_{SL}–d_{C\beta}$ is computed.

5) Steps 1-4 are repeated 10,000 times and the values $d_{SL}–d_{C\beta}$ are plotted as a histogram (Figure 1D).

This analysis of the difference between d_{SL} and $d_{C\beta}$ showed that (d_{SL}+2.5Å) ≥ $d_{C\beta}$ ≥ (d_{SL}–12.5Å) (Figure 1D).

σ_{SL}, which is the experimentally determined standard deviation in d_{SL}, is a measure of the magnitude of the spin label's motion, or, its static distribution relative to the Cβ. Since an increased magnitude of motion increases the ambiguity of the derived $d_{C\beta}$, σ_{SL} is added as an additional allowance to the restraint which gives (d_{SL}+σ_{SL}+2.5Å) ≥ $d_{C\beta}$ ≥ (d_{SL}–σ_{SL}–12.5Å).

Development of a model to translate EPR spin label solvent accessibility into structural restraints

Obtaining a structural restraint from EPR spin label solvent accessibility is accomplished by building a consensus linear regression relation of e_{SL} to $e_{C\beta}$ in a three step procedure:

1) Using the crystal structure of T4-lysozyme, $e_{C\beta}$ was computed for all residues with an e_{SL} (Table 2). A linear regression was fit to a plot of $e_{C\beta}$ of a residue versus the corresponding e_{SL}, yielding the relation $e_{C\beta}$ = (0.71–e_{SL})•24.23 with a correlation coefficient of –0.80. Using this relation to calculate the number of neighbors for a residue gives $e_{C\beta}^{fit}$ for that residue. The standard deviation ($\sigma_{C\beta}$) of $e_{C\beta}$ from $e_{C\beta}^{fit}$ was calculated to be 1.65.

2) For those residues in αA-crystallin that have an experimentally determined accessibility (Table 4), $e_{C\beta}$ was determined using the comparative model of αA-crystallin. In addition, the equation from 1) above was used to calculate the number of neighbors, $e_{C\beta}^{fit}$, for each residue. Amino acids were excluded from a linear fitting of $e_{C\beta}$ versus e_{SL} when $|e_{C\beta}–e_{C\beta}^{fit}| > 2 \bullet \sigma_{C\beta}$. This procedure was necessary in order to exclude amino acids in αA-crystallin that show reduced experimental accessibility due to intermolecular contacts with other αA-crystallin units in the oligomeric protein (Figure 2F, H). Fitting a linear regression to the remaining data gives a relation similar to that seen for T4-lysozyme: $e_{C\beta} = (0.72–e_{SL}) \bullet 21.63$ with a correlation coefficient of –0.87.

3) Combining the data for both proteins in a single consensus linear regression model yields $e_{C\beta} = (0.76–e_{SL}) \bullet 20.87$ with a correlation coefficient of –0.83 (Figure 2D, and Figure 2H).

Implementation of structural restraints for de novo structure determination

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

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

Rosetta folding simulations

De novo model generation using Rosetta was performed in four steps:

Step 1) The protein is folded using RosettaNMR with EPR distance restraints to guide the simulation (Bowers, Strauss et al. 2000; Rohl and Baker 2002). In this step, amino acid side chains are embraced in a single super-atom – a “centroid” (Simons, Kooperberg et al. 1997).

Step 2) Choosing the models with lowest energy and best agreement with experimental restraints prunes the large number of ~500,000 models from Step 1 to ~10,000.

Step 3) Models obtained from Step 2 are refined to high-resolution. High-resolution refinement is used to distinguish between the best models as to which model is likely the structure that is closest to native based on energy. High-resolution refinement is described as follows:

After replacing side chain centroids with full-atom side chain representations from a backbone dependent rotamer library (Dunbrack and Karplus 1993), an iterative protocol of all-atom gradient minimization and side chain repacking is repeated eight times. The details of the protocol are published elsewhere (Bradley, Misura et al. 2005; Misura and Baker 2005). No restraints were used during these refinement simulations in order to fully leverage the discriminative power of the Rosetta energy function (Kuhlman and Baker 2000; Bradley, Malmstrom et al. 2005; Bradley, Misura et al. 2005; Misura and Baker 2005; Misura, Chivian et al. 2006). The protocol is implemented in the Rosetta software package.

Step 4) Models from Step 3 are again filtered for good agreement with the experimental restraints.

Specific standard Rosetta procedures were used which are described in detail elsewhere (Simons, Kooperberg et al. 1997; Simons, Ruczinski et al. 1999; Bowers,

Strauss et al. 2000; Bonneau, Strauss et al. 2001; Rohl and Baker 2002; Meiler, Bradley et al. 2003; Rohl, Strauss et al. 2004; Bradley, Malmstrom et al. 2005). Secondary structure predictions were obtained from the primary sequence of the C-terminal 107 amino acids of T4-lysozyme and the C-terminal 88 amino acid primary sequence of αA-crystallin using Jufo (Meiler, Müller et al. 2001; Meiler and Baker 2003), PsiPred (Jones 1999), and Sam (Karplus, Sjolander et al. 1997). All T4-lysozyme and αA-crystallin homologues were excluded from the protein database prior to the search for overlapping nine amino acid fragments of similar sequence which match the predicted secondary structure. A library of 200 fragments for each position was built.

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

Acknowledgements

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

Table 1 T4-lysozyme EPR distance restraints in comparison with crystal structure distances.

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

[a] indices of spin labeled amino acids with respect to the crystal structure.
[b] Cβ distance as reported in the crystal structure.
[c] spin label distance as observed by EPR.
[d] standard deviation as observed by EPR.
[e] maximum Cβ atom distance predicted by cone model.
[f] minimum Cβ atom distance predicted by cone model.

Table 2 T4-lysozyme EPR solvent accessibility in comparison with crystal structure.

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

[a] indices of spin labeled amino acids with respect to the crystal structure.

[b] number of Cβ atom neighbors in the crystal structure.

[c] spin label accessibility as observed by EPR (Sompornpisut, Mchaourab et al. 2002).

[d] number of Cβ atom neighbors predicted by the consensus linear regression relation.

Table 3 αA-crystallin EPR distance restraints in comparison with comparative model.

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

Table 3 continued

119-132	21.9	23.0	4.0	29.5	6.5	(Koteiche and McHaourab 1999)
125-144	4.0	7.0	4.0	13.5	0.0	(Koteiche and McHaourab 1999)
127-142	13.1	20.0	4.0	26.5	3.5	new data
128-144	5.0	4.0	4.0	10.5	0.0	(Koteiche and McHaourab 1999)
128-146	10.4	13.0	4.0	19.5	0.0	(Koteiche and McHaourab 1999)
130-144	9.7	6.0	1.0	9.5	0.0	(Koteiche and McHaourab 1999)
131-133	8.0	7.0	1.0	10.5	0.0	(Koteiche and McHaourab 1999)
131-134	12.3	17.0	4.0	23.5	0.5	(Koteiche and McHaourab 1999)
132-135	11.5	16.0	5.0	23.5	0.0	(Koteiche and McHaourab 1999)
132-142	8.8	6.0	1.0	9.5	0.0	(Koteiche and McHaourab 1999)
132-144	16.1	9.0	5.0	16.5	0.0	(Koteiche and McHaourab 1999)

[a] indices of spin labeled amino acids with respect to the protein sequence.

[b] Cβ atom distance in comparative model.

[c] spin label distance as observed by EPR.

[d] standard deviation as observed by EPR.

[e] maximum Cβ atom distance predicted by cone model.

[f] minimum Cβ atom distance predicted by cone model.

Table 4 : αA-crystallin EPR solvent accessibility in comparison comparative model.

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

[a] indices of spin labeled amino acids with respect to the protein sequence.

[b] number of Cβ atom neighbors in the crystal structure.

[c] spin label accessibility as observed by EPR (Koteiche, Berengian et al. 1998; Koteiche and McHaourab 1999).

[d] number of Cβ atom neighbors predicted from the consensus linear regression relation.

CHAPTER III

ROSETTA-EPR: ROTAMER LIBRARY FOR SPIN LABEL STRUCTURE AND DYNAMICS

This chapter is based on the manuscript to be published of the same title.

Summary

An increasingly used parameter in structural biology is the measurement of distances between spin labels bound to a protein. One limitation to these measurements is the unknown position of the spin label relative to the protein backbone. To overcome this drawback, we introduce a rotamer library of the methanethiosulfonate spin label (MTSSL) into the protein modeling program Rosetta. Spin label rotamers have been derived from conformations observed in crystal structures of spin labeled T4 lysozyme and molecular dynamics simulations. Rosetta's ability to accurately recover spin label conformations and EPR measured distance distributions was evaluated against 19 experimentally determined MTSSL labeled structures of T4 lysozyme and the membrane protein LeuT and 73 distance distributions from T4 lysozyme and the membrane protein MsbA. In the protein core, the correct spin label conformation (X_1 and X_2) is recovered in 99.8% of trials. In surface positions 53% of the trajectories agree with crystallized conformations in X_1 and X_2. This level of recovery is on par with Rosetta performance for the 20 natural amino acids. In addition, Rosetta predicts the distance between two spin labels with a mean error of 4.4 Å. The width of the experimental distance distribution, which reflects the flexibility of the two spin labels, is predicted with a mean error of 1.3 Å. Modeling MTSSL at this level of accuracy moves towards atomic-detail refinement of protein structures based on experimental EPR distance restraints.

Introduction

Electron paramagnetic resonance (EPR) can be applied to both large and membrane proteins (MPs). Thereby EPR opens an avenue to study the structure and dynamics of proteins which are often difficult to study with X-ray crystallography or nuclear magnetic resonance (NMR). EPR in conjunction with site directed spin labeling (SDSL) allows specific inter-residue distances to be routinely measured up to 60Å (Hubbell and Altenbach 1994; Rabenstein and Shin 1995; Borbat, McHaourab et al. 2002; Czogalla, Pieciul et al. 2007) and can reach up to 80Å (Jeschke, Bender et al. 2004; Jeschke and Polyhach 2007). The limitation of EPR in its application to protein structure determination is that the distances are measured between unpaired electrons in the nitroxide group of the spin label side chain. The most widely used methanethiosulfonate spin label (MTSSL) projects from the backbone of the protein. It has five rotatable bonds (X_1 - X_5) with an a priori unknown conformation between the Cα of the protein backbone and the unpaired electron at the midpoint of the N-O bond. Without the knowledge of the spin label conformation it is difficult to directly relate the distance between the unpaired electrons to a distance between its anchor points on the protein backbone. This task becomes even more challenging in solvent exposed positions on the protein surface with little spatial restriction. Here the spin label will adopt an ensemble of conformations with comparable free energies (Figure 7A). In result a broad distance distribution for the unpaired electrons is observed in the EPR measurement (Polyhach, Bordignon et al. ; Rabenstein and Shin 1995; Chiang, Borbat et al. 2005)

Previous computational methods have been developed to determine correct spin label conformations (Sale, Sar et al. 2002; Fajer, Li et al. 2007) and structurally interpret EPR distance distributions (Sale, Song et al. 2005) within a protein environment. While generally successful, these techniques relied upon computationally intense molecular

dynamics, Monte Carlo searches, or combinations of the two, in order to effectively sample the necessary conformational space available to the spin label probe. The algorithms focused on the local environment around the spin label assuming a rigid protein backbone in order to make the calculation computationally tractable, potentially missing preferred rotamers

Libraries of likely conformations of spin labels (rotamers) have been previously applied for explicit modeling of MTSSL. A rotamer is a likely side chain conformation with a specific set of chi angles derived from statistical analysis of the Protein Data Bank (PDB) (Dunbrack 2002). An initial library of 62 rotamers (Jeschke and Polyhach 2007) was expanded to 98 (Hilger, Polyhach et al. 2009) and then to approximately 200 rotamers (Polyhach, Bordignon et al.) in order to capture the allowable conformational space of the spin label. The rotamer libraries in the latter study were derived from molecular dynamics calculations of spin label flexibility. These methods accurately predicted a) conformations of MTSSL seen in experimentally determined soluble structures and b) measured distance distributions between spin labels in doubly mutated soluble proteins.

Further, a knowledge-based potential has been introduced (Hirst, Alexander et al. ; Alexander, Bortolus et al. 2008) that in combination with coarse-grained potentials and sparse EPR distance restraints can be used to determine protein topology. Instead of an atomic detail model of the spin label it converts the experimental spin label distance into a probability distribution of Cβ distances. While efficient in determining the protein fold with RosettaEPR, the potential lacks detail needed for high-resolution structure refinement.

The objective of the present work is to extend RosettaEPR with an atomic detail representation of the spin label that aligns with the Rosetta “rotamer” approach for rapid sampling of protein side chain degrees of freedom (Kuhlman and Baker 2000). The

ability of Rosetta to recover native rotamers has been demonstrated for protein structure prediction (Bradley, Misura et al. 2005; Misura and Baker 2005; Alexander, Bortolus et al. 2008) and protein design (Kuhlman, Dantas et al. 2003). The present study extends the amino acid rotamer libraries used by Rosetta to include MTSSL. The rotamer library for MTSSL is derived from the experimentally and computationally observed correlated preferences of the side chain dihedral angles. Consequently, the library consists of only 54 conformations. The incorporation of MTSSL into RosettaEPR enables modeling of the spin label in a wide range of Rosetta protocols such as atomic detail refinement (Tsai, Bonneau et al. 2003; Misura and Baker 2005) and membrane protein modeling (Ganguly, Weiner et al. ; Van Eps, Preininger et al. ; Barth, Schonbrun et al. 2007). After initial placement of the spin label rotamer the Rosetta full atom potential enables sampling of off-rotamer conformations thereby limiting the number of initial rotamers needed. RosettaEPR optimizes all other protein side chains and backbone degrees of freedom in parallel with the spin label thereby capturing structural perturbations caused by the spin label. RosettaEPR makes the technology readily available to the EPR community through RosettaCommons free non-commercial licensing.

The current study details the development of Rosetta's MTSSL rotamer library and demonstrates: a) Rosetta's ability to sample MTSSL conformations experimentally observed in 19 structures of the soluble protein T4 lysozyme and the membrane protein LeuT; b) Rosetta's ability to recover the experimental probability distribution for a measured EPR distance in T4 lysozyme and the membrane protein MsbA; and c) the unbiased cross-validation of the cone model parameters (Hirst, Alexander et al. ; Alexander, Bortolus et al. 2008).

Results

MTSSL rotamer library

Sixteen structures of T4 lysozyme with single MTSSL mutations (Langen, Oh et al. 2000; Guo, Cascio et al. 2007; Guo, Cascio et al. 2008; Fleissner, Cascio et al. 2009), and one with a double MTSSL mutation (Langen, Oh et al. 2000), have been determined experimentally by x-ray crystallography, allowing 21 low energy conformations of the MTSSL side chain to be observed (Table 5). The labels in the double mutant K65R1/R80R1 are structurally independent and do not interact (Langen, Oh et al. 2000), so for the purposes of this study will be considered separate individual single mutants. Here, the convention of Lovell et al. (Lovell, Word et al. 2000) is used to denote X_1 and X_2 angles; $X_1 = 0$ when S_γ eclipses the backbone nitrogen (Figure 7A). Additionally, “m”, “p”, and “t” indicate dihedral angles of -60°, +60°, and 180°, respectively. Tombolato et al. (Tombolato, Ferrarini et al. 2006) defines X_5 as $S_\delta – C – C = C$, which is the convention used here (Figure 7A). Although most of the mutations are on exposed helical sites, crystal structures for one core position (Guo, Cascio et al. 2007) and exposed loop residues (Fleissner, Cascio et al. 2009) have been determined. This experimental knowledge base provides the necessary foundation for building a rotamer library for MTSSL.

Note that a rotamer not only captures likely conformations for all X-angles but also their respective interdependences, i.e. how likely a certain combination of X-angles is observed. The relatively small number of spin label conformations observed experimentally forbids a statistical analysis of all interdependences between X_1 - X_5, in particular as many experimental structures lack information on X_4- and X_5 - angles. Assuming just three conformations for each of the $X_{1,2,4,5}$-angles and two for X_3, 162 conformations need to be considered. While some of those can be excluded for internal clashes, the number of possible conformations is still much larger than the 21 experimental conformations available. On the order of 500 experimental structures

Table 5 Experimentally determined MTSSL conformations for single mutants of t4-lysozyme. *Mutant* indicates the residue of t4-lysozyme which was mutated to the MTSSL side chain. Subscripts denote the protein subunit from the crystal structure asymmetric unit as indicated in the PDB file. *Temp.* gives the temperature at which the crystal was formed. *Environ.* gives the environment in which the residue lies: on the surface of the protein (surface); within the core of the protein (core); at the contact point of two a crystallographic subunits (crystal contact). *SSE Type* gives the type of secondary structure element on which the mutated residue sits. *Rotamer* indicates the X_1 and X_2 angles observed for the spin label in the crystal structure according to the m, t, p convention of (Lovell, Word et al. 2000). The X angles observed in the crystal structure are shown in their respective columns. Blank columns indicate the X angles were not resolved. *PDB ID* is the Protein Data Bank accession identifier of the crystal structure, if available. *Ref* provides the primary citation for the crystal structure.

Mutant	Temp. (K)	Environ.	SSE Type	Rotamer	X_1(°)	X_2(°)	X_3(°)	X_4(°)	X_5(°)	PDB ID	Ref.
R080	298	surface	helix	{m, m}	-74	-66					(Langen, Oh et al. 2000)
R119	100	surface	helix	{m, m}	-50	-50					(Langen, Oh et al. 2000)
R119	100	surface	helix	{t,p}	175	54					(Langen, Oh et al. 2000)
K065	298	crystal contact	helix	{t,p}	153	89	53				(Langen, Oh et al. 2000)
V075	100	crystal contact	helix	{m,t}	-73	173	91	95			(Langen, Oh et al. 2000)
T115	100	surface	helix	{m, m}	-81	-57	-92	76	98	2IGC	(Guo, Cascio et al. 2007)
T115/R119A	100	surface	helix	{m, m}	-77	-33				2OU9	(Guo, Cascio et al. 2007)

Table 5 continued

T115	298	surface	helix	{m,m}	-94	-28				2OU8	(Guo, Cascio et al. 2007)
T115	298	surface	helix	{t,m}	163	-63				2OU8	(Guo, Cascio et al. 2007)
L118	100	core	helix		-104	32	88	54	107	2NTH	(Guo, Cascio et al. 2007)
A041	100	crystal contact	helix	{t,p}	-175	57	86			2Q9D	(Guo, Cascio et al. 2008)
S044[a]	100	crystal contact	helix	{m,m}	-83	-58	-95	76	-86	2Q9E	(Guo, Cascio et al. 2008)
S044[b]	100	crystal contact	helix	{m,m}	-85	-55	-96	71	-78	2Q9E	(Guo, Cascio et al. 2008)
S044[c]	100	surface	helix	{t,m}	173	-96				2Q9E	(Guo, Cascio et al. 2008)
A082	100	surface	loop	{m,m}	-68	-56	101			1ZYT	(Fleissner, Cascio et al. 2009)
V131	100	surface	helix	{m,m}	-69	-60				2CUU	(Fleissner, Cascio et al. 2009)
V131	100	surface	helix	{t,p}	175	80				2CUU	(Fleissner, Cascio et al. 2009)
V131	291	surface	helix	{m,m}	-75	-57				3G3V	(Fleissner, Cascio et al. 2009)

Table 5 continued

V131	291	surface	helix	{t,p}	175	83				3G3V	(Fleissner, Cascio et al. 2009)
T151	100	surface	helix	{m, m}	-83	-72				3G3X	(Fleissner, Cascio et al. 2009)
T151	291	surface	helix	{m, m}	-82	-72				3G3W	(Fleissner, Cascio et al. 2009)

resolving all X- angles would be needed to build a complete rotamer library from a knowledge base. Therefore, we follow a hybrid approach deriving likely (X_1, X_2) combinations from experimental structures. Possible conformations for X_3 are taken from the quantum chemical studies (Tombolato, Ferrarini et al. 2006) which agree closely with crystallographic data. X_3 is decoupled from X_1 and X_2, i.e. all combinations of X_3 with (X_1, X_2) pairs will be considered. Combinations of X_4 and X_5 are derived from quantum chemical studies (Tombolato, Ferrarini et al. 2006) as these X- angles are resolved in only four experimental structures. We expect to update this rotamer library as additional experimental structures of the spin label become available.

Only four (X_1, X_2) combinations of m, t, and p have been experimentally observed: {m, m}, {m, t}, {t, p}, and {t, m} (Figure 7B). One conformation of MTSSL observed in the core of the protein (Guo, Cascio et al. 2007) is excluded from consideration from the rotamer library as it cannot be classified into the “m”, “t”, or “p” categories described above. As it was observed only once, it remains unclear if this conformation represents a low energy state of the spin label in isolation or is induced by packing interaction in the protein core. While a single conformation is insufficient to perform the statistical analysis needed for creation of a rotamer, Rosetta relaxation protocols will be capable of

modeling off-rotamer conformations starting from one of the rotamers provided (read below). Quantum chemical calculations have shown that also the {t, t} conformation, not yet seen in any experimental structure, is sterically allowed for sites on an exposed poly-alanine helix (Tombolato, Ferrarini et al. 2006). Therefore, the {m, m}, {m, t}, {t, p}, {t, m}, and {t, t} conformations will be represented in the current rotamer library as the average angle observed for each pair (Figure 7C, Table 6).

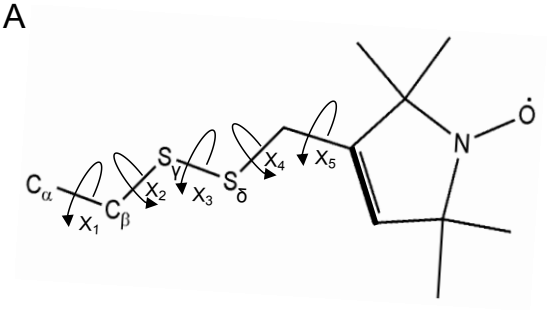
X_3 is experimentally and computationally observed to adopt an angle of ±90°, independent of X_1 and X_2. As a result, both states will be considered for each of the five sets of X_1 and X_2 angles (Figure 7C). In the instance where X_3 is 53°, the crystal structure reveals several favorable contacts in the crystal lattice that presumably overcome the unfavorable energy of the distortion (Langen, Oh et al. 2000). This X_3 angle was not considered in the rotamer library.

X_4 and X_5 have been observed in only five and four of the crystal structures, respectively. Due to the small sample size for (X_4,X_5) combinations the values predicted from quantum chemical calculations will be used (Tombolato, Ferrarini et al. 2006). The calculations predict a correlation between X_4 and X_5 where the highest probability conformers are: a) when X_4 is 180°, X_5 is ±77°; b) when X_4 is -75°, X_5 is either -8° or +100°; c) when X_4 is +75°, X_5 is either 8° or -100° (Figure 7C). Key surface interactions of mutant T115 at 100K ($T115^{100}$, superscripts will denote temperature) and core packing of mutant L118 alter X_4 and X_5, 76° and 98° for T115 and 54° and 107° for L118 (Guo, Cascio et al. 2007). These values were not considered in the rotamer library, though if additional structures show these to be frequently observed conformations, they will be added.

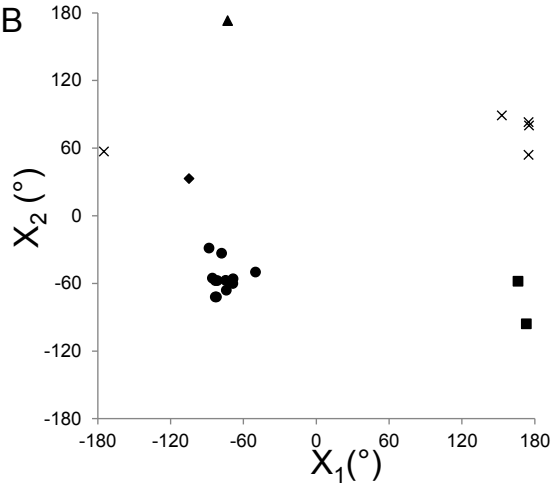
Taking into account all combinations of the X angles, there are 60 possible rotamers (5x2x3x2=60). However, these 60 rotamers include some conformations which contain intramolecular clashes. After removing conformations with internal atomic clashes and

minimization to alleviate minor clashes (please see Methods section for more details), 54 rotamers form the initial MTSSL rotamer library for RosettaEPR (Figure 7D).

A



B



C

X_1	-76 (10)	-73	170 (3)	173 (11)	180
X_2	-56 (13)	173	-77 (19)	73 (14)	180

X_3	90	-90

X_4	180		-75		75	
X_5	-77	77	100	-8	-100	8

D

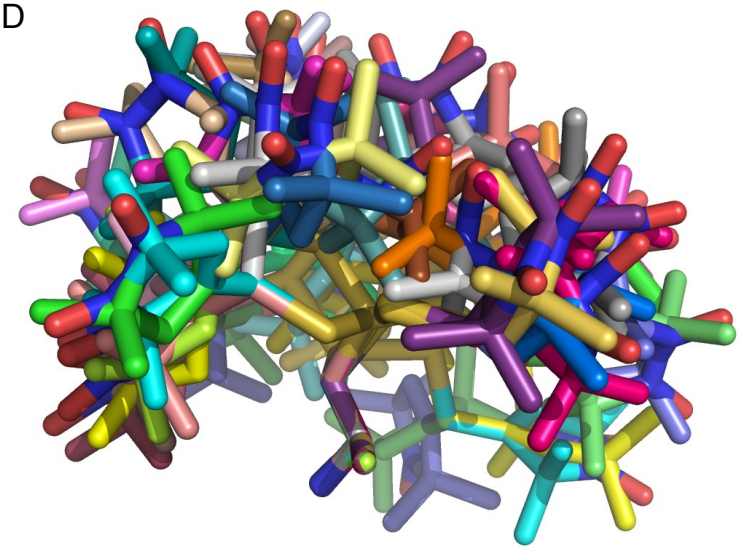


Figure 7 Characteristics of the MTSSL rotamer library.
A.) Designation of the five rotatable bonds in the methanethiosulfonate spin label (MTSSL) side chain. X1 is defined with the backbone nitrogen atom. X5 is defined by the doubly bonded carbon atom (bold) (Tombolato, Ferrarini et al. 2006) (Guo, Cascio et al. 2008). B.) Combinations of MTSSL X1 and X2 angles observed in T4 lysozyme crystallographically. {m, t} =▲; {m,m} = ●; (MOE) = ■; {t,p} = x. The diamond (♦) denotes what is observed at core site mutant L118; excluding this point, four groups of X1 and X2 combinations are observed. C.) Combinations of X angles used in the MTSSL rotamer library. X1 and X2 are correlated and there are five combinations possible. X3 is not correlated with any other X angle and there are two possible conformations of X3. X4 and X5 are correlated such that for each X4 angle, there are two possible X5 angles. Enumerating the possible combinations gives $\mathbf{5 \times 2 \times 3 \times 2 = 60}$ total possible rotamer conformations. Numbers in parentheses give standard deviations, if available. D.) After removing conformations with internal clashes, 54 rotamers remain in the library.

Table 6 Combinations of X_1 and X_2 leading to the combinations contained in the rotamer library.
Only the {t,t} combination is not observed experimentally and solely predicted from computational methods. The bottom two rows show the average and standard deviation, respectively.

{m,m}		{t,p}		{t,m}		{m,t}		{t,t}	
-68	-56	175	80	166	-58	-73	173	180	180
-69	-60	175	83	173	-96				
-81	-57	175	54						
-88	-29	153	89						
-78	-33	185	57						
-83	-57								
-85	-55								
-75	-57								
-82	-72								
-83	-72								
-50	-50								
-74	-66								
-76	-56	173	73	170	-77	-73	173	180	180
10	13	11	14	3	19	0	0	0	0

Ability of Rosetta to recover experimentally observed spin label conformations

MTSSL mutants of the soluble T4 lysozyme protein (17 mutants) and the LeuT membrane protein (2 mutants, Table 7) were used to demonstrate the ability of Rosetta to recover conformations of spin labels experimentally observed. For each mutant, approximately 1,000 independent relaxation trajectories were conducted and the percentage of models finding the experimentally observed X angles was calculated (Table 8). Values within ±30° were considered correct (Guo, Cascio et al. 2008). The

percentages are computed such that preceding X angles must be correct before a more distal angle can be counted as correct. For example, Rosetta predicts the crystallized X_1 angle of T4 lysozyme mutant T151[100] 100% of the time and predicts both, the experimental X_1 and X_2 angles, 51% of the time correctly. If there is more than one empirical conformation, a model rotamer is counted as correct if it matches any experimentally observed conformation.

Table 7 Experimentally determined MTSSL conformations for single mutants of LeuT. *Mutant* indicates the residue of LeuT which was mutated to the MTSSL side chain. Subscripts denote the protein subunit from the crystal structure asymmetric unit as indicated in the PDB file. *Temp.* gives the temperature at which the crystal was formed. *Environ.* gives the environment in which the residue lies: on the surface of the protein (surface); within the core of the protein (core); at the contact point of two a crystallographic subunits (crystal contact). *SSE Type* gives the type of secondary structure element on which the mutated residue sits. *Rotamer* indicates the X_1 and X_2 angles observed for the spin label in the crystal structure according to the m, t, p convention of (Lovell, Word et al. 2000). The X angles observed in the crystal structure are shown in their respective columns. *PDB ID* is the Protein Data Bank accession identifier of the crystal structure. *Ref* provides the primary citation for the crystal structure.

Mutant	Temp. (K)	Environ.	SSE Type	Rotamer	X_1(°)	X_2(°)	X_3(°)	X_4(°)	X_5(°)	PDB ID	Ref.
F177	100	Surface	Helix	{m,m}	-69	-57	107	103	-24	3MPN	(Kroncke, Horanyi et al. 2010)
I204	100	Surface	Helix	{m,m}	-69	-59	-87	-71	-95	3MPQ	(Kroncke, Horanyi et al. 2010)

Excluding crystal contact sites, Rosetta samples the correct rotamer for all of the remaining fourteen structures. X_1 and X_2 are correctly predicted in nine out of fourteen cases with at least 50% frequency. In seven out of twelve cases for T4 lysozyme, Rosetta recovers all experimentally observed X angles at least 50% of the time. On average for the fourteen mutants of T4 lysozyme and LeuT, recovery of experimentally observed X_1 and X_2 occurs in 53% of sampling trajectories (Figure 8, Figure 9). X_1 –X_5 is observed only four times, and Rosetta samples frequently the observed angles for site L118 (see below). The other three sites are surface sites (see below).

Table 8 Ability to recover experimentally observed conformations of MTSSL.
top) T4 lysozyme bottom) LeuT. For each crystallographically observed single mutant, the percentage of Rosetta relaxation trajectories that recover experimental conformations. *Mutant is* the site which was mutated. Superscripts indicate the temperature at which the crystal was formed. Subscripts indicate the component of the crystallographic asymmetric unit as indicated in the PDB file. *Environ.* gives the environment in which the residue lies: on the surface of the protein (surface); within the core of the protein (core); at the contact point of two crystallographic subunits (crystal contact). The remaining X angle columns denote the percentage of models out of the 1000 relaxation trajectories that recover the experimental conformation. The X angle recovery is dependent on previous X angles, e.g. X_2 cannot be correct unless X_1 is correct. Blank columns indicate those X angles were not resolved crystallographically. The average recovery of non-crystal contact sites is given.

T4 Lysozyme Mutant	**Environ.**	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**
L118	core	99.8	99.8	99.8	99.8	99.8
$S044_C$	surface	67.6	0.1			
R080	surface	100.0	100.0			
A082	surface	61.8	61.8	61.8		
$T115^{100}$	surface	79.3	63.6	63.6	2.4	2.4
$T115^{298}$	surface	95.0	19.9			
T115/R119A	surface	24.4	1.1			
R119	surface	74.9	1.2			
$V131^{100}$	surface	100.0	99.9			
$V131^{291}$	surface	99.9	99.2			
$T151^{100}$	surface	100.0	51.0			
$T151^{291}$	surface	98.6	78.1			
Mean		83.4	56.3	75.1	51.1	51.1
A041	crystal contact	60.2	56.0	46.3		
$S044_A$	crystal contact	13.7	0.0	0.0	0.0	0.0
$S044_B$	crystal contact	0.8	0.0	0.0	0.0	0.0
K065	crystal contact	34.0	33.2	0.0		
V075	crystal contact	1.2	1.2	1.2	0.2	
LeuT Mutant	**Environ.**	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**
F177	Surface	10.0	2.5	0.2	0.2	0.2
I204	Surface	77.8	60.2	3.9	2.8	2.8
mean	43.9	31.4	2.1	1.5	1.5	1.5

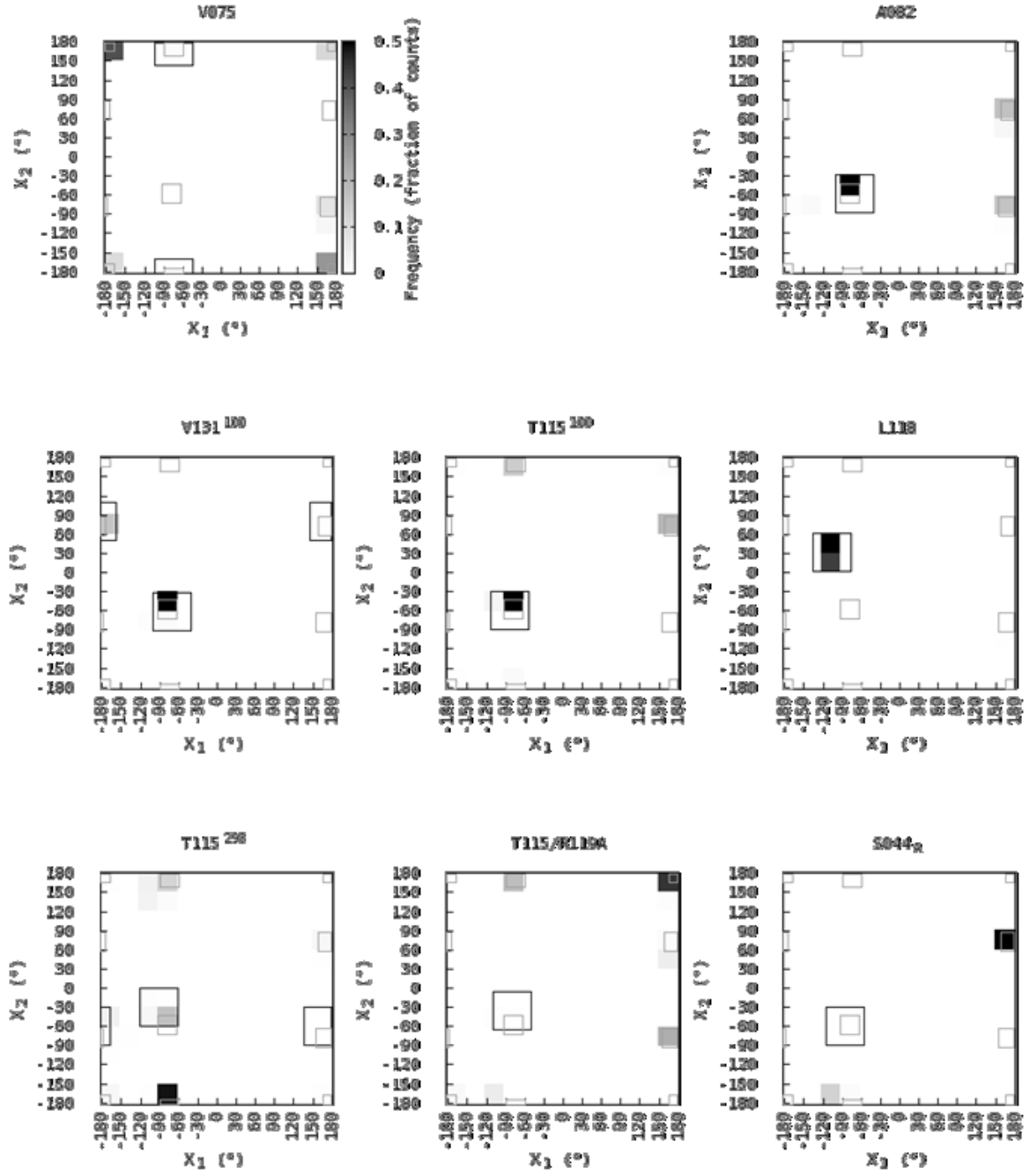


Figure 8 All experimentally observed MTSSL X1 and X2 angles for single mutants of t4-lysozyme.

Squares with dark lines indicate the experimentally observed X_1 and X_2 values ±30°. Squares with light grey lines indicate combinations of X_1 and X_2 which are contained in the rotamer library. The frequency with which combinations of X_1 and X_2 which are sampled by Rosetta for each single mutant are given according to grey scale with white areas never being sampled and darker areas being sampled more frequently.

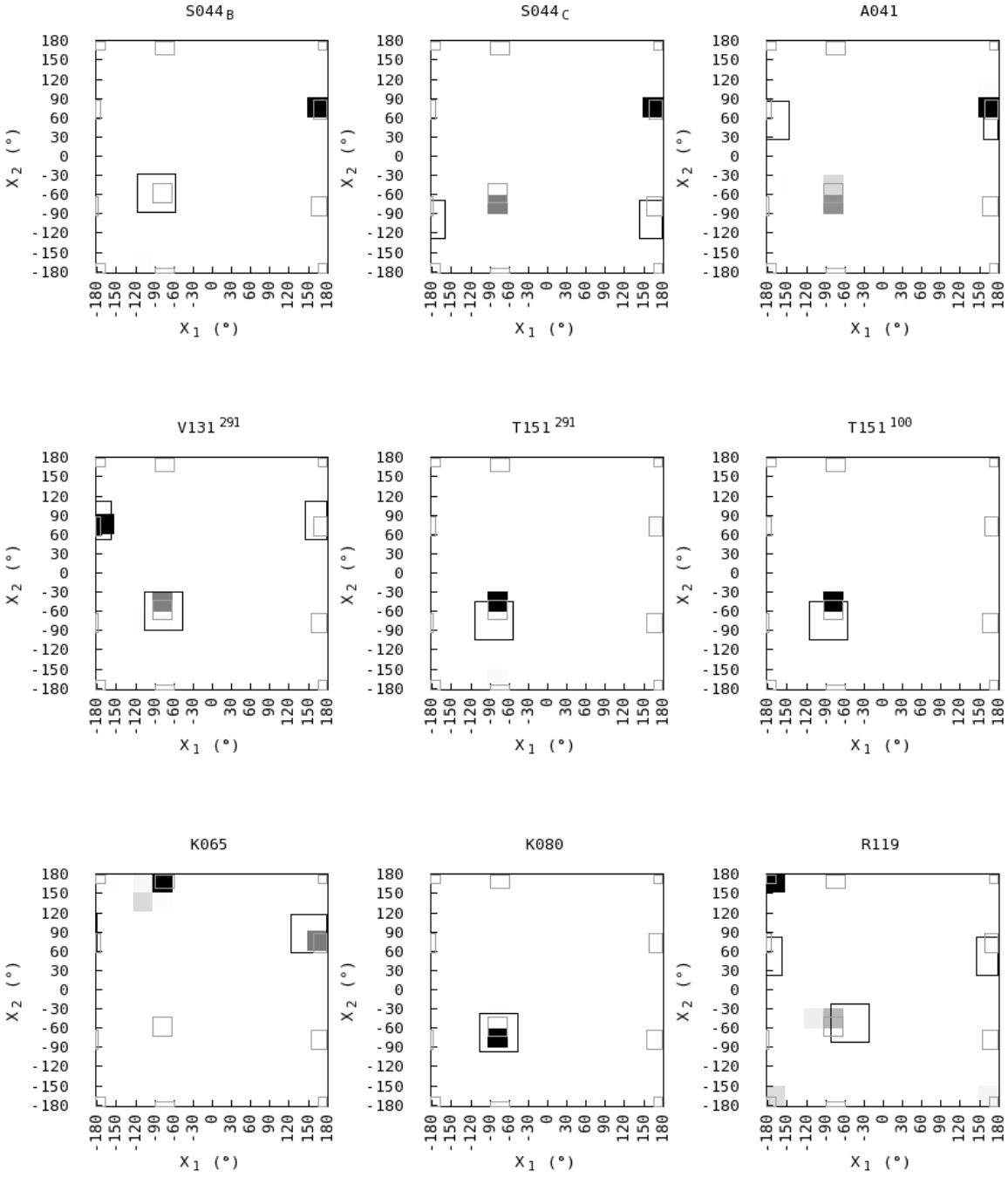


Figure 8 continued.

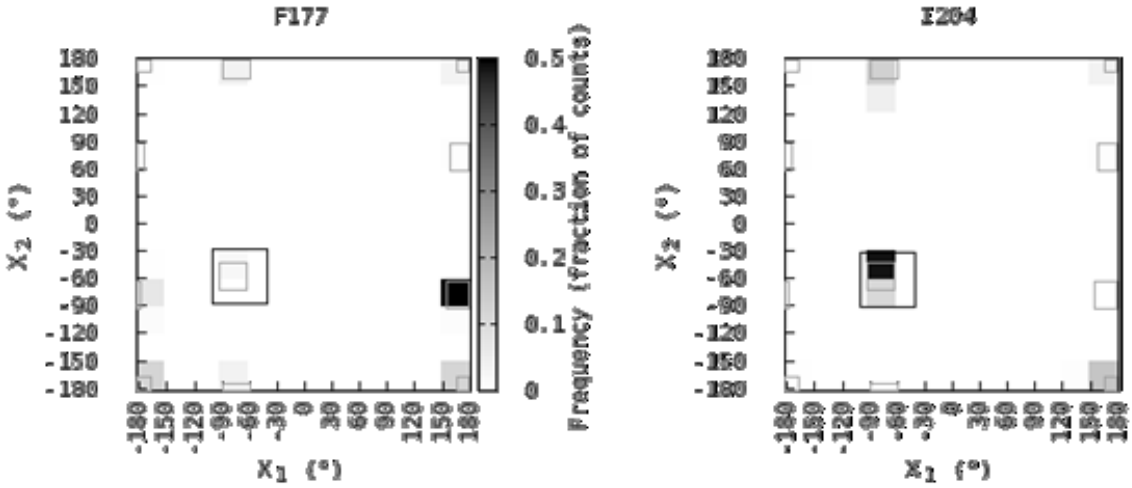


Figure 9 All experimentally observed MTSSL X1 and X2 angles for single mutants of LeuT. Squares with dark lines indicate the experimentally observed X_1 and X_2 values ±30°. Squares with light grey lines indicate combinations of X_1 and X_2 which are contained in the rotamer library. The frequency with which combinations of X_1 and X_2 which are sampled by Rosetta for each single mutant are given according to grey scale with white areas never being sampled and darker areas being sampled more frequently.

In the only mutant at a buried site L118, Rosetta recovers the experimentally observed X angles 99.8% of the time. The pocket in which the spin label resides greatly restricts the number of possible non-clashing conformations (Figure 10A). The crystallized X_1 and X_2 angles are distorted from the expected values due to the steric constraints of the pocket. In spite of the X_1 and X_2 not being in the rotamer library, Rosetta's potentials are able to accurately drive the spin label to adopt the correct conformation starting from one of the rotamers.

Surface mutants allow the spin label the possibility to adopt more conformations than core sites due to the reduced number of surrounding residues. In result, Rosetta finds often multiple low-energy conformations for spin labels. This results in three scenarios: a) Rosetta almost exclusively (greater than 75%) samples the experimental X angles for four out of the thirteen surface mutants (Figure 10B); b) Rosetta sometimes (approximately 50%) samples the observed rotamers for two out of the thirteen surface

mutants (Figure 10C).; and c) Rosetta seldom (less than 20%) samples the experimental conformations for seven out of the thirteen surface mutants (Figure 10D). Three of these seven cases involve the instances where X_1 –X_5 are observed, making it difficult for Rosetta to find the experimental conformation for all the degrees of freedom. In the other four cases, only X_1 and X_2 are observed so it is difficult to determine what, if any, interactions lead Rosetta to frequently differ from the experimentally observed conformations.

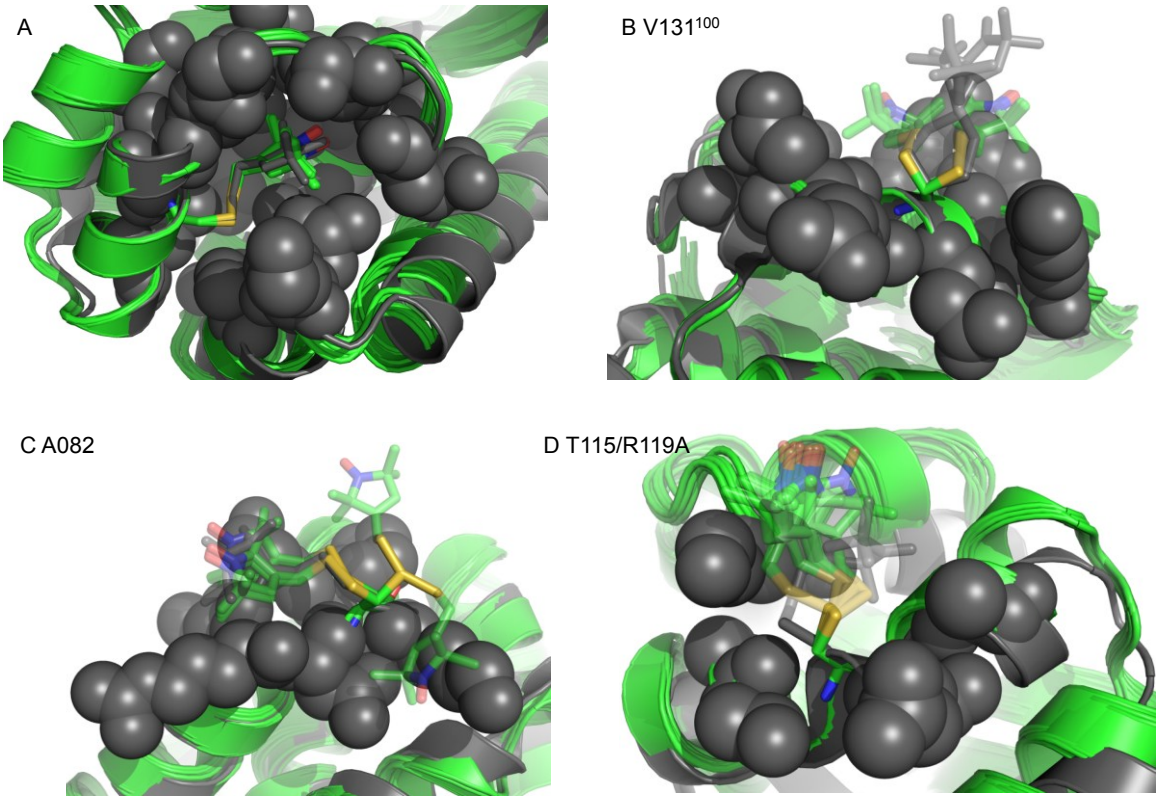


Figure 10 Ten best scoring Rosetta models (green) overlayed with the crystal structure (grey) for four examples of MTSSL mutated sites on T4 lysozyme.
Crystallographically observed X angles are shown solid, while atoms and X angles not experimentally seen are translucent. A) Rosetta's ability to recover a crystallographically observed spin label conformation at buried site 118 in T4 lysozyme. Spheres are used to indicate the buried nature of the site. B.) Two conformations of X1 and X2 were experimentally observed for single mutant site V131100. Rosetta models frequently sample these two conformations of X1 and X2. C) X1, X2, and X3 were experimentally observed for mutant A082. Several of the top ten conformations by Rosetta score sample these X angles, while other conformations are also sampled with a lower frequency. D) One conformation of X1 and X2 was observed for mutant T115/R119A. None of the ten best Rosetta models by score sample the experimental conformation.

With the exception of one mutant (A041), Rosetta is unable to successfully recover the observed X angles at crystal contact sites. The X angles of A041 are recovered with approximately the same frequency as the one of the surface mutants. Of the other spin labels placed at crystal contact sites, Rosetta samples all experimental X angles of only V075 and does so only 0.2% of the time (see Discussion).

Ability of Rosetta to recover experimental distance distributions

Fifty-eight EPR measured distance distributions have been collected for the T4 lysozyme protein (Borbat, McHaourab et al. 2002; Alexander, Bortolus et al. 2008; Kazmier 2010), including twelve new measurements. Additionally, nine EPR distance measurements of less than 70Å in transmembrane segments of the membrane protein MsbA in the apo-open and ten in the AMP-PNP bound state have previously been collected (Zou, Bortolus et al. 2009). These data provide an opportunity to test Rosetta's ability to recover experimental distance distributions. Such distributions can be roughly characterized as an average distance (μ_{EPR}) and a standard deviation (σ_{EPR}). Each spin labeled double mutant model for T4 lysozyme and MsbA was subjected to 2000 and about 1000 independent relaxation trajectories within Rosetta, respectively.

Table 9 Statistical measures of how well Rosetta recovers μ EPR and σ EPR for T4 lysozyme and MsbA double mutants.
when using the best 200 and 100 models by Rosetta score, respectively. The mean absolute error (MAE), root mean square deviation (RMSD), and correlation coefficient (R) of the mean distances for the ensembles, $\mu_{Rosetta}$, are calculated compared to the experimental mean distances. The same measures are calculated for the standard deviation of the Rosetta ensemble distance distributions, $\sigma_{Rosetta}$, compared to corresponding experimental standard deviations. Combined and CombinedCβ give values calculated from all T4 lysozyme and MsbA double mutants for spin label distances and Cβ distances, respectively.

	$\mu_{Rosetta}$			$\sigma_{Rosetta}$		
	MAE	RMSD	R	MAE	RMSD	R
T4lysozyme	3.5	4.5	0.92	0.9	1.1	0.56
MsbA apo-open	6.8	7.6	0.53	2.5	3.5	0.67
MsbA AMP-PNP bound	7.0	10.2	0.72	2.6	3.5	0.24
Combined	4.4	6.1	0.89	1.3	2.0	0.58
$Combined^{C\beta}$	6.1	7.1	0.93	2.7	3.2	0.55

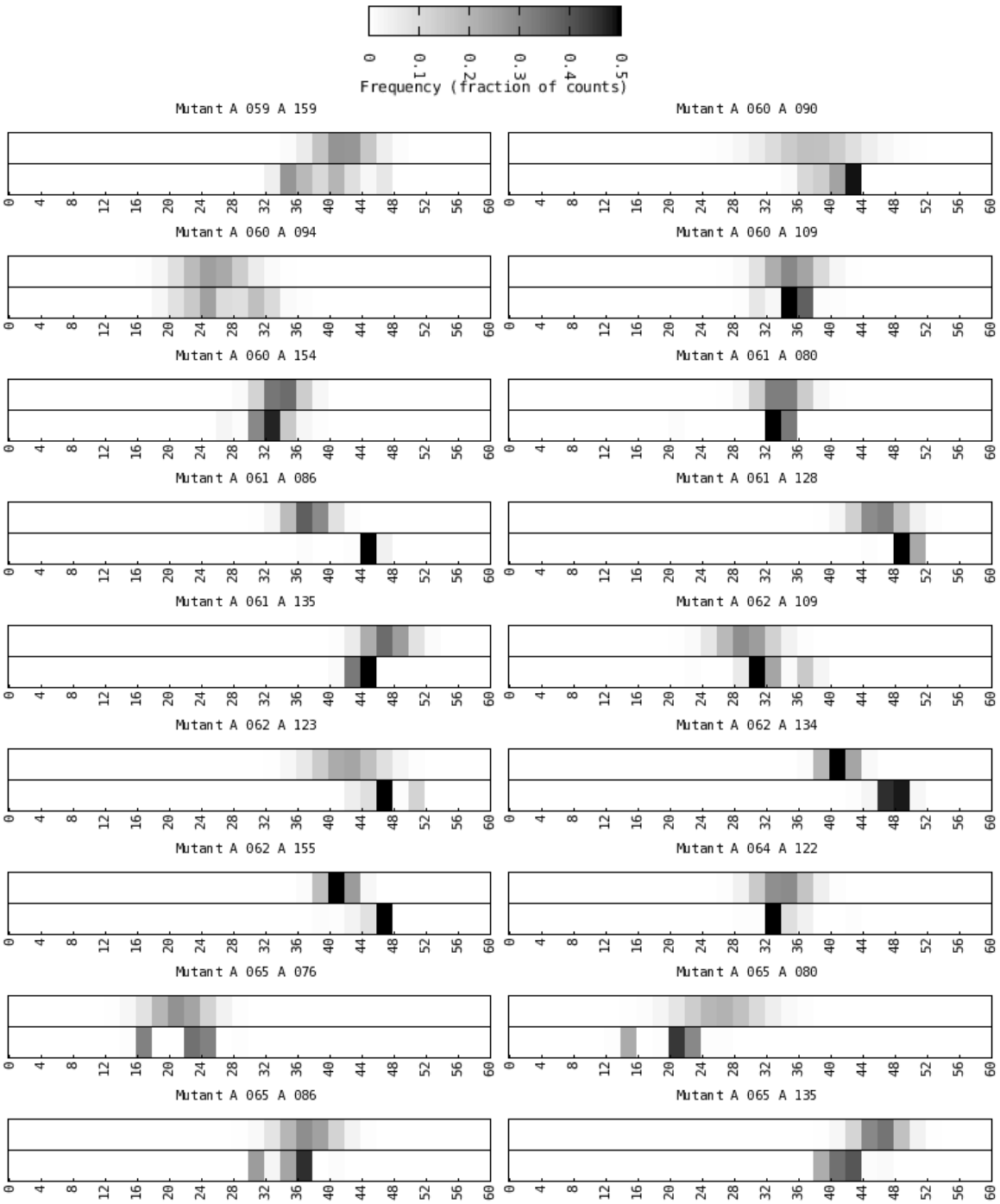


Figure 11 Heat maps for 58 double mutants of t4-lysozyme.
Shown are Gaussian distributions given by experimentally measured mean and standard deviation parameters compared with distance distributions recovered by Rosetta from the top 200 models according to Rosetta score. *Experimental* distance distributions are the *top bar* and Rosetta distributions are the bottom bar for each pair of heat maps. Distances are given in Angstroms, and the probability of observing a distance is defined by grayscale. Mutants 131/154, 131/151, 140/147, 116/131 were excluded from statistical analysis but are shown here for completeness.

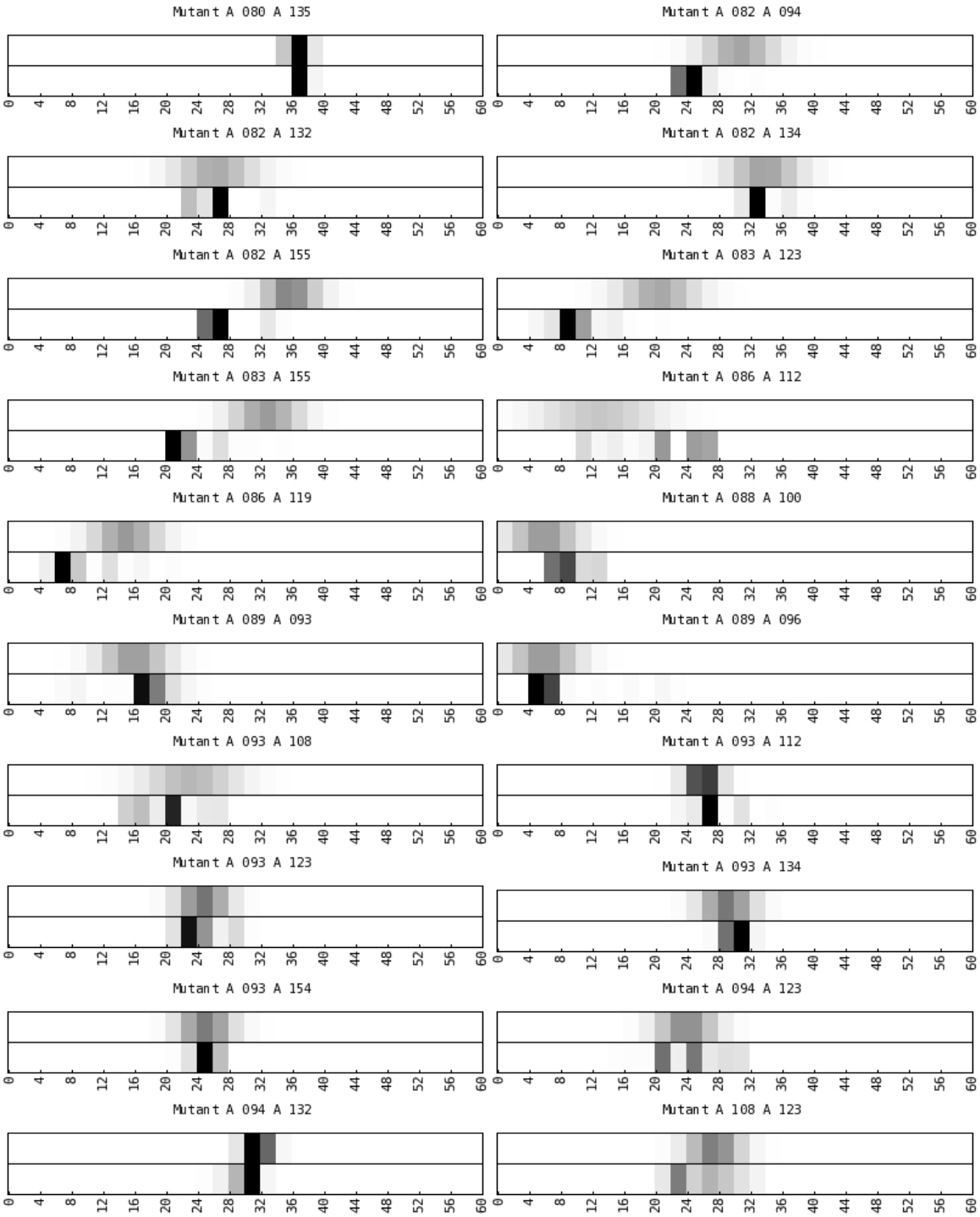


Figure 11 continued.

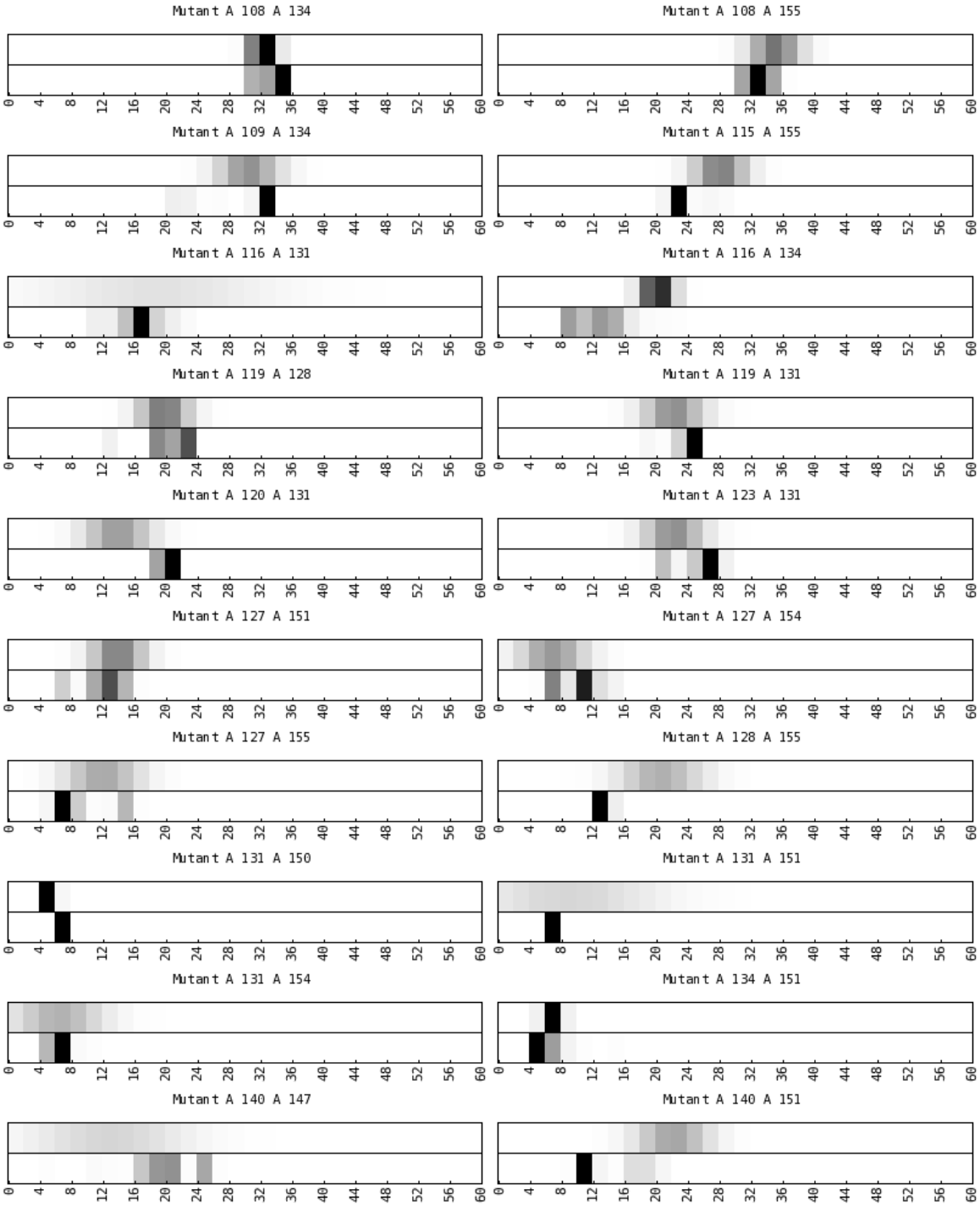


Figure 11 continued.

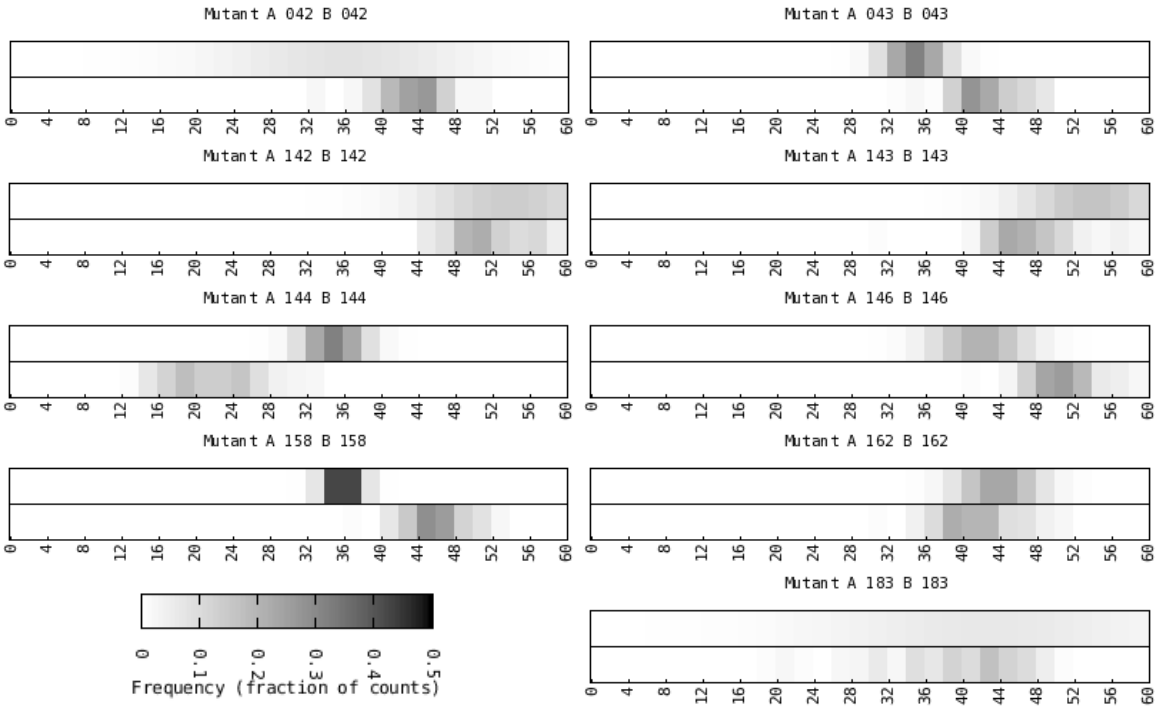


Figure 12 Heat maps for 9 double mutants of MSBA in the apo-open state.
Shown are Gaussian distributions given by experimentally measured mean and standard deviation parameters compared with distance distributions recovered by Rosetta from the top 100 models according to Rosetta score. Experimental distance distributions are the top bar and Rosetta distributions are the bottom bar for each pair of heat maps. Distances are given in Angstroms, and the probability of observing a distance is defined by grayscale.

The mean ($\mu_{Rosetta}$) and standard deviation ($\sigma_{Rosetta}$) of the inter-spin label distance was then calculated for the best 200 and 100 models according to Rosetta score for T4 lysozyme and MsbA, respectively. Four T4 lysozyme double mutants (131/154, 131/151, 140/147, 116/131) were excluded from analysis due to a high uncertainty in the accuracy of the measurement as determined by instances where the standard deviation is greater than 50% of the measured distance. The midpoint of the N-O bond was used as the location of the unpaired electron (Polyhach, Bordignon et al.).

Across all distance distributions, Rosetta achieves a mean absolute error (MAE) for $\mu_{Rosetta}$ versus μ_{EPR} of 4.4 Å (Table 9, Figure 11, Figure 12, Figure 13). This compares to a MAE of 6.1 when Cβ atoms are used to approximate the position of the spin label, indicating that Rosetta is able to provide

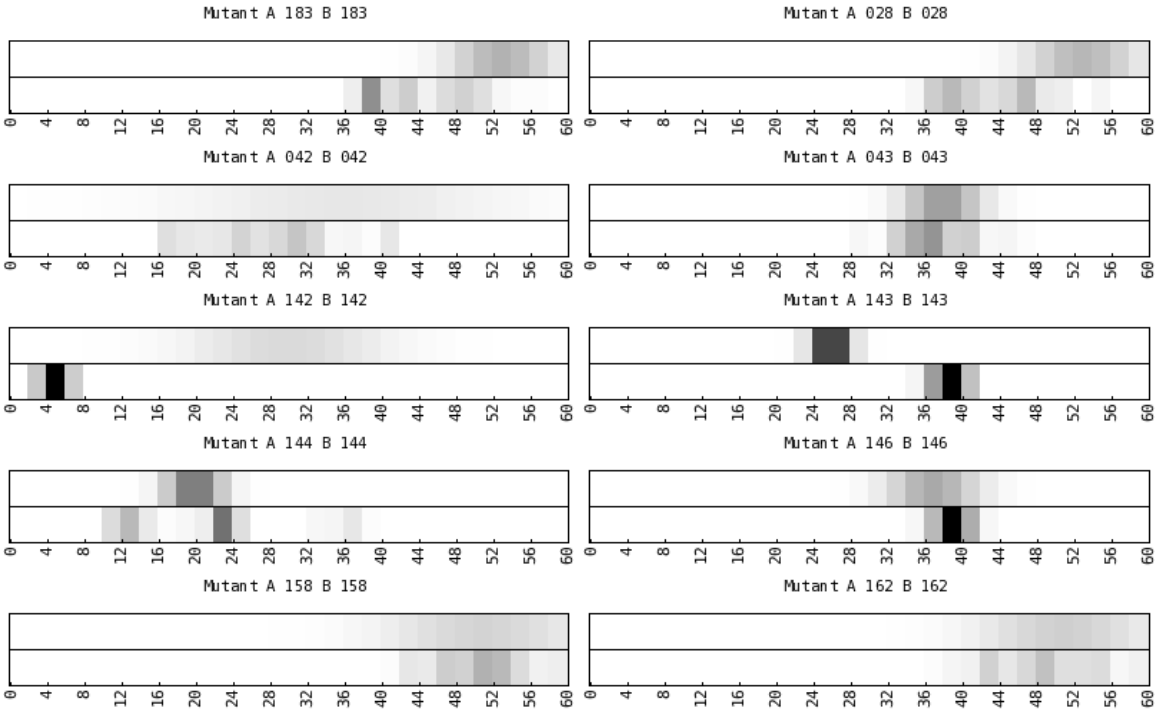


Figure 13 Heat maps for 10 double mutants of MSBA in the AMP-PNP bound state. Shown are Gaussian distributions given by experimentally measured mean and standard deviation parameters compared with distance distributions recovered by Rosetta from the top 100 models according to Rosetta score. Experimental distance distributions are the top bar and Rosetta distributions are the bottom bar for each pair of heat maps. Distances are given in Angstroms, and the probability of observing a distance is defined by grayscale same as Figure 12.

additional, more accurate information compared to a simple Cβ approximation for the spin label. On the T4 lysozyme dataset, the MAE for $\mu_{Rosetta}$ compared to μ_{EPR} is 3.5 Å (Figure 14A *circles*,Table 10). This is an improvement over simply using Cβ atoms, which gives a MAE of 5.7 Å (Table 13). For the MsbA dataset, the MAE for $\mu_{Rosetta}$ compared to μ_{EPR} is 6.8 Å (Figure 14A *crosses*, Table 11) and 7.0 Å (Figure 14A *triangles*, Table 12) for the apo open and AMP-PNP bound states, respectively. This offers a 0.4 Å improvement in MAE for the AMP-PNP bound state when compared to using Cβ distances (Table 14, Table 15).

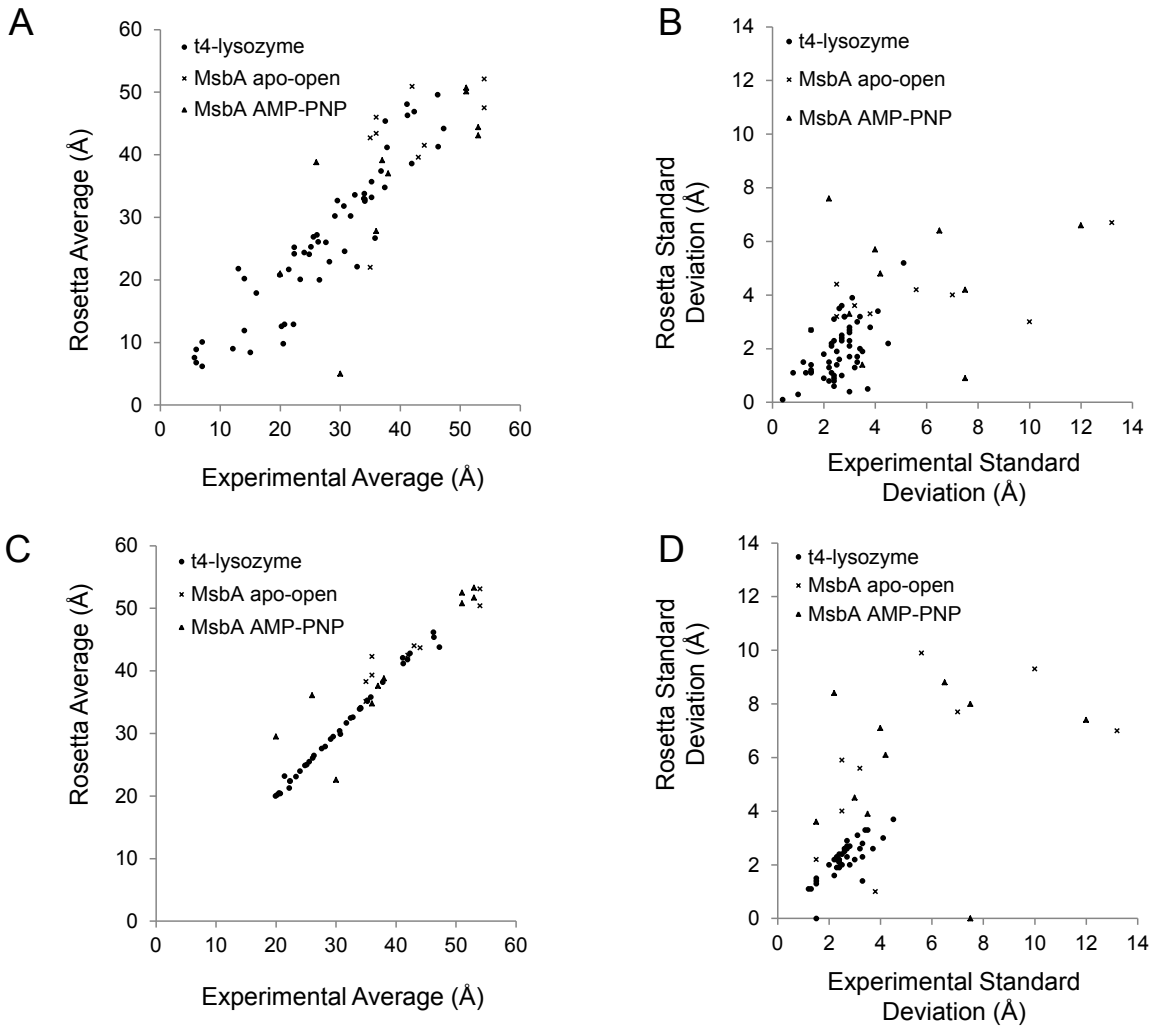


Figure 14 Plots of the average distance and standard deviation of ensembles of T4 lysozyme and MsbA double mutant distance distributions sampled by Rosetta versus the experimentally determined mean and standard deviation.
A and B) The ensembles of the best 200 (for T4 lysozyme) and 100 (for MsbA) models by Rosetta score. C) and D) The ensembles of Rosetta models determined by fitting the models to the experimental distance distributions.

The standard deviation of the distribution of distances determined in an EPR distance measurement (σ_{EPR}) indicates the breadth of conformations of MTTSSL and of the backbone sampled by the ensemble of labeled proteins present during the experiment. The standard deviation for the distribution of distances determined by Rosetta ($\sigma_{Rosetta}$) for all double mutants achieves a MAE to σ_{EPR} of 1.3 Å (Table 9). The MAE of $\sigma_{Rosetta}$ across the T4 lysozyme dataset is 0.9 Å (Figure 14B *circles*,Table 10),

compared to MAE of 2.4 Å if Cβ are used to approximate the spin label position (Table 13). For the MsbA datasets in the apo-open and AMP-PNP bound states, $\sigma_{Rosetta}$ has an MAE of 2.5 Å (Figure 14B *crosses*, Table 11) and 2.6 Å (Figure 14 B *triangles*, Table 12), respectively. Compared to using Cβ approximations, $\sigma_{Rosetta}$ is better in MAE by 0.6 Å and 1.1 Å for the apo-open and AMP-PNP bound states of MsbA, respectively (Table 14, Table 15).

Broad distributions of distances measured for MsbA in the apo-open and AMP-PNP bound states make it difficult for Rosetta to recover μ_{EPR} and σ_{EPR} as accurately as is done for T4 lysozyme. The average σ_{EPR} over the nineteen MsbA measurements is 5.3 as opposed to 2.6 for the T4 lysozyme distributions, and the distributions can contain multiple peaks spread out over a wide range of distances. This is indicative of significant backbone fluctuations independent of spin label conformation. Rosetta's difficulty with reproducing μ_{EPR} and σ_{EPR} for MsbA therefore arises a) due to the difficulty in summarizing broad complex distributions into a mean and standard deviation and b) because the relaxation protocol is not expected to produce large backbone changes.

Table 10 The average (μ) and standard deviation (σ) of inter-spin label distance distributions for double mutants (AA1 and AA2) of t4-lysozyme
Calculations are from the best 200 Rosetta models according to score and from EPR experiment, respectively. The deviation of Rosetta from experiment in terms μ and σ is also given for each double mutant. The bottom four rows show the mean deviation, standard deviation of the deviation, RMSD, and the correlation coefficient (R) of Rosetta with experiment. Double mutants 116-131, 131-151, 131-154, and 140-147 are not included in the statistics calculations.

AA_1	AA_2	$\mu_{Rosetta}$	$\sigma_{Rrosetta}$	μ_{EPR}	σ_{EPR}	$\lvert\mu_{Rosetta} - \mu_{EPR}\rvert$	$\lvert\sigma_{Rrosetta} - \sigma_{EPR}\rvert$
59	159	38.6	3.6	41.9	2.7	3.3	0.9
60	90	41.2	2.2	37.8	4.5	3.4	2.3
60	94	26.9	3.9	25.5	3.1	1.4	0.8
60	109	35.7	1.6	35.2	2.6	0.5	1.0
60	154	32.6	1.8	34.1	2.0	1.5	0.2
61	80	33.8	1.3	34.0	2.2	0.2	0.9
61	86	45.4	0.9	37.5	2.0	7.9	1.1
61	128	49.6	0.6	46.2	2.4	3.4	1.8
61	135	44.2	0.8	47.2	2.2	3.0	1.4
62	109	32.7	2.5	29.5	2.7	3.2	0.2
62	123	46.9	1.7	42.3	3.3	4.6	1.6
62	134	48.1	1.1	41.1	1.5	7.0	0.4
62	155	46.3	1.2	41.2	1.5	5.1	0.3

Table 10 continued

64	122	32.9	1.4	34.1	2.5	1.2	1.1
65	76	21.7	3.2	21.4	2.8	0.3	0.4
65	80	20.0	2.8	26.5	3.8	6.5	1.0
65	86	34.8	2.3	37.4	2.7	2.6	0.4
65	135	41.3	1.5	46.3	2.2	5.0	0.7
80	135	37.4	0.3	36.8	1.0	0.6	0.7
82	94	24.6	1.5	30.7	3.3	6.1	1.8
82	132	26.1	1.9	26.3	3.5	0.2	1.6
82	134	33.0	1.3	33.9	3.2	0.9	1.9
82	155	26.7	1.9	35.8	2.5	9.1	0.6
83	123	9.8	2.0	20.5	3.4	10.7	1.4
83	155	22.1	2.1	32.8	3.0	10.7	0.9
86	112	21.8	5.2	13.0	5.1	8.8	0.1
86	119	8.4	2.7	15.0	3.0	6.6	0.3
88	100	8.9	1.7	6.0	3.0	2.9	1.3
89	93	17.9	2.6	16.0	3.0	1.9	0.4
89	96	6.8	2.8	6.0	3.0	0.8	0.2
93	108	20.1	3.4	23.3	4.1	3.2	0.7
93	112	27.2	1.4	26.1	1.5	1.1	0.1
93	123	24.1	2.2	24.8	2.3	0.7	0.1
93	134	30.2	0.8	29.1	2.4	1.1	1.6
93	154	25.3	0.9	25.1	2.4	0.2	1.5
94	123	24.4	3.5	24.0	2.6	0.4	0.9
94	132	30.2	1.1	31.7	1.3	1.5	0.2
108	123	26.0	3.1	27.6	2.4	1.6	0.7
108	134	33.6	1.5	32.4	1.2	1.2	0.3
108	155	33.2	1.1	35.2	2.3	2.0	1.2
109	134	31.8	3.2	30.6	2.8	1.2	0.4
115	155	22.9	1.0	28.2	2.4	5.3	1.4
116	134	12.6	2.7	20.2	1.5	7.6	1.2
119	128	20.8	2.1	19.9	2.3	0.9	0.2
119	131	24.2	1.0	22.3	2.7	1.9	1.7
120	131	20.2	0.4	14.0	3.0	6.2	2.6
123	131	25.2	2.4	22.3	2.7	2.9	0.3
127	151	11.9	2.3	14.0	2.4	2.1	0.1
127	154	10.1	2.3	7.0	3.0	3.1	0.7
127	155	9.0	3.2	12.1	3.4	3.1	0.2
128	155	12.9	0.5	20.7	3.7	7.8	3.2
131	150	7.6	0.1	5.7	0.4	1.9	0.3
134	151	6.2	1.1	7.0	0.8	0.8	0.3
140	151	12.9	3.0	22.2	3.3	9.3	0.3
μ						3.5	0.9
σ						2.9	0.7
RMSD						4.5	1.1
R						0.92	0.56

Table 11 The average (μ) and standard deviation (σ) of inter-spin label distance distributions for double mutants (AA_1 and AA_2) of MSBA in the apo open state.
Calculations are from the best 100 Rosetta models according to score and from EPR experiment, respectively. The deviation of Rosetta from experiment in terms μ and σ is also given for each double mutant. The bottom four rows show the mean deviation, standard deviation of the deviation, RMSD, and the correlation coefficient (R) of Rosetta with experiment.

AA_1	AA_2	$\mu_{Rosetta}$	$\sigma_{Rrosetta}$	μ_{EPR}	σ_{EPR}	$\lvert\mu_{Rosetta} - \mu_{EPR}\rvert$	$\lvert\sigma_{Rrosetta} - \sigma_{EPR}\rvert$
42	42	43.4	3.0	36.0	10.0	7.4	7.0
43	43	42.7	3.2	35.0	2.5	7.7	0.7
142	142	52.1	4.0	54.0	7.0	1.9	3.0
143	143	47.5	4.2	54.0	5.6	6.5	1.4
144	144	22.0	4.4	35.0	2.5	13.0	1.9
146	146	50.9	3.3	42.0	3.8	8.9	0.5
158	158	46.0	2.7	36.0	1.5	10.0	1.2
162	162	41.5	3.6	44.0	3.2	2.5	0.4
183	183	39.6	6.7	43.0	13.2	3.4	6.5
μ						6.8	2.5
σ						3.5	2.4
RMSD						7.6	3.5
R						0.53	0.66

Table 12 The average (μ) and standard deviation (σ) of inter-spin label distance distributions for double mutants (AA1 and AA2) of MSBA in the AMP-PNP bound state.
Calculations are from the best 100 Rosetta models according to score and from EPR experiment, respectively. The deviation of Rosetta from experiment in terms μ and σ is also given for each double mutant. The bottom four rows show the mean deviation, standard deviation of the deviation, RMSD, and the correlation coefficient (R) of Rosetta with experiment.

AA_1	AA_2	$\mu_{Rosetta}$	$\sigma_{Rrosetta}$	μ_{EPR}	σ_{EPR}	$\lvert\mu_{Rosetta} - \mu_{EPR}\rvert$	$\lvert\sigma_{Rrosetta} - \sigma_{EPR}\rvert$
28	28	43.1	4.8	53.0	4.2	9.9	0.6
42	42	27.8	6.6	36.0	12.0	8.2	5.4
43	43	37.0	3.3	38.0	3.0	1.0	0.3
142	142	5.0	0.9	30.0	7.5	25.0	6.6
143	143	38.8	1.2	26.0	1.5	12.8	0.3
144	144	21.0	7.6	20.0	2.2	1.0	5.4
146	146	39.1	1.4	37.0	3.5	2.1	2.1
158	158	50.7	4.2	51.0	7.5	0.3	3.3
162	162	50.1	6.4	51.0	6.5	0.9	0.1
183	183	44.4	5.7	53.0	4.0	8.6	1.7
μ						7.0	2.6
σ						7.4	2.3
RMSD						10.2	3.5
R						0.72	0.24

Table 13 Using Cβ atoms to approximate the position of the spin label, the average (μ) and standard deviation (σ) of inter-spin label distance distributions for double mutants (AA1 and AA2) of t4-lysozyme.
Calculations are from the best 200 Rosetta models according to score and from EPR experiment, respectively. The deviation of Rosetta from experiment in terms μ and σ is also given for each double mutant. The bottom four rows show the mean deviation, standard deviation of the deviation, RMSD, and the correlation coefficient (R) of Rosetta with experiment.

AA_1	AA_2	$\mu_{Rosetta}^{C\beta}$	$\sigma_{Rosetta}^{C\beta}$	μ_{EPR}	σ_{EPR}	$\left\|\mu_{Rosetta}^{C\beta}-\mu_{EPR}\right\|$	$\left\|\sigma_{Rosetta}^{C\beta}-\sigma_{EPR}\right\|$
59	159	33.5	0.2	41.9	2.7	8.4	2.5
60	90	36.6	0.2	37.8	4.5	1.2	4.3
60	94	28.1	0.6	25.5	3.1	2.6	2.5
60	109	31.0	0.3	35.2	2.6	4.2	2.3
60	154	34.1	0.4	34.1	2.0	0.0	1.6
61	80	28.3	0.1	34.0	2.2	5.7	2.1
61	86	36.5	0.1	37.5	2.0	1.0	1.9
61	128	42.6	0.3	46.2	2.4	3.6	2.1
61	135	39.9	0.3	47.2	2.2	7.3	1.9
62	109	27.0	0.3	29.5	2.7	2.5	2.4
62	123	41.3	0.2	42.3	3.3	1.0	3.1
62	134	35.5	0.3	41.1	1.5	5.6	1.2
62	155	34.8	0.2	41.2	1.5	6.4	1.3
64	122	33.6	0.2	34.1	2.5	0.5	2.3
65	76	16.5	0.1	21.4	2.8	4.9	2.7
65	80	22.1	0.1	26.5	3.8	4.4	3.7
65	86	30.9	0.1	37.4	2.7	6.5	2.6
65	135	36.1	0.2	46.3	2.2	10.2	2.0
80	135	26.6	0.2	36.8	1.0	10.2	0.8
82	94	23.0	0.1	30.7	3.3	7.7	3.2
82	132	19.9	0.2	26.3	3.5	6.4	3.3
82	134	25.3	0.2	33.9	3.2	8.6	3.0
82	155	27.5	0.1	35.8	2.5	8.3	2.4
83	123	14.2	0.3	20.5	3.4	6.3	3.1
83	155	23.9	0.2	32.8	3.0	8.9	2.8
86	112	11.2	0.3	13.0	5.1	1.8	4.8
86	119	10.3	0.3	15.0	3.0	4.7	2.7
88	100	8.3	0.3	6.0	3.0	2.3	2.7
89	93	12.1	0.1	16.0	3.0	3.9	2.9
89	96	8.6	0.2	6.0	3.0	2.6	2.8
93	108	20.9	0.4	23.3	4.1	2.4	3.7
93	112	21.7	0.2	26.1	1.5	4.4	1.3
93	123	18.5	0.1	24.8	2.3	6.3	2.2
93	134	23.0	0.2	29.1	2.4	6.1	2.2
93	154	15.6	0.2	25.1	2.4	9.5	2.2
94	123	17.4	0.2	24.0	2.6	6.6	2.4
94	132	18.2	0.2	31.7	1.3	13.5	1.1
108	123	21.3	0.2	27.6	2.4	6.3	2.2
108	134	21.4	0.2	32.4	1.2	11.0	1.0
108	155	25.2	0.3	35.2	2.3	10.0	2.0
109	134	20.4	0.2	30.6	2.8	10.2	2.6
115	155	22.5	0.2	28.2	2.4	5.7	2.2
116	134	11.3	0.3	20.2	1.5	8.9	1.2
119	128	8.7	0.2	19.9	2.3	11.2	2.1
119	131	12.0	0.2	22.3	2.7	10.3	2.5
120	131	8.6	0.1	14.0	3.0	5.4	2.9
123	131	13.2	0.1	22.3	2.7	9.1	2.6
127	151	11.6	0.3	14.0	2.4	2.4	2.1
127	154	6.8	0.3	7.0	3.0	0.2	2.7
127	155	10.3	0.4	12.1	3.4	1.8	3.0

Table 13 continued

128	155	13.9	0.2	20.7	3.7	6.8	3.5
131	150	9.1	0.1	5.7	0.4	3.4	0.3
134	151	10.5	0.3	7.0	0.8	3.5	0.5
140	151	16.6	0.1	22.2	3.3	5.6	3.2
μ						5.7	2.4
σ						3.3	0.9
RMSD						6.6	2.5
R						0.93	0.10

Table 14 Using Cβ atoms to approximate the position of the spin label, the average (μ) and standard deviation (σ) of inter-spin label distance distributions for double mutants (AA_1 and AA_2) of MSBA in the apo open state.
Calculations are from the best 100 Rosetta models according to score and from EPR experiment, respectively. The deviation of Rosetta from experiment in terms μ and σ is also given for each double mutant. The bottom four rows show the mean deviation, standard deviation of the deviation, RMSD, and the correlation coefficient (R) of Rosetta with experiment.

AA_1	AA_2	$\mu^{C\beta}_{Rosetta}$	$\sigma^{C\beta}_{Rosetta}$	μ_{EPR}	σ_{EPR}	$\left\|\mu^{C\beta}_{Rosetta}-\mu_{EPR}\right\|$	$\left\|\sigma^{C\beta}_{Rosetta}-\sigma_{EPR}\right\|$
42	42	31.8	2.4	36	10	4.2	7.6
43	43	33.4	1.8	35	2.5	1.6	0.7
142	142	40.3	4.3	54	7	13.7	2.7
143	143	37.8	3.3	54	5.6	16.2	2.3
144	144	29.1	3.4	35	2.5	5.9	0.9
146	146	40.4	3.2	42	3.8	1.6	0.6
158	158	36.9	1.9	36	1.5	0.9	0.4
162	162	37.1	1.3	44	3.2	6.9	1.9
183	183	33.2	2.3	43	13.2	9.8	10.9
μ						6.8	3.1
σ						5.2	3.5
RMSD						8.5	4.6
R						0.67	0.14

Table 15 Using Cβ atoms to approximate the position of the spin label, the average (μ) and standard deviation (σ) of inter-spin label distance distributions for double mutants (AA_1 and AA_2) of MSBA in the AMP-PNP bound state.
Calculations are from the best 100 Rosetta models according to score and from EPR experiment, respectively. The deviation of Rosetta from experiment in terms μ and σ is also given for each double mutant. The bottom four rows show the mean deviation, standard deviation of the deviation, RMSD, and the correlation coefficient (R) of Rosetta with experiment.

AA_1	AA_2	$\mu_{Rosetta}^{C\beta}$	$\sigma_{Rosetta}^{C\beta}$	μ_{EPR}	σ_{EPR}	$\left\|\mu_{Rosetta}^{C\beta}-\mu_{EPR}\right\|$	$\left\|\sigma_{Rosetta}^{C\beta}-\sigma_{EPR}\right\|$
28	28	40.9	0.7	53	4.2	12.1	3.5
42	42	24.4	2.4	36	12	11.6	9.6
43	43	29.4	2.5	38	3	8.6	0.5
142	142	18.7	0.9	30	7.5	11.3	6.6
143	143	26.5	0.8	26	1.5	0.5	0.7
144	144	21.5	1.7	20	2.2	1.5	0.5
146	146	25.6	1.1	37	3.5	11.4	2.4
158	158	43.6	1.7	51	7.5	7.4	5.8
162	162	48.2	2.2	51	6.5	2.8	4.3
183	183	46.7	0.6	53	4	6.3	3.4
μ						7.4	3.7
σ						4.2	2.8
RMSD						8.5	4.7
R						0.91	0.40

Fitting of Rosetta models to experimental distance distributions indicates RosettaEPR is robust enough to sample within all experimental distances probability distributions

For thirty-eight of the T4 lysozyme (Borbat, McHaourab et al. 2002; Alexander, Bortolus et al. 2008; Kazmier 2010) and all nineteen of the MsbA (Zou, Bortolus et al. 2009) experimental double mutant EPR measurements, distance probability distributions were available. These data sets allow the models generated for each double mutant by Rosetta to be used in a fitting procedure to determine, out of these models, an ensemble that can accurately reproduce the experimental distance distribution.

The 2000 models for each double mutant of T4 lysozyme and the top 1000 models by Rosetta score for each mutant of MsbA in the apo-open and AMP-PNP bound states were used to find an ensemble reproducing the corresponding distance distribution. After this procedure and across all double mutants, the MAE of the average distance calculated from the Rosetta ensemble, $\mu_{Rosetta}^{fitted}$, compared to μ_{EPR} is 1.1 Å (Table 16). For T4 lysozyme double mutants, the MAE of $\mu_{Rosetta}^{fitted}$ is 0.3 Å (Table 17),

Table 16 Statistical measures of how well Rosetta recovers μ_{EPR} and σ_{EPR} for T4 lysozyme and MsbA double mutants after selecting relaxed structures to match the experimental distance distributions.
Entries are similar to Table 9.

	$\mu_{Rosetta}^{fitted}$			$\sigma_{Rosetta}^{fitted}$		
	MAE	RMSD	R	MAE	RMSD	R
T4-lysozyme	0.3	0.7	1.00	0.4	0.6	0.80
MsbA apo-open	2.1	2.9	0.95	2.5	3.1	0.59
MsbA AMP-PNP bound	3.3	5.0	0.90	3.0	3.8	0.14
Combined	1.1	2.5	0.97	1.2	2.1	0.63

compared to 3.5 Å when models are selected solely by Rosetta score. The MAE of $\mu_{Rosetta}^{fitted}$ for the apo-open and AMP-PNP bound states of MsbA drops to 2.1 Å (Table 18) and 3.3 Å (Table 19), compared to 6.8 Å and 7.0 Å, respectively.

The standard deviation calculated from ensembles of Rosetta models selected to fit the corresponding distance distribution, $\sigma_{Rosetta}^{fitted}$, for T4 lysozyme double mutants achieves an MAE of 0.4 Å to σ_{EPR} compared to 0.9 Å when models are selected by Rosetta score alone. For double mutants of MsbA, the MAE of $\sigma_{Rosetta}^{fitted}$ in the apo-open and AMP-PNP bound states are 2.5 Å and 3.0 Å, respectively, which is not an improvement over selecting models strictly by score.

Instead of attempting to summarize the shape of distance distributions with μ and σ, using a measure to compare the entire distribution (cumulative Euclidean distance, see Methods) can more accurately describe the improvement in Rosetta's ability recover the distributions of T4 lysozyme and MsbA after fitting (Figure 15, Figure 16, Figure 17). For T4 lysozyme double mutants, the error in the ensembles of Rosetta models is reduced by an average of 87% (Table 20). Although $\sigma_{Rosetta}^{fitted}$ was not sensitive to improvements in the agreement between Rosetta and experimental distance distributions for MsbA, comparison of the distributions show an average reduction in error of 62% (Table 21) and 54% (Table 22) for the apo-open and AMP-PNP bound states, respectively. Over all

double mutants, the error is reduced by an average of 77% using an average ensemble size of 18 relaxed structures.

Table 17 The average (μ) and standard deviation (σ) of inter-spin label distance distributions for double mutants (AA_1 and AA_2) of t4-lysozyme as calculated from the best ensemble of Rosetta models fitted to the experimental distance probability distribution.
This fitted μ and σ is compared with μ and σ from experiment. The deviation of Rosetta from experiment in terms μ and σ is also given for each double mutant. The bottom four rows show the mean deviation, standard deviation of the deviation, RMSD, and the correlation coefficient (R) of Rosetta with experiment.

AA_1	AA_2	$\mu_{Rosetta}^{fitted}$	$\sigma_{Rosetta}^{fitted}$	μ EPR	σ EPR	$\left\|\mu_{Rosetta}^{fitted}-\mu_{EPR}\right\|$	$\left\|\sigma_{Rosetta}^{fitted}-\sigma_{EPR}\right\|$
59	159	41.8	2.7	41.9	2.7	0.1	0.0
60	90	38.2	3.7	37.8	4.5	0.4	0.8
60	94	25.5	3.1	25.5	3.1	0.0	0.0
60	109	35.2	2.5	35.2	2.6	0.0	0.1
60	154	34.1	2.0	34.1	2.0	0.0	0.0
61	128	46.2	2.1	46.2	2.4	0.0	0.3
61	135	43.8	1.6	47.2	2.2	3.4	0.6
62	109	29.5	2.9	29.5	2.7	0.0	0.2
62	123	42.8	2.8	42.3	3.3	0.5	0.5
62	134	42.1	0.0	41.1	1.5	1.0	1.5
62	155	41.2	1.5	41.2	1.5	0.0	0.0
64	122	34.0	2.4	34.1	2.5	0.1	0.1
65	76	23.2	2.7	21.4	2.8	1.8	0.1
65	135	45.4	2.2	46.3	2.2	0.9	0.0
82	94	29.9	2.3	30.7	3.3	0.8	1.0
82	132	26.5	3.3	26.3	3.5	0.2	0.2
82	134	33.9	2.6	33.9	3.2	0.0	0.6
82	155	35.8	2.0	35.8	2.5	0.0	0.5
83	123	20.5	3.3	20.5	3.4	0.0	0.1
83	155	32.6	2.2	32.8	3.0	0.2	0.8
93	108	23.1	3.0	23.3	4.1	0.2	1.1
93	112	26.1	1.3	26.1	1.5	0.0	0.2
93	123	24.9	2.2	24.8	2.3	0.1	0.1
93	134	29.1	2.2	29.1	2.4	0.0	0.2
93	154	25.0	2.1	25.1	2.4	0.1	0.3
94	123	24.0	2.6	24.0	2.6	0.0	0.0
94	132	31.7	1.1	31.7	1.3	0.0	0.2
108	123	27.6	2.4	27.6	2.4	0.0	0.0
108	134	32.5	1.1	32.4	1.2	0.1	0.1
108	155	35.3	1.9	35.2	2.3	0.1	0.4
109	134	30.4	2.0	30.6	2.8	0.2	0.8
115	155	27.9	1.9	28.2	2.4	0.3	0.5
116	134	20.2	1.4	20.2	1.5	0.0	0.1
119	128	20.0	2.3	19.9	2.3	0.1	0.0
119	131	22.4	2.3	22.3	2.7	0.1	0.4
123	131	22.4	2.6	22.3	2.7	0.1	0.1
128	155	20.4	2.6	20.7	3.7	0.3	1.1
140	151	21.3	1.4	22.2	3.3	0.9	1.9

Table 17 continued.

μ		0.3	0.4
σ		0.6	0.4
RMSD		0.7	0.6
R		0.996	0.80

Table 18 The average (μ) and standard deviation (σ) of inter-spin label distance distributions for double mutants (AA1 and AA2) of MSBA in the apo open state as calculated from the best ensemble of Rosetta models fitted to the experimental distance probability distribution.
This fitted μ and σ is compared with μ and σ from experiment. The deviation of Rosetta from experiment in terms μ and σ is also given for each double mutant. The bottom four rows show the mean deviation, standard deviation of the deviation, RMSD, and the correlation coefficient (R) of Rosetta with experiment.

AA_1	AA_2	$\mu_{Rosetta}^{fitted}$	$\sigma_{Rosetta}^{fitted}$	μ_{EPR}	σ_{EPR}	$\left\|\mu_{Rosetta}^{fitted}-\mu_{EPR}\right\|$	$\left\|\sigma_{Rosetta}^{fitted}-\sigma_{EPR}\right\|$
42	42	42.3	9.3	36.0	10.0	6.3	0.7
43	43	38.3	5.9	35.0	2.5	3.3	3.4
142	142	53.1	7.7	54.0	7.0	0.9	0.7
143	143	50.4	9.9	54.0	5.6	3.6	4.3
144	144	35.1	4.0	35.0	2.5	0.1	1.5
146	146	42.5	1.0	42.0	3.8	0.5	2.8
158	158	39.3	2.2	36.0	1.5	3.3	0.7
162	162	43.7	5.6	44.0	3.2	0.3	2.4
183	183	44.0	7.0	43.0	13.2	1.0	6.2
μ						2.1	2.5
σ						2.0	1.8
RMSD						2.9	3.1
R						0.95	0.59

Table 19 The average (μ) and standard deviation (σ) of inter-spin label distance distributions for double mutants (AA_1 and AA_2) of MSBA in the AMP-PNP bound state as calculated from the best ensemble of Rosetta models fitted to the experimental distance probability distribution.
This fitted μ and σ is compared with μ and σ from experiment. The deviation of Rosetta from experiment in terms μ and σ is also given for each double mutant. The bottom four rows show the mean deviation, standard deviation of the deviation, RMSD, and the correlation coefficient (R) of Rosetta with experiment.

AA_1	AA_2	$\mu_{Rosetta}^{fitted}$	$\sigma_{Rosetta}^{fitted}$	μ_{EPR}	σ_{EPR}	$\left\|\mu_{Rosetta}^{fitted}-\mu_{EPR}\right\|$	$\left\|\sigma_{Rosetta}^{fitted}-\sigma_{EPR}\right\|$
28	28	51.7	6.1	53.0	4.2	1.3	1.9
42	42	34.8	7.4	36.0	12.0	1.2	4.6
43	43	38.8	4.5	38.0	3.0	0.8	1.5
142	142	22.6	0.0	30.0	7.5	7.4	7.5
143	143	36.1	3.6	26.0	1.5	10.1	2.1
144	144	29.5	8.4	20.0	2.2	9.5	6.2
146	146	37.6	3.9	37.0	3.5	0.6	0.4
158	158	50.8	8.0	51.0	7.5	0.2	0.5
162	162	52.5	8.8	51.0	6.5	1.5	2.3
183	183	53.3	7.1	53.0	4.0	0.3	3.1
μ						3.3	3.0
σ						3.8	2.3
RMSD						5.0	3.8
R						0.90	0.14

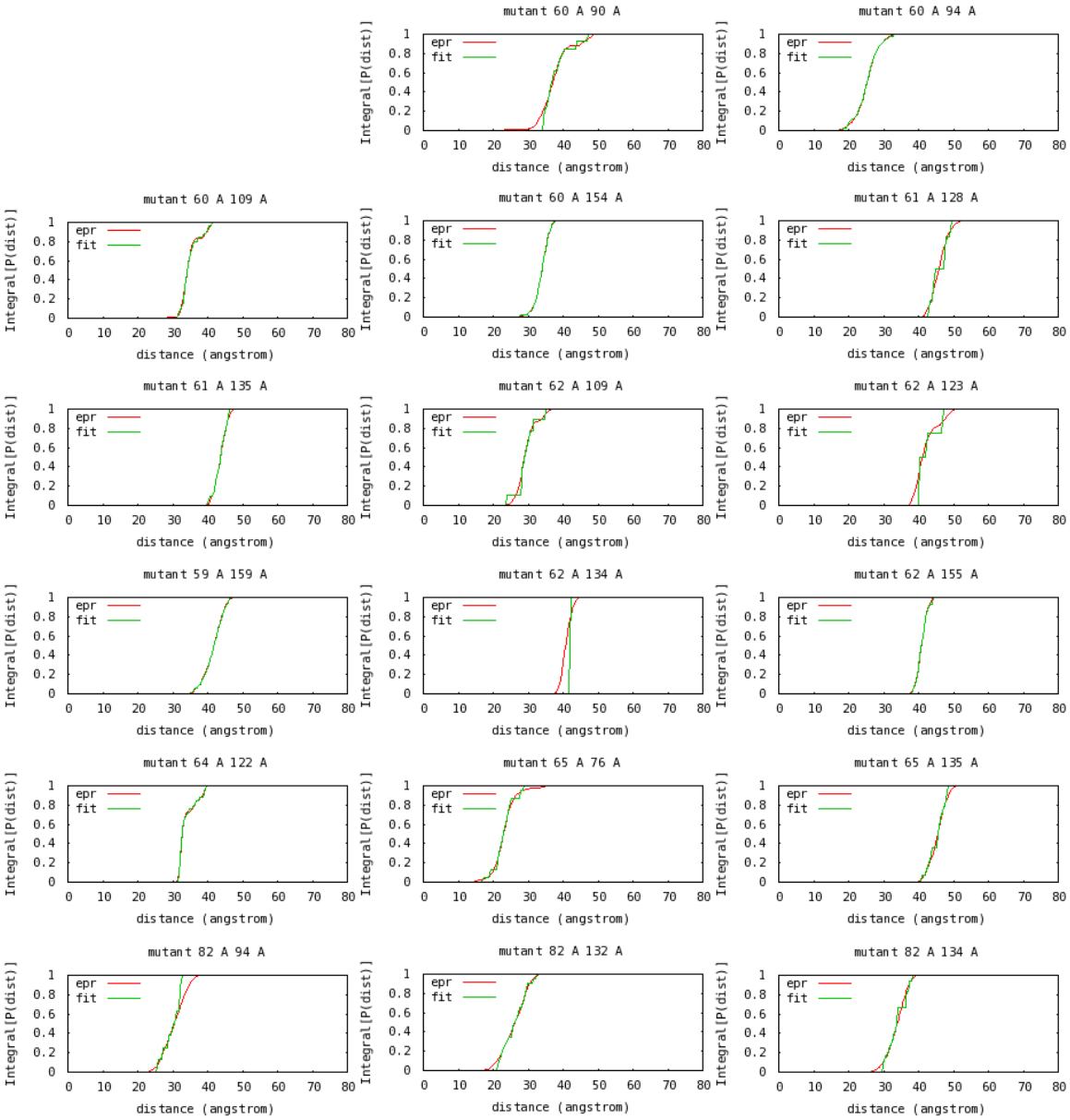


Figure 15 Agreement between experimental distance probability distributions and an ensemble of Rosetta models fitted to the experimental distribution for 38 double mutants of t4-lysozyme. Curves show the integral of the probability up to a given distance.

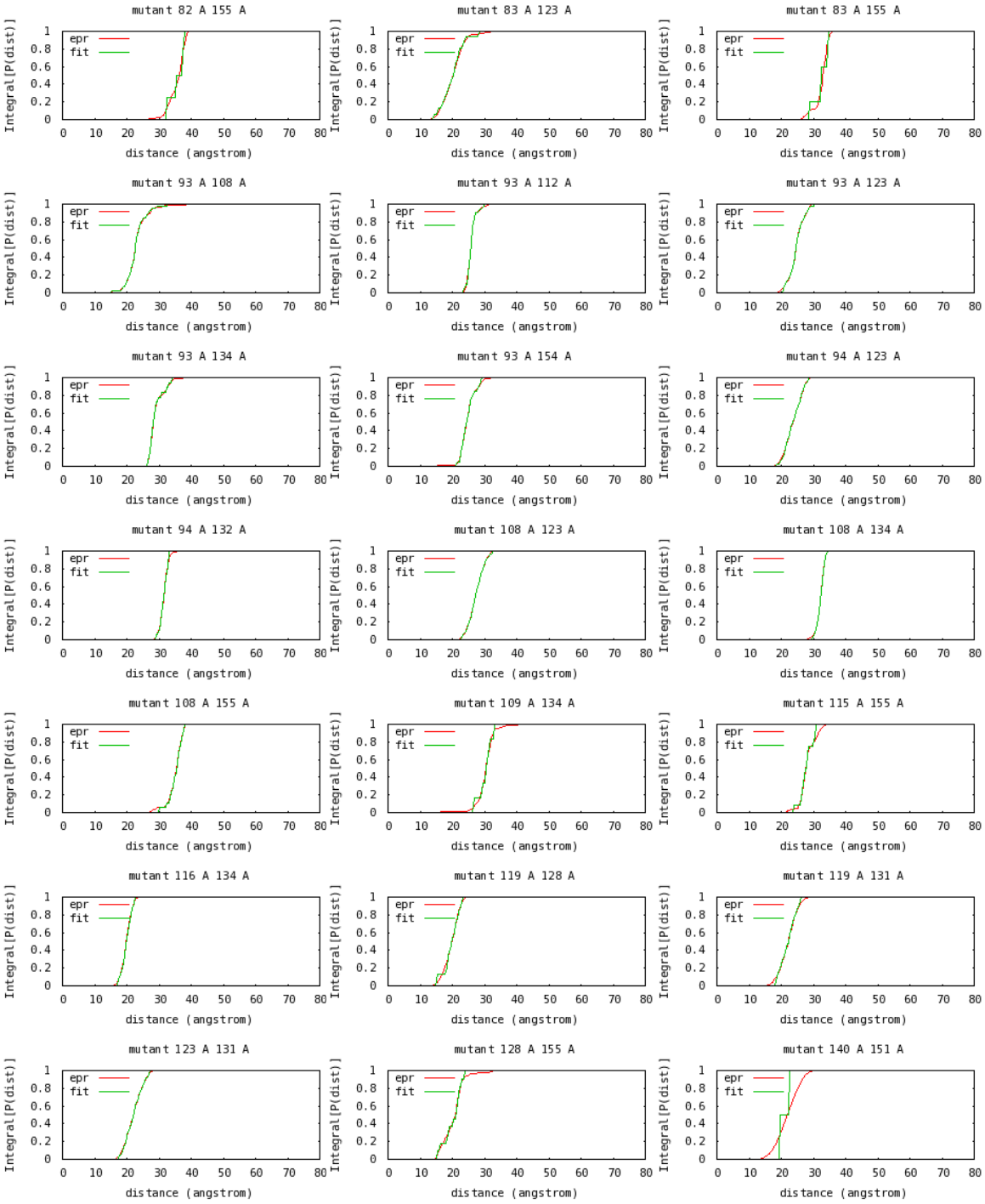


Figure 15 continued.

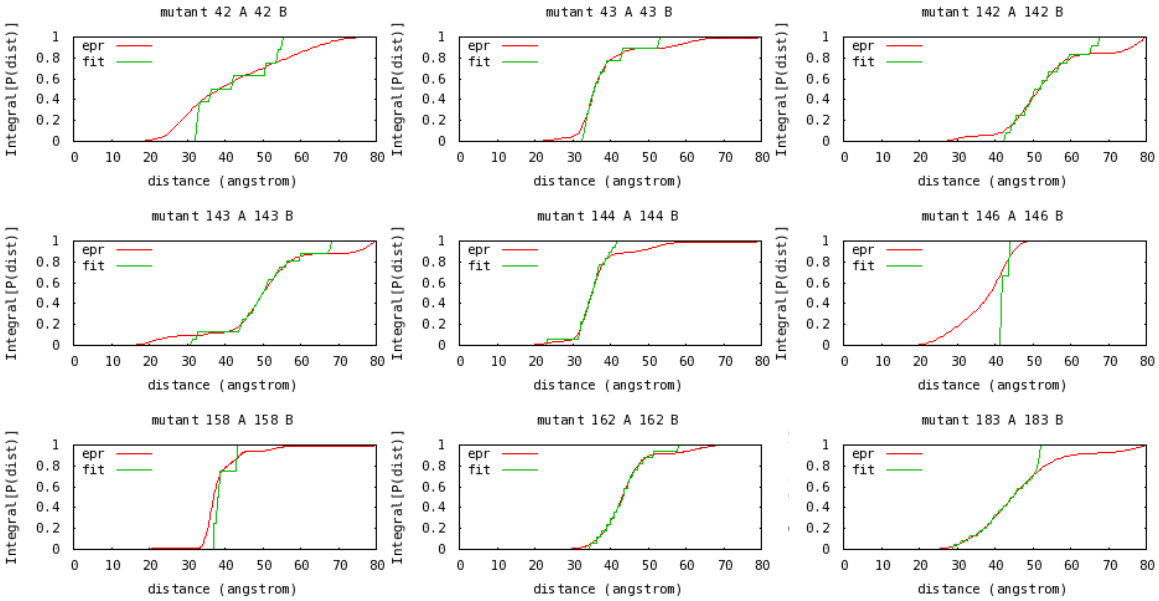


Figure 16 Agreement between experimental distance probability distributions and an ensemble of Rosetta models fitted to the experimental distribution for double mutants of MSBA in the apo-open state. Curves show the integral of the probability up to a given distance.

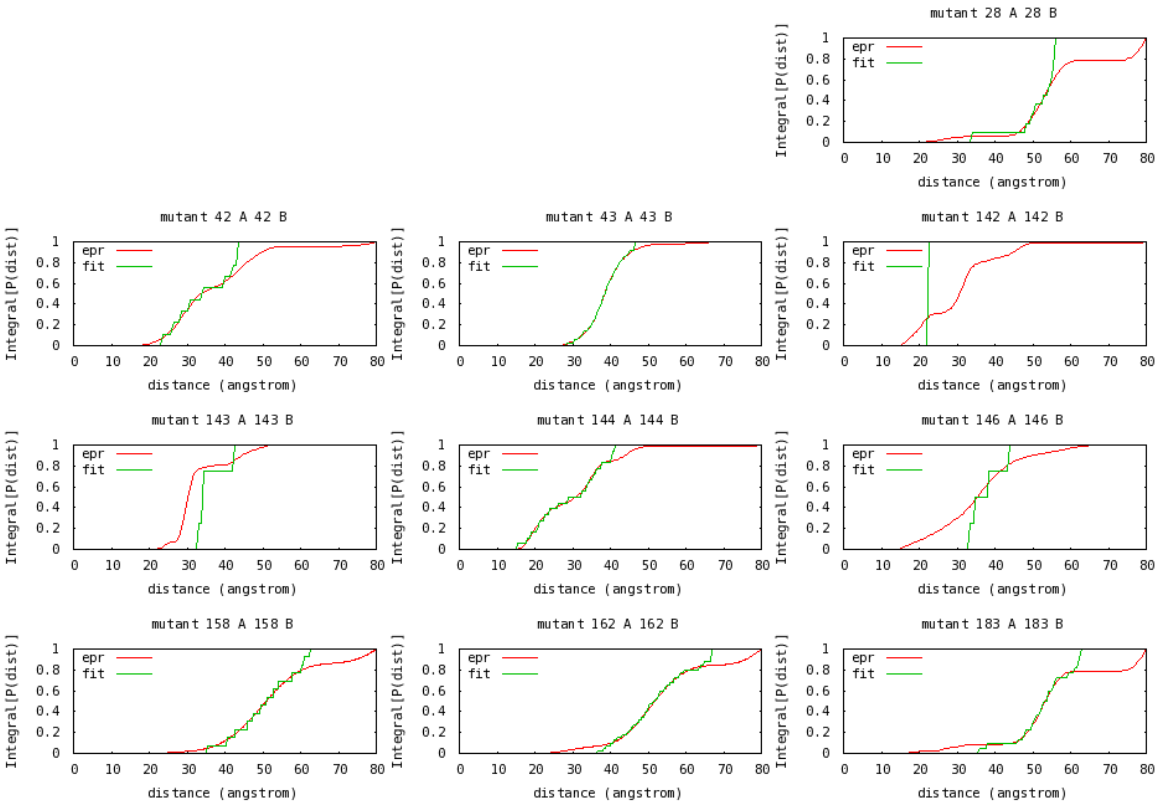


Figure 17 Agreement between experimental distance probability distributions and an ensemble of Rosetta models fitted to the experimental distribution for double mutants of MSBA in the AMP-PNP bound state. Curves show the integral of the probability up to a given distance.

Table 20 For t4-lysozyme, the cumulative Euclidean disagreement values of the best 200 relaxed structures by Rosetta score (Top 200 Disagreement) and an ensemble of Rosetta models selected to fit the experimental (Fitted Disagreement).
The disagreement is calculated between the distance distribution obtained from the Rosetta models and the corresponding experimental distance distribution, with 0 being perfect agreement. The size of the fitted ensemble is also provided (Size). The amount that the disagreement is reduced as a percentage of starting disagreement (Percent Disagreement Reduction) is calculated as $(Top200Disgreement - FittedDisagreement)/Top200Disagreement * 100$

AA_1	AA_2	Top 200 Disagreement	Fitted Disagreement	Size	Percent Disagreement Reduction
59	159	0.130	0.005	50	96.0
60	90	0.156	0.031	13	80.5
60	94	0.056	0.005	43	91.3
60	109	0.081	0.008	25	89.6
60	154	0.077	0.004	39	94.8
61	128	0.172	0.031	6	81.9
61	135	0.048	0.010	20	78.8
62	109	0.143	0.024	9	83.4
62	123	0.201	0.045	4	77.4
62	134	0.294	0.103	1	64.9
62	155	0.242	0.006	15	97.6
64	122	0.063	0.006	25	90.4
65	76	0.063	0.014	23	78.0
65	135	0.186	0.019	14	90.0
82	94	0.232	0.045	8	80.7
82	132	0.077	0.015	20	79.9
82	134	0.091	0.022	18	76.1
82	155	0.329	0.030	4	91.0
83	123	0.315	0.011	16	96.5
83	155	0.362	0.036	5	90.0
93	108	0.102	0.006	39	94.5
93	112	0.080	0.008	31	90.6
93	123	0.057	0.006	50	89.1
93	134	0.106	0.007	18	93.2
93	154	0.062	0.008	19	86.8
94	123	0.039	0.004	34	89.8
94	132	0.089	0.010	17	89.3
108	123	0.076	0.004	30	95.0
108	134	0.077	0.003	35	96.6
108	155	0.122	0.010	18	91.8
109	134	0.113	0.022	6	80.5
115	155	0.237	0.023	12	90.4
116	134	0.256	0.005	15	98.0
119	128	0.047	0.014	24	71.5
119	131	0.102	0.011	21	89.0
123	131	0.118	0.005	23	95.5
128	155	0.268	0.016	11	93.9
140	151	0.274	0.077	2	71.9
μ		0.146	0.019	20	87.3
σ		0.091	0.020	13	8.1

Table 21 For MsbA in the apo-open state, the cumulative Euclidean disagreement values of the best 200 relaxed structures by Rosetta score (Top 200 Disagreement) and an ensemble of Rosetta models selected to fit the experimental (Fitted Disagreement).
The disagreement is calculated between the distance distribution obtained from the Rosetta models and the corresponding experimental distance distribution, with 0 being perfect agreement. The size of the fitted ensemble is also provided (Size). The amount that the disagreement is reduced as a percentage of starting disagreement (Percent Disagreement Reduction) is calculated as $(Top200Disgreement - FittedDisagreement)/Top200Disagreement * 100$

AA_1	AA_2	Top 200 Disagreement	Fitted Disagreement	Size	Percent Disagreement Reduction
42	42	0.209	0.103	8	50.9
43	43	0.188	0.039	9	79.2
142	142	0.107	0.065	12	39.6
143	143	0.135	0.055	16	58.9
144	144	0.333	0.040	18	88.0
146	146	0.348	0.176	3	49.4
158	158	0.266	0.080	4	70.0
162	162	0.072	0.022	17	68.8
183	183	0.139	0.070	24	49.7
μ		0.200	0.072	12	61.6
σ		0.093	0.043	7	15.0

Table 22 For MsbA in the AMP-PNP state, the cumulative Euclidean disagreement values of the best 200 relaxed structures by Rosetta score (Top 200 Disagreement) and an ensemble of Rosetta models selected to fit the experimental (Fitted Disagreement).
The disagreement is calculated between the distance distribution obtained from the Rosetta models and the corresponding experimental distance distribution, with 0 being perfect agreement. The size of the fitted ensemble is also provided (Size). The amount that the disagreement is reduced as a percentage of starting disagreement (Percent Disagreement Reduction) is calculated as $(Top200Disgreement - FittedDisagreement)/Top200Disagreement * 100$

AA_1	AA_2	Top 200 Disagreement	Fitted Disagreement	Size	Percent Disagreement Reduction
28	28	0.308	0.136	11	55.7
42	42	0.173	0.065	9	62.3
43	43	0.077	0.017	34	78.0
142	142	0.499	0.240	1	52.0
143	143	0.256	0.167	4	34.5
144	144	0.187	0.033	18	82.5
146	146	0.184	0.124	4	32.5
158	158	0.114	0.067	13	41.1
162	162	0.097	0.059	40	38.9
183	183	0.268	0.105	22	60.8
μ		0.216	0.101	16	53.8
σ		0.119	0.064	12	16.6

Validation of implicit spin label cone model parameters

The introduction of a full atom representation of MTSSL within Rosetta allows the explicit description of the ensemble of conformations accessible to spin labels attached to various sites on a protein. The previously published spin label cone-model implicitly described the ensemble of conformations using uniform parameters applied to all sites (Hirst, Alexander et al. ; Alexander, Bortolus et al. 2008). It defined an effective position for the spin label (SL_{ef}) as the positional average of all possible spin label locations as it projects from the protein backbone. The "cone model" assumes the allowable spin label positions are contained within a cone with a defined opening angle ($\angle^{max} SL_A C_\beta SL_B = 90°$, Figure 18 A), which corresponds to the maximum observed angle between any two spin labels with vertex C_β. The cone model also assumes the cone is oriented at a random angle with respect to the protein backbone ($\angle SL_{ef} C_\beta C_\alpha = 120°$, Figure 18 B). Lastly, as a trigonometric result of $\angle^{max} SL_A C_\beta SL_B$ and the length of the spin label tether (8.5Å), the cone model defines a distance from the C_β to the SL_{ef} ($D_{C_\beta}^{SL_{ef}} = 6Å$, Figure 18 C).

The Rosetta rotamer library was used to explicitly compute the cone model parameters and compare with the original assumptions. Residues at 162 exposed sites on the primarily α-helical T4 lysozyme (PDBid 2LZM) and β-strand chitinase (PDBid 2CWR) (Nakamura, Mine et al. 2008) proteins were computationally mutated to create 162 single spin labeled mutants. Each of these mutants was subjected to 500 independent Rosetta relaxation trajectories in order to obtain an ensemble of allowable spin label conformations at each site.

The parameters calculated from the Rosetta ensembles are comparable to the original cone model parameters (Table 23). The distribution of $\angle^{max} SL_A C_\beta SL_B$ values shows a mean 103° with standard deviation of 50° (Figure 19A).

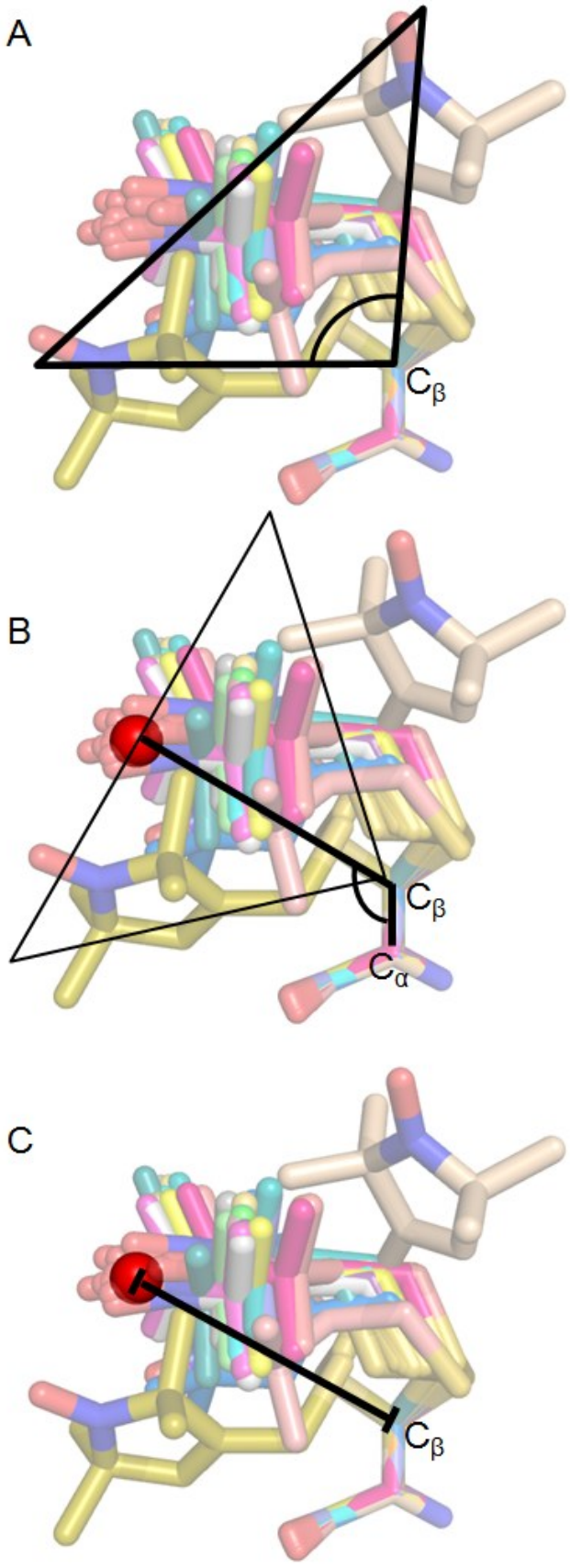

A
C_β
B
C_β
C_α
C
C_β

Figure 18 Visual description of the three parameters that define the cone model and their relation to the full atom representation of the spin label.
The effective spin label position, SL_{ef}, is the average position of the midpoint of the N-O bond vector. In B.) and C.) the SL_{ef} position is represented as a red sphere. A.) $\angle^{max}\boldsymbol{SL_A C_\beta SL_B}$ is the opening angle of the cone and is calculated as the widest angle observed between two MTSSL conformations obtained from Rosetta. B.) $\angle\boldsymbol{SL_{ef} C_\beta C_\alpha}$ is the angle defined by the C_α, C_β, and SL_{ef} positions, and gives information on the allowable tilt angles of the cone. C.) $\boldsymbol{D_{C_\beta}^{SL_{ef}}}$ is the distance from the C_β to the SL_{ef} position.

For$\angle SL_{ef} C_\beta C_\alpha$, the Rosetta distribution shows a mean of 111° and a standard deviation of 63° (Figure 19 B). The values of $D_{C_\beta}^{SL_{ef}}$ sampled by Rosetta have a mean of 6.3 Å and a standard deviation of 1.2 Å (Figure 19 C).

Figure 20 displays a comparison of D_{SL} - $D_{C\beta}$ statistics for the initial cone model (Alexander, Bortolus et al. 2008) with an updated cone model computed using the currently calculated parameters. D_{SL} is a distance between two spin labels, as approximated by the cone model. $D_{C\beta}$ is the distance between the C_β atoms of the residues containing the spin labels. With the increased length of $D_{C_\beta}^{SL_{ef}}$ and the decreased $\angle SL_{ef} C_\beta C_\alpha$ compared to initial values, there is an increased fraction of D_{SL} - $D_{C\beta}$ values between 10 Å and 12 Å. However, the small difference in the curves demonstrates the robustness of the cone model to small deviations in the parameters.

Table 23 Comparison of the parameters used by the cone model (Alexander, Bortolus et al. 2008) of a spin label and the values recovered by the rotamer library for the single mutants at 99 sites on a primarily alpha-helical protein and 63 sites on a beta-strand protein.
The mean and the standard deviation of the distributions for each parameter obtained from Rosetta for the 162 sites are given.

Parameter	Cone Model	Rosetta Explicit Spin Label Model	
		Mean	Standard Deviation
$\boldsymbol{D_{C_\beta}^{SL_{ef}}}$ (Å)	6.0	6.3	1.2
$\angle\boldsymbol{SL_{ef} C_\beta C_\alpha}$ (°)	120	111	63
$\angle^{max}\boldsymbol{SL_A C_\beta SL_B}$ (°)	90	103	50

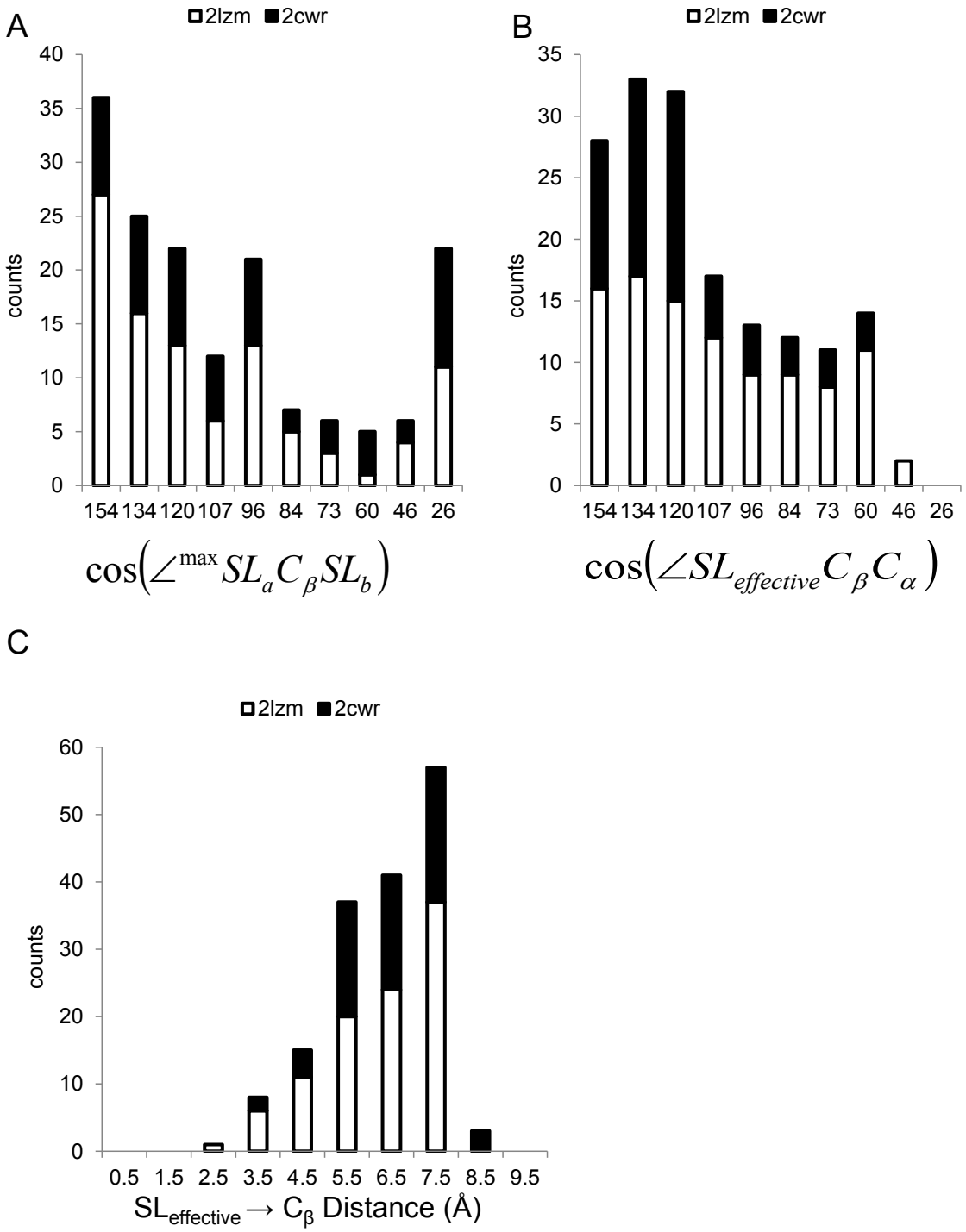


Figure 19 Distributions of the parameters that define the "cone model" as determined by Rosetta using the rotamer library full atom representation of MTSSL.

Shown are the frequencies with which given values of A.) $\angle^{max}\boldsymbol{SL_A C_\beta SL_B}$ B.) $\boldsymbol{D}_{C_\beta}^{SL_{ef}}$, and C.) $\angle\boldsymbol{SL_{ef} C_\beta C_\alpha}$ are observed by Rosetta at 162 singly labeled MTSSL sites on primarily alpha-helical and beta-strand proteins.

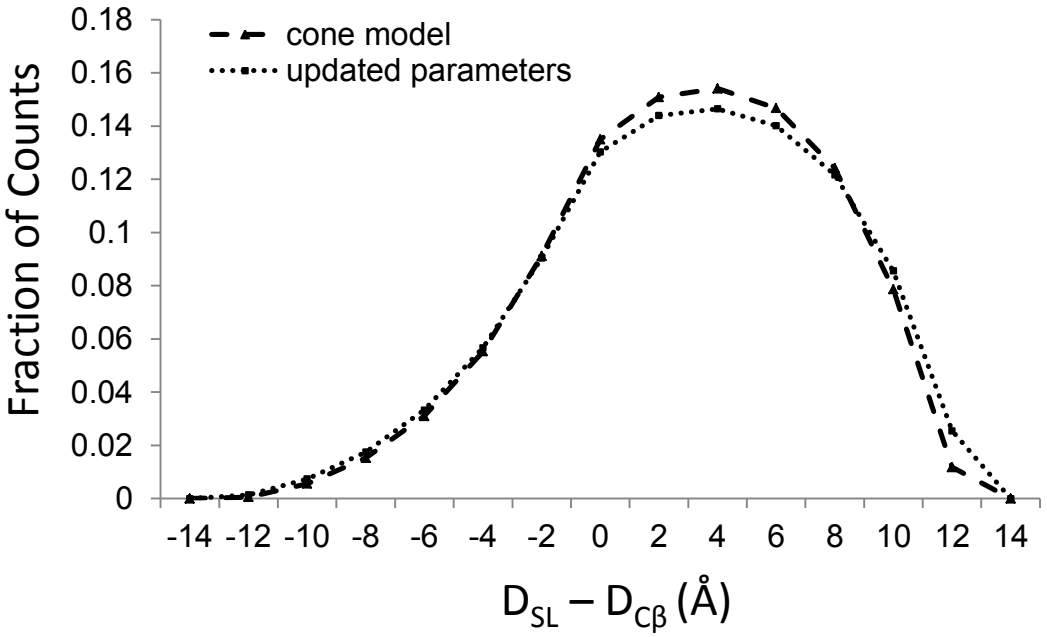


Figure 20 Statistics on the frequency with DSL - DCβ is observed for the initial (Alexander, Bortolus et al. 2008) cone model parameters (cone model) and the updated parameters calculated from RosettaEPR (updated parameters).
D_{SL} is a distance between two spin labels, where each has been randomly oriented and approximated by the corresponding cone model parameters. $D_{C\beta}$ is the distance between the C_β atoms of the residues containing the spin labels. The frequency is given on the y-axis as the fraction of observed D_{SL} - $D_{C\beta}$ values falling within a given bin.

Discussion

The RosettaEPR spin label rotamer library leverages experimentally observed and computationally predicted correlations between X angles of MTSSL. A rotamer library reduces the search space in order to produce a biologically probable conformation. Such efficiency allows RosettaEPR to sample in parallel with the spin label all other protein side chains and backbone degrees of freedom, rather than being restricted to a rigid protein structure. All-atom refinement of the protein structure allows determination of off-rotamer spin label conformations and can capture structural perturbations caused by the spin label.

RosettaEPR rotamer library combines experimentally determined spin label conformations with quantum chemical calculations

The present knowledge-base of experimentally observed MTSSL conformations is small. Therefore the current rotamer library supplements experimentally observed (X_1, X_2) combinations with computationally predicted X_{3-5} angles. Specifically, the (X_1, X_2) {t, t} rotamer has not yet been experimentally observed but was added to the rotamer library based on quantum chemical calculations (Tombolato, Ferrarini et al. 2006). X_3 was considered to be ±90°which is in agreement with both, experimental values and quantum chemical calculations (Tombolato, Ferrarini et al. 2006). Conformations for X_4 and X_5 were determined experimentally only four times for the soluble T4 lysozyme protein. This initial rotamer therefore relies on quantum chemical calculations alone (Tombolato, Ferrarini et al. 2006). As additional crystal structures of MTSSL become available, especially for membrane proteins, the rotamer library will be extended to take into account an expanded experimental knowledge-base. The immediate advantage of atomic level verification of EPR experiments outweighs the initially limited knowledge-based rotamer library.

RosettaEPR spin label library is robust enough for use in a wide range of modeling protocols of proteins

Compared with a systematic search of larger rotamer libraries, the RosettaEPR rotamer library is limited to a relatively small number of 54 discrete conformers which maximizes efficiency of the conformational search and enable parallel optimization of additional protein degrees of freedom. This approach is balanced by sampling off-rotamer conformations in all atom refinement protocols. Further, Rosetta systematically samples close-to-rotamer conformations by varying (X_1, X_2) by one standard deviation. The number of spin label rotamers aligns with the number of rotamers seen for large amino acid side chains (Arg, Lys 81 (Shapovalov and Dunbrack 2011)) which have been demonstrated to be sufficient for atomic-detail structure determination (Bradley, Misura et al. 2005; Alexander, Bortolus et al. 2008; Krivov, Shapovalov et al. 2009) (Qian,

Raman et al. 2007). The success of the approach is demonstrated by a) recovery of the off-rotamer experimental conformation of T4 lysozyme mutant L188, b) Rosetta's ability to sample all experimentally observed conformations of MTSSL in soluble T4 lysozyme and the membrane protein LeuT, and c) the ability of the Rosetta models to accurately fit the experimental EPR distance distributions.

RosettaEPR samples experimentally observed spin label conformations on the surface and in the protein core for soluble and membrane proteins

RosettaEPR samples all experimentally observed conformations of MTSSL at core and surface sites at least in some trajectories. However, RosettaEPR also samples alternative conformations sometimes with a higher frequency and superior energy to the experimentally observed conformation. A combination of reasons is expected to contribute to this result: a) the spin label samples multiple and additional conformations of similar free energy in solution that are not observed in the crystal. This notion is supported by the frequent uncertainty in reconstructing spin labels on the surface of proteins as displayed by lack of coordinates beyond X_3. b) The RosettaEPR energy function ranks different conformations of the spin label incorrectly with respect to each other. This is expected on the protein surface given the close free energy of such conformations, the approximations inherent to the pair-wise decomposable Rosetta energy function (Kuhlman and Baker 2000), and the lack of specific treatment of partial covalent interactions the nitroxide group might engage the protein in.

RosettaEPR poorly samples the experimental conformations of MTSSL at crystal contact sites. Each protein component of the asymmetric unit was relaxed in Rosetta independently, i.e. not in the presence of the other copies in the crystal. Therefore, such performance is expected because the spin label conformations are significantly influenced by non-biologically relevant crystal contact interactions that are not present in

examination of the rotamers in RosettaEPR (Langen, Oh et al. 2000; Guo, Cascio et al. 2007; Guo, Cascio et al. 2008; Fleissner, Cascio et al. 2009).

RosettaEPR reproduces specific dynamics seen for spin labels

RosettaEPR achieves an MAE of 4.4 Å for predicting experimental EPR distances. This compares favorably to usage of the C_β distances as an approximation for the spin label (MAE = 6.1 Å). The cone model fits the difference between spin label distance and C_β distance to a set of experimental data (Hirst, Alexander et al. ; Alexander, Bortolus et al. 2008). It minimizes the RMSD between experimental and predicted distance to 4.7 Å which is comparable to the explicit treatment of the spin label in RosettaEPR. This indicates the power of a simple linear correlation between spin label and C_β distances. However, the cone model inherently assumes the same conformational sampling, σ, for all spin labels independent of labeling site which is also represented by the standard deviation of the distance difference distribution (4.7 Å). The standard deviation of the experimental distance distributions are reproduced much more closely by the atomic-detail representation of the spin label with a RMSD of 2.0 Å. Thereby, explicit treatment of the spin label provides information on the actual conformational sampling of MTSSL.

By selecting ensembles of models from RosettaEPR specifically to reproduce experimental EPR distance probability distributions, the accuracy of RosettaEPR is further improved. RosettaEPR can sample within all of the experimental distance probability distributions. This indicates the range of sampling with the rotamer library is not the limiting factor in RosettaEPR's ability to reproduce spin label dynamics. For double mutants where sampling within the experimental probability distribution is infrequent, a more accurate scoring function could focus sampling to produce smoother, more accurate fits to the distributions.

Comparison with previous methods

RosettaEPR recovers native X_1 and X_2 of MTSSL with a frequency similar to Rosetta's ability to recover arginine and lysine X_1 and X_2. Over a dataset of 129 proteins, Rosetta recovered native X_1 and X_2 of arginine and lysine 60-65% of the time (Wang, Schueler-Furman et al. 2005). Though this is a slightly higher percentage than observed for MTSSL, the fraction of exposed positions in the MTSSL dataset is large, which would account for the reduced accuracy of RosettaEPR.

RosettaEPR's rotamer recovery is slightly less accurate than the side chain prediction method SCWRL4 (Krivov, Shapovalov et al. 2009) in recovery for X_1 and X_2 (70%) and X_1 –X_4 (36%) in arginine and lysine side chains across buried and exposed sites in 379 protein structures. However, as X_1 and X_2 recovery is calculated for arginine and lysine at increasingly exposed positions, the performance of SCWRL4 more closely aligns with RosettaEPR's X_1 and X_2 recovery for MTSSL. This is important because thirteen of the fourteen MTSSL single mutants at non-crystal contact sites occur at surface positions.

In T4 lysozyme, single mutants A082 and L118 for the study of an MTSSL rotamer library (Polyhach, Bordignon et al.). This study was also successful in predicting the experimentally observed conformations at these sites. However, for L118, the population of rotamers predicted to be buried within the cavity as observed in the experimental structure is 99.8% for RosettaEPR versus 52% for the previous study. Without additional experimental data, it is difficult to determine which is more accurate.

A previous attempt at recovering the average distance of an EPR double mutant measurement have a reported mean error of 3.0 Å over twenty-seven distances measured in troponin C, the troponin complex and the KcsA channel (Sale, Song et al. 2005). Rosetta EPR achieves MAE of 4.4 Å over all seventy-three EPR distances for T4 lysozyme and MsbA, and 3.5 Å for fifty-eight T4 lysozyme distances specifically. Differences in accuracy are mitigated by the differences in the protein systems and size

of the datasets. In addition, the previous study made no attempt at reproducing σ_{EPR}, which RosettaEPR can recover to a MAE of 1.3 Å.

The utility of fitting an ensemble of structures to EPR distance data has been demonstrated for the transmembrane domain IX of the Na^+/proline transporter PutP of *Escherichia coli* ((Hilger, Polyhach et al. 2009)). This single transmembrane span has a helix-loop-helix motif. MTSSL rotamers and backbone ψ, φ were varied to produce an RMSD of 1.00 Å of the models to experimental mean distances. This compares favorably to the 0.7 Å RMSD achieved by RosettaEPR over the thirty-eight T4 lysozme distributions and 2.5 Å when all fifty-seven distributions (T4 lysozyme and MsbA) are considered.

Verification of cone model parameters

The distribution of $\angle^{max} SL_A C_\beta SL_B$ observed indicates that the width of the spin label conformational ensemble (the opening angle of the cone) can vary widely across different sites on a protein. The original cone model parameter of $\angle^{max} SL_A C_\beta SL_B$ = 90° falls within one standard deviation of the $\angle^{max} SL_A C_\beta SL_B$ distribution average. The distribution of $\angle SL_{ef} C_\beta C_\alpha$ obtained by Rosetta indicates that the ensemble can be tilted closely towards the backbone, indicative of the spin label hugging the surface of the protein. Given the hydrophobic nature of the MTSSL side chain, it is likely the spin label would exhibit such behavior. The average $\angle SL_{ef} C_\beta C_\alpha$ value calculated from RosettEPR of 111° matches closely with the original parameter of 120°. The distance between the effective spin label position and the corresponding C_β, $D_{C_\beta}^{SL_{ef}}$, was originally proposed in the cone model to be 6.0Å. The distribution obtained by RosettaEPR indicates that $D_{C_\beta}^{SL_{ef}}$ value is on average slightly longer at 6.3 Å. The $D_{C_\beta}^{SL_{ef}}$ is related to $\angle^{max} SL_A C_\beta SL_B$ as an increasing width of the ensemble will produce a decreasing $D_{C_\beta}^{SL_{ef}}$. The fact that the

average $D_{C_\beta}^{SL_{ef}}$ is slightly longer than what would be expected given the average $\angle^{max} SL_A C_\beta SL_B$ is due to the population of MTSSL ensembles with a small width.

Overall we find the cone model parameters accurate within the error of the experiment. It is apparent that while the cone model rather accurately captures distances, experimental distance deviations are not adequately represented with a unified model. Through the atomic-detail description of spin labels during structure prediction, this study overcomes one critical limitation of the cone model. The cone model was derived by observing spin label distances over many independent experiments. Spin label pairs in very different structural and dynamical states were folded into a single probability distribution. This probability distribution encompasses uncertainty over the precise conformation of the spin label and its dynamics, convoluting both contributions. Its allowable distance range is therefore inherently too wide. The model is very effective in medium-resolution modeling due to its speed and due to omitting explicit modeling of side chains – an approach that is widely used at this stage. At the same time it reaches its limitations in atomic-detail refinement of the models – for example restraints were not employed for atomic-detail refinement in our previous research on de novo folding of proteins from EPR restraints (Hirst, Alexander et al. ; Alexander, Bortolus et al. 2008).

Potentially, RosettaEPR could yield insight into the environmental factors that determine the disorder of the spin label at a site. Such a scenario could occur as the database of crystallographically observed spin label conformations grows, allowing for an improved scoring function describing the interactions of the nitroxide with its environment. With an accurate description of the nitroxides behavior, a refined cone model would allow for the quick verification of a putative model or structure.

Conclusion

RosettaEPR can recover and sample experimentally observed conformations of the MTSSL spin label on single mutants of T4 lysozyme and the membrane protein LeuT. This has not been previously demonstrated with such a large dataset or with a membrane protein. RosettaEPR's ability to reproduce EPR distance distributions has also not previously been observed for the number and complexity of measurements. Such a method will be a powerful tool for investigating the structure and dynamics of proteins. Compared to proprietary approaches developed specifically for computational spin label investigations, RosettaEPR is easily disseminated among the EPR scientific community within the Rosetta protein structure prediction suite.

Experimental Procedures

Development of MTSSL Rotamer Library

The non-canonical methanesulfonothioate spin label residue was created in the Molecular Operating Environment (MOE). The Pymol Molecular Graphics System (PyMOL) was then used to create 60 rotamers taking into account all the possible combinations of the canonical X angles as elaborated in the Results section. The potential energy of each rotamer was calculated for use as an indicator of which rotamers contained intramolecular clashes. The potential energy was calculated in MOE using the "Potential" function with the default MMFF94x force field. The rotamers were sorted by energy. Ten rotamers were determined to have clashes because a large increase in potential energy (54.9%) for the most energetically favorable of the ten rotamers separated them from the other 50 rotamers. Outside of these ten rotamers, the largest potential energy increase was 10%. The ten rotamers were subject to energy minimization in MOE using the "MM" function in an attempt to rescue each rotamer in the event that small changes to the X angles could relieve the clash. After minimization, the

potential energy of eight of the ten rotamers was minimized into the regime of the other 50 rotamers. In addition to a reduction in potential energy, the eight minimized rotamers were also filtered by the amount of change in each X angle such that no X angle changed by more than 30°. Four of the eight rotamers met this criterion. As a result, the total rotamer library contains 54 conformations of MTSSL.

Single Mutant MTSSL Conformational Sampling

Each of the crystal structures of T4 lysozyme singly labeled with MTSSL were downloaded from the Protein Data Bank (PDB) (Berman, Westbrook et al. 2000). The PDB accession identifiers (PDB IDs) are 2IGC, 2OU8, 2OU9, and 2NTH (Guo, Cascio et al. 2007), 2Q9D and 2Q9E (Guo, Cascio et al. 2008), and 1ZYT, 2CUU, 3G3V, 3G3W, and 3G3X (Fleissner, Cascio et al. 2009) (See Supplemental Table 1 for identification of the mutant for each PDB file). Mutants R080, R119, K065, and V075 (Langen, Oh et al. 2000) were not available to download from the PDB website. Therefore, the single mutants for these were computationally created from the T4 lysozyme crystal structure with PDB ID 2LZM (Weaver and Matthews 1987). In order to create the cys-less sequence (Matsumura and Matthews 1989), which was used for these four single mutant crystal structures, cysteine residues 54 and 97 were computationally mutated to threonine and alanine, respectively. All computational mutations were done using the Rosetta Fixed Backbone Design application (Kuhlman, Dantas et al. 2003). Each crystallized protein structure, including those involving crystal contacts, was relaxed (see below) in Rosetta individually without the presence of any other crystallographic subunits. The starting protein structures were subjected to 1000 independent relaxation trajectories in Rosetta, which were then used for analysis on Rosetta's ability to recover experimentally observed conformations.

For single MTSSL mutants of LeuT, the two experimental structures downloaded were 3MPN and 3MPQ ((Kroncke, Horanyi et al. 2010)). These structures were relaxed by Rosetta in 1015 trajectories.

Double Mutant MTSSL Conformational Sampling

A pseudo wild type starting structure was created as described above whereby cysteine residues 54 and 97 of PDB ID 2LZM were computationally mutated to threonine and alanine, respectively. Next, structures for 58 double mutants were created from this pseudo wild type starting structure. Forty-six of these mutants have been previously described (Borbat, McHaourab et al. 2002; Alexander, Bortolus et al. 2008; Kazmier 2010) with twelve double mutants described below. All computational mutations were done using the Rosetta Fixed Backbone Design application. Each of these fifty-eight double mutants was subjected to 2000 independent relaxation trajectories in Rosetta. For each relaxation trajectory, the distance between the final conformations of the two spin labels was calculated, where the unpaired electron is taken to be at the midpoint of the N-O bond. The set of distances from the top 200 of models by Rosetta score was used as the distance distribution for each mutant, and compared against the corresponding experimental distance distributions. Double mutants 131/154, 131/151, 140/147, and 116/131 were excluded from analysis because the standard deviation of the distance measurement was determined to be greater than 50% of the distance. The experimental distance distributions for double mutants 119/128, 119/131, 123/131, and 140/151 were reanalyzed for this study using Tikhonov regularization (Chiang, Borbat et al. 2005), producing means and standard deviations of the distributions which differ slightly from the originally published values (Alexander, Bortolus et al. 2008).

Nineteen previously published EPR distances measured in the transmembrane region of MsbA (Zou, Bortolus et al. 2009) were used for this study. Computational double mutants were created from PDB ID 3B60 (Ward, Reyes et al. 2007) for the AMP-

PNP closed state and from the full atom structure of the open state provided from (Zou, Bortolus et al. 2009) based on PDB ID 3B5X (Ward, Reyes et al. 2007). Cysteine residues 88 and 315 were mutated to alanine, resulting in the pseudo wild type used for creating computational double MTSSL mutants. All double mutants were relaxed at least 1000 times in Rosetta and the top 100 models by Rosetta score were used as the distance distribution for each mutant.

Three statistical values are used to compare Rosetta to EPR experiment. The mean absolute error (MAE) is calculated as $MAE = \frac{|model-exp|}{\#\ of\ values}$. The root mean square deviation (RMSD) is calculated as RMSD = $\sqrt{\frac{\sum(model-exp)^2}{\#values}}$. The correlation coefficient (R) is also used for comparison of Rosetta to experiment.

Rosetta Relaxation and Computational Mutant Protocols

The standard Rosetta refinement protocol (Bradley, Misura et al. 2005; Misura and Baker 2005) was used to relax the T4 lysozyme protein structures and determine MTSSL conformations. For MsbA and LeuT, the relaxations took place using the membrane specific potentials of Rosetta ((Barth, Schonbrun et al. 2007)). During relaxation all side chains are repacked and small perturbations of the backbone occur. This means that the starting conformations of side chains do not impact the final rotamers chosen. A single Rosetta relaxation trajectory takes about 15 minutes on an Intel Xeon W3570 3.2 GHz processor for T4 lysozyme. Please see Supplemental Experimental Procedures for the specific command line flags used.

The fixed backbone design application of Rosetta was used to introduce MTSSL at desired sites in the benchmark proteins. The protocol does not allow any backbone optimization and all other side chains were held fixed in their native conformation. So, only the conformation of the specific mutated residue was optimized, which was sufficient because the mutants later underwent Rosetta relaxation. The application takes

approximately one minute to run on an Intel Xeon W3570 3.2 GHz processor. Please see Supplemental Experimental Procedures for specific command line flags used.

Fitting of Rosetta Generated Ensembles to Experimental EPR Distance Distributions

Fifty-seven experimental EPR distance distributions analyzed by Tikhonov regularization were used as the dataset for finding Rosetta generated ensembles that give spin-label to spin-label distance distributions similar to experiment: thirty-eight from T4 lysozyme and nineteen from MsbA. For each T4 lysozyme double mutant, all 2000 relaxation models were possible constituents of the matching sub-ensemble. For MsbA, the top 1000 models according to Rosetta score were available for fitting. A Monte Carlo process of adding or removing models and allowing only favorable moves was used to determine the matching sub-ensembles. Agreement between the EPR measured and Rosetta recovered distance distributions calculated from the sub-ensemble was measured by the cumulative Euclidian distance $d(p,q) = \sqrt{\sum_{i=0}(\sum_{u=0} p_u - \sum_{u=0} q_u)^2}$ (Kamarainen, Kyrki et al. 2003), where *p* and q give the probability of a given distance bin, and *u* and *i* are iterations over the distance bins. This value $d(p,q)$ is normalized by the number of bins summed over, N, such that $d^{norm}(p,q) = \sqrt{\frac{d(p,q)^2}{N}}$.

Derivation of implicit spin label cone model parameters

The primarily alpha-helical T4 lysozyme pseudo-wild type starting structure and the primarily beta-strand chitinase (PDB ID 2CWR (Nakamura, Mine et al. 2008) were used as the basis to determine the implicit model parameters. Single mutations introducing MTSSL were computationally created for the two proteins at residues having a neighbor count (Durham, Dorr et al. 2009) less than ten. 63 and 99 sites met this neighbor count criteria for T4 lysozyme and 2CWR, respectively. Each of these single mutants was subjected to 500 independent relaxation trajectories in Rosetta.

For each single mutant, the effective spin label position, SL_{ef}, was calculated as the average of all the observed positions of the N-O bond midpoints on the nitroxide moiety of the spin label. In order to determine SL_{ef}, the backbone C_α, H_α, C, N, and CB atoms of the spin label were used to superimpose the 500 structures for each mutant. Superimposition was done using the "fit" command in Pymol . SL_{ef} was then calculated for each single mutant along with the corresponding $\angle SL_{ef} C_\beta C_\alpha$ and $D_{C_\beta}^{SL_{ef}}$ parameters. Also, $\angle^{max} SL_A C_\beta SL_B$ was determined for each single mutant after superimposition, by calculating all pairwise $\angle SL_A C_\beta SL_B$ for the 500 models and finding the maximum value observed.

These updated parameters for the cone model were then used to simulate spin-spin label distances, D_{SL}, in multiple proteins. 4379 single chains from soluble proteins filtered by PISCES (Wang and Dunbrack 2003) for not more than 25% sequence identity and resolution of at most 2.0 Å were used to calculate the distances. These spin label distances, D_{SL}, were then compared to the distance between the C_β atoms of the residues containing the spin labels, $D_{C\beta}$. A histogram describing the difference between D_{SL} and $D_{C\beta}$, was then calculated.

Acknowledgements

This work is supported by grant R01-GM080403 from the National Institute of General Medical Sciences. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. N.A. is funded by NIH NIMH Award Number F31-MH08622.

CHAPTER IV

INTERACTION OF A G PROTEIN WITH AN ACTIVATED RECEPTOR OPENS THE INTERDOMAIN INTERFACE IN THE ALPHA SUBUNIT

This chapter is based on (Van Eps, Preininger et al.).

Summary

In G protein signaling, an activated receptor catalyzes GDP/GTP exchange on the G_α subunit of a heterotrimeric G protein, leading to activation of the subunit. In an initial step, receptor interaction with G_α acts to allosterically to trigger GDP release from a binding site located between the nucleotide binding domain and a helical domain, but the molecular mechanism is unknown. In this study, Site Directed Spin Labeling and Double Electron Electron Resonance spectroscopy are employed to reveal a large-scale separation of the domains, providing a direct pathway for nucleotide escape. The inter-domain opening is coupled to receptor binding *via* the C-terminal helix of G_α, the extension of which is a high-affinity receptor binding element.

Introduction

The α-subunit (G_α) of heterotrimeric G proteins ($G_{\alpha\beta\gamma}$) mediates signal transduction in a variety of cell signaling pathways (Tesmer 2010). Multiple conformational states of G_α are involved in the signal transduction pathway shown in Figure 21A. In the inactive state, the G_α subunit contains a bound GDP [G_α(GDP)] and has a high affinity for $G_{\beta\gamma}$. When activated by an appropriate signal, a membrane-bound G-protein coupled

receptor (GPCR) binds the heterotrimer in a quaternary complex, leading to the dissociation of GDP and formation of an "empty complex" [$G_{\alpha}(0)_{\beta\gamma}$], which subsequently binds GTP. The affinity of G_{α}(GTP) for $G_{\beta\gamma}$ is dramatically reduced relative to G_{α}(GDP), resulting in functional dissociation of active G_{α}(GTP) from the membrane-bound complex. The active Gα(GTP) subsequently binds downstream effector proteins to trigger a variety of regulatory events, depending on the particular system. Thus, the GPCR acts to catalyze GDP/GTP exchange via an empty complex. Crystallographic (Noel, Hamm et al. 1993; Coleman, Berghuis et al. 1994; Lambright, Noel et al. 1994; Wall, Coleman et al. 1995; Lambright, Sondek et al. 1996; Coleman and Sprang 1998), biochemical (Higashijima, Ferguson et al. 1987), and biophysical (Oldham, Van Eps et al. 2006; Van Eps, Oldham et al. 2006; Oldham, Van Eps et al. 2007) studies have elucidated details of the conformational states of G_{α} that correspond to the discrete steps indicated in Figure 21A, but the mechanism by which receptor interaction leads to release of the bound GDP from G_{α} and the structure of the empty complex remain a major target of research in the field.

The G_{α} subunit has two structural domains, namely a nucleotide binding domain and a helical domain that partially occludes the bound nucleotide (Figure 21B). From the initial G_{α} crystal structure in 1993, Noel et al. (Noel, Hamm et al. 1993) recognized that nucleotide release would probably require an opening between the two domains in the empty complex, but in the intervening 18 years there has been little compelling experimental support for this idea. Nevertheless, some constraints on the general topology of the complex are known. For example, numerous studies indicate that the C terminus of G_{α} is bound tightly to the receptor in the empty complex (Oldham, Van Eps et al. 2006). In addition, the N-terminal helix of G_{α} is associated with $G_{\beta\gamma}$ and with the membrane via N-terminal myristoylation (Linder, Pang et al. 1991; Resh 1999). Together, these constraints fix the position of the nucleotide domain with respect to the

membrane. The helical domain is connected to the nucleotide domain through two flexible linkers, and linker 1 (switch I) undergoes conformational changes upon receptor binding (Oldham, Van Eps et al. 2007). These observations provided the motivation to look for relative motion of the two G_{α} domains during formation of the empty complex.

For this purpose, Site Directed Spin-Labeling (SDSL) and Double Electron Electron Resonance (DEER) spectroscopy were employed to measure distances between pairs of spin labels, with one label in each domain. Distances were measured for each state of $G_{\alpha i}$ along the activation pathway using activated rhodopsin (R*) as the GPCR. The results indicate that receptor-catalyzednucleotide exchange in G proteins requires a large-scale reorientation of domains in the G protein α-subunit.

Results and Discussion

Using SDSL and DEER spectroscopy, distances were measured for each state of $G_{\alpha i}$ along the activation pathway using activated rhodopsin (R*) as the GPCR. In these experiments, the R1 nitroxideside chain (Figure 22) was introduced via cysteine substitution mutagenesis into the background of $G_{\alpha i}$ with reactive cysteines removed, HexaI ($G_{\alpha i}$ HI) (Medkova, Preininger et al. 2002). Figure 21B shows the set of sites from which pairs were selected and the five specific interdomain distances investigated.

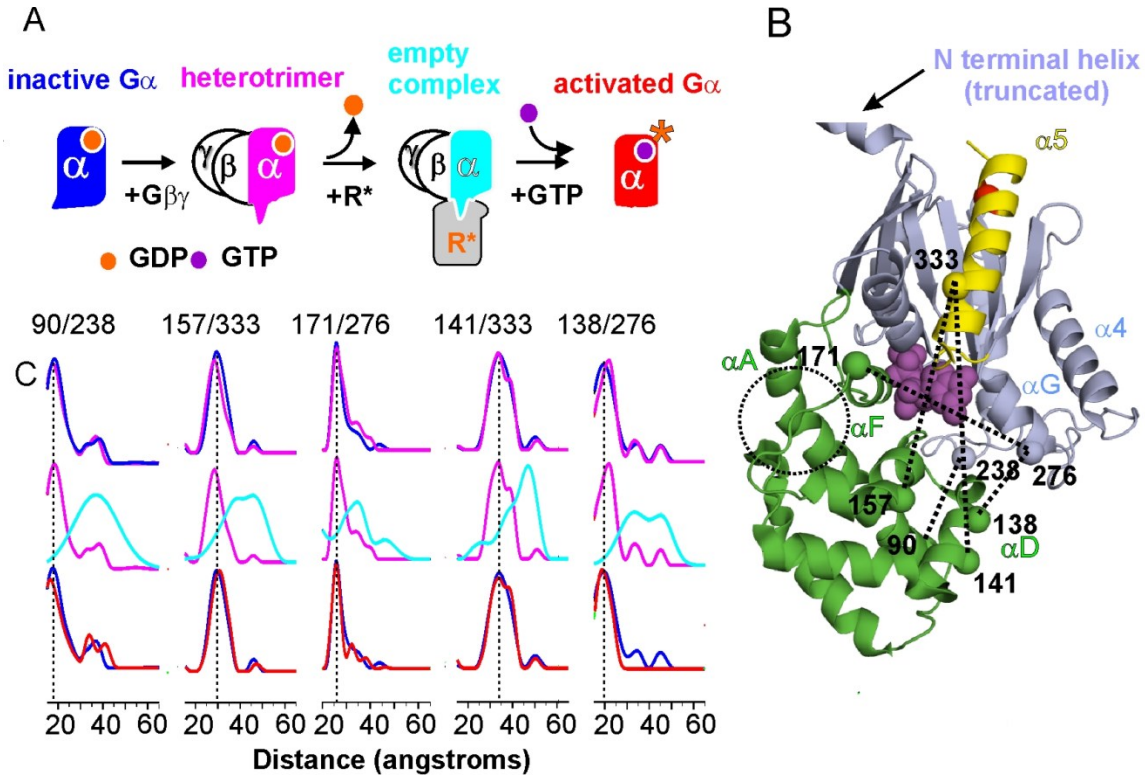


Figure 21 Receptor activation of G proteins leads to a separation between domains. **(A)** The pathway of Gα activation *via* activated rhodopsin (R*). The alpha subunit is color coded to denote the four different states investigated by SDSL/DEER spectroscopy. (**B**) Ribbon model of Gαi (pdb 1GP2). The helical and nucleotide binding domains are colored green and light blue, respectively, and GDP is shown as magenta spheres. Relevant secondary structural elements are noted for reference. The C-terminal helix α5 is colored yellow; six disordered residues at the C-terminus are not shown. The N-terminal helix is truncated for convenience. Sites from which R1 nitroxide side chains were selected pair wise for distance measurements are indicated by spheres; dotted traces indicated specific distances measured for each state in (A). **(C)** Distance distributions for the indicated doubly spin-labeled mutants. The top panel compares Gαi(GDP) and Gαiβγ(GDP); the center panel compares Gαiβγ(GDP) and Gαiβγ(0); the lower panel compares Gαi(GDP) and Gαi(GTP); traces are color coded to match states in (A).

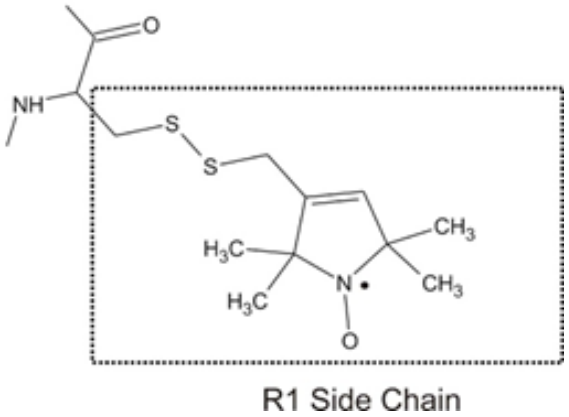


Figure 22 The nitroxides R1 side chain.

All doubly spin-labeled proteins bind to R* to an extent similar to the $G_{\alpha i}$ HI parent protein as shown in direct endpoint binding assays (Figure 23). In addition, they are all functional with respect to receptor-mediated nucleotide exchange, although mutants 138R1/276R1 and 157R1/333R1 have, respectively, about 40% and 55% of the receptor-catalyzed nucleotide exchange rate of the parent $G_{\alpha i}$ HI protein (Figure 23). The reduced rates suggest that the residues involved are important in modulating receptormediated nucleotide exchange. In crystal structures of the inactive protein, residues Asn157 and Glu276 are involved in side chain H bonding and electrostatic interactions, respectively, and mutation of these may influence local conformation.

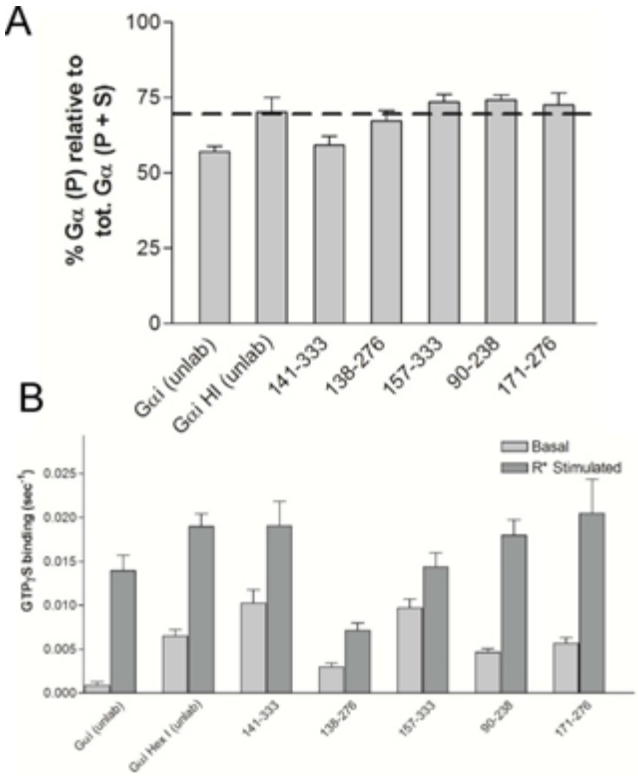


Figure 23 Binding and functional assays for doubly labeled G protein.
(A) Binding of doubly spin-labeled mutants to rhodopsin in disc membranes. (B) Basal and receptor catalyzed nucleotide exchange rates for the doubly spin-labeled mutants. Assays were performed as described in Methods.

DEER spectroscopy relies on magnetic dipolar interactions between spin labels to measure interspin distances in the range of ≈17–60 Å (Pannier, Veit et al. 2000), (Jeschke 2002). Of particular importance is the ability to resolve multiple distances and the widths of the distributions. Figure 21C compares the distance probability distributions for the five transdomain R1 pairs in each of the four states of $G_{\alpha i}$, i.e., $G_{\alpha i}$(GDP), $G_{\alpha i}(GDP)_{\beta\gamma}$, $G_{\alpha i}(0)_{\beta\gamma}$, and $G_{\alpha i}$(GTP). For each pair, the measured most probable distances for $G_{\alpha i}$(GDP) and $G_{\alpha i}(GDP)_{\beta\gamma}$ agree well with expectations from the crystal structures (Coleman, Berghuis et al. 1994; Wall, Coleman et al. 1995; Coleman and Sprang 1998) and models of the R1 side chain (Fleissner, Cascio et al. 2009). In all cases there is little difference between $G_{\alpha i}$(GDP) and $G_{\alpha i}(GDP)_{\beta\gamma}$.

Upon photoactivation of rhodopsin and formation of the R*•$G_{\alpha i}(0)_{\beta\gamma}$ complex, there is a remarkable increase in each interspin distance, with increases being as large as 20 Å (at 90⁄238) (for details, see Supporting Information section and Figure 24 and

Figure 25). Moreover, there is a dramatic increase in width of each distribution as well as multiple distances in most cases. It is of interest that distances present in the $G_{\alpha i}(0)_{\beta\gamma}$ distributions correspond approximately to minor populations already present in $G_{\alpha i}$(GDP) and $G_{\alpha i}(GDP)_{\beta\gamma}$, suggesting that activation may shift an existing equilibrium. Although the exact widths of the distributions in $G_{\alpha i}(0)_{\beta\gamma}$ may not be well determined in each case, they are clearly broader than possible from multiple rotamers of R1, suggesting spatial disorder of the G_{α} protein in the empty-pocket state of the activated complex (see Supporting Information section). Finally, addition of GTPγS restores a state with a most probable distance and width of distribution similar to the GDP bound state. This is in agreement with expectations from GTPγS bound crystal structures (Coleman, Berghuis et al. 1994).

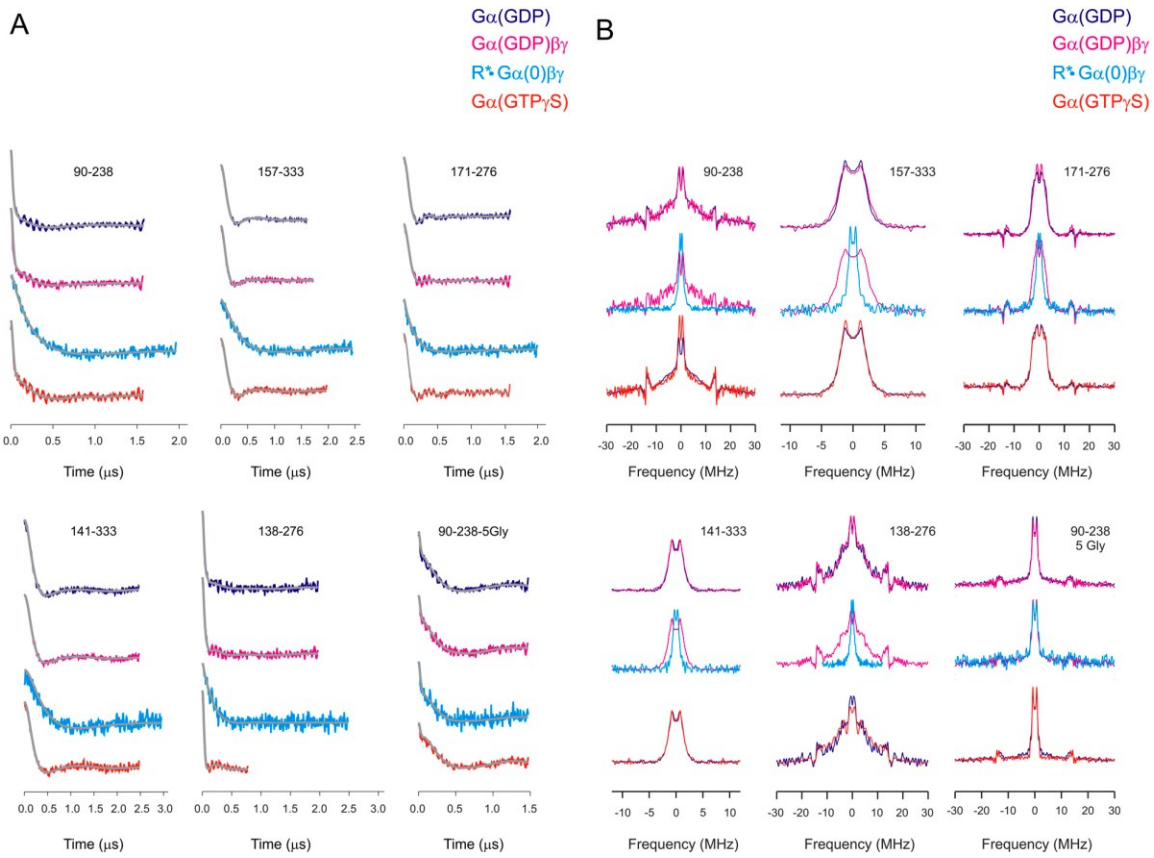


Figure 24 Individual EPR spectra along the activation pathway.
(A) Background corrected dipolar evolution data for each double-labeled mutant along the activation pathway. Gray traces show fits to each individual dipolar evolution. (B) Fourier transformation of the dipolar evolution data given in A yields the dipolar spectra in B. The data are shown for eachspin-labeled double mutant along the activation pathway.

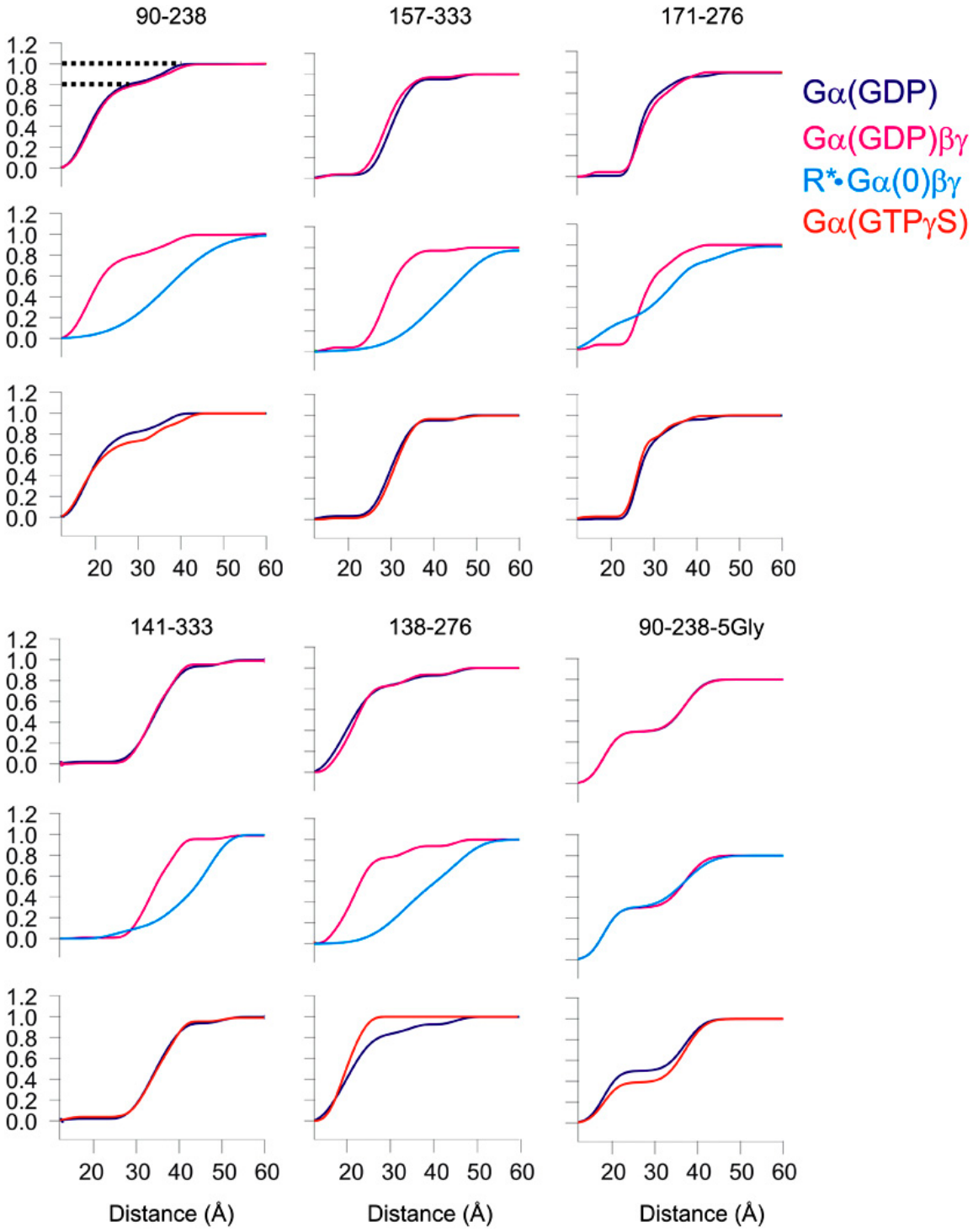


Figure 25 Normalized integral representations of the distance distributions shown in Figure 21C of the main text.

Such representations are particularly useful for visually estimating the relative populations of the distances. This is illustrated, for example, in the top panel of the 90–238 mutant; the major population is about 80%. The most probable distance for a population is estimated from the midpoint of the transition.

The EPR spectra of R1 residues at the sites shown in Figure 21B have little or no changes upon receptor activation (Figure 26). This result, taken together with the very large distance changes observed, ensure that the detected distance increases reflect global domain movement rather than simple R1 side chain rearrangements due to changes in local environment. Collectively, the data strongly support a model for a $G_{\alpha i}(0)_{\beta\gamma}$ in which the helical domain is displaced relative to the nucleotide domain in the heterotrimer, and in which the structure is highly flexible with respect to the relative domain orientations.

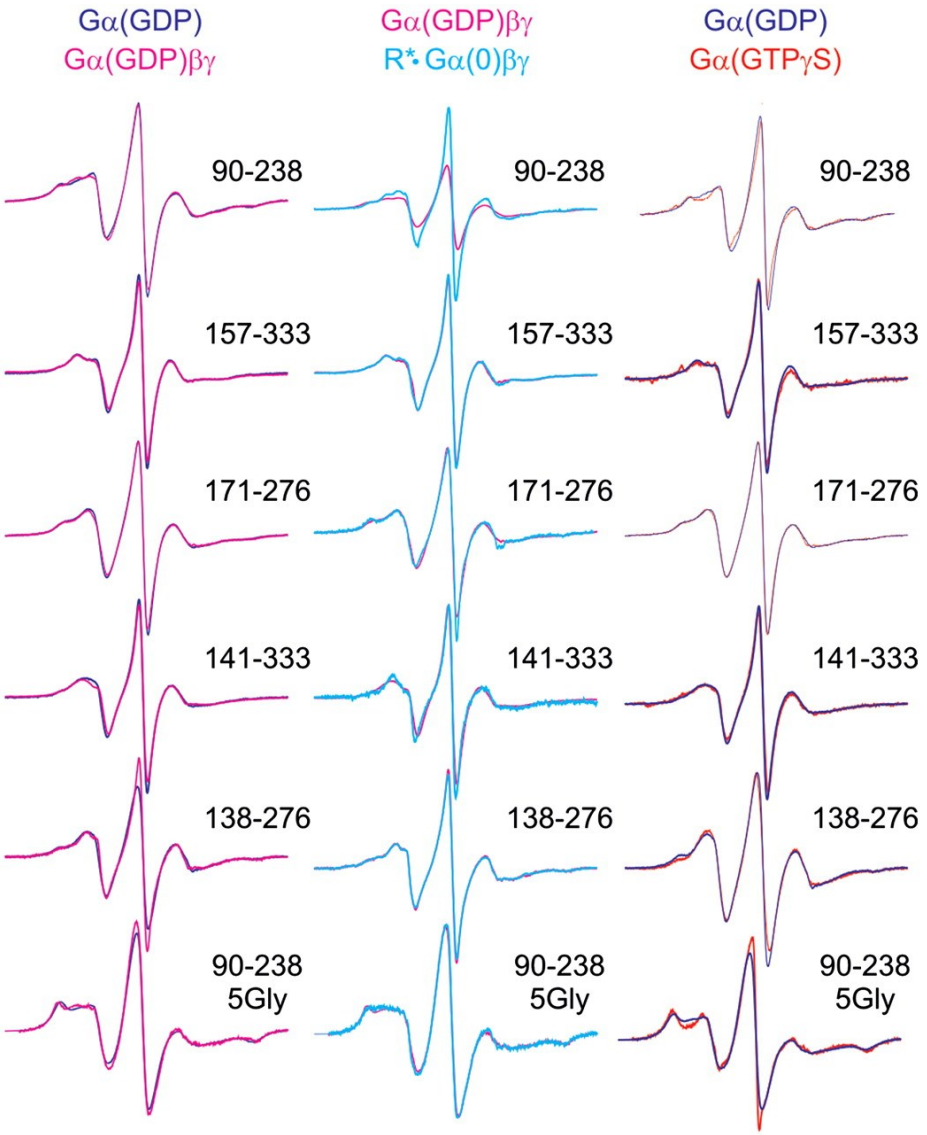


Figure 26 CW EPR spectra of the spin-labeled double mutants in Gαi at the indicated states along the activation pathway.
(Left) Compares EPR spectra of the doubly labeled Gαi(GDP) and Gαi(GDP)βγ mutants; (Middle) compares Gαi(GDP)βγ and R ∗ •Gαi(0)βγ; (Right) compares Gαi(GDP) and Gαi(GTP).

To visualize the domain opening, a model of the empty complex on the receptor was constructed that is consistent with the DEER and other available experimental data (see Supporting Information text). To generate the model, the heterotrimeric Gi was docked with the photoreceptor using crystal structures of $G_{\alpha i}(GDP)_{\beta\gamma}$ (Lambright, Sondek et al. 1996) (Wall, Coleman et al. 1995) and opsin in complex with the high-affinity Gαt C-terminal peptide (Scheerer, Park et al. 2008). The Gαi C-terminal helix was fused with

the high affinity Gα C-terminal peptide bound to opsin (for details, see Supporting Information text and Figure 27 and Figure 28), which provided a convenient starting point for the model (Scheerer, Heck et al. 2009). The myristoylated N-terminal amphipathic helix was placed parallel to the membrane surface and the heterotrimer oriented such that both the myristoyl group and the nearby farnesylated C terminus of the Gγ-subunit can be inserted into the membrane; together these hydrophobic interactions cooperatively drive membrane binding of the intact heterotrimer (Herrmann, Heck et al. 2006).

```
Gαt/Gαi chimera (1GOT)   1 MGAGASAEEK----HSRELEKKLKEDAEKDARTVKLLLLGAGESGKSTIVKQXKIIHQDG   56
                             MG   SAE+K     S+ +++ L+ED EK AR VKLLLLGAGESGKSTIVKQ KIIH+ G
Gαi rat                  1 MGCTLSAEDKAAVERSKMIDRNLREDGEKAAREVKLLLLGAGESGKSTIVKQMKIIHEAG   60

Gαt/Gαi chimera (1GOT)  57 YSLEECLEFIAIIYGNTLQSILAIVRAXTTLNIQYGDSARQDDARKLXHXADTIEEGTXP  116
                             YS EEC ++ A++Y NT+QSI+AI+RA   L I +GD+AR DDAR+L   A   EEG
Gαi rat                 61 YSEEECKQYKAVVYSNTIQSIIAIIRAMGRLKIDFGDAARADDARQLFVLAGAAEEGFMT  120

Gαt/Gαi chimera (1GOT) 117 KEXSDIIQRLWKDSGIQACFDRASEYQLNDSAGYYLSDLERLVTPGYVPTEQDVLRSRVK  176
                              E + +I+RLWKDSG+QACF+R+ EYQLNDSA YYL+DL+R+  P Y+PT+QDVLR+RVK
Gαi rat                 121 AELAGVIKRLWKDSGVQACFNRSREYQLNDSAAYYLNDLDRIAQPNYIPTQQDVLRTRVK  180

Gαt/Gαi chimera (1GOT) 177 TTGIIETQFSFKDLNFRXFDVGGQRSERKKWIHCFEGVTAIIFCVALSDYDLVLAEDEEX  236
                             TTGI+ET  F+FKDL+F+ FDVGGQRSERKKWIHCFEGVTAIIFCVALSDYDLVLAEDEE
Gαi rat                 181 TTGIVETHFTFKDLHFKMFDVGGQRSERKKWIHCFEGVTAIIFCVALSDYDLVLAEDEEM  240

Gαt/Gαi chimera (1GOT) 237 NRXHESXKLFDSICNNKWFTDTSIILFLNKKDLFEEKIKKSPLTICYPEYAGSNTYEEAG  296
                             NR HES KLFDSICNNKWFTDTSIILFLNKKDLFEEKIKKSPLTICYPEYAGSNTYEEA
Gαi rat                 241 NRMHESMKLFDSICNNKWFTDTSIILFLNKKDLFEEKIKKSPLTICYPEYAGSNTYEEAA  300

Gαt/Gαi chimera (1GOT) 297 NYIKVQFLELNXRRDVKEIYSHXTCATDTQNVKFVFDAVTDIIIKENLKDCGLF        350
                              YI+ QF +LN R+D KEIY+H TCATDT+NV+FVFDAVTD+IIK NLKDCGLF
Gαi rat                 301 AYIQCQFEDLNKRKDTKEIYTHFTCATDTKNVQFVFDAVTDVIIKNNLKDCGLF        354
```

Figure 27 A BLAST sequence alignment of Gαi and the Gαt/Gαi chimera of 1GOT, which was used in comparative modeling.
The sequence alignment features a single gap (red) within the N-terminal α-helix of the protein. The $G_{\alpha i}$ region (residues 216-294 of the 1GOT sequence) is shown in orange. The α-helical domain is shown in green. The C-terminal helix and 11 residues of the opsin bound peptide are shown in yellow and blue, respectively.

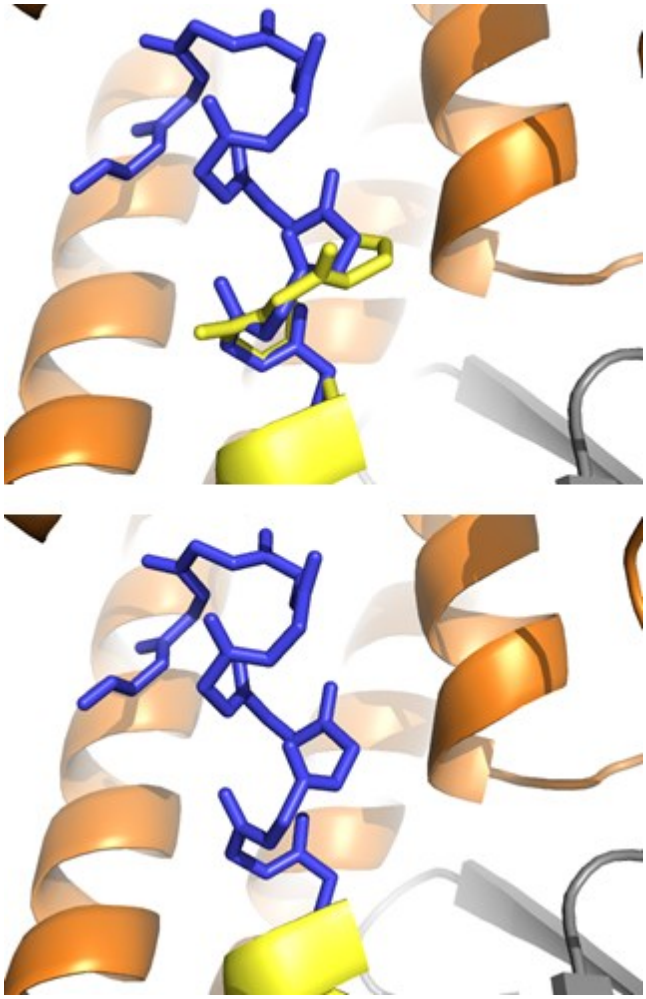


Figure 28 Superimposition of transducin's C-terminal helix with the opsin-bound peptide ligand.
(A) The opsin structure is shown as orange ribbon with the eleven residue C-terminal peptide of transducin as blue sticks (PDB 3DQB). The C-terminus of the α-subunit of Gαi-1GOT in yellow has been superimposed so that residues 344-347 overlap with the first four residues of the peptide. (B) The residues from the peptide are merged with Gαi-1GOT by replacing residues 344-347 of the α- subunit with the first four residues of the peptide.

The procedure required chain breaks within the linker regions of the α-subunit (between residues 59–60 and 184–185) and resulted in clashes in loop regions within the heterotrimer that were then resolved through loop reconstruction and model relaxation in Rosetta (Hirst, Alexander et al. ; Kaufmann, Lemmon et al. 2010). A rigid body docking protocol was executed to find placements of the helical domain consistent with the DEER distance restraints (Supporting Information text, Figure 29, and Table 24). An ensemble of models was found to be in agreement with the experimental distances from DEER data, consistent with the increase in width of the distance distributions (Figure 30). The model that agrees best with the most probable distances from DEER data (Figure 32B) fulfills all distance restraints within the error of the experiment and involves an approximately 8-Å motion of the helical domain away from the nucleotide domain as well as an approximately 29° rotation relative to its starting position (Figure 31).

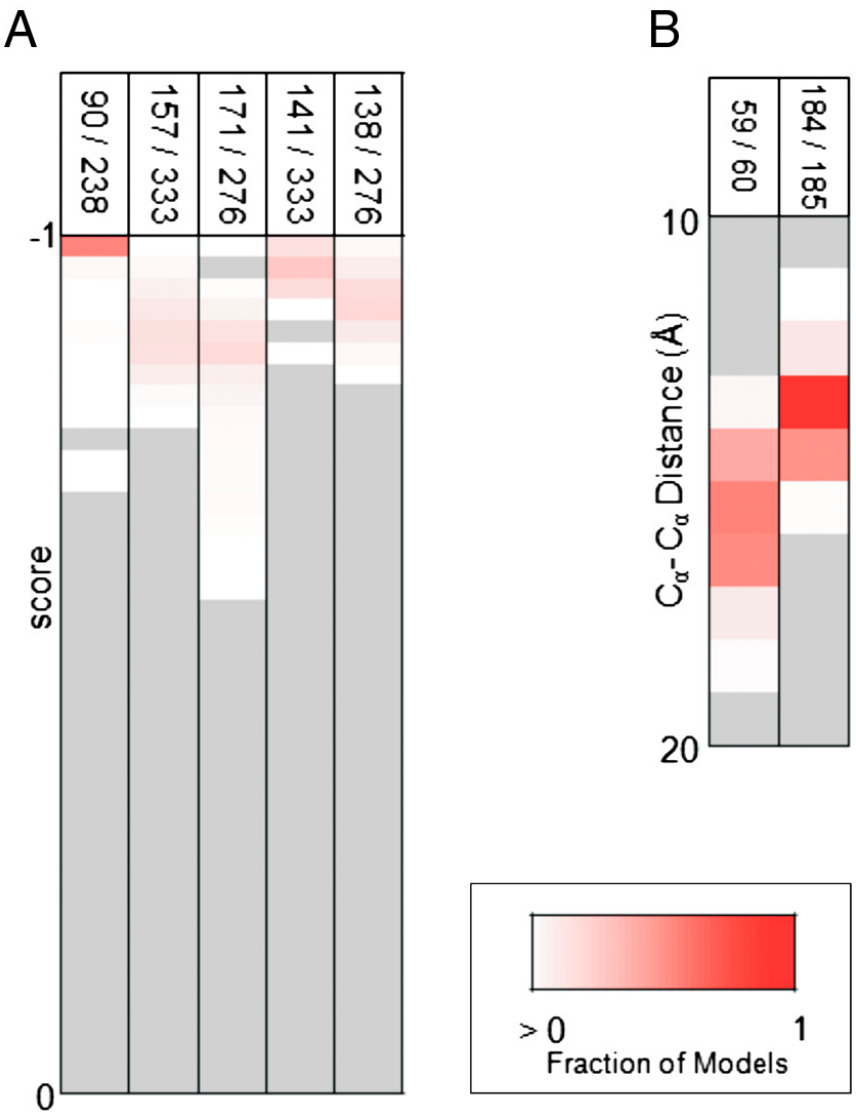


Figure 29 The 1,000 models with repositioned helical domain filtered by EPR-score and chain break distance.
(A) The models were scored for agreement with the distance measurements according to the knowledge-based potential of Hirst et al. (Hirst, Alexander et al.) The potential provides a score between -1 (perfect agreement) to zero (no agreement). Shown is the fraction of models for which a given score is observed for each EPR measurement. (B) It is important that the helical domain does not move too far from the initial position of its cut points from the rest of the α-subunit. Shown is the fraction of models with which a given Cα - Cα distance is observed for the two cut points. In both (A) and (B), grey areas have counts of zero.

Table 24 Agreement of the Gαi-1GOT model after Rosetta loop building and relaxation with experimentally measured EPR distances.
The EPR distances in the unbound and bound state are the most probable distance. The distances measured in structures are measured between Cβ atoms. Distances for the free heterotrimer were calculated using the experimental crystal structure (PDB 1GOT). Distances for the bound to activated receptor structure were calculated using the Gαi-1GOT model. The distance agreement between the model in the bound state and the EPR measurement in the bound state is calculated according to the knowledge-based scoring potential (KBP) (Hirst, Alexander et al.). Perfect agreement would be -1.0 and no agreement would be 0.0.

Mutant:	90 / 238	157 / 333	171 / 276	141 / 333	138 / 276
EPR experiment:					
Free heterotrimer	19 Å	28 Å	26 Å	33 Å	20 Å
Bound to activated receptor	37 Å	45 Å	34 Å	46 Å	34 Å
Distance change	18 Å	17 Å	8 Å	13 Å	14 Å
Structures:					
Free heterotrimer	11 Å	25 Å	23	32 Å	16 Å
Bound to activated receptor	32 Å	40 Å	25 Å	41 Å	29 Å
Distance change	21 Å	15 Å	2 Å	9 Å	13 Å
Agreement between experiment and model according to KBP	-0.96	-0.96	-0.71	-0.96	-0.97

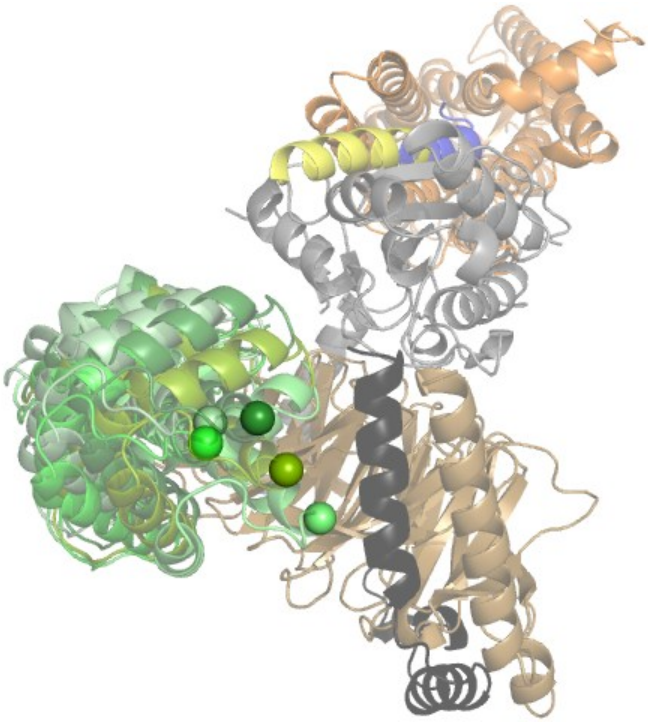


Figure 30 The 1,000 models resulting from repositioning the helical domain were hierarchically clustered.
Using a distance cutoff between clusters of 2.0 Å results in five cluster centers. Residue 90 is shown as spheres as a guide to the eye in distinguishing the different orientations of the helical domain. The cluster centers show relatively similar placements of the helical domain

.

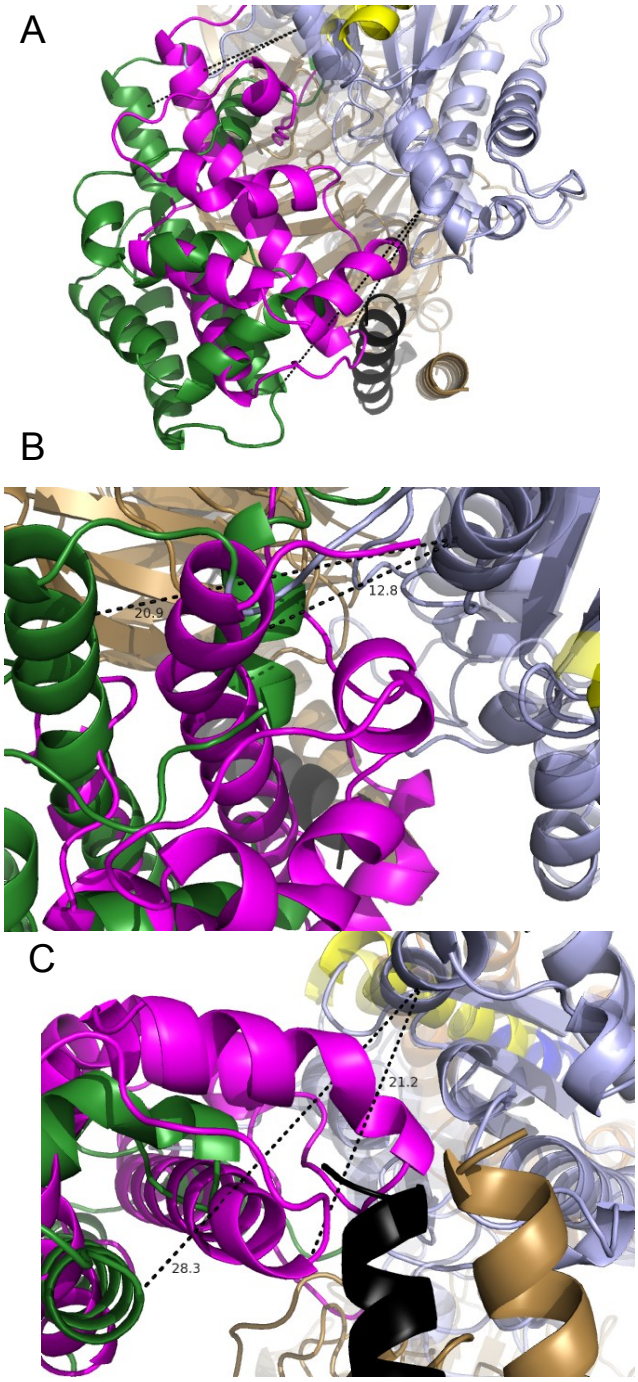


Figure 31 Shown is the position of the helical domain in the unbound heterotrimer as determined in the structure PDB 1GOT.
(magenta; rest of heterotrimer is translucent). Also, shown is the position of the helical domain in the Gαi-1GOT structure (green). The relative positions of the helical domains were determined by aligning residues in the α-subunit not within the helical domain. (A) The Cα - Cα distances at opposite ends of the helical domain are calculated in order to demonstrate the extent of the movement captured by the docking protocol. The distances were calculated between residues 51 and 66 (top), and 90 and 277 (bottom). The Gαi-1GOT coordinates were used for the reference (i.e. not moving) residues that are outside the helical domain (residues 51 and 277). The helical domain rotates 29°. (B) The change in distance of residue 66 from residue 51 is 8.1 Å. (C) The change in distance of residue 90 from residue 277 is 7.1 Å.

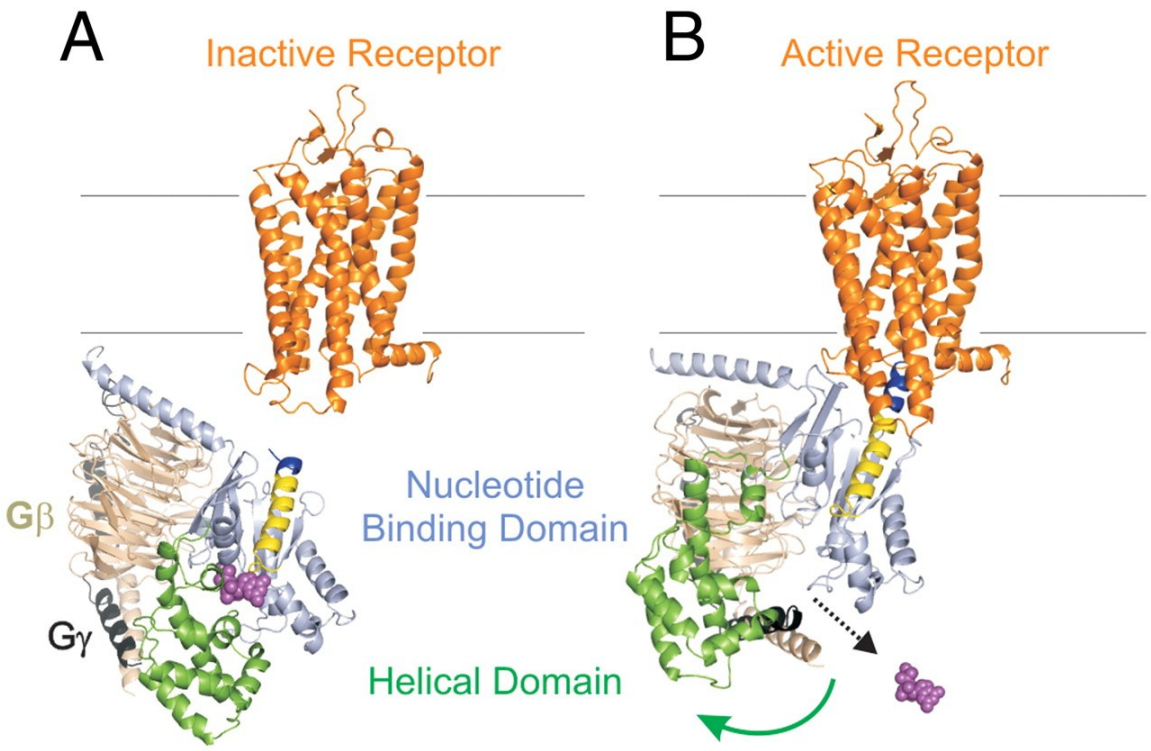


Figure 32 A model showing the opening of the interdomain cleft in formation of the empty complex.
(A) The inactive receptor (1U19.pdb) and inactive G protein (see Supporting Information text), with color coding as in Figure 21. (B) Model of the complex with active receptor (3DQB.pdb) showing the reorientation of the helical domain.

The model shown in Figure 32 incorporates a constraint gleaned from an interesting feature of the Gαi structure. In the structure, helix αA has a pronounced kink (dotted circle, Figure 21B) that is not due to proline or glycine residues in the sequence. Rather, the strained kink may be stabilized by a three-element network of packing interactions between the α5/β6 turn, the αF helix, and the helix αA. Previous results showed that receptor interaction with Gαi moves the α5/β6 turn, a change that could weaken the three-element interaction and trigger kink relaxation, thus moving the body of the helical domain relative to the nucleotide domain. Coupling between α5 and αF was suggested by several Gαi proteins that act as functional mimetics of the receptor bound state (Preininger, Funk et al. 2009). Kink relaxation is incorporated into the preliminary model

of Figure 32, but the actual relative movement of the helical domain shown in the figure does not depend on this mechanism, which will be examined in future studies.

The C terminus of G_α is a critical interaction site between the G protein and the receptor (Hamm, Deretic et al. 1988; Martin, Rens-Domiano et al. 1996; Marin, Krishna et al. 2002; Oldham, Van Eps et al. 2006) as illustrated in the model of Figure 32. Previous studies demonstrate that the C terminus undergoes a disorder-to-order transition upon binding to activated receptors, inducing structural changes that are important for efficient GDP release (Dratz, Furstenau et al. 1993; Kisselev, Kao et al. 1998; Van Eps, Anderson et al. 2010). $G_{\alpha i}$ with a flexible 5-glycine linker inserted at the base of the α5 helix (at residue 343, Figure 33A) binds to R* but eliminates a receptor-mediated movement of this helix, increases basal exchange, and uncouples nucleotide exchange from binding (Natochin, Moussaif et al. 2001; Oldham, Van Eps et al. 2006). We have introduced the same 5-glycine insertion into the interdomain pair, R90R1/E238R1. Figure 33 shows the distance distribution for the various states of $G_{\alpha i}$, to be compared with those of the parent protein shown in Figure 21C. Remarkably, the 5-Gly insertion results in a bimodal distance distribution in all states, the components of which correspond approximately to the open and closed positions of the helical domain. However, the distribution for the population at longer distances (approximately 40 Å) is substantially sharper than that in Figure 21C. Apparently, the perturbation of α5 by the insertion uncouples movement of the helical domain from receptor interaction. Although additional studies would be required to characterize the states of the insertion mutant, the result suggests a critical role of the C terminus in allosteric communication from the receptor to helical domain opening and the nucleotide binding pocket.

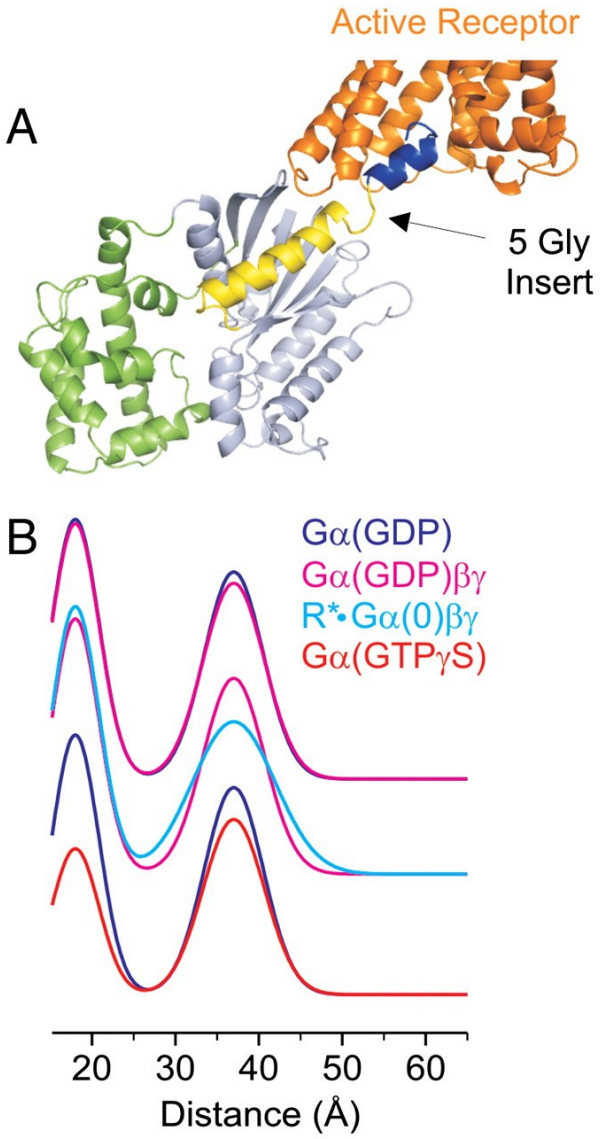


Figure 33 A 5-Gly insertion in α5 of Gαi uncouples domain opening from receptor binding. (A) Ribbon model of Gαi(GDP) showing the location of the 5-Gly insertion between residues 343–344; additional residues (345–354, blue ribbon) from the opsin/peptide crystal structure (3DQB.pdb) were added after the insert to suggest the subunit bound to activated rhodopsin. (B) Distance distributions of 90R1/238R1 compared for Gαi(GDP) and Gαi(GDP)βγ (Top), Gαi(GDP)βγ and R ∗ •Gαi(0)βγ (Middle), and Gαi(GDP) and Gαi(GTP) (Lower). The 5-Gly insert bearing the 90R1/238R1 double mutation binds to R* in native disc membranes to approximately the same extent as the GαiHI parent.

Is the domain rearrangement required for GDP release? To address this question, the two domains were cross-linked, disallowing the domain opening. For this purpose, a bifunctional, thiol-directed bis-maleimide was selected to cross-link cysteine residues in the R90C-E238C protein, based on the predicted proximity between these thiols in the $G_{\alpha i}$(GDP) protein (Figure 34A). Cross-linking resulted in a $G_{\alpha i}(GDP)_{\beta\gamma}$-protein competent to bind activated receptors to approximately the same extent as the parent protein (Figure 34B). Moreover, the cross-linked protein undergoes aluminum fluoride-dependent conformational changes (Figure 34C, Inset) consistent with an active, properly folded protein. On the other hand, this protein exhibited severely impaired rates of receptor-mediated nucleotide exchange as compared to either the parent or uncross-linked protein (Figure 34C), demonstrating the essential nature of the domain separation in receptor-mediated G-protein activation. The basal nucleotide exchange rate was only slightly reduced (Figure 34C), suggesting an effect specific to receptor-mediated nucleotide release, the slow step in G-protein activation.

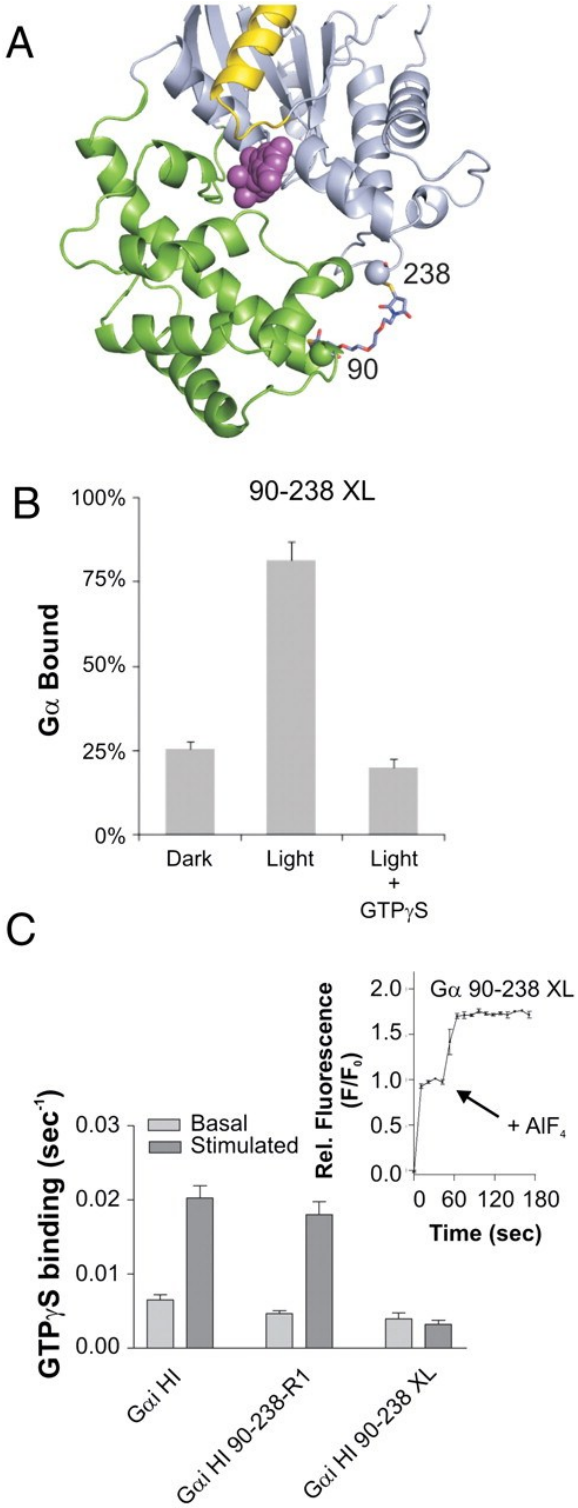


Figure 34 Cross-linking of the helical and nucleotide domains of a R90C-E238C Gαi double mutant.

(A) Model of the bis-maleimide interdomain cross-linker; the color code is as in Figure 21. (B) Binding of the cross-linked mutant to rhodopsin in disc membranes. (C) Basal and receptor-stimulated nucleotide exchange rates for the bis-maleimido cross-linked (XL) Gαi. For comparison, the $G_{\alpha i}$HI and R90R1/E238R1 nucleotide exchange rates are shown. (Inset) Tryptophan fluorescence changes of the XL $G_{\alpha i}$ subunit upon aluminum fluoride addition.

Conclusions

This study demonstrates that the result of G-protein interaction with an activated receptor is propagated allosterically to reorient the distant helical domain of $G_{\alpha i}$, opening the domain interface in formation of a flexible ternary receptor–G-protein complex. Preventing the large interdomain movement through cross-linking markedly reduces the rate of catalyzed nucleotide exchange, demonstrating the crucial role of the interdomain opening in receptor-mediated G-protein activation. Although the detailed mechanism is currently under further investigation, this domain opening would be predicted to reduce the GDP binding energy as interactions are lost upon opening of the domain interface. Together these changes help broaden our understanding of the conformational changes in the G protein that lead to GDP release, the slow step in G-protein activation.

Methods

Membrane binding assays

The ability of wild-type and $G_{\alpha i}$ proteins containing the side chain R1 (Figure 22) to bind rhodopsin was tested as described previously (Preininger, Parello et al. 2008). Gαi (5 μM) subunits were preincubated with Gβγ (10 μM) subunits on ice for 10 min. Then, in the dark, rhodopsin (50 μM) within native membranes was added to the heterotrimetric G protein in a buffer containing 50 mM Tris (pH 8.0), 100 mM NaCl, 1 mM MgCl2 and incubated on ice for 5 min. For dark measurements, reaction mixtures were protected from light for the rest of the procedure. Light activated samples, as well as light activated samples with GTPγS (100 μM), were incubated on ice for 30 min. The membranes in each treatment (dark, light, and light plus GTPγS) were pelleted by centrifugation at 20;000 × g for 1 h at 4 °C, and supernatants were removed from pellets. For the dark samples, supernatants were removed under dim red light. The supernatants and pellets

of each treatment were boiled and resolved by SDS-PAGE. The protein samples were visualized with Coomassie blue and quantified by densitometry using a BioRad Multimager. Each sample was evaluated by comparison of the amount of Gαi subunits in pellet (P) or supernatant (S) to the total amount of Gαi subunits (P+S) in both treatments and expressed as a percentage of the total Gαi protein. Results (Figure 23) are averages from at least three independent experiments.

Comparative Model of the Heterotrimeric G-Protein Transducin with Gαi Sequence.

The structure of the heterotrimeric G-protein transducin (PDB ID code 1GOT) was used as a template. The heterotrimeric protein consists of three subunits, α, β, and γ, and has GDP bound. The α-subunit (chain A) of the protein is a chimera of $G_{\alpha t}$ of bovine and $G_{\alpha i}$ of rat. A comparative model was constructed that consists entirely of the $G_{\alpha i}$ rat sequence using the sequence alignment shown in Figure 27. The sequence alignment shows an extension of the N-terminal α-helix by one winding (four-residue gap) that was built in the comparative model as a straight α-helix. The Rosetta side chain construction algorithm (Kuhlman, Dantas et al. 2003) was then used to convert the appropriate residues of 1GOT into $G_{\alpha i}$ sequence, yielding a comparative model termed $G_{\alpha i}$-1GOT. The command line options used were :

fixbb.linuxgccrelease -database -in:file:s -out:file:fullatom –resfile -out:prefix.

Superposition of the Transducin C-Terminal Helix with the Opsin- Bound Peptide Ligand.

The structure of G-protein coupled receptor opsin in complex with the C-terminal 11 residues of the α-subunit of the G-protein heterotrimer (PDB ID code 3DQB) was fused with the comparative model $G_{\alpha i}$-1GOT. Specifically, residues 344–347 in the α-subunit of the $G_{\alpha i}$-1GOT structure overlap in sequence with the first four residues of the peptide ligand in 3DQB (Figure 28). Using these four overlapping residues, the heterotrimer was positioned relative to the receptor. This defines an initial position of the heterotrimer relative to the receptor. As already described by Scheerer et al. (Scheerer, Park et al.

2008), this procedure positions portions of the heterotrimer in the membrane core in a nonphysical way. In order to resolve the penetration of the heterotrimer into the membrane core, rotations of portions of the heterotrimer are performed at two pivot points. Subunits β and γ are rotated along with the N-terminal helix and switch-2 region of the α-subunit such that the resulting position of the N-terminal helix is approximately parallel with the membrane (40° rotation). A second rotation of 15° of the heterotrimer is applied at the junction of the 3DQB peptide and C-terminal helix of Gαi-1GOT, moving the N-terminal helix parallel with the membrane. The combination of these two rotations creates a physically realistic model that removes the β-, γ-subunits from the membrane core, places the N-terminal amphipathic helix parallel to the membrane surface, and puts the N terminus in a location that allows the alkyl chain of the myristoyl group and the nearby farnesylated C terminus of the γ-subunit to penetrate the membrane. The procedure results in chain breaks within the α-subunit and minor clashes in loop regions within the heterotrimer that are resolved via the Rosetta loop building protocol.

α-Helical Domain Docking

EPR distance measurements display a reorientation of the helical domain of the α-subunit when the heterotrimer binds to the receptor (Figure 21). In order to capture this conformational motion, the α-helical domain was detached from the rest of the α-subunit by introduction of chain breaks between residues 59/60 and 184/185 of chain A of the Gαi-1GOT structure. Next, a rigid body docking protocol was executed to sample possible placements of the helical domain with respect to the α-subunit. A total of 140,000 structures were created using Rosetta (Gray, Moughon et al. 2003). The starting position of the α-helical domain was initially perturbed by up to 1.5 Å and 4° rotation. During docking trajectories translations of up to 0.05 Å and rotations of up to 2.5° were performed in a stepwise procedure. The command line flags used follow:

docking_protocol.linuxgccrelease -in:file:s start.pdb -out:nstruct 100 -docking:dock_pert 1.5 4 -docking:dock_mcm_trans_magnitude 0.05 - docking:dock_mcm_rot_magnitude 2.5 –out:overwrite

Filtering of α-Helical Domain Docking Models

Docking models were filtered for agreement with EPR distance data after docking. Agreement with the EPR distance restraints is calculated according to the knowledge-based potential given by Hirst et al. (Hirst, Alexander et al.). Agreement can be expressed with a value between 0 (no agreement) and −1 (perfect agreement, Figure 29). In addition to the EPR distances, a filter was applied to ensure the chain break created at the cut points can be resolved through remodeling a minimal number of residues around the cut points. This filter minimizes the distances between residues 59/60 and 184/185 of the α-subunit of Gαi-1GOT (Figure 29B). The 1,000 models that pass both filters undergo a clustering analysis (Figure 30), and the cluster center that agrees best with the experimental data is used for all further analysis (Table 24). This model shows a translation of approximately 8 Å and a rotation of 29° of the α-helical domain compared to its starting position. The increased width in the distance distributions obtained from EPR spectroscopy (Figure 21C) suggests a flexible relative orientation of the helical domain with respect to the heterotrimer in the receptor-bound state. The ensemble of 1,000 models in agreement with the EPR data might reflect part of this spatial disorder. A single model was selected to facilitate discussion of the general movement of the α-helical domain, as it is consistent between all models (Figure 30). We conclude that this movement is well defined by the experimental data. Additional experimental measurements will be necessary to determine the parameters of the spatial disorder. Rosetta loop building (Wang, Bradley et al. 2007) and relaxation protocols (Misura and Baker 2005) were utilized in order to reconnect the helical domain back to the rest of the α-subunit and refine the complex within the Rosetta energy

functions. In addition, the αA helix (α-subunit residues 63–90) is unkinked in the model of the activated heterotrimer–receptor complex solely for demonstrative purposes of a possible mechanism of leverage for generating the helical domain movement.

CHAPTER V

A ROTATION OF THE C TERMINAL HELIX CONNECTS BINDING OF THE GI PROTEIN AND ACTIVATED RECEPTOR TO DISASSOCIATION OF HELICAL DOMAIN AND GDP RELEASE

This chapter is based on the manuscript in preparation of the same name.

Summary

We have developed a unified model of R*-Gi interaction that reconciles the β2AR-Gs structure determined by X-ray crystallography with experimental data from EPR spectroscopy, hydrogen/deuterium exchange, and fluorescence. An *in silico* analysis of energetic changes upon interaction with the activated receptor links the binding event to displacement of the helical domain and GDP release. The 67^{o} rotation and 3.6Å movement of the α5 helix weakens its interaction with the α1 helix and with the β sheet scaffold. The unwinding of the N-terminal turn of the α5 helix induces a conformational change in the β6-α5 loop that pushes on the αG helix. Both changes loosen the interaction between GTPase and helical domain ultimately triggering its release from the GTPase domain and opening of the GDP binding pocket. Loosening of the β6-α5 loop decreases further the affinity for GDP. Further, we have performed calculations specifically designed to determine structural dynamics of the helical domain in the receptor-bound state. Specifically we compute an ensemble of 10 structures that reproduces closely the distribution of distances measured in EPR DEER experiments. The helical domain movement is large. The space sampled by the helical domain is markedly different from the orientation seen in the crystal structure of β2AR-Gs, identical with our previous model (PNAS, 2011, Van Eps et al), and EM structures. We further

validate the unified model through comparison with accessibility information deduced from new and previously reported EPR CW, fluorescence, and deuterium exchange measurements. The present model integrates all experimental data into a model of conformational states associated with G protein activation pin-pointing energetic contributions to activation. It thereby provides a roadmap for future experimental studies of this process.

Results

An ensemble of helical domain positions consistent with EPR distance restraints

The pool of docked models was used to find a subset of models that can reproduce the distance probability distributions of five EPR DEER distance measurements(Van Eps, Preininger et al. 2011). A Monte Carlo Metropolis (MCM) simulation was used to select subsets of models that collectively reproduce the distance distributions observed in the EPR DEER experiments. The cone model (Alexander, Al-Mestarihi et al. 2008; Hirst, Alexander et al. 2011) was used to convert CB-CB distances measured in the models to EPR DEER distance probability distributions. For a given ensemble of models, these probability distributions are compared with the EPR DEER measurement using the cumulative EUCLIDIAN distance (Kamarainen, Kyrki et al. 2003). The overall score of a given ensemble of models is the sum of the scores for the five EPR distance measurements. 1000 independent Monte Carlo simulations were conducted. The ensemble of ten structures with the best agreement score constitutes the ensemble of the R*-Gi complex. Figure 35 compares the experimentally observed distance distributions with the distance distributions of the final ensemble model of the R*-Gi complex. Figure 36 illustrates the space sampled by the helical domain.

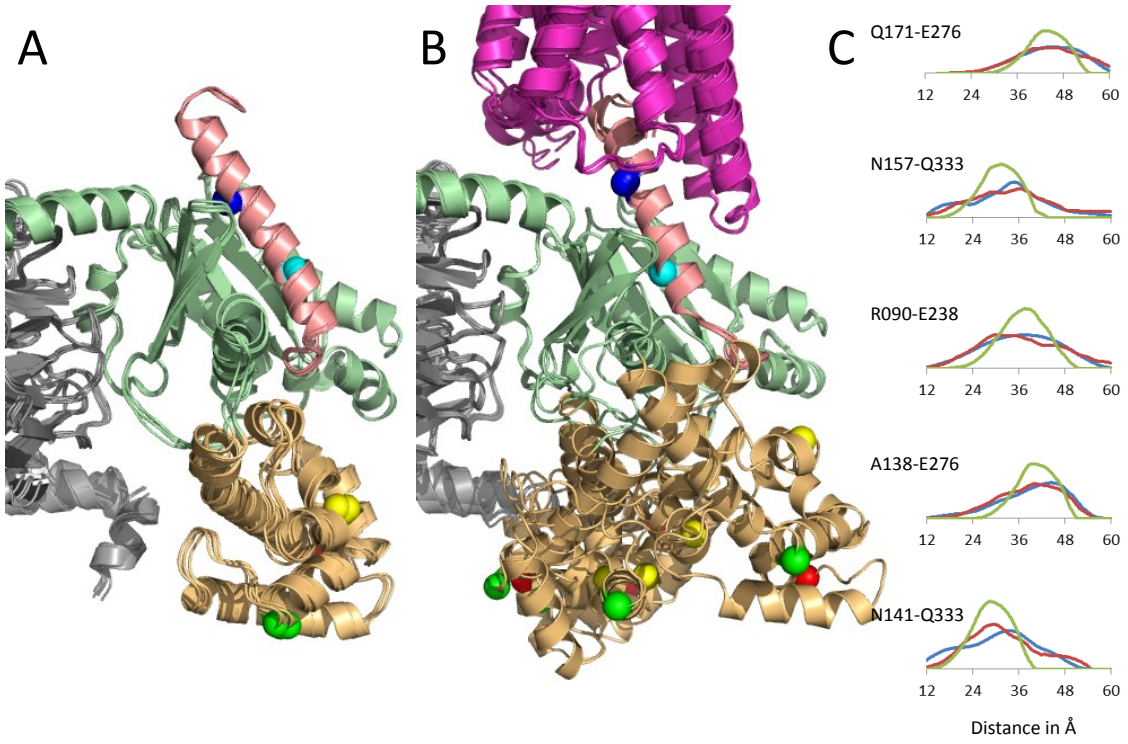


Figure 35 Placement of helical domain and rotation of a5 as observed by EPR measurements.
Panel (A) shows Gai in the basal state. Panel (B) displays the Gai bound to activated receptor R*. To illustrate motion landmark residues are colored: L092 (red), E122 (green), D158 (yellow), V335 (cyan), I343 (blue). Panel (C) compares the experimental distance distribution as observed in EPR DEER measurements (blue) with the predicted distribution computed from the ensemble mode of the R*-Gi complex (red). In green we show the distance distribution of our previous model which reproduces average distance accurately but not the distance distribution (Van Eps, Preininger et al. 2011).

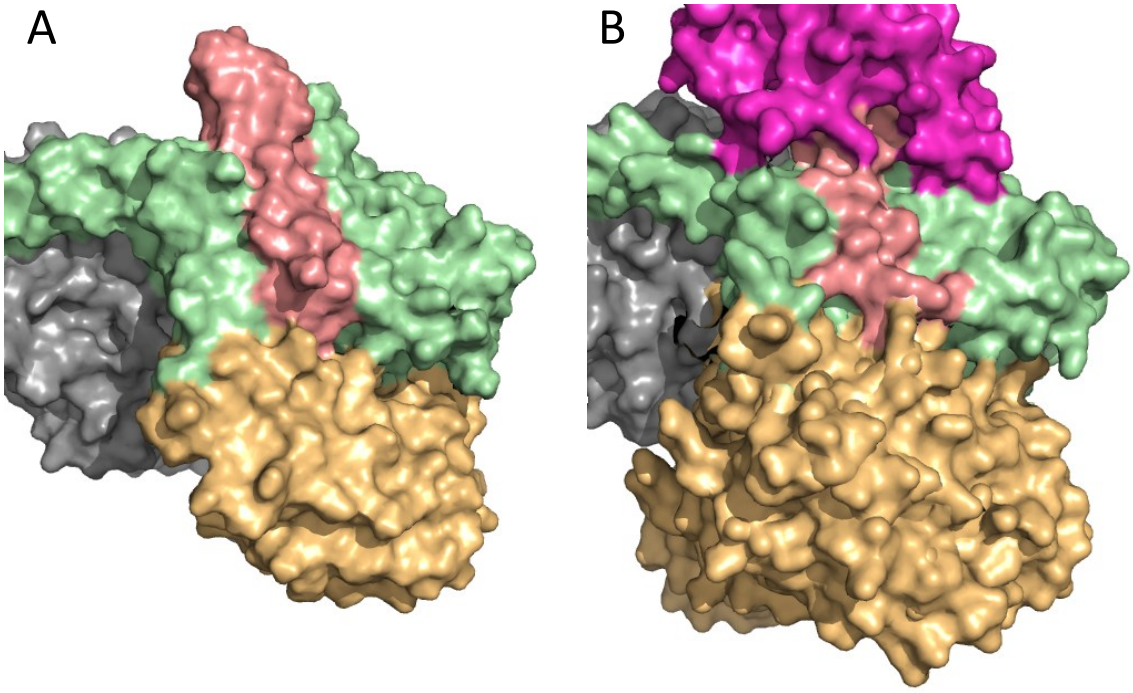


Figure 36 Space occupied by helical domain.
Panel (A) shows Gai in the basal state. Panel (B) displays the Gai bound to activated receptor R*.

Agreement of model with accessibility/mobility data from CW EPR, Fluorescence, and H/D exchange experiments

Solvent accessibility in the ensembles is calculated using a neighbor count measure that has been optimized to correlate with relative solvent accessible surface area (rSASA) (Durham, Dorr et al. 2009). To calculate the accessibility change between the unbound and bound states predicted by the model all pairwise neighbor count changes were calculated. The average, standard deviation, and Z-score was calculated and classified into five groups (strong increase: $\Delta NV > 3$, moderate increase: $3 > \Delta NV > 1$, about equal: $1 > \Delta NV > -1$, moderate decrease: $-1 > \Delta NV > -3$, strong decrease: $-3 > \Delta NV$). For H/D-exchange data, values are averaged over the length of the peptide of interest. The magnitude and direction of accessibility change was compared with the experimental data which had been manually classified in five corresponding groups (Table 25 and Table 26). We generally find that the predicted changes in accessibility track the experimental data (Figure 37). The correlation coefficients are 0.51 for the fluorescence measurements, 0.49 for the CW EPR measurements, and 0.55 for the H/D-exchange data. A notable difference is observed for CW EPR and fluorescence data in switch 1 and 2 regions. Experimentally observed changes in these regions often contradict each other suggesting the introduction of large labels into the interface of Gα and Gβγ might perturb the structure.

Table 25 Agreement of unified model with changes in accessibility observed EPR CW and fluorescence measurements

entity	amino acid	CW EPR	fluorescence	Δ exposure		(Z-score)	comment
a1	V050	-1	0	-0.3	±0.2	(2)	
helical	Q171	1	1	1.4	±0.5	(3)	
switch1	V179	1	-1	-1.1	±0.4	(3)	
switch1	K180	2	-1	-1.0	±0.7	(1)	
switch1	T182	0	-1	0.9	±0.4	(2)	
switch1	I184	0	-1	0.8	±0.6	(1)	
GTPase	E186	2	-1	-0.5	±0.2	(3)	
GTPase	T187	0	N/A	-0.4	±0.1	(3)	
GTPase	F191	2	-1	-0.6	±0.1	(6)	
GTPase	L194	-1	-1	-0.3	±0.1	(3)	
switch2	S206	1	-2	3.1	±0.2	(13)	
switch2	K209	0	-1	0.0	±0.0	(0)	
switch2	W211	0	0	0.1	±0.1	(1)	
switch2	C214	1	0	0.0	±0.1	(0)	
switch2	G217	-1	-1	0.6	±0.2	(3)	
aG	L273	2	N/A	0.2	±0.1	(3)	
a4	A300	-1	N/A	0.4	±0.0	(8)	
GTPase	E318	-2	N/A	-2.1	±0.1	(26)	
GTPase	Y320	-2	N/A	-1.2	±0.0	(30)	
a4-b6	T321	1	N/A	-0.7	±0.1	(14)	
b6-a5	K330	-2	N/A	-0.7	±0.1	(5)	
b6-a5	N331	-2	N/A	-0.5	±0.1	(7)	
a5	Q333	0	N/A	0.3	±0.0	(7)	
a5	F334	-1	N/A	-0.9	±0.0	(21)	
a5	T340	0	N/A	0.4	±0.1	(6)	
a5	V342	-2	N/A	-0.8	±0.1	(7)	
a5	I344	-2	N/A	-1.0	±0.1	(9)	
a5	K349	-2	N/A	-0.8	±0.0	(21)	

Table 26 Agreement of unified model with changes in accessibility observed H/D exchange measurements

entity	amino acid	H/D exchange	Δ exposure		(Z-score)	comment
GTPase	A033-Q038	2	-0.3	±0.3	(1)	
a1	A041-Q052	2	0.6	±1.0	(1)	
a1/helical	M053-I081	1	0.1	±1.5	(0)	
helical	I082-A087	0	-0.3	±0.5	(1)	
helical	F095-A099	1	0.0	±0.4	(0)	
helical	A101-A104	1	0.1	±0.1	(1)	
helical	Q106-L110	0	0.3	±0.7	(0)	
helical	A111-A114	0	0.7	±1.9	(0)	
helical	E116-M119	0	-1.1	±1.7	(1)	
helical	E122-L123	0	0.7	±0.1	(5)	
helical	V126-G135	0	0.2	±0.5	(0)	
helical	I127-F140	0	0.1	±0.4	(0)	
helical	L148-A153	2	0.3	±0.7	(0)	
helical	N157-I168	0	-0.4	±0.9	(0)	
helical	T170-L175	1	1.7	±1.4	(1)	
helical/switch1	T177-V185	2	0.1	±2.7	(0)	
GTPase	H188-F196	1	-0.7	±0.6	(1)	
switch2	F199-C214	1	0.5	±1.9	(0)	
GTPase	G217-A220	0	-0.1	±0.9	(0)	
b4-a3	A226-L232	2	2.7	±2.0	(1)	
b4-a3/ GTPase	A235-L249	1	0.4	±0.8	(1)	
GTPase	S252-W258	0	-0.4	±0.5	(1)	
GTPase	F259-S263	0	-0.7	±0.6	(1)	
GTPase	I264-L268	0	-0.3	±0.5	(1)	
aG	N269-L273	2	0.7	±0.4	(2)	
aG	F274-I278	1	0.2	±0.6	(0)	
aG	E289-Y290	0	-0.8	±0.2	(5)	
aG	A291-N294	0	-0.8	±1.8	(0)	
GTPase	T295-Y302	0	-0.2	±0.6	(0)	
GTPase	E297-Y302	0	-0.2	±0.6	(0)	
GTPase	Q304-E308	0	-1.1	±1.0	(1)	
a4-b5	L310-Y320	1	-2.5	±2.1	(1)	
GTPase	H322-A326	1	1.4	±1.8	(1)	
b6-a5	D328-Q333	2	0.7	±1.7	(0)	
a5	F334-T340	0	-1.0	±1.3	(1)	
a5	D341-F354	-2	-2.7	±1.8	(1)	

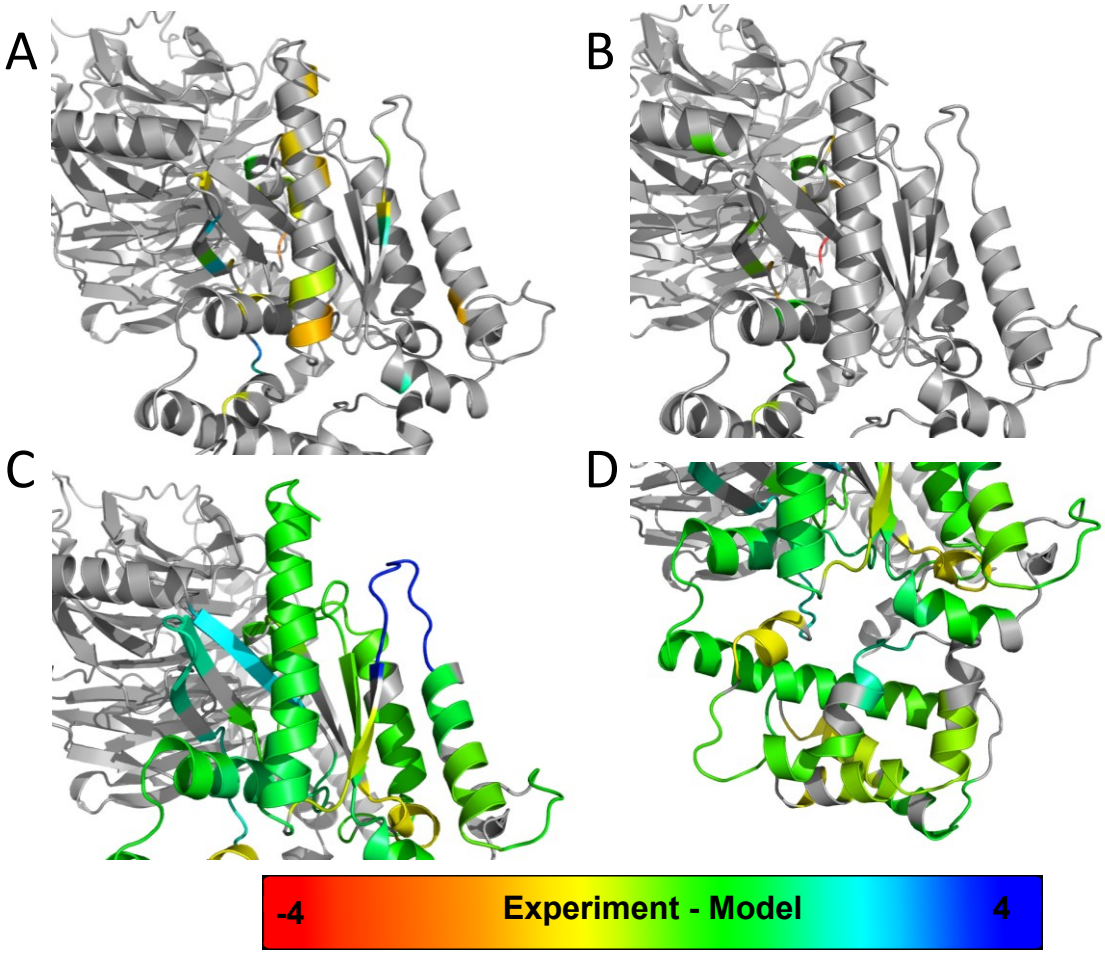


Figure 37 Agreement of unified model with changes in accessibility observed in EPR CW, fluorescence, and deuterium exchange measurements.
Panel (A) illustrates CW EPR experimental data in C-terminus | Gai interface. Experimentally observed changes were classified into five groups from strong decrease (-2) to strong increase (+2). Average amino acid accessibility changes were classified likewise into five groups from strong decrease (-2) to strong increase (+2). Plotted is the difference, i.e. yellow and green colors indicate good agreement of model and experiment. Panel (B) illustrates Fluorescence experimental data. Panel (C) displays H/D exchange in C-terminus | Gai interface. Panel (D) illustrates H/D exchange in helical domain | Gai interface. Note that no perfect correlation is expected as (1) experiments capture additional aspects beyond amino acid exposure and (2) exposure is estimated from the CB position alone.

Targeted energetic analysis of selected interfaces using Rosetta

The stabilizing interactions between key interfaces in Gαi were examined before and after receptor binding. Specifically, we studied four interfaces: Gαi-helical domain|Gαi-GTPase domain interface (Figure 38), GDP| Gαi-GTPase domain interface (Figure 39), R*|Gαi-GTPase domain interface (Figure 40), and C-terminal helix h5|Gαi-GTPase domain (Figure 41). The Rosetta ΔΔG protocol (Kortemme, Kim et al. 2004) was used to determine all interactions that contribute to the stabilization of these interfaces before and after receptor binding. The resulting ΔΔG values are broken down on a per-residue basis to identify hot-spots (Table 27). Looking at the difference in the ΔΔG values between the unbound and bound states for a given interface indicate changes in key interactions. Calculations were conducted over the ensembles for the unbound and bound states to compute mean ΔΔG and standard deviations. Only statistically significant contributors (Z-score larger than 2) that are large ($|\Delta\Delta G| > 0.5$ REU) were considered for further analysis.

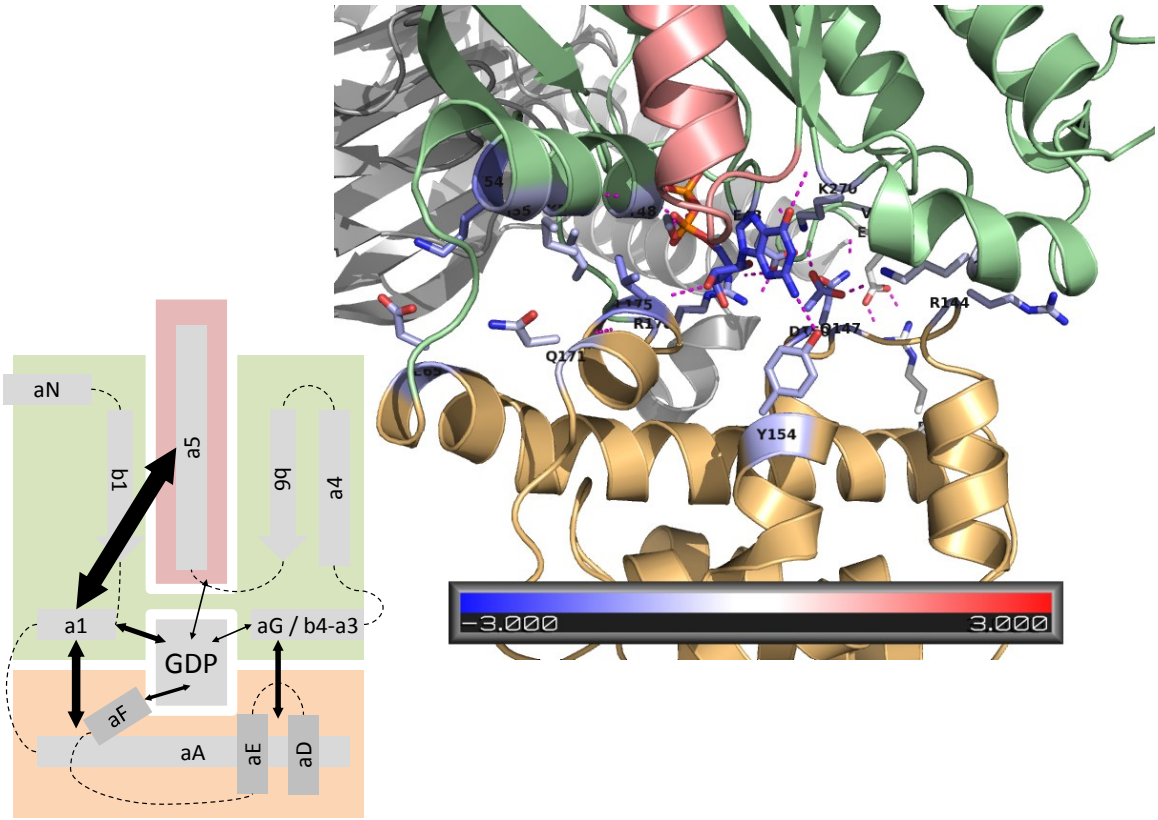


Figure 38 Energetics of helical domain | Gai interface in free Gai.
Residues are colored by the interaction energy in Rosetta Energy Units (REU) from red (repulsive) over white (neutral) to blue (attractive). Attractive polar interactions between the helical domain and Gai are indicated by dotted, magenta lines. Residues that contribute more than 0.5 REU are displayed as sticks and labeled.

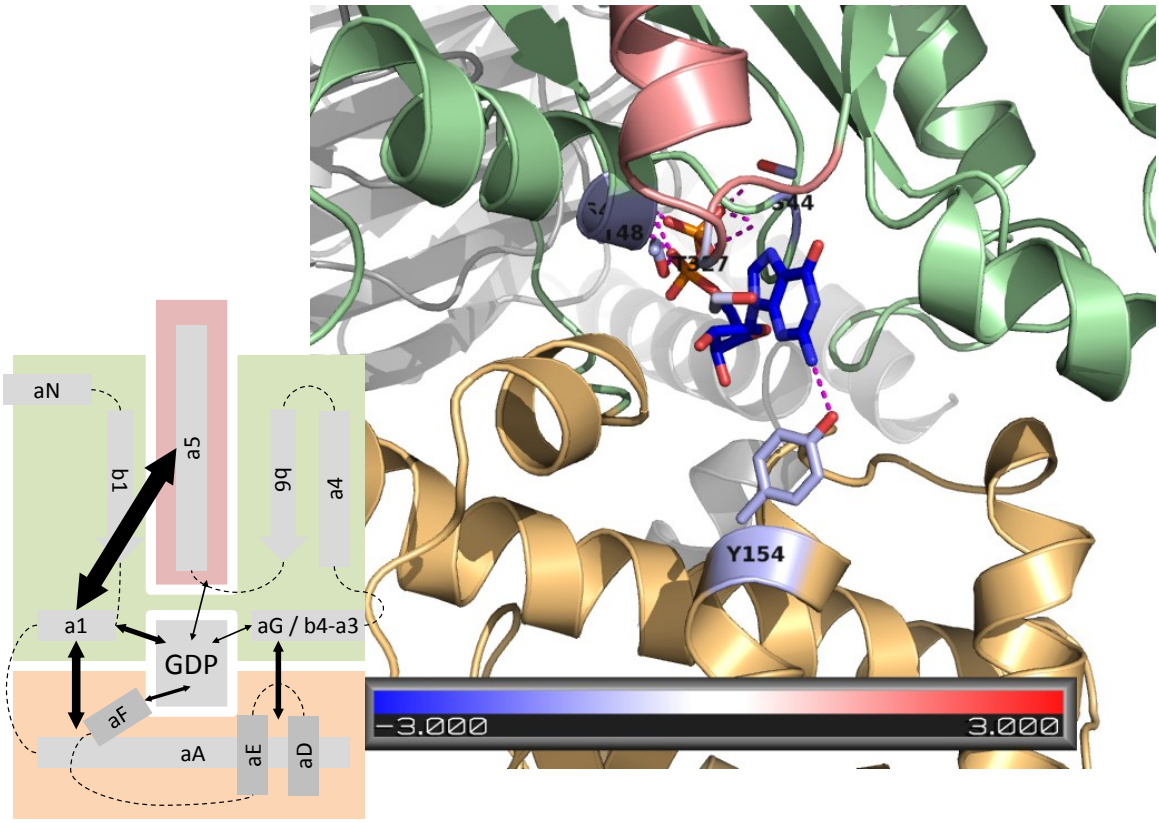


Figure 39 Energetics of GDP | Gai interface in free Gai.
Residues are colored by the interaction energy in Rosetta Energy Units (REU) from red (repulsive) over white (neutral) to blue (attractive). Attractive polar interactions between GDP and the Gαi are indicated by dotted, magenta lines. Residues that contribute more than 0.5 REU are labeled are displayed as sticks and labeled.

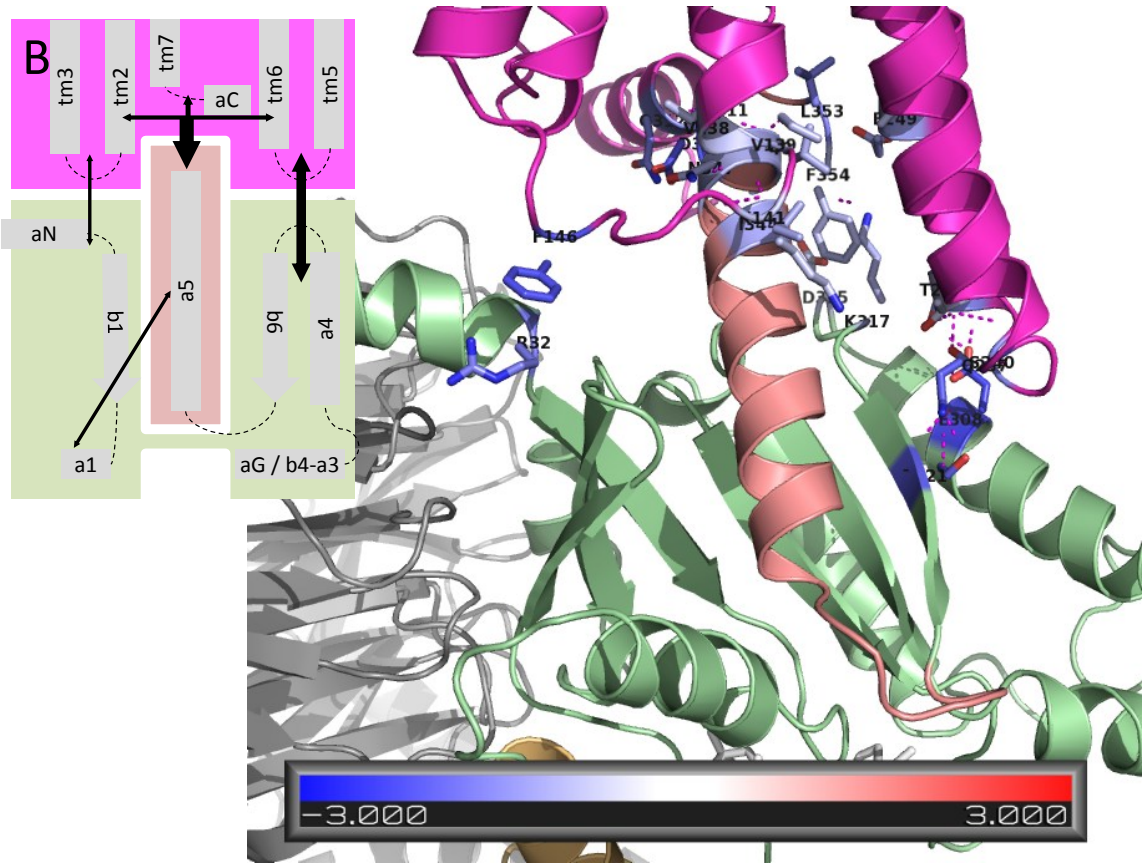


Figure 40 Energetics of R* | Gai interface in the R*-Gai complex.
Residues are colored by the interaction energy in Rosetta Energy Units (REU) from red (repulsive) over white (neutral) to blue (attractive). Attractive polar interactions between R* and Gai are indicated by dotted, magenta lines. Residues that contribute more than 0.5 REU are labeled are displayed as sticks and labeled.

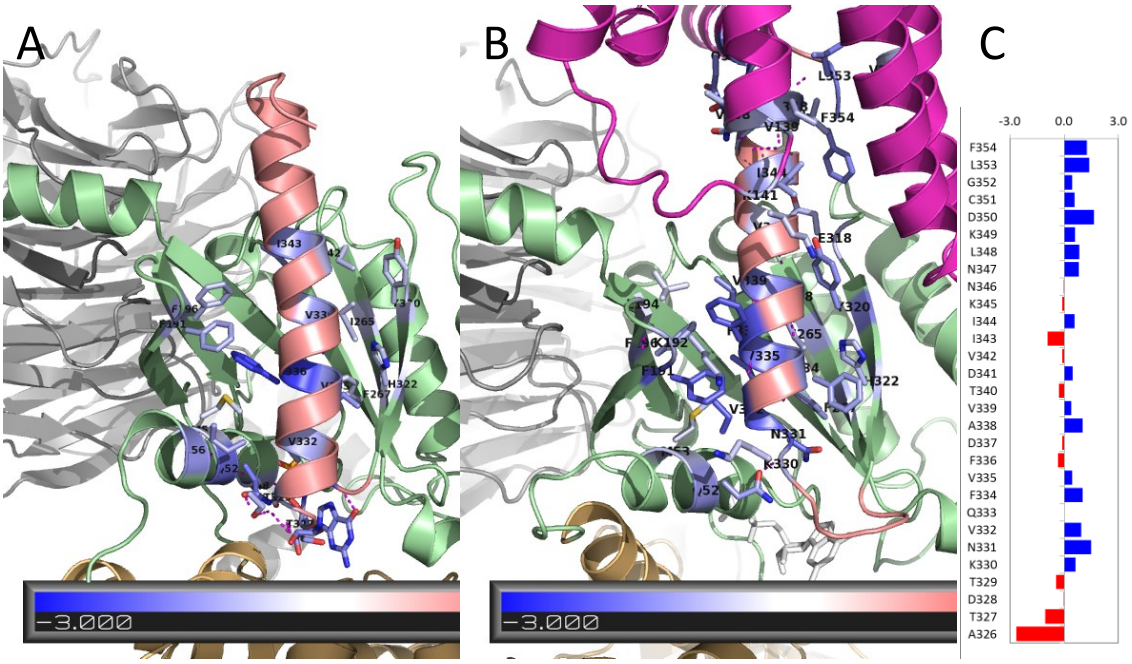


Figure 41 Energetics of C-terminus | (R*-)Gai interface in free Gai (A) and the R*-Gai complex (B).
Residues are colored by the interaction energy in Rosetta Energy Units (REU) from red (repulsive) over white (neutral) to blue (attractive). Attractive polar interactions are indicated by dotted, magenta lines. Residues that contribute more than 0.5 REU are displayed as sticks and labeled. Panel (C) plots energy change of C-terminal residues (b6-a5 and a5) upon receptor binding. A blue color indicates stabilzation, a red color indicates destabilization.

Table 27 Interaction energies across selected interfaces in free and R* bound Gai

free Gai					R*-Gai complex				
	GDP | Gai interface					R* | Gai interface			
entity	amino acid	energy in REU		(Z-score)	entity	amino acid	energy in REU		(Z-score)
GDP		5.1	±0.2	(21)	R* IL2	V138	0.6	±0.0	(21)
a1	S044	1.0	±0.0	(27)	R* IL2	V139	0.8	±0.0	(19)
a1	S047	0.8	±0.0	(24)	R* IL2	K141	0.6	±0.0	(15)
a1	T048	0.9	±0.1	(19)	R* IL2	F146	2.0	±0.1	(27)
helical	Y154	0.8	±0.0	(25)	R* IL3	Q237	1.8	±0.1	(15)
b6-a5	T327	0.5	±0.0	(21)	R* IL3	S240	1.3	±0.0	(47)
a1	cumulative	3.1			R* IL3	T242	1.4	±0.1	(26)
helical	cumulative	0.8			R* IL3	T243	0.5	±0.0	(19)
b6-a5	cumulative	0.9			R* IL3	E249	1.0	±0.0	(24)
aG	cumulative	1.0			R* IL3	V250	0.6	±0.0	(15)
					R* aC	K311	1.0	±0.2	(5)
	helical domain | Gai interface				R* aC	Q312	1.3	±0.2	(8)
entity	amino acid	energy in REU		(Z-score)	R*	cumulative	17.3		
helical	E065	0.9	±0.1	(8)	aN-b1	R032	1.4	±0.1	(22)
helical	R090	0.5	±0.4	(1)	a4-b6	E308	2.1	±0.0	(49)
helical	R144	0.8	±0.1	(15)	a4-b6	D315	0.7	±0.2	(3)
helical	Q147	1.2	±0.1	(9)	a4-b6	K317	0.8	±0.1	(9)
helical	D150	1.5	±0.2	(8)	a4-b6	T321	1.7	±0.1	(15)
helical	Y154	0.8	±0.0	(21)	a5	I344	0.9	±0.1	(20)
helical	Q171	0.5	±0.2	(3)	a5	N347	0.8	±0.0	(21)
helical	L175	1.3	±0.1	(16)	a5	L348	0.8	±0.1	(16)
helical	R178	1.0	±0.1	(15)	a5	D350	1.8	±0.1	(19)
helical	cumulative	10.1			a5	C351	0.6	±0.0	(19)
a1	E043	1.0	±0.1	(16)	a5	L353	1.4	±0.1	(23)
a1	T048	0.7	±0.1	(14)	a5	F354	0.7	±0.1	(12)
a1	K051	0.8	±0.1	(9)	aN-b1	cumulative	2.2		
a1	K054	1.4	±0.1	(19)	a4-b6	cumulative	5.7		
a1	I055	0.7	±0.2	(4)	a5	cumulative	8.1		
b4-a3	V233	0.5	±0.0	(15)					
b4-a3	E238	0.6	±0.4	(1)					
aG	K270	0.8	±0.0	(21)					
aG	K277	0.5	±0.2	(2)					
GDP		2.0	±0.1	(18)					
a1	cumulative	5.5							
b4-a3	cumulative	2.5							
aG	cumulative	1.8							

	a5 | Gai interface					a5 | R*-Gai interface			
entity	amino acid	energy in REU		(Z-score)	entity	amino acid	energy in REU		(Z-score)
b6-a5	A326	2.4	±0.1	(39)	b6-a5	A326	-0.2	±0.1	(5)
b6-a5	T327	1.1	±0.1	(13)	b6-a5	T327	0.0	±0.0	n.d.
b6-a5	D328	0.4	±0.0	(39)	b6-a5	D328	0.4	±0.3	(2)
b6-a5	T329	0.8	±0.0	(40)	b6-a5	T329	0.4	±0.2	(2)
b6-a5	K330	0.0	±0.0	n.d.	b6-a5	K330	0.6	±0.3	(3)
b6-a5	N331	0.2	±0.1	(2)	b6-a5	N331	1.7	±0.1	(21)
a5	V332	0.9	±0.0	(27)	a5	V332	1.8	±0.1	(13)
a5	Q333	0.5	±0.0	(32)	a5	Q333	0.5	±0.0	(19)
a5	F334	0.0	±0.0	n.d.	a5	F334	1.0	±0.1	(12)
a5	V335	1.1	±0.1	(9)	a5	V335	1.6	±0.1	(34)
a5	F336	2.2	±0.1	(34)	a5	F336	1.9	±0.0	(46)
a5	D337	0.2	±0.0	(13)	a5	D337	0.1	±0.0	(5)
a5	A338	0.2	±0.0	(14)	a5	A338	1.2	±0.0	(35)
a5	V339	1.0	±0.1	(21)	a5	V339	1.4	±0.0	(43)
a5	T340	0.5	±0.0	(17)	a5	T340	0.2	±0.0	(6)
a5	D341	0.0	±0.0	n.d.	a5	D341	0.5	±0.1	(5)
a5	V342	0.6	±0.1	(7)	a5	V342	0.5	±0.0	(13)
a5	I343	1.1	±0.0	(42)	a5	I343	0.2	±0.0	(9)
a5	I344	0.3	±0.2	(2)	a5	I344	0.9	±0.1	(20)
a5	K345	0.0	±0.0	n.d.	a5	K345	-0.1	±0.2	(1)
a5	N346	0.1	±0.1	(1)	a5	N346	0.1	±0.0	(9)
a5	N347	0.0	±0.0	n.d.	a5	N347	0.8	±0.0	(21)
a5	L348	0.0	±0.0	n.d.	a5	L348	0.8	±0.1	(16)

Table 27 continued

a5	K349	0.1	±0.1	(1)	a5	K349	0.7	±0.5	(2)
a5	D350	0.2	±0.0	(4)	a5	D350	1.8	±0.1	(19)
a5	C351	0.0	±0.0	n.d.	a5	C351	0.6	±0.0	(19)
a5	G352	0.0	±0.0	n.d.	a5	G352	0.5	±0.0	(28)
a5	L353	0.0	±0.0	n.d.	a5	L353	1.4	±0.1	(23)
a5	F354	0.0	±0.0	n.d.	a5	F354	1.3	±0.1	(20)
b6-a5	cumulative	5.0			b6-a5	cumulative	2.9		
a5	cumulative	9.0			a5	cumulative	19.4		
a1	T048	0.8	±0.0	(23)	a1	Q052	1.0	±0.1	(16)
a1	Q052	1.5	±0.0	(39)	a1	M053	0.6	±0.0	(25)
a1	M053	0.5	±0.1	(10)	GTPase	F191	1.8	±0.1	(14)
a1	I056	1.1	±0.0	(32)	GTPase	K192	0.8	±0.0	(21)
GTPase	F191	0.9	±0.0	(21)	GTPase	L194	0.5	±0.1	(8)
GTPase	F196	0.8	±0.1	(16)	GTPase	F196	1.0	±0.0	(24)
GTPase	I265	0.6	±0.0	(28)	GTPase	I265	0.8	±0.1	(15)
GTPase	F267	0.6	±0.1	(7)	GTPase	F267	0.9	±0.1	(19)
GTPase	Y320	0.6	±0.1	(10)	GTPase	E318	0.8	±0.3	(2)
GTPase	H322	0.7	±0.1	(9)	GTPase	Y320	1.2	±0.2	(7)
					GTPase	H322	0.8	±0.1	(12)
					R* IL2	V138	0.6	±0.0	(21)
					R* IL2	V139	0.8	±0.0	(19)
					R* IL2	K141	0.6	±0.0	(15)
					R* IL3	E249	0.7	±0.1	(16)
					R* IL3	V250	0.6	±0.0	(15)
					R* aC	K311	0.7	±0.1	(9)
					R* aC	Q312	1.4	±0.2	(9)
GDP		1.4	±0.1	(11)	GDP		0.2	±0.0	(5)
a1	cumulative	5.0			a1	cumulative	2.2		
GTPase	cumulative	6.4			GTPase	cumulative	10.3		
R*	cumulative	0.0			R*	cumulative	8.6		

Basal state: Gαi-helical domain|Gαi-GTPase and GDP| Gαi-GTPase domain interface

The helical domain is held in place by interactions of a1 (E043, T048, K051, K054, I055) with aA (E65) and aF (Q171, L175, 5.5 REU, Figure 38, Table 27). The helical domain is also fixed by electrostatic interactions of aG (K270, K277) and b4-a3 (V233, E238) loops with aA (R090), aD-aE loop (R144, Q147, D150) and aF-b2 loop (R178, 4.3 REU). Lastly, the interface is stabilized by contact between GDP and aD-aE loop (Y154, 2.0 REU). The total interaction energy is approximately 10.1 REU. GDP is stabilized through interactions with a1 (S044, S047, T048, 3.1 REU), the helical domain (Y154, 0.8 REU), and b6-a5 loop (T327, 0.5 REU). The total interaction energy is approximately 5.1 REU (Figure 39, Table 27).

Receptor-bound state: R|Gαi-GTPase domain interface*

Upon interaction with the receptor the high-affinity peptide (I344, N347, L348, D350, C351, L353, F354) is bound through TM3 (V138, V139, K141), TM6 (E249, V250), and TM7-aC loop (K311, Q312, 8.1 REU, (Figure 40, Table 27). Further, intracellular loop 2 (F146) is interacting with aN-b1 loop at R(32, 2.2 REU). The extended intracellular loop 3 (Q237, S240, T242, T243) interacts with a4 (E308), a4-b6-loop (D315, K317), and b6 (T321, 5.7 REU). The total interaction energy is approximately 17.3 REU.

Rewiring of the h5|Gαi-GTPase domain interface upon receptor interaction

The C-terminal helix a5 (V332, Q333, V335, F336, A338, V339, T340, V342, I343) interacts favorably with the β-sheet of the GTPase domain (F191, F196, I265, F267, Y320, H322, 6.4REU) and a1 (M053, I056, 2.5REU, (Figure 41A, Table 27). The b6-a5 loop (A326, T327, T329) interacts with a1 (T048, Q052, 2.5 REU) and GDP (1.4 REU).

Upon interaction with the activated receptor (Figure 41B) a5 (I344, N347, L348, K349, D350, C351, G352, L353, F354) experiences an attraction of 8.6 REU. This attractive interaction moves a5 3.6Å towards the receptor and triggers a rotation of 67°. The N-terminal winding of the helix melts. Helix a5 interaction with the β-sheet of the

GTPase domain is modified and strengthened (F191, K192, L194, F196, I265, F267, E318, Y320, H322, 10.3 REU). Reversely, interaction with a1 (Q052, M053, 2.2 REU) and GDP (0.2 REU) are significantly weakened supporting the melting of a1 and the release of the helical domain and GDP.

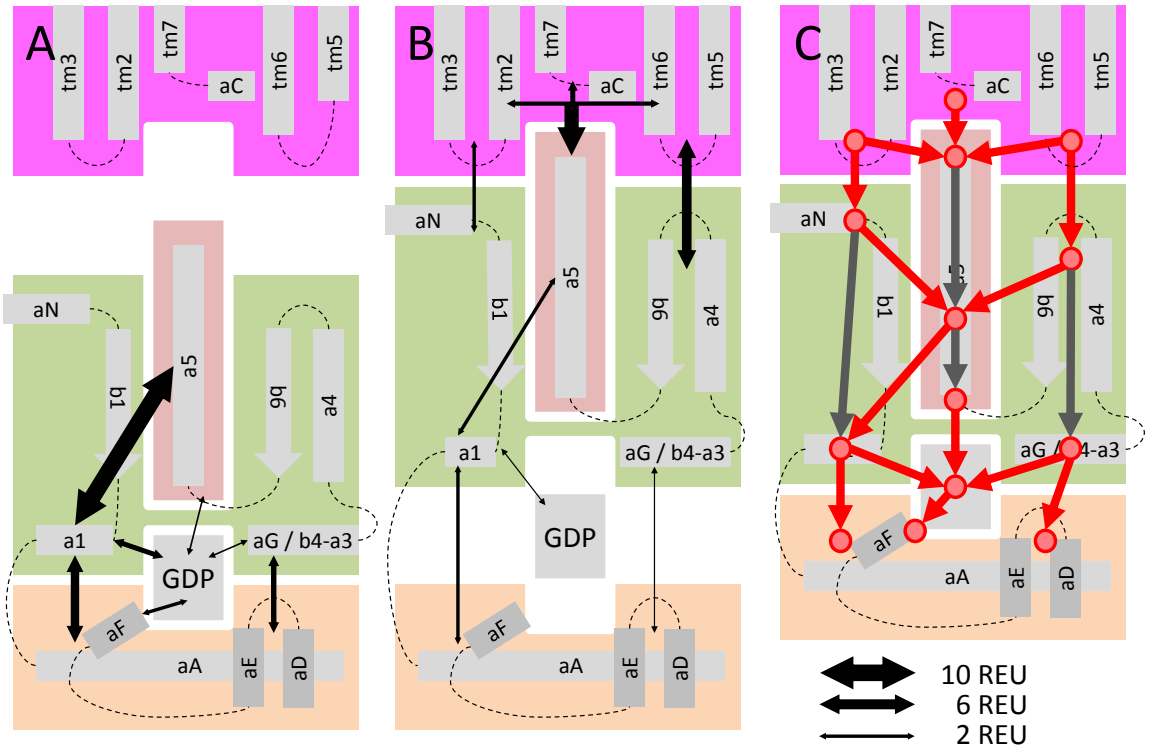


Figure 42 Energetic basis of signal transduction during Gai interaction with activated receptor R*.

The receptor is violet, the helical domain is orange, the GTPas domain is blue, a5 is red, and GDP is green. The thickness of arrows indicates magnitude of stabilization. Panel (A) displays interactions across interfaces in free Gai. Panel (B) displays interactions across interfaces in the R*-Gai complex. Panel (C) combines panels (A) and (B) highlighting signal transduction due to changes in interaction strength across interfaces (red arrows) and rigid body motion (grey arrows).

Discussion

The challenges associated with crystallization of membrane proteins require substantial perturbation of the native system. These perturbations together with the crystal lattice interactions can even alter the crystallized structure to conformations not present or minor in the native conformational ensemble. These perturbations from the native conformation are often localized to regions of the protein particularity flexible or proximal to the site of perturbation. Often the system crystallized is a homolog of the system of interest. At the same often structure and dynamics of such systems is studied with orthogonal methods such as EPR, NMR, fluorescence, cryo-Electron Microscopy, or H/D exchange. Next to observing agreement or disagreement of the crystal structure with these data the challenge is to construct a unified model that reconciles the crystal structure with the available experimental data. Ideally, such a hybrid methods model reflects structural dynamics of the associated state, is affiliated with a confidence level derived from the experimental restraints, and is verified through independent experimental data. Recent successes in modeling membrane proteins accurately suggest that computational methods become capable of adding atomic detail in regions where experimental information is not at high resolution. The model will not claim to be correct in all its detail but consistent with the current state of knowledge. The power of such a model is that it presents an atomic-detail hypothesis of the structure and dynamics thereby creating a roadmap for future experimental studies that can verify of reject parts of the model. In an iterative fashion a completely verified atomic-detail of the system can be constructed.

The recent determination of the crystal structure of the β2 adrenergic receptor–Gs protein complex provides atomic-detail insight into the interaction of a G-protein with an active G-protein coupled receptor (GPCR). While the availability of this experimental structure is a milestone in understanding the structural determinants of this critical

interaction in GPCR signaling, it is also obvious that crystallization conditions selected one low energy conformation from a potentially wide ensemble of structurally dynamic state. Specifically the introduction of T4-Lysozyne into an extracellular loop and complexation of the nanobody Nb35 perturb the system. In this specific case, it was suggested that the location of the dissociated helical domain of Gα is not accurately reflected in the experimental structure.

Energetic basis of signal transduction during Gai interaction with activated receptor R.*

Figure 42 summarizes these findings in a scheme: the activated receptor R* interacts with Gai through three major pathways: the high affinity C-terminal peptide of a5 (~50% of energetic contribution), the interaction of R* IL 3 with a4-b6 loop (~33% of the energetic contribution) and through the interaction of R* IL 2 with the aN-b1 loop (~17% of the energetic contribution). The consequences of binding of the high-affinity peptide include a 3.6Å and a 67° rotation of a5. This shifts a5 from one energetically favorable interaction with the β-sheet of the GTPase domain into a second state that is energetically even slightly more favorable, a process that is captured by the crystallographic snapshots and was previously deduced from CW-EPR mobility studies (Van Eps, Oldham et al. 2006). Figure 42c highlights that the interaction between I343 and the β-sheet is destabilized. This residue was part of a hydrophobic cluster with F191 and F196, an interaction that is weakened. F191 and F196 as well as the outmost tips of the β-sheet at F191, K192 and E318, Y320 rearrange and engage in improved contacts with a5.

The energetic stabilization of the C-terminus of a5 and its interaction with the β-sheet drive the conformational reorientation of this helix and trigger destabilization of its N-terminus. The N-terminal segment of the helix (K330, N331, F334, A338) unwinds and is destabilized upon receptor binding. Loosening its attractive interactions with GDP. More importantly, an exquisitely strong interaction of a5 with a1 is also weakened leading to its

structural destabilization as also indicated by the absence of crystallographic coordinates for the C-terminus of a1 in the crystal structure of the GPCR-Gαiβγ protein complex (PDB 3SN6 (Rasmussen, Devree et al. 2011)). The structural destabilization of a1 and GDP are propagated to the helical domain contributing to its release.

The second major contributing factor towards stabilization of the interface between Gαi-helical domain|Gαi-GTPase is a network of polar interactions between the b5-aG-a4 loop and the b4-a3 loop on the one side and residues in the aA helix, aD-aE loop, and aF-b2 loop on the other side. These interactions need to be broken for the helical domain to be released. We hypothesize that two mechanisms contribute to this event: firstly, unwinding of the N-terminal winding of a5 lengthens the b6-a5 loop. The loop adopts a different conformation and requires extra space making it push on aG. Secondly, we hypothesize that the interaction of R* IL 3 with a4-b6 loop is propagated to aG and possibly the b4-a3 loop as our models show a 20° rotation that shifts the C-terminus of aG accompanied with conformation changes in the respective loop regions.

Once released the helical domain samples a wide but well-defined space distinct form the location observed in the crystal structure of the GPCR-Gαiβγ protein complex (PDB 3SN6 (Rasmussen, Devree et al. 2011)) but preliminarily seems consistent with cryo-EM studies (Westfield, Rasmussen et al. 2011). The model shows also an attractive interaction between R* IL 2 with the aN-b1 loop. This signal could be propagated via the GTPase domain β-sheet to a5, a1, or the switch regions – a process difficult to track given the small amplitude of this interaction.

Conclusions

We have developed a hybrid model using β2AR-Gs structure and our published EPR data. We observe a 3.6Å shift and 68^{o} rotation of the α5 helix. We have performed energetic analysis of this rotation which demonstrated that the conformational change is

energetically feasible. Computational analysis of the hybrid model shows the rotation is accompanied by flexing of the β-sheet and is transmitted to the β6-α5 loop, the a1 and aG helices, and GDP. Disruption of interactions of these entities with the helical domain causes its separation. The hybrid model presented is consistent with published and new experimental data from a variety of sources. We have determined an ensemble of models that match the experimental data, resulting in a wider but well defined conformational space sampled by the helical domain compared to (Van Eps, Preininger et al. 2011). The helical domain is in a different orientation than in (Rasmussen, DeVree et al. 2011). Our model is similar to placements observed for the low resolution cryo-EM structure. This model integrates all published data and provides a detailed energetic pathway for signal transduction between activated receptor and Gi protein. It thereby creates a pathway to elucidate the structural and energetic determinants of signal transduction in Gai interaction with activated receptor R*.

Methods

Our strategy includes construction of a unified comparative model for the interaction of activated rhodopsin with Gi (R*-Gi) that is consistent with available experimental data. Next, we systematically compare interactions across key interfaces within this model to the free form of Gi. Thereby we identify “hot-spot” residues that contribute to stabilizing both states. We map how these key interactions are altered when Gi interacts with the activated rhodopsin.

Receptor unbound model of Gi

A comparative model of Gαiβγ was constructed based on the PDB coordinates 1GOT (Lambright, Sondek et al. 1996; Van Eps, Preininger et al. 2011). Missing residues were reconstructed using kinematic loop closure(Mandell, Coutsias et al. 2009). The model of the receptor unbound state was then subjected to 100 independent

relaxation trajectories that iterate between backbone perturbation, fast side chain optimization using a rotamer library (Bower, Cohen et al. 1997), and all atom gradient minimization in ROSETTA full-atom force field (Bradley, Misura et al. 2005). The ten models with lowest ROSETTA energy form the conformational ensemble representing Gαiβγ in the receptor unbound state. GDP was present throughout all steps of the protocol.

Receptor bound model of Gαiβγ consistent with experimental data

The crystal structure of the GPCR-Gαiβγ protein complex (PDB 3SN6 (Rasmussen, Devree et al. 2011)) is used as the template for constructing a comparative model for the rhodopsin bound state of Gαiβγ. The sequence of metarhodopsin, bovine β1 and γ1, and Gαi were threaded on the 3SN6 crystal structure. Receptor sequence was aligned using structure-structure alignment of 3SN6 with the structure of metarhodopsin 3PQR (Choe, Kim et al. 2011). A blast sequence alignment was used to align Gβγ. For the alpha subunit, the published sequence alignment between the Gαs and Gαi was used. For each chain, ROSETTA kinematic loop closure (Mandell, Coutsias et al. 2009) is used to construct missing coordinates. After loop construction, the model was relaxed in ROSETTA 46 times. To accommodate the receptor, the relaxation utilized ROSETTA's full atom membrane potential (Barth, Schonbrun et al. 2007; Barth, Wallner et al. 2009). The ten models with lowest ROSETTA energy were used as the starting point for the comparative model of the R*-Gi complex.

Exploring possible locations of the helical domain

The placement of the helical domain in PDB 3SN6 is inconsistent with EPR DEER experiments for the R*-Gi complex (Rasmussen, Devree et al. 2011; Van Eps, Preininger et al. 2011). These measurements display widened distance distributions between residues in the helical and GTPase domain consistent with the notion that the helical domain is flexible and explores a wide range of conformations. Crystallization

induces one placement for the helical domain that is not observed in the EPR DEER experiments. Therefore, we explored possible positions of the helical domain of Gαi upon receptor binding through rigid body docking (Gray, Moughon et al. 2003). The comparative model of the receptor bound state is used as the starting point for each of 743 docking trajectories that created non-clashing models. During the docking protocol the helical domain (residues 63 to 177) is separated from the rest of the nucleotide binding domain by removing linking residues 58-62 and 178-185. Both linker regions are reconstructed after docking and before each of these models was relaxed in the ROSETTA full atom energy membrane potential. This protocol resulted in a pool of 739 models of the receptor bound state with different positions of the helical domain.

CHAPTER VI

EPR RESTRAINT GUIDED MEMBRANE PROTEIN STRUCTURE PREDICTION WITH BCL FOLD

This work is based on the manuscript in preparation of the same title.

Summary

For many membrane proteins the determination of topology remains a challenge for methods like X-ray crystallography and NMR spectroscopy. Electron paramagnetic resonance (EPR) spectroscopy has evolved as an alternative technique to study structure and dynamics of membrane proteins, typically after the overall topology has been elucidated. *De novo* determination of a membrane protein topology from EPR spectroscopic data is hindered by sparseness of structural restraints. Every restraint requires the production of a dedicated sample with spin labels introduced at specific sites. Further, the two spin labels project from the protein backbone into an unknown spatial position, making it difficult to relate the measured distance to distances on the protein backbone that constrain protein topology. The present study demonstrates membrane protein topology determination using EPR distance and accessibility measurements for a small fraction of amino acids. The algorithm assembles secondary structure elements (SSEs) in the membrane with a Monte Carlo Metropolis (MCM) algorithm quickly enumerating possible topologies. Likely models are selected based on agreement with EPR restrains and on agreement with a knowledge-based energy function. Twenty membrane proteins of up to 312 residues and three symmetric homo-multimer proteins with up to 595 residues are used to test the algorithm. The RMSD100 value of the most accurate model is better than 8 Å for all 23 proteins, better than 6 Å for

18 of the 23 cases, and better than 4 Å for 9 of the 23 cases demonstrating the algorithms ability to sample the native topology when distance and accessibility restraints are used. For 19 out of 23 proteins models with the correct topology can be selected by agreement with the EPR distance and accessibility restraints.

Introduction

Membrane protein structure determination continues to be a challenge. An estimated 60% of pharmaceutical therapies target membrane proteins, and about 22% of proteins are membrane proteins (Overington, Al-Lazikani et al. 2006). However, only 2% of proteins deposited in the Protein Data Base are classified as membrane proteins (Berman, Westbrook et al. 2000; Tusnady, Dosztanyi et al. 2004). Protein structures are typically determined to atomic resolution using X-ray crystallography or nuclear magnetic resonance (NMR). However membrane proteins provide challenges for both techniques (Bill, Henderson et al. 2011). It is difficult to obtain quantities of purified membrane protein sufficient for both, X-ray crystallography and NMR. The two-dimensional nature of the membrane complicates crystallization in a three-dimensional crystal lattice. In order to obtain crystals, the target protein is often subjected to non-native-like environments and/or modifications such as stabilizing sequence mutations (Tate and Schertler 2009; Tate 2010). Many membrane proteins continue to be too large for structure determination by NMR spectroscopy (Kang and Li 2011). Even if the target itself is not too large, the membrane mimic adds significant additional mass to the system (Kim, Howell et al. 2009). Despite wonderful successes in determining the structure of high-profile targets, membrane protein structure determination will continue to lack behind and not be a routine process for years to come.

Electron paramagnetic resonance (EPR) spectroscopy in conjunction with site-directed-spin-labeling provides a powerful alternative technique for probing structural

aspects of membrane proteins (Hubbell and Altenbach 1994; Dong, Yang et al. 2005; Czogalla, Pieciul et al. 2007). Advantages of EPR spectroscopy include that the protein can be studied in a native-like environment. Also, EPR measurements need a relatively small sample amount, and EPR can be used to study large targets. Although EPR is a versatile tool for probing membrane protein structure it has its own challenges: at least one unpaired electron (spin label) needs to be introduced into the protein. Typically, this requires mutation of all cysteine residues to either alanine or serine, introduction of one or two cysteines at the desired labeling sites, coupling to the thiol-specific nitroxide spin label MTSSL, and functional characterization of the protein. As a result, datasets from EPR spectroscopy are sparse containing only a fraction of measurements per residue in the target protein. EPR is not a high-throughput technique.

EPR provides two categories of structural information important to membrane protein topology: First, EPR can provide information about the local environment of the spin label (Altenbach, Marti et al. 1990; Koteiche, Berengian et al. 1998; Koteiche and McHaourab 1999; Zou and McHaourab 2009). The accessibility of the spin label to either oxygen or NiEDDA probe molecules indicates the degree of burial of the spin label within the protein. Importantly, because the distinct partition of oxygen or NiEDDA between membrane and soluble phase the ratio between oxygen or NiEDDA accessibility is indicative of depth in the membrane. Accessibility measurements are typically done in a sequence scanning fashion. This provides an accessibility profile over a large portion of the sequence (Altenbach, Yang et al. 1996; Lietzow and Hubbell 2004). The accessibility profile tracks the periodicity of secondary structure elements (SSEs) as individual measurements rise and fall according the periodic exposure and burial of residues. The exposed face of a SSE can be determined (Salwinski and Hubbell 1999), a task that is difficult within the hydrophobic environment of the membrane.

Secondly, when two spin labels are introduced EPR can measure inter-spin label distances, routinely of up to 60 Å through the DEER experiment (Borbat, McHaourab et al. 2002; Jeschke and Polyhach 2007). EPR distance measurements have been demonstrated on several large membrane proteins including MsbA (Zou, Bortolus et al. 2009), rhodopsin (Altenbach, Kusnetzow et al. 2008), and LeuT (Claxton, Quick et al.). Given the sparseness of data, EPR has been frequently used to probe different structural states of proteins (Chakrapani, Sompornpisut et al. ; Vásquez, Sotomayor et al. 2008). Changes in distances and accessibilities track regions of the protein that move when converting from one state into another. Such investigations rely upon an already determined experimental structure to define the protein topology and provide a scaffold to map changes observed via EPR spectroscopy.

One critical limitation for *de novo* protein structure determination from EPR spectroscopic data is that measurements are made on the spin label while information of the placement of backbone atoms is needed to define the protein fold. For distance measurements, this introduces an uncertainty in relating the distance measured between the two spin labels to a distance between points in the backbone of the protein. This uncertainty, defined as the difference between the distance between the spin labels and the distance between the corresponding C_β atoms is up to 12 Å (Hirst, Alexander et al. ; Alexander, Bortolus et al. 2008).

To address this limitation we introduced a cone model which provides a knowledge-based probability distribution for the C_β atom distance given an EPR-measured spin label distance. We demonstrated that 25 distance restraints in conjunction with the cone model and the structure prediction algorithm Rosetta are sufficient to determine the topology of soluble proteins such as T4-lysozyme and αA-crystallin. Using just 25 EPR measured distances for T4-lysozyme, Rosetta was able to provide models matching the experimentally determined structure to atomic detail including backbone and side-chain

placement. It was further demonstrated that selection of restrain by information content reduces the number of required restraints to one-third. We developed an algorithm for selection of optimal restraints. (Hirst, Alexander et al. ; Alexander, Bortolus et al. 2008; Yang, Ramelot et al. 2010). These studies demonstrate that *de novo* prediction methods can supplement EPR data sufficiently to allow structure elucidation of a target.

De novo membrane protein structure prediction was demonstrated with Rosetta using twelve proteins with multiple transmembrane spanning helices (Yarov-Yarovoy, Schonbrun et al. 2006). The method was generally successful for the membrane topology for small proteins. The results of the study suggest that sampling of large membrane topologies requires methods that directly sample structural contacts between sequence distance regions of the protein.

For this purpose we developed an algorithm that assembles protein topologies from SSEs termed BCL::Fold (Karakas, Woetzel et al. 2012). The omission of loop regions in the initial protein folding simulation allows sampling of structural contacts between sequence distance regions and thereby rapidly enumerates all likely protein topologies. A knowledge-based potential guides the algorithm towards physically realistic topologies. The algorithm is particularly applicable for the determination of membrane protein topologies as trans-membrane spans are dominated by regularly ordered SSEs (in preparation). Loop regions and amino acid side chains can be added in later stages of modeling the structure. The algorithm was tested in conjunction with medium resolution density maps achieving models accurate at atomic detail in favorable cases.

The present study combines EPR distance and accessibility restraints with the BCL::Fold SSE assembly methodology for the prediction of membrane protein topology. In a first step we introduce scores specific to EPR distances and accessibilities and demonstrate their ability to enrich for accurate models. In a second test we assemble 20 monomeric and three multimeric membrane proteins guided by the EPR restraints.

Results

Compilation of benchmark set

Twenty-three membrane proteins of known structure were used to demonstrate the ability of EPR specific scores to improve sampling during protein structure prediction (Table 28). Twenty of the proteins are monomers ranging in size from 100 to 300 residues. One protein (2L35) has two chains, although the second chain is a single transmembrane span. Three of the proteins are symmetric multimeric proteins of up seven subunits containing up to a total of 595 residues. 2000 independent structure prediction trajectories are conducted for each protein without restraints, with distance restraints, and with distance and accessibility restraints.

Table 28 Membrane proteins and residues used for benchmarking.
. Proteins 1BL8, 2IUB, and 2OAU are symmetric homo-multimers with 4, 5, and 7 subunits, respectively. Each subunit is denoted by a separate chain. 2L35 has two chains, with chain B being a single transmembrane span.

PDB ID	TM Segments	Chain	Residues	Number of Residues	reference
1PY6	7	A	5-231	227	(Faham, Yang et al. 2004)
1PY6*	4	A	77-199	123	(Faham, Yang et al. 2004)
1OCC	5	C	71-261	191	(Tsukihara, Aoyama et al. 1996)
1PV6	6	A	1-190	190	(Abramson, Smirnova et al. 2003)
1J4N	3	A	4-119	116	(Sui, Han et al. 2001)
2BS2	5	C	21-237	217	(Madej, Nasiri et al. 2006)
2BL2	4	A	12-156	145	(Murata, Yamato et al. 2005)
2BG9	3	A	211-301	91	(Unwin 2005)
1IWG	5	A	330-497	168	(Murakami, Nakashima et al. 2002)
1RHZ	5	A	23-188	166	(Berg, Clemons et al. 2004)
1KPL	7	A	31-233	203	(Dutzler, Campbell et al. 2002)
1U19	7	A	33-310	278	(Okada, Sugihara et al. 2004)
2KSF	4	A	396-502	107	(Maslennikov, Klammt et al. 2010)
2L35	3	A;B	1-63; 1-32	95	(Call, Wucherpfennig et al. 2010)
2KSY	7	A	1-223	223	(Gautier, Mott et al. 2010)
3KCU	7	A	29-280	252	(Wang, Huang et al. 2009)
2IC8	6	A	91-272	182	(Wang, Zhang et al. 2006)
3P5N	6	A	10-188	189	(Zhang, Wang et al. 2010)
3KJ6	7	A	35-346	312	(Bokoch, Zou et al. 2010)
2K73	4	A	1-164	164	(Zhou, Cierpicki et al. 2008)
1BL8	8	A; B; C; D	25-114	360	(Doyle, Cabral et al. 1998)
2IUB	10	A; B; C; D; E	294-310 + 327-345	180	(Eshaghi, Niegowski et al. 2006)
2OAU	21	A; B; C; D; E; F; G	27-111	595	(Steinbacher, Bass et al. 2007)

Simulation of missing EPR restraints

EPR distance and accessibility restraints were simulated where needed to obtain datasets for each of the 23 proteins. Distance restraints were simulated for all proteins, and accessibility restraints were simulated for all proteins except for the multimeric proteins where published accessibility data were used for the multimer proteins (Vásquez, Sotomayor et al. 2008) (Dalmas, Cuello et al. ; Perozo, Cortes et al. 1998). Accessibility restraints were simulated by calculating the neighbor vector value (Durham, Dorr et al. 2009) for residues within SSEs of each protein. This value was considered to be an oxygen accessibility measurement.

Distance restraints were simulated using a restraint selection algorithm (Kazmier, Alexander et al.) which distributes measurements across all SSE. It also favors measurements between residues that are far apart in sequence. One restraint was generated for every 0.2 residues within predicted SSEs. Distances are calculated between the Cβ atoms; for glycine the HA2 atom is used. To simulate a likely distance observed in an actual EPR experiment, the distance is adjusted by an amount selected randomly from the probability distribution of observing a given D_{SL}-$D_{C\beta}$ value (Hirst, Alexander et al.). In order to reduce the possibility of bias arising from restraint selection, ten independent restraint sets were generated. For the three multimer proteins, the protocol was the same except only 0.1 restraints per residue within predicted SSEs were selected.

Translating EPR accessibilities into structural restraints

EPR accessibility measurements are typically made in a sequence scanning fashion over a large portion of the target protein. While each individual accessibility measurement can be affiliated with an elevated error, the overall pattern of accessibilities tracks reliably exposure of the SSE to solvent or membrane. Therefore, the approach for developing an EPR accessibility score takes advantage of this. The

exposure moment of a window of amino acids is defined as $E_w = \sum_{n=1}^{N} e_n s_n$, where N is the number of residues in the window, e_n is the exposure value of residue *n*, and s_n is the normalized vector from the C_α atom to the C_β atom of residue *n*. This equation was inspired by the hydrophobic moment as previously defined (Eisenberg, Weiss et al. 1984). The exposure moment calculated from solvent accessible surface area SASA has been previously shown to approximate the moment calculated from EPR accessibility measurements (Salwinski and Hubbell 1999).

During *de novo* protein structure prediction the protein is represented only by its backbone atoms hampering calculation of SASA. Further, calculation of SASA from an atomic-detail model would be computationally prohibitive for a rapid scoring function in *de novo* protein structure prediction. Therefore, the neighbor vector approximation for SASA is used (Durham, Dorr et al. 2009). The exposure moment is calculated for overlapping windows of length seven for α-helices and four for β-strands. The score is computed as E_{orient} = -½ cos (θ) where θ is the torsion angle between the exposure moments. This procedure assigns a score of -1 is given if θ = 0° and a score of 0 if θ = 180°.

It has previously been demonstrated that the burial of sequence segments relative to other segments can be determined from the average accessibility values measured for that stretch of sequence (Chakrapani, Cuello et al. 2008). To capture this information, the magnitude of the exposure moment for overlapping residue windows is determined from the model structure and from the measured accessibility. The Pearson correlation is then calculated between the rank order magnitudes of the structural versus experimental moments. This gives a value between -1 which indicates the structural and exposure magnitudes are oppositely ordered, to 1, which means the structural and exposure magnitudes are ordered equivalently. The score E_{magn} is obtained by negating

the resulting Pearson correlation value so that matching ordering will get a negative score and be considered favorable.

Translating EPR distances into structural restraints

The knowledge-based score for EPR distances previously reported is used score agreement of models with distance restraints (Hirst, Alexander et al.), D_{EPR}. This score spans a range of D_{SL}-$D_{C\beta}$ between ±12 Å. D_{SL} is the EPR measured distance between two spin labels; $D_{C\beta}$ is the distance between the corresponding Cβ on the residues of interest; D_{SL}-$D_{C\beta}$ is the difference between these two distances.

In addition we found it beneficial to add an attractive potential on either side of D_{EPR} to provide an incentive for the MCM minimization to bring structures within the range of D_{EPR}. These attractive potentials use a cosine function to transition between a most unfavorable score of 0 and a most favorable score of -1. The potentials stretch from D_{SL}-$D_{C\beta}$ values of +30 Å and -30 Å to the first values of D_{SL}-$D_{C\beta}$ where D_{EPR} can provide scoring information.

Summary of folding protocol

The protein structure prediction protocol is based on the protocol of BCL::Fold for soluble proteins (Karakas, Woetzel et al. 2012). The method assembles SSEs in space, drawing from a pool of predicted SSEs. A Monte Carlo energy minimization with the Metropolis criteria is used to search for models with favorable energies. Models are scored after each Monte Carlo step using knowledge-based potentials describing optimal SSE packing, radius of gyration, amino acid exposure, and amino acid pairing, loop closure geometry, secondary structure length and content, and penalties for clashes (Woetzel, Karakaş et al. 2012).

The algorithm was adapted for membrane protein folding. An additional score is used which favors orthogonal placement of SSEs relative to the membrane (SSE_{align}). All moves introduced for soluble proteins are used (Karakas, Woetzel et al. 2012). In

addition we include perturbations that optimize the placement of the protein in the membrane such as translation of individual SSEs orthogonally to the membrane plane as well as rigid body translation and rotation of the entire protein.

The assembly of the protein structure is broken down into five stages of sampling with large structural perturbation moves that can alter the topology of the protein. Each of the five stages lasts for a maximum of 1000 Monte Carlo steps. If an energetically improved structure has not been generated within the previous 400 Monte Carlo steps, the minimization for that stage will cease. Over the course of the five assembly stages, the weight of clashing penalties in the total score is ramped as 0, 125, 250, 375, 500. The weight of the SSE_{align} score is 8.

Following the five stages of protein assembly, a structural refinement stage takes place. This stage lasts for a maximum of 2000 Monte Carlo steps and will terminate sooner if an energetically improved model is not sampled within the previous 400 steps. The refinement stage consists of small structural perturbations which will not drastically alter the topology of the protein model. After the refinement stage, residues missing from the model are added in a loop building protocol to produce a complete backbone model of the protein. The protocol is based on cyclic coordinate descent (Canutescu and Dunbrack 2003).

Summary of benchmark setup

To test the influence of EPR restraints, each protein was folded in the absence of restraints, with just distance restraints, and with distance and accessibility restraints. To test the influence of secondary structure prediction accuracy (see Methods section), the experiment will be repeated with optimal secondary structure elements derived from the experimental structure. 1000 models will be created for each of the benchmark proteins in independent MCM folding trajectories. EPR specific scores are used during the five assembly and one refinement stages of structure prediction. The EPR distance scores

have a weight of 500 during the first assembly stage. The weight of the accessibility score is 5.0 during all assembly and refinement stages using either pool.

After 1000 models have been generated for each protein, the models are filtered according to EPR distance score. The top 10% of models resulting from the structure prediction protocol for each of the SSE pools are selected for a second round of minimization starting from these models. Structure prediction trajectories not using EPR distance restraints do not undergo the second round of minimization.

EPR specific scores select for accurate models of membrane proteins

The ability of EPR specific scores D_{EPR}, E_{magn} and E_{orien} to select for accurate models is tested by calculating enrichment values for structure prediction trials of twenty three membrane proteins (Table 29). The enrichment of a score indicates how well the score identifies a protein model that is truly accurate as being accurate and is given by the fraction of correctly identified accurate models over the fraction of accurate to inaccurate models (see Methods for details) (Woetzel, Karakaş et al. 2012). Accurate is defined as models with an RMSD100 (Carugo and Pongor 2001) less than 8.0 Å. Datasets for each protein are chosen to contain 10% accurate models. Therefore, if a score correctly identifies all accurate models as being accurate, a perfect enrichment would give a value of 10.0.

Enrichments are computed for the protein models created without experimental restraints. The enrichment for D_{EPR} is greater than 2.0 for all proteins with models better than 8.0 Å RMSD100 (Table 29). Across the twenty three proteins, D_{EPR} achieves an average enrichment of 4.02. The average enrichment of E_{orient} and E_{magn} is 0.47 and 0.87, respectively (Table 29). These numbers indicate that both scores do not select for native-like models. We attribute this performance to the accuracy measure used: correct topology. At a cutoff of RMSD100 8.0 Å a model will have the correct overall topology. While D_{EPR} provides information on the overall correct placement of SSEs relative to one

another, E_{orient} and E_{magn} can be considered more local scores important for refining orientation but not placement of SSEs. Using a more stringent cutoff of RMSD100 of 5.0 Å, E_{orient} and E_{magn} provide enrichments greater than 1.0 for the majority of proteins where models were produced at an accuracy of 5.0 Å RMSD100 (Table 30). Therefore both scores are expected to in conjunction with D_{EPR} improve sampling of protein models at low RMSD100 values.

Table 29 Using a cutoff of 8.0 Å RMSD100, the average of the enrichment of three EPR specific scores across 23 membrane proteins.
Proteins below the thick line are multimeric proteins. Enrichments are cross validated ten times per protein and averaged. Random subsets of models for a protein are used to achieve datasets where 10% of the models have an RMSD100 value to the experimental structure of less than 8.0 Å. The enrichment calculations are done for each of the ten restraint datasets. Values shown are the average cross validation value across the ten restraint datasets. DEPR is the enrichment achieved by the knowledge-based EPR distance potential (Hirst, Alexander et al.). Eorient is the enrichment for the accessibility score using the exposure moment orientation. Emagn is the enrichment for the accessibility score taking into account the magnitude of the exposure moment.

PDBID	# models below 8.0 Å RMSD100	D_{EPR}	E_{orient}	E_{magn}
1IWG	155	4.83	0.96	1.76
1J4N	88	5.25	1.19	0.68
1KPL	1	4.50	0.00	0.00
1OCC	25	5.37	0.00	0.04
1PV6	3	5.50	0.00	0.00
1PY6	12	5.82	0.00	0.25
1PY6*	438	2.53	0.70	0.92
1RHZ	5	2.28	0.00	0.00
1U19	4	5.65	0.00	0.00
2BG9	550	3.23	1.27	0.62
2BL2	313	5.95	1.59	2.11
2BS2	2	5.25	0.00	0.00
2IC8	4	6.43	0.00	1.00
2K73	24	4.03	0.42	0.54
2KSF	376	2.67	0.64	0.72
2KSY	5	6.22	0.00	1.80
2L35	233	2.29	0.26	0.10
3KCU	0	0.00	0.00	0.00
3KJ6	0	0.00	0.00	0.00
3P5N	3	4.20	2.33	3.33
1BL8	48	2.31	0.60	0.71
2IUB	847	4.30	0.63	3.77
2OAU	234	3.83	0.32	1.62

Table 30 Using a cutoff of 5.0 Å RMSD100, the average of the enrichment of three EPR specific scores across 23 membrane proteins.
Proteins below the thick line are multimeric proteins. Enrichments are cross validated ten times per protein and averaged. Random subsets of models for a protein are used to achieve datasets where 10% of the models have an RMSD100 value to the experimental structure of less than 8.0 Å. The enrichment calculations are done for each of the ten restraint datasets. Values shown are the average cross validation value across the ten restraint datasets. DEPR is the enrichment achieved by the knowledge-based EPR distance potential (Hirst, Alexander et al.). Eorient is the enrichment for the accessibility score using the exposure moment orientation. Emagn is the enrichment for the accessibility score taking into account the magnitude of the exposure moment.

	# models below 5.0Å RMSD100	D_{EPR}	σ	Upper Penalty	σ	Lower Penalty	σ	E_{orient}	σ	E_{magn}	σ
1IWG	1	9.20	1.96	1.40	2.38	8.70	2.59	9.00	3.00	4.00	4.90
1J4N	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1KPL	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1OCC	31	7.10	4.35	0.80	2.00	7.90	3.55	0.00	0.00	2.00	4.00
1PV6	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1PY6	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1PY6*	31	3.75	1.95	1.46	0.31	4.57	2.36	1.29	0.00	1.29	0.00
1RHZ	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1U19	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2BG9	232	3.46	0.11	1.27	0.01	3.28	0.06	1.62	0.04	1.04	0.09
2BL2	19	6.63	3.05	1.90	0.85	7.16	3.13	2.16	0.44	3.26	1.12
2BS2	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2IC8	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2K73	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2KSF	32	5.50	3.07	2.04	1.58	5.45	2.64	0.00	0.00	0.34	0.29
2KSY	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2L35	74	4.50	2.16	1.54	0.29	4.01	2.31	0.38	0.18	0.18	0.12
3KCU	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3KJ6	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3P5N	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1BL8	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2IUB	389	5.09	0.38	1.86	0.02	1.94	0.27	0.01	0.02	6.69	0.31
2OAU	6	7.05	1.37	6.18	1.84	4.10	1.44	1.50	0.50	5.00	1.05

Using EPR specific scores during membrane protein structure prediction improves sampling accuracy

Folding trajectories using distance restraints sample the correct topology more frequently. This is shown by the shifts to lower RMSD100 values in distributions showing the frequency with which a given RMSD100 value is sampled for a given protein (Figure 43). Additionally, by using accessibility restraints in combination with distance restraints, the sampling of models at the lowest RMSD100 values is further increased.

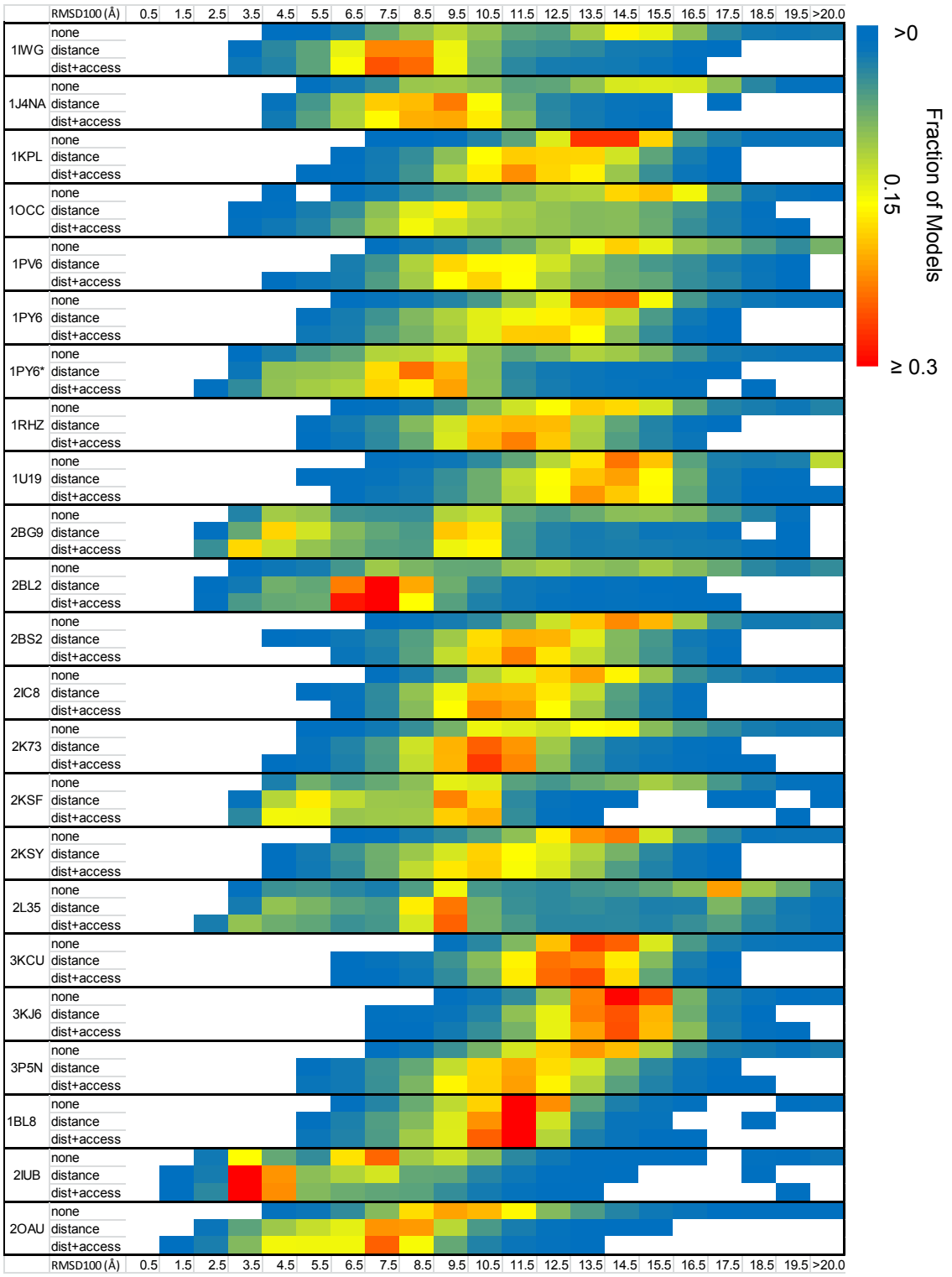


Figure 43 Accuracies of protein models created without restraints (none), with distance restraints (distance), and with distance and accessibility restraints (dist+access).
2000 models were generated for each protein for each restraint scenario and the accuracy of each model relative to the native structure was determined. Accuracy was calculated as the RMSD100 value between the model and native structure. The frequency with which a given RMSD100 value is achieved for each protein is denoted by the heat map colors. The frequency is given by the fraction of models out of 2000.

By using EPR specific scores, not only is the frequency increased with which higher accuracy models are created, but best models achieve an accuracy not sampled in the absence of EPR data (Table 31, Figure 43). For each protein, the five models sampled with the best RMSD100 values are used to determine ability to sample accurate models by taking their RMSD100 value average, μ_5. Using the best five models by RMSD100 provides a more consistent measure of sampling ability compared to looking at the single best because of the random nature of the structure prediction protocol. Across all proteins, μ_5 is, on average, 6.43 Å when EPR specific scores are not used. When adding scores for distance and then both distance and accessibility, the average μ_5 value drops to 5.05 Å and 4.84 Å, respectively. The improvement of μ_5 can be compared on a per-protein basis also (Figure 44). When EPR distance restraints are used during structure prediction, μ_5 is improved by an average of 1.38 Å per protein compared to not using any EPR data. When both distance and accessibility scores are used, μ_5 is improved by an average of 1.59 Å per protein compared to not using any EPR data.

The multimeric proteins are have an average μ_5 of 3.34 Å. Although the multimer proteins are the only ones where real EPR accessibility data was used, the behavior of their improvement when using accessibility scores during structure prediction is similar to the proteins using simulated accessibility data. This is demonstrated comparing μ_5 on a per-protein basis between using distance restraints and using distance plus accessibility restraints during structure prediction. The average improvement of μ_5 for the multimer proteins (which use real accessibility data) is 0.30 Å (standard deviation 0.15), whereas the average improvement of μ_5 for the other proteins (which use simulated accessibility data) is 0.20 Å (standard deviation 0.47).

Table 31 Ability of EPR specific scores to improve sampling.
Results are shown of protein structure prediction benchmarks for twenty three membrane proteins under three different conditions: a) without the use of EPR data (None) b) with the use of EPR distance data (Distance) c) using both EPR accessibility and distance data (Distance+Accessibility). For each scenario and protein, the five most accurate models according to RMSD100 to the native structure are determined. These five RMSD100 values are averaged (μ_5) and the standard deviation calculated (σ_5). The RMSD100 value of the most accurate model sampled is also provided (best).

	None			Distance			Distance + Accessibility		
PDB	μ_5	σ_5	best	μ_5	σ_5	best	μ_5	σ_5	best
1IWG	5.46	0.71	4.25	3.79	0.23	3.38	3.44	0.09	3.36
1J4N	5.97	0.16	5.78	4.77	0.06	4.65	4.49	0.08	4.39
1KPL	8.95	0.72	7.65	6.93	0.17	6.71	6.47	0.43	5.85
1OCC	6.43	0.87	4.71	4.45	0.46	3.62	3.69	0.26	3.28
1PV6	7.87	0.21	7.56	6.16	0.11	6.01	5.17	0.21	4.92
1PY6	6.91	0.34	6.57	5.67	0.24	5.22	5.34	0.09	5.22
1PY6*	4.15	0.22	3.85	3.66	0.09	3.51	2.94	0.19	2.62
1RHZ	7.57	0.52	6.64	5.84	0.31	5.25	6.17	0.23	5.81
1U19	7.49	0.45	7.00	6.36	0.24	5.98	6.54	0.30	6.11
2BG9	3.32	0.13	3.20	2.99	0.08	2.91	2.41	0.09	2.27
2BL2	3.91	0.30	3.38	3.45	0.25	2.97	2.95	0.11	2.82
2BS2	8.28	0.28	7.97	5.55	0.50	4.61	6.51	0.17	6.19
2IC8	7.71	0.28	7.29	6.33	0.24	5.87	6.65	0.27	6.11
2K73	6.65	0.46	5.74	5.76	0.26	5.24	5.36	0.45	4.66
2KSF	4.43	0.09	4.29	3.83	0.08	3.71	3.45	0.03	3.42
2KSY	7.39	0.46	6.55	4.98	0.36	4.47	5.01	0.21	4.68
2L35	3.83	0.18	3.57	3.36	0.15	3.13	2.79	0.02	2.75
3KCU	9.47	0.18	9.29	7.43	0.41	6.67	7.82	0.47	6.99
3KJ6	9.83	0.33	9.20	8.12	0.58	7.12	8.44	0.63	7.56
3P5N	8.06	0.39	7.63	5.82	0.27	5.35	5.60	0.20	5.34
1BL8	6.73	0.22	6.48	5.77	0.19	5.56	5.38	0.14	5.21
2IUB	2.64	0.10	2.53	2.54	0.46	1.62	2.45	0.41	1.63
2OAU	4.73	0.36	4.01	2.60	0.14	2.45	2.19	0.16	2.00

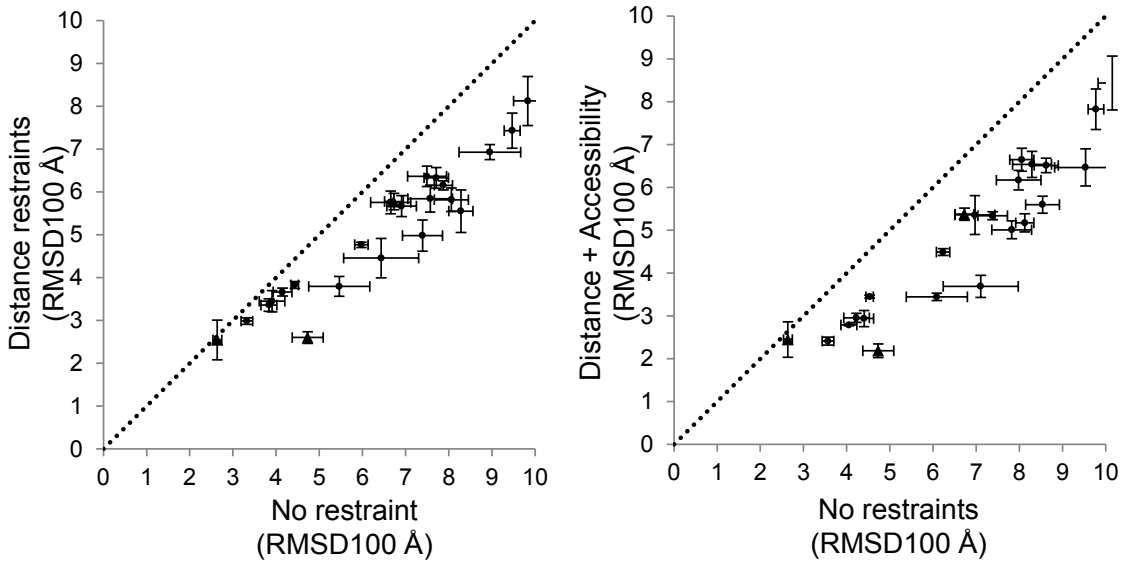


Figure 44 Sampling of most accurate models when not using EPR data compared to using left) EPR distance data and right) EPR distance and accessibility data.
Points are the average of the most accurate five models according to RMSD100. Error bars show the standard deviation of RMSD100 within these five models. The dotted diagonal line is provided as a guide to the eye and indicates what would be no change.

With the exception of 3KCU and 3KJ6, all proteins achieve a μ_5 value of under 7.0 Å RMSD100. This indicates the placement of the transmembrane spanning regions follow the native structure (Figure 45). The 278 residues of the GPCR rhodopsin (1U19) with 7 transmembrane spanning helices are predicted to a μ_5 value of 6.54 Å. The other proteins with over 200 residues and seven transmembrane spanning segments are predicted to similar or better accuracies, with the 227 residues of 1PY6 being predicted to a μ_5 value of 3.44 Å. The multimer protesin 2IUB and 2OAU are predicted to μ_5 accuracies of 2.45 Å and 2.19 Å, respectively. 2IUB has ten transmembrane segments in five subunits for a total of 180 residues. 2OAU has twenty one transmembrane segments in seven subunits for a total of 595 residues.

Although the multimer protein 1BL8 has a μ_5 value of 5.38 Å, this is more than double the μ_5 value for the two other multimer proteins. Inspection of 3KCU, 3KJ6, and 1BL8 indicates that inaccuracy in the SSE pool leads to less accurate predictions for these proteins than other proteins. Large strands are predicted for these proteins that are not observed in the native structure (Figure 45).

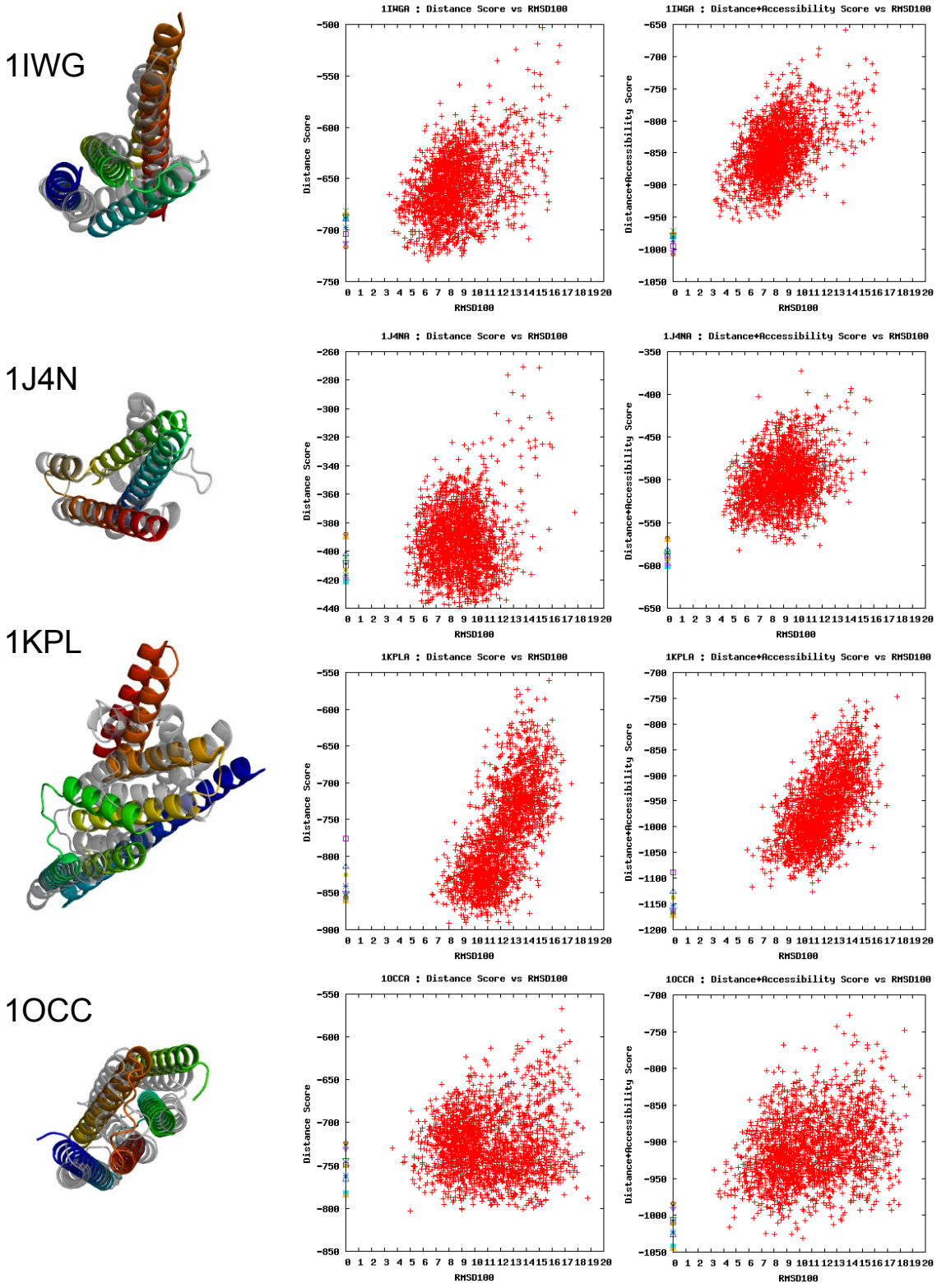


Figure 45 Structure predication results for twenty three membrane proteins. Shown is the best model sampled according to RMSD100 value. In grey is the native structure and colored in rainbow according to sequence is the predicted model. Also shown is a plot comparing the EPR distance score to the corresponding RMSD100 value for each of the 2000 models created for every protein (middle plot) for prediction trajectories using EPR distance

scores. An analogous plot is shown using the sum of the EPR distance and accessibility scores for prediction trajectories done using these scores (right plot). For the monomer proteins, the native structure is scored using EPR distance restraints or distance and accessibility restraints for each of the ten restraint datasets and shown in the middle and right plots (multiple colored points at RMDS100 value of 0.0).

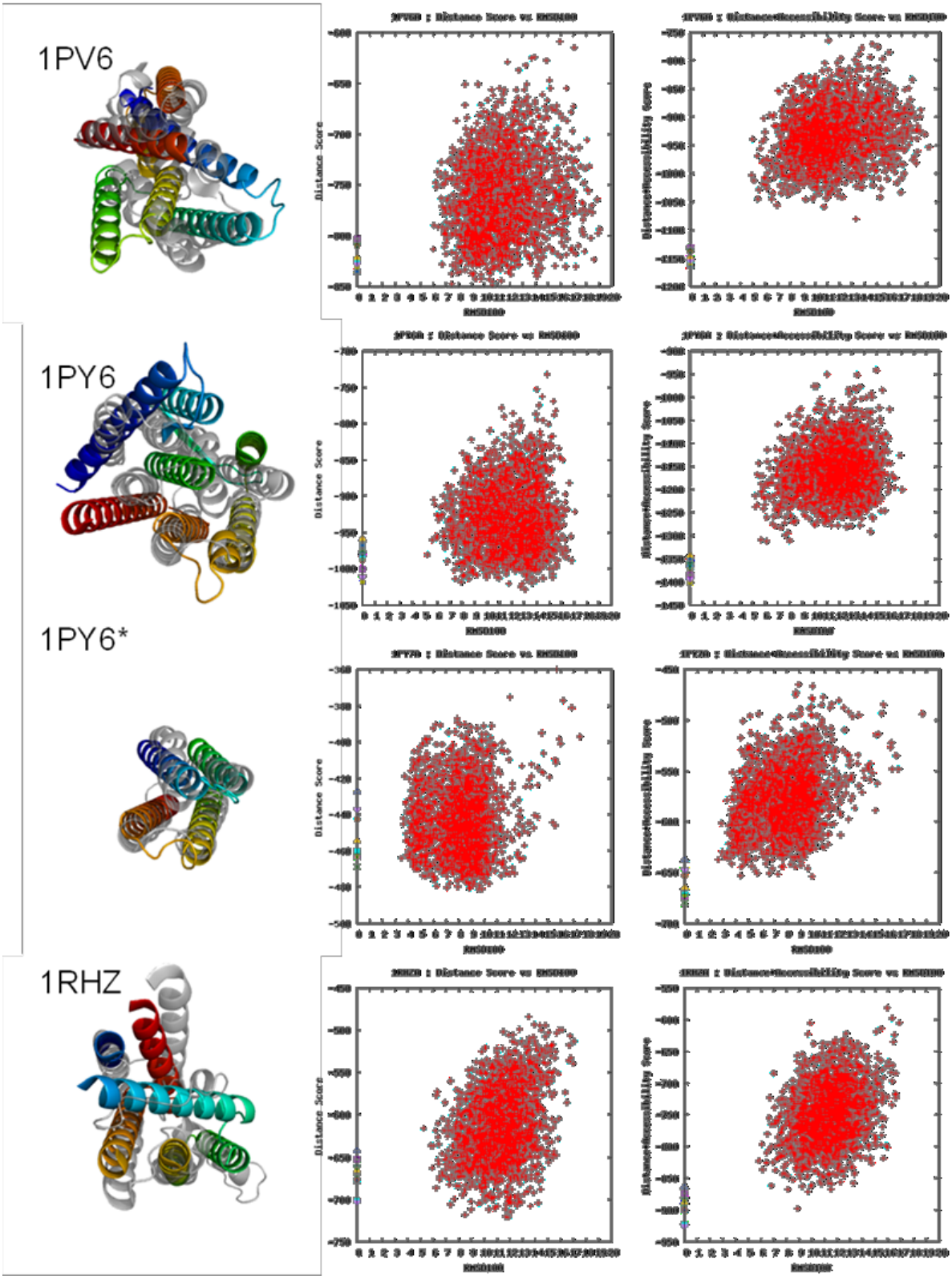
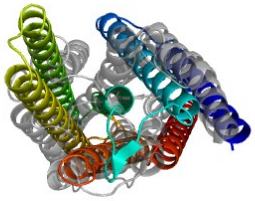
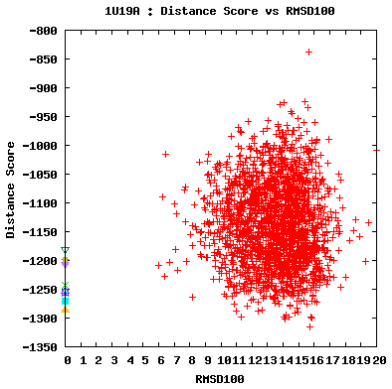
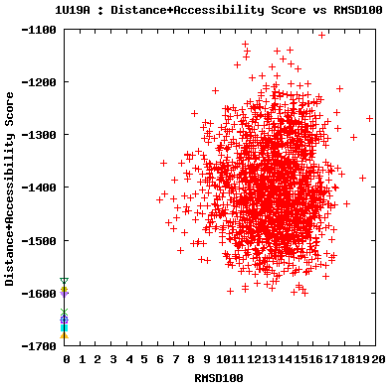


Figure 45 continued.

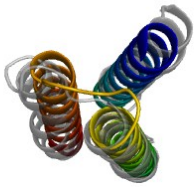
1U19

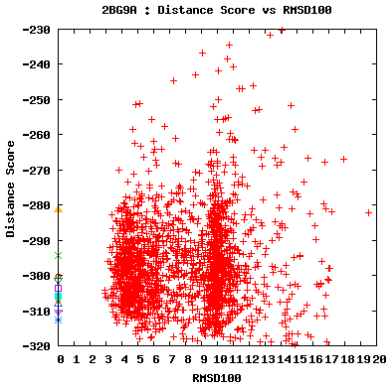


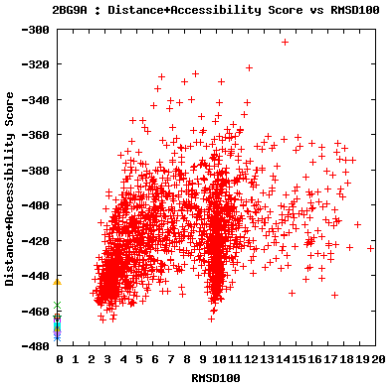




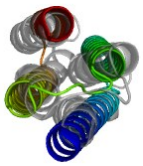
2BG9

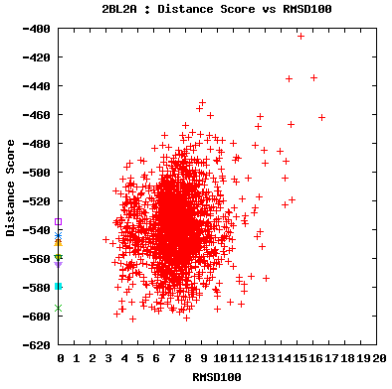


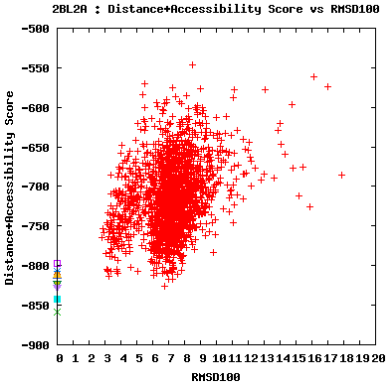




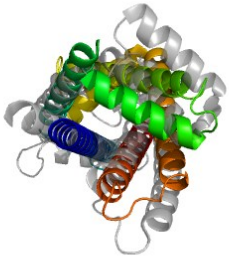
2BL2

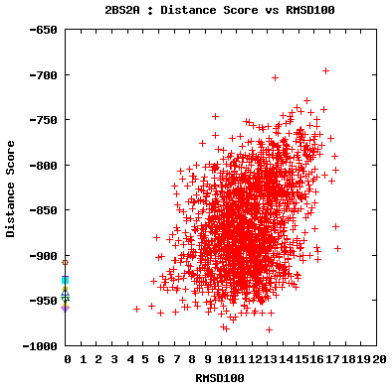






2BS2





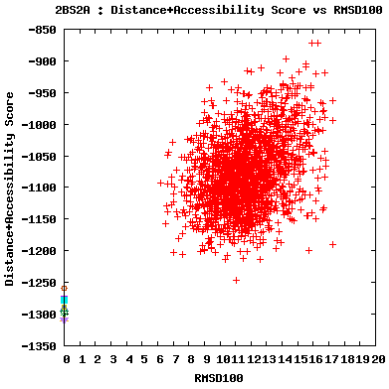
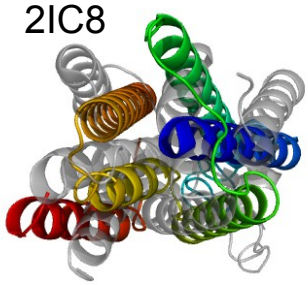
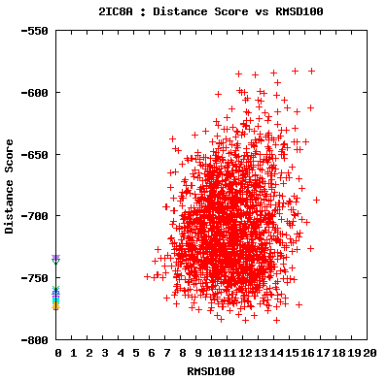
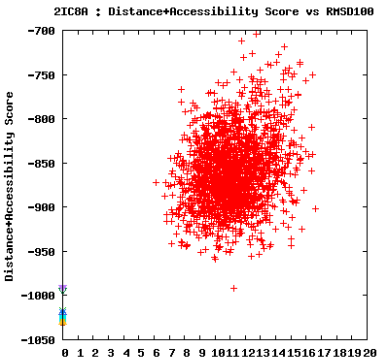


Figure 45 continued

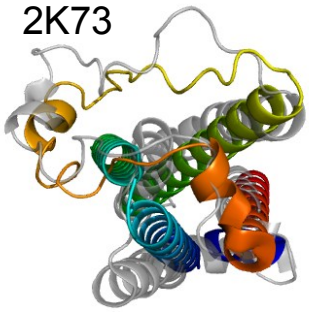
2IC8

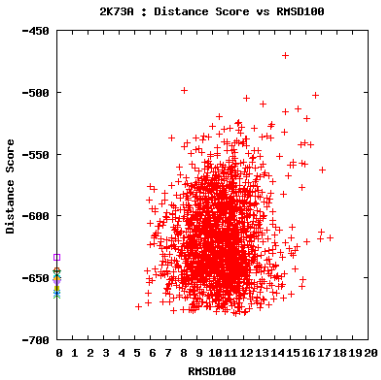


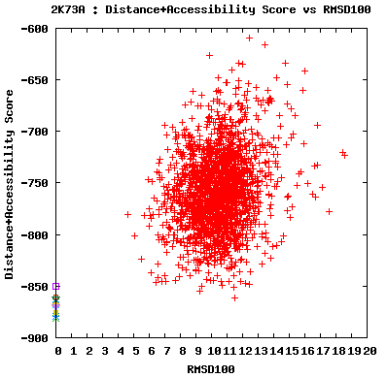




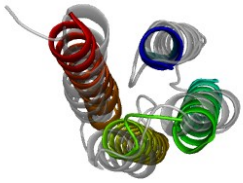
2K73

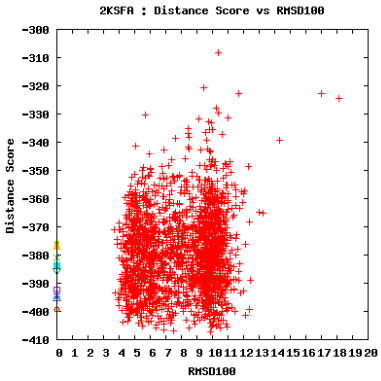


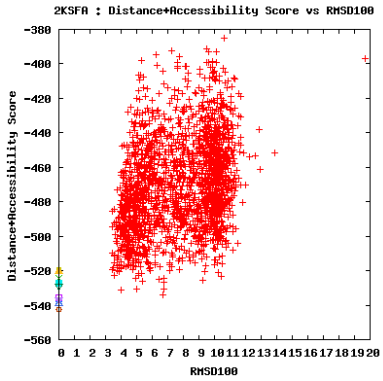




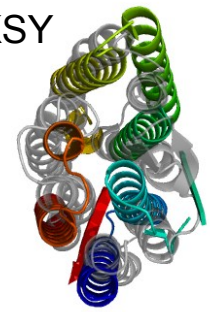
2KSF

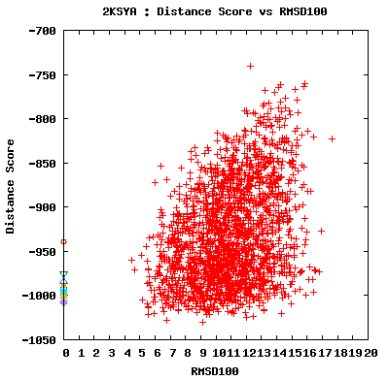






2KSY





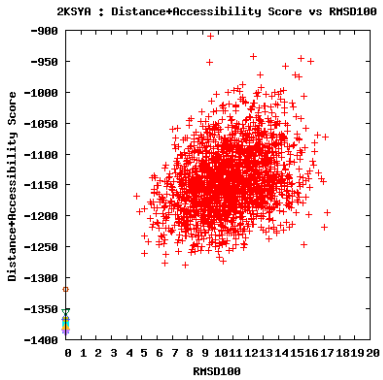
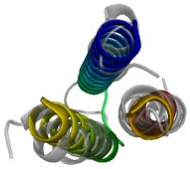
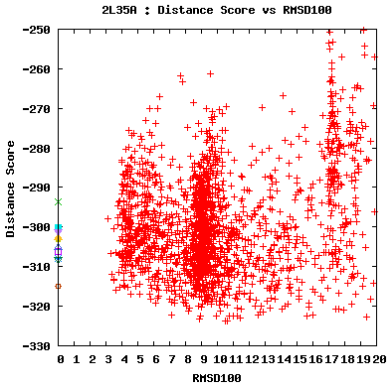
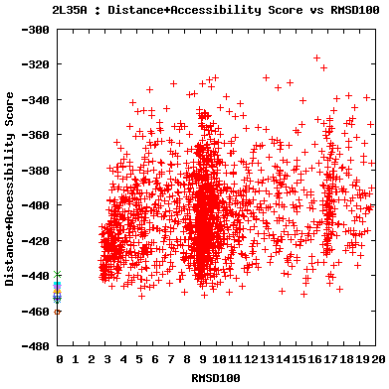


Figure 45 continued.

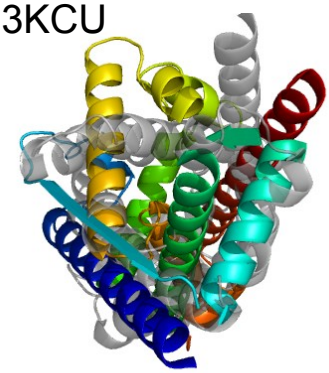
2L35

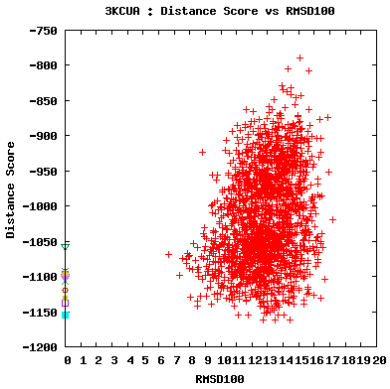


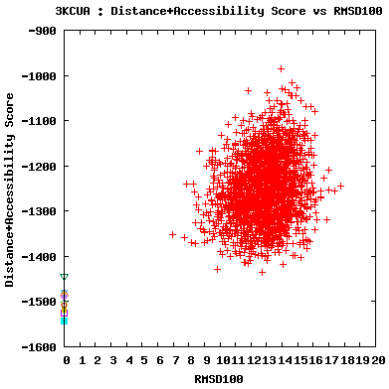




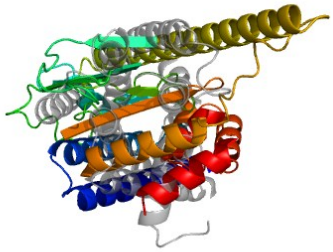
3KCU

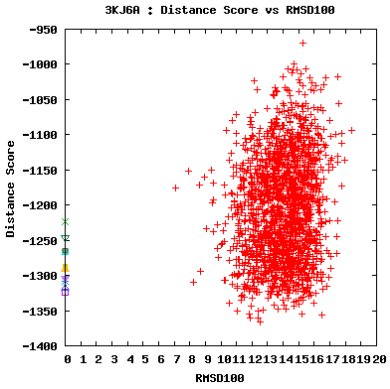


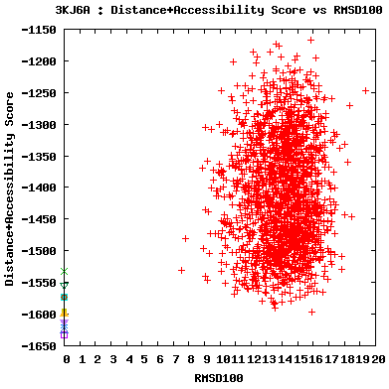




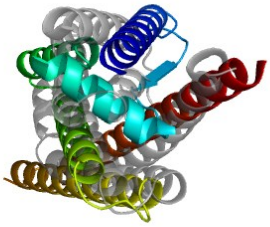
3KJ6

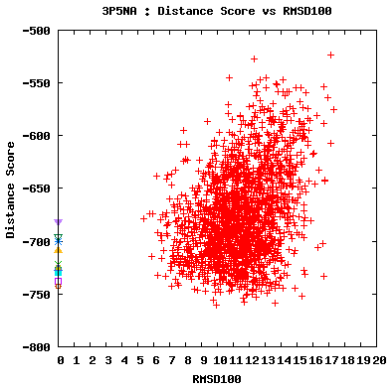






3P5N





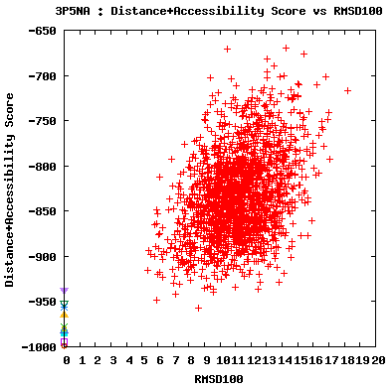
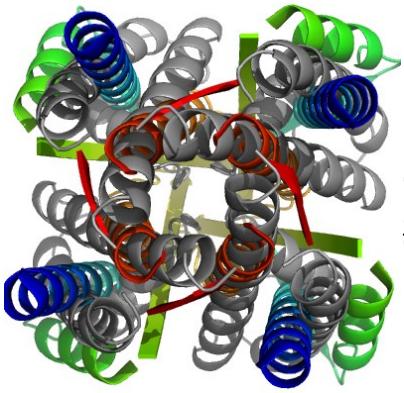
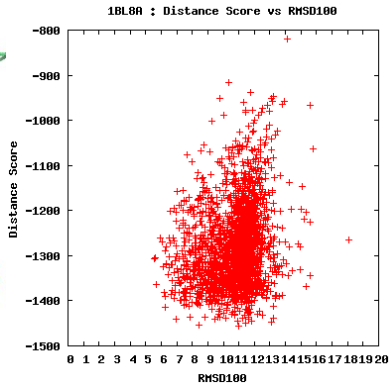
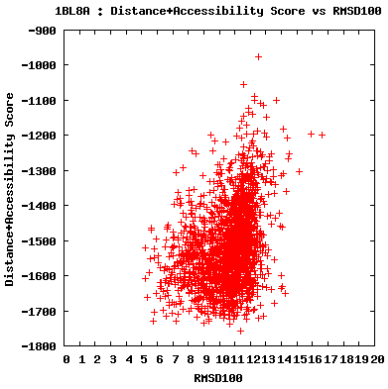


Figure 45 continued.

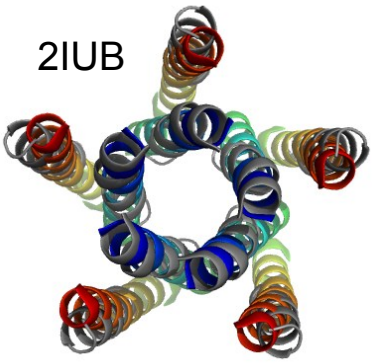
1BL8

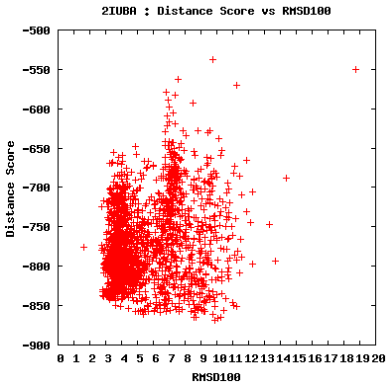


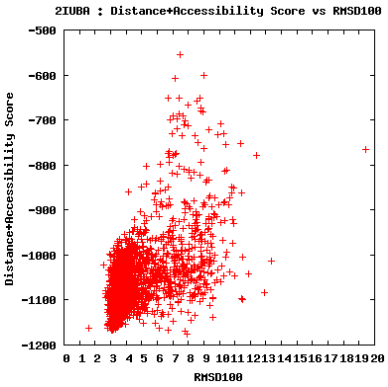




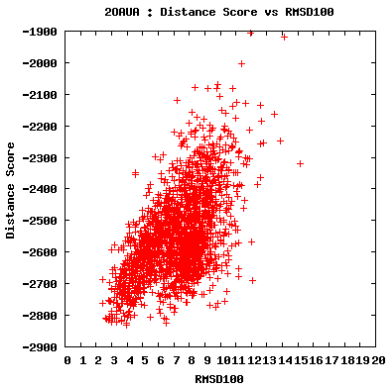
2IUB







2OAU



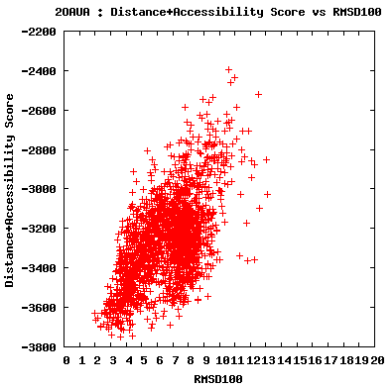


Figure 45 continued.

EPR specific scores allow selection of accurate models

Plots of the EPR specific D_{EPR} score for a model versus the RMSD100 value to the native for that model indicate there is not a strong correlation between the EPR score and accuracy (Figure 45). Additionally, the scores of the generated models reach the score of the native structure, indicating that the structural information from the distance restraints has been exhausted. When the accessibility scores are used in addition to the EPR distance score, the most accurate models can be selected by score for some of the proteins, such as 2OAU, 2IUB, and 2BG9. In addition, the score of the native structure is not as readily reached when the accessibility score is used in conjunction with the distance score, suggesting that more accurate models will be sampled as the score is optimized.

For the twenty monomeric proteins, the 1% best models by the EPR score used during structure prediction were selected. For models generated during structure prediction without EPR scores, the weighted sum of the knowledge-based scores was used to select the best 1% of models. For models generated during structure prediction with EPR distance and accessibility scores, the unweighted sum of the scores was used to select the best 1% of models. The magnitude of the EPR distance score is approximately three times larger than the accessibility score. This ensures that the EPR distance score, which provides information on the topology of the protein, is fulfilled. With the distance score being fulfilled, the accessibility score will then select for models with correct local interactions.

Using EPR distance and accessibility scores allows finding one of the best 1% of models according to RMSD100 value within the top 1% of models by score more frequently than using knowledge-based potentials alone (Table 32). Since 2000 models are generated for each protein with and without EPR restraints, the top 1% will be 20 models. When the knowledge-based potentials are used to select the best 1% of models

by score, one of the twenty best models by RMSD100 is found for 12 out of 20 proteins. When EPR distance and accessibility scores are used to select the best 1% of models by score, one of the twenty best models by RMSD100 is found for 16 out of 20 proteins. For two out of the four cases where the distance and accessibility scores fail to select one of the best 20 models by RMSD100, the best model by RMSD100 that is selected has values of 5.29 Å and 4.83 Å. The other two proteins sample a model with an RMSD100 value 2-3 Å better than the best in the top 1% by score.

For the three multimeric proteins, the EPR distance and accessibility scores were also used to determine the best 1% by score for each protein. This was done in a two-step process. First, the best 10% of models according to EPR distance score were selected. Second, the best twenty models according to EPR accessibility score are then taken as the best 1% of models according to the EPR specific scores. Using this protocol one of the best 1% of models by RMSD100 is recovered for each of the proteins. This compares favorably to when the knowledge-based scores are used to select the best 1% of models by score, which does not recover one of the best 1% of models by RMSD100 value for any of the three proteins.

Overall, EPR distance and accessibility scores select one of the twenty best models by RMSD100 within the top 1% by score for 19 out of 23 proteins. The protocols used for selecting the best 1% of models for the multimer was attempted with the monomer proteins, and the protocol used to select the best 1% of models for monomer proteins was attempted with the multimer proteins. In both cases, the recovery of one of the best 20 models by RMSD100 value was reduced; in the case of the multimer proteins, none of the three proteins did so successfully.

Table 32 Ability of EPR specific scores to pick the most accurate models sampled.
Results are shown of protein structure prediction benchmarks for twenty three membrane proteins under three different conditions: a) without the use of EPR data (None) b) with the use of EPR distance data (Distance) c) using both EPR accessibility and distance data (Distance+Accessibility). The best 20 models by score are selected as described in the Methods section. Then, the most accurate model within this set is determined according to RMSD100 to the native. The rank according to RMSD100 of the determined most accurate model is reported (Rank). This number would ideally be 1 for all proteins, indicating that selecting by score allows the most accurate model to be found. The RMSD100 value of the determined model is also reported (RMSD100).

	None		Distance		Distance + Accessibility	
PDB	**Rank**	**RMSD100**	**Rank**	**RMSD100**	**Rank**	**RMSD100**
1IWG	49	7.15	175	6.09	12	3.84
1J4N	704	11.82	160	6.43	38	5.29
1KPL	9	10.05	54	8.56	2	6.25
1OCC	1	4.71	2	4.45	2	3.60
1PV6	6	8.18	6	6.37	8	5.76
1PY6	3	6.74	1	5.22	2	5.27
1PY6*	10	4.54	9	3.85	4	3.11
1RHZ	28	9.06	146	8.43	1	5.81
1U19	1	7.00	4	6.43	2	6.38
2BG9	473	6.81	188	4.20	5	2.51
2BL2	1	3.38	78	4.47	12	3.22
2BS2	52	10.47	25	7.07	5	6.64
2IC8	50	9.20	109	8.31	72	8.13
2K73	1	5.74	85	7.45	59	7.27
2KSF	1	4.29	9	4.00	1	3.42
2KSY	25	8.84	8	5.52	1	4.68
2L35	183	6.33	281	5.54	310	4.83
3KCU	6	9.77	7	7.92	14	8.85
3KJ6	15	10.87	49	10.98	2	7.79
3P5N	2	7.76	10	6.50	13	6.09
1BL8	118	8.20	3	3.71	4	3.69
2IUB	33	6.64	1053	6.92	1	2.00
2OAU	35	13.28	36	10.06	8	8.19

Discussion

The use of EPR data has been demonstrated to aid in the prediction of membrane protein structures. EPR specific scores were coupled with the protein structure prediction method BCL::Fold. BCL::Fold assembles predicted SSEs in space without explicitly modeling the SSE connecting loop regions. This allows for rapid sampling of complex topology that is not easily achieved when an intact backbone must be maintained. By adding EPR specific scores into the knowledge-based scoring function, sampling of accurate structures is increased.

EPR accessibility scores are important for sampling the accurate membrane protein structures

EPR accessibility scores were previously used in conjunction with the Rosetta protein structure prediction algorithm (Alexander, Bortolus et al. 2008). The scores were applied in a benchmark to predict the structures of the small soluble proteins T4-lysozyme and αA-crystallin. The improvement in sampling more accurate models was compared between prediction trajectories using an EPR distance score and trajectories using an EPR distance score coupled with an accessibility score. For T4-lysozyme and αA-crystallin, using the accessibility score did not show a significant improvement in the accuracy of models sampled. This was attributed to the simple environments of soluble proteins: exposed to solvent, or buried in the core of the protein. These environments could be easily predicted by the knowledge-based potentials of Rosetta.

Membrane proteins are subjected to a more complex set of possible environments. Any given residue can reside buried in the protein or exposed. If a residue is exposed, it can be exposed to lipid within the membrane or solvent. If the protein has some porous nature, a residue can be solvent-exposed within the membrane (Dalmas, Cuello et al.). Such a complex interplay of environments will not be as easily distinguished by knowledge-based potentials. Here it has been demonstrated that using EPR accessibility

information consistently improves the models sampled at the best accuracies. The current benchmark has been done using oxygen accessibility measurements to help determine which face of SSEs should be buried or exposed. However, measurements of NiEDDA can be also readily be used when available to further help orient SSEs (Chakrapani, Cuello et al. 2008) (Dalmas, Cuello et al.).

The differing environments around membrane proteins means that measured EPR accessibility values will are complicated by the environment. Oxygen preferentially partitions into the membrane core, while NiEDDA is hydrophilic and prefers soluble environments. Consequently, the exposure of burial of a residue is not the only determining factor in the measured EPR accessibility value. The potential magnitude of an oxygen accessibility measurement will be greater towards the membrane core. This highlights an advantage of using the exposure moment to orient SSEs. The moment is independent of the relative magnitudes of individual accessibility values. This simplifies the comparison of the structure-based moment and the experimentally moment calculated from experiments.

Currently accessibility measurements are not used to help determine the depth a SSE should be placed within the membrane. However, it has been shown that membrane depths for residues in transmembrane segments can be determined from the combination of NiEDDA and oxygen accessibility measurements (Altenbach, Greenhalgh et al. 1994). The present study has demonstrated that such restraints may not be necessary for accurately determining the topology of membrane proteins, as given predicted transmembrane segments, it placement of the segment within the membrane is already narrowly constrained.

EPR distance scores improve the accuracy of topologies predicted for membrane proteins

EPR distance measurements are associated with large uncertainties in relating the measured spin label – spin label distance into backbone distances. In spite of this, EPR distance measurements provide important data on membrane protein structures (Claxton, Quick et al.) (Zou, Bortolus et al. 2009; Zou and McHaourab 2009) (Altenbach, Kusnetzow et al. 2008). In the present study, it has been shown that EPR distance data can significantly increase the frequency with which the correct topology of a membrane protein is sampled by the bcl::Fold protein structure prediction method. In addition, models are sampled with accuracy higher than achieved without EPR distance data. This is important because as the correct topologies are sampled with higher accuracy, models start to reach the point where they can be subjected to atomic detail refinement to further increase their accuracy (Barth, Schonbrun et al. 2007).

The EPR distance data used for the present study is simulated from known experimental structures. However, considerable effort was put forth to ensure that the simulated data mimics what would be obtained from a true EPR experiment, so that any results are not biased by the simulated data. The previously published method for selecting distance restraints was used to create ten different datasets per protein (Kazmier, Alexander et al.). This ensures results are not biased by a particularly selected dataset. Previously, the uncertainty in the difference between spin label distances and the corresponding C_β distance (D_{SL} - $D_{C\beta}$) was accounted for in simulated distance restraints by adding a random value between 12.5 and -2.5 (Kazmier, Alexander et al.). Here, the probability of observing a given D_{SL} - $D_{C\beta}$ is used to determine the amount that should be added to the Cβ-Cβ distance measured from the experimental structure.

Using a method developed for soluble proteins to select restraints for membrane proteins is not necessarily ideal. The constraints already imposed upon membrane proteins by the membrane suggest that methods for selecting restraints for membrane proteins should be needed. One such strategy could be to measure distances between

transmembrane segments on the same side of the membrane, with the assumption that transmembrane helices are mostly rigid, parallel structures. Initial attempts at updating the restraint selection method with membrane protein specific considerations did not produce significantly better results than the previously published method. Additional work is needed to account for topologically important SSEs that do not span the membrane, as well take into account the deviations of transmembrane segments from ideal geometries.

EPR specific scores allow selection of accurate models

Previous work has shown that selection and filtering of models by EPR distance scores improves overall accuracy of the pool of remaining models (Hirst, Alexander et al.; Kazmier, Alexander et al. ; Alexander, Bortolus et al. 2008). The current iterative protocol leverages these findings by taking the best models according to EPR distance score from initial protein structure prediction trajectories and feeding them into a second round of structure prediction. After the structure prediction protocol is complete, one of the most accurate models by RMSD100 can be frequently recovered by EPR score. Further refinement of models, for instance using the Rosetta high resolution refinement protocol for membrane proteins (Barth, Schonbrun et al. 2007), would allow even more accurate selection of the most native like models by score.

Two different protocols were utilized when selecting the best models by restraint score. One was specific for multimeric proteins and the other was specific for the monomeric proteins in the benchmark. Separate protocols can be justified due to the disparate challenges in predicting symmetric multimeric proteins versus single subunit proteins. Structure prediction of symmetric multimers is two-fold: the subunit structure needs to be correctly predicted and the interface between the subunits needs to be identified. If the interface between the subunits is not correctly predicted there will be significant inaccuracies in exposure for large portions of each subunit. Multimer proteins

were filtered first for agreement by EPR distance information. Second, the remaining models were filtered according to EPR accessibility data. This allows the information in the accessibility restraints to be fully leveraged to identify models with exposure profiles matching the native structure. This process was facilitated by the fact that the multimer proteins were sampled to very high accuracies (< 3 Å RMSD100 for 2OAU and 2IUB), within the realm where the accessibility data is most useful.

Improved secondary structure predictions will increase the accuracy of predicted structures

The pool of SSEs used to assemble the membrane protein topologies is the most important determinant in successfully predicting the membrane proteins' structure. The two monomeric proteins which were not successfully sampled below 7 Å RMSD100 show that the models were trying to be constructed with large strand elements not seen in the native structure. Although the multimer protein 1BL8 was sampled to under 6 Å RMSD100, the topology of predicted models also contained large stand elements not observed in the native model. The SSE pools are created in order to reduce the possibility of missing a SSE, which is generally a successful approach as demonstrated by 20 out of the 23 membrane proteins and previously for soluble proteins (Karakas, Woetzel et al. 2012). No secondary structure prediction techniques developed specifically for membrane proteins is currently available. As a result, the helical transmembrane span prediction software Octopus (Viklund and Elofsson 2008) is used in conjunction with Jufo and PsiPred (Jones 1999). Jufo and PsiPred provide predictions for SSEs that do not necessarily span the membrane and therefore won't be predicted by Octopus. Improved secondary structure prediction methods will benefit membrane protein structure prediction. In addition, it has been demonstrated that the pattern of accessibility values for measurements along a sequence follow the periodicity of the SSE on which they are measured (Lietzow and Hubbell 2004; Zou, Bortolus et al. 2009;

Zou and McHaourab 2009). Measured accessibility profiles could therefore be used to inform the pool of SSEs used for structure prediction.

Conclusion

Membrane protein structure determination will be significantly aided by the use of EPR experimental information coupled with structure prediction methods. The present work has introduced EPR specific scores for use in membrane protein structure prediction. Scores for incorporating EPR accessibility measurements have not previously been described for use within structure prediction methods. The ability of EPR data to improve the sampling of native-like topologies and the importance of EPR accessibility data for obtaining the most accurate models was demonstrated. Further, the EPR specific scores allow the selection of close-to-native models at the end of the structure prediction protocol.

Methods

Structure prediction protocol

The protein structure prediction protocol is based of the protocol of bcl::Fold for soluble proteins (Karakas, Woetzel et al. 2012). The method assembles SSEs in space, drawing from a pool of predicted SSEs. A Monte Carlo energy minimization with the Metropolis criteria is used to search for models with favorable energies. Models are scored after each Monte Carlo step using knowledge-based potentials describing optimal SSE packing, radius of gyration, amino acid exposure, and amino acid pairing, loop closure geometry, secondary structure length and content, and penalties for clashes (Woetzel, Karakaş et al. 2012). For membrane protein structure prediction an additional score is used which favors orthogonal placement of SSEs relative to the membrane (SSE_{align}). The assembly of the protein structure is broken down into five

stages of sampling with large structural perturbation moves that can alter the topology of the protein. The moves are similar to those described for soluble proteins (Karakas, Woetzel et al. 2012) and include translations and rotations of SSEs as well as swapping positions of SSEs within the current model and between the model and the SSE pool. Membrane specific perturbations include translation of SSEs orthogonally to the membrane to find optimal placement as well as rigid body translation and rotation of the entire protein to find the optimal placement within the membrane. Each of the five stages lasts for a maximum of 1000 Monte Carlo steps. If an energetically improved structure has not been generated within the previous 400 Monte Carlo steps, the minimization for that stage will cease. Over the course of the five assembly stages, the weight of clashing penalties in the total score is ramped as 0, 125, 250, 375, 500. The weight of the SSE_{align} score is 8.

Following the five stages of protein assembly, a structural refinement stage takes place. This stage lasts for a maximum of 2000 Monte Carlo steps and will terminate sooner if an energetically improved model is not sampled within the previous 400 steps. The refinement stage consists of small structural perturbations which will not drastically alter the topology of the protein model: swapping of SSEs is not allowed and the amount of translation and rotation per move is reduced. After the refinement stage, residues missing from the model are added in a loop building protocol to produce a complete structure of the protein. The protocol is based on cyclic coordinate descent (Canutescu and Dunbrack 2003). 1000 models are generated per structure prediction protocol.

For each protein, two sets of SSE pools are generated for use during structure assembly. The first SSE pool consists of the transmembrane spanning helices as predicted by Octopus (Viklund and Elofsson 2008). The second SSE pool contains elements predicted by Octopus as well as SSEs predicted from sequence by Jufo and PsiPred (Jones 1999). Using these two SSE pools, the structure prediction protocol is

independently conducted twice: a) once using the SSE pool containing predictions from Octopus, Jufo, and Psipred (“full pool”) and b) once emphasizing the predictions by Octopus (“Octopus pool”). Emphasis is placed on Octopus predictions by using only the Octopus generated SSE pool during the first two stages of assembly. During last three stages of structure assembly, the SSEs predicted from Jufo and PsiPred are added to the pool. This allows for better coverage of SSEs within the structure, since Octopus only predicts transmembrane spanning helices. Since the structure prediction protocol is carried out twice (once using the full pool and once using the Octopus pool), a total of 2000 models are created for each protein.

EPR specific scores are used during the five assembly and one refinement stages of structure prediction. The EPR distance scores have a weight of 500 during the first assembly stage. When the full pool is used, the weight is maintained at 500 until the refinement stage when it is reduced to 1.0. When the Octopus pool is used, the EPR distance score weight is 5 during assembly stages 2-5, and 0.5 during the refinement stage. The weight of the accessibility score is 5.0 during all assembly and refinement stages using either pool.

After 2000 models have been generated for each protein, the models are filtered according to EPR distance score. The top 10% of models resulting from the structure prediction protocol for each of the SSE pools are selected for a second round of minimization. This means that 100 models resulting from the full pool prediction protocol and 100 models from the Octopus pool prediction protocol are taken to a second round. The second round occurs as described above, the only difference being that the minimization uses the SSE placements of a given protein as a starting point. For each starting structure 10 models are created, giving a total of 2000 models being produced during the second round of minimization. Structure prediction trajectories not using any EPR distance restraints do not undergo the second round of minimization.

The procedures described above are conducted to produce models created without EPR data, models created using EPR distance data, and models created using EPR distance and accessibility data.

Simulating restraints

EPR distance and accessibility restraints were simulated where needed to obtain datasets for each of the 23 proteins. Distance restraints were simulated for all proteins, and accessibility restraints were simulated for all proteins except for the multimer proteins. Published accessibility data were used for the multimer proteins (Vásquez, Sotomayor et al. 2008) (Dalmas, Cuello et al.) (Perozo, Cortes et al. 1998). The oxygen accessibility measurements were used during protein structure prediction.

Accessibility restraints were simulated by calculating the neighbor vector value (Durham, Dorr et al. 2009) for residues within SSEs of each protein. This value was considered to be an oxygen accessibility measurement.

Distance restraints were simulated using a restraint selection algorithm (Kazmier, Alexander et al.) which attempts to distribute measurements across all SSEs. It also favors measurements between residues that are far apart in sequence. One restraint was generated for every 0.2 residues within predicted SSEs. As the predicted SSE in an SSE pool can contain multiple variations for SSEs, random non-overlapping sets of SSEs were used. This implicitly selects for more likely configurations of non-overlapping sets because these sets will be present more frequently in the SSE pool and therefore be selected more frequently than less likely sets. After the desired measurements are selected by the method, the experimental structure is used to calculate the distances. Distances are calculated between first side chain atoms; for glycine the HA2 atom is considered the first side chain atom. To take into account the uncertainty introduce d in a real EPR distance measurement, an amount is added or subtracted to the distance determined from the experimental structure. The amount is randomly selected from the

probability distribution of observing a given D_{SL}-$D_{C\beta}$ value biased by the probability (Hirst, Alexander et al.). In order to reduce the possibility of bias arising from a single restraint generation trajectory, ten independent restraint sets were generated for use. For the three multimer proteins, the protocol was the same except 0.1 restraints per residue within SSEs were selected.

Translating EPR accessibilities into structural restraints

EPR accessibility measurements are typically made in a sequence scanning fashion over a large portion of the target protein. Therefore, the approach for developing an EPR accessibility score takes advantage of this. The exposure moment of a window of amino acids is defined as $E_w = \sum_{n=1}^{N} e_n s_n$, where N is the number of residues in the window, e_n is the exposure value of residue *n*, and s_n is the normalized vector from the C_α atom to the C_β atom of residue *n*. This closely follows the hydrophobic moment for a sequence previously defined (Eisenberg, Weiss et al. 1984). The moment calculated for solvent accessible surface area has been shown to approximate the moment calculated from EPR accessibility measurements (Salwinski and Hubbell 1999). However, atomic-detail solvent accessible surface area is too computationally intense to be used during *de novo* protein structure prediction. Therefore, the neighbor vector approach to solvent accessible surface area approximation is used (Durham, Dorr et al. 2009) for calculating exposure moments from structure. The exposure moment is calculated for overlapping windows of length seven for helices and four for strands. To evaluate the agreement of a protein model with EPR data, the moment is calculated from the structure and from the experimental measurements. For each window a score between -1 and 0 is calculated by a cosine function where a score of -1 is given if the angle between the experimental and structural moments is zero and a score of 0 is given if the angle is 180°.

It has previously been demonstrated that the burial of sequence segments relative to other segments can be determined from the average accessibility values measured for

that stretch of sequence (Chakrapani, Cuello et al. 2008). To capture this information, the magnitude of the exposure moment for overlapping residue windows is determined from the model structure and from the measured accessibility. The Pearson correlation is then calculated between the rank order magnitudes of the structural versus experimental moments. This gives a value between -1 which indicates the structural and exposure magnitudes are oppositely ordered, to 1, which means the structural and exposure magnitudes are ordered equivalently. The score is obtained by negating the resulting Pearson correlation value so that matching ordering will get a negative score and be considered favorable.

EPR specific distance scores

The knowledge-based score for EPR distances previously reported is used score agreement of models with distance restraints (Hirst, Alexander et al.), D_{EPR}. This score spans a range of D_{SL}-$D_{C\beta}$ between ±12 Å. D_{SL} is the EPR measured distance between two spin labels; $D_{C\beta}$ is the distance between the corresponding Cβ on the residues of interest; D_{SL}-$D_{C\beta}$ is the difference between these two distances. An attractive potential on either side of D_{EPR} provides incentive for the Monte Carlo minimization to bring structures within the range of D_{EPR}. These attractive potentials use a cosine function to transition between a most unfavorable score of 0 and a most favorable score of -1. The potentials stretch from D_{SL}-$D_{C\beta}$ values of +30 Å and -30 Å to the first values of D_{SL}-$D_{C\beta}$ where D_{EPR} can provide scoring information.

Calculating EPR score enrichments

The enrichment is calculated as $enrichment = \frac{TP}{TP+FN} * \frac{P+N}{P}$, where *P* (positive) is the number of models under 8 Å RMSD100 (Carugo and Pongor 2001), *N* (negative) is the number of models above 8 Å RMSD100, *TP* (true positive) is the number of models with an RMSD100 under 8 Å and have a score rank in the top 10%, and *FN* (false negative)

is the number of models with an RMSD100 under 8 Å but a score not ranking within the top 10%. All models with an RMSD100 under 8 Å are therefore contained within the quantity $TP + FN$, and $\frac{TP}{TP+FN}$ indicates the fraction of accurate models the score can correctly identify. Ideally this value would be 1.0. The quantity $\frac{P+N}{P}$ indicates the ratio of all models to models which are under 8 Å RMSD100. The value of $\frac{P+N}{P}$ is fixed at 10. Therefore, the perfect enrichment value will be 10.0. No enrichment would be a value of 1.0, and an enrichment value between 0.0 and 1.0 indicates the score selects against accurate models. $\frac{P+N}{P}$ is fixed at 10 by selecting the subset of models from a protein structure prediction run which has 10% of models under 8 Å. This enrichment calculation process is repeated ten times to remove any bias from selecting a particular model subset. Enrichment is calculated for each of the 23 proteins using each EPR specific score. The models used for enrichment are those models which were created without using any EPR data, and the process is completed for each of the ten distance restraint datasets.

CHAPTER VII

CONCLUSIONS

The body of work in this dissertation presents the development of an array of computational methods for incorporating EPR data with protein structure prediction techniques.

The use of sparse EPR data sets with Rosetta allowed the prediction of the structure of the small soluble protein T4-lysozyme to an accuracy of 1 Å RMSD to the experimental structure as calculated over the residues in the helical domain. It was further demonstrated that only eight distance restraints are needed to significantly improve the accuracy of structure predictions by Rosetta. The information content of a distance restraint was formalized as the sequence separation divided by the Euclidean distance. This provided the rational for the development of an algorithm that selects the optimal EPR restraints that should be measured in order to most efficiently define the topology of a protein (Kazmier, Alexander et al.).

The incorporation of EPR distance restraints with Rosetta was preceded by the development of a model for the conformational dynamics of the MTS spin label. This "cone model" hypothesized that the conformational ensemble sampled by a spin label could be approximated by a cone. The cone model defined an effective position for the spin label, $SL_{effective}$, at the center of the cone. Statistics were calculated by placing two $SL_{effective}$ at positions on an ellipse in random orientations satisfying the geometric constraints of the cone model. This simple setup simulated EPR distance measurements on a protein. The juncture of $SL_{effective}$ with the ellipse was taken to be an effective C_β position. 'The statistic collected was the frequency of observing a given difference between two distance of the two spin labels and the distance of the two C_β atoms (D_{SL} –

$D_{C\beta}$). These statistics were compared to an experimentally observed frequency distribution for T4-lysozyme and α-crystlalin. The result showed that the cone model reproduced the range of $D_{SL} - D_{C\beta}$. The significance of this was that it allowed measured EPR distances to be related to the backbone of a protein within an error margin without the need for explicitly modeling the spin label at full atom detail. This is critical for *de novo* protein structure prediction which relies on computationally inexpensive methods to sampled the vast array of topologies accessible to a protein sequence. The cone model has since been refined to not only reproduce the range of $D_{SL} - D_{C\beta}$ observed experimentally but also the probability of observing a given $D_{SL} - D_{C\beta}$ value (Hirst, Alexander et al.).

A spin label rotamer library was created and incorporated with Rosetta's to reproduce experimentally observed spin label conformations and dynamics. The rotamer library consists of conformations observed experimentally in structures of singly spin labeleled T4-lysozyme solved by X-ray crystallography. The rotamer library also contained conformations predicted by molecular dynamics studies (Tombolato, Ferrarini et al. 2006; Tombolato, Ferrarini et al. 2006). Using the rotamer library, Rosetta was able to sample an experimentally observed conformation of a spin label in a buried site of T4-lysozyme which was not included in the rotamer library. Incorporating the rotamer library with Rosetta makes full atom modeling of MTSSL available to the variety of structure prediction protocols available within Rosetta. It also makes such modeling widely available to the scientific community using Rosetta. Other rotamer libraries are based on molecular dynamics studies and contain a large number (200) of rotamers (Polyhach, Bordignon et al.). Future work to further the development would include updating the rotamer library as more structures of MTSSL become available. In addition, development of an improved scoring function for the nitroxide moiety on the spin label could improve the accuracy of predictions.

A structure was predicted for the complex of the GPCR rhodopsin bound to its G-protein transducin. No experimental structures existed for the complex at the time. The structure incorporated EPR distance information to also predict the conformation of the G-protein helical domain relative to the G-protein nucleotide binding domain. The EPR data showed that a the helical domain undergoes a large conformational change upon binding to the receptor. Such a large domain motion in a system of this size would be difficult to study with molecular dynamics (Fenwick, Esteban-Martin et al. 2011). Therefore the docking protocol of Rosetta was used to rapidly sample potential conformations of the helical domain relative to the rest of the complex. Upon filtering for agreement with the experimental distances and clustering analysis, the resulting structures indicated the helical domain opens away from the nucleotide binding domain. This provides a hypothesized mechanism for release of GDP. Soon after the model was published, an experimental structure for the complex of B_2 adrenergic GPCR and the G-protein Gs was published (Rasmussen, DeVree et al. 2011). Although experimental conditions needed to make the complex crystallize make it difficult to compare certain hypotheses put forth by the model, one interesting component is that the helical domain is not well resolved in the experimental structure. This provided the motivation to create a second generation model based on the experimental structure to further refine predicted dynamics of the helical domain. Basing the model on the experimental structure allowed specific residue interactions to be investigated with Rosetta. Future work on modeling the GPCR-G-protein complex could focus on the β and γ domains of the G-protein. The original model predicted these would also undergo conformational changes, that were not observed in the crystal structure. However, a nanobody needed for crystallization sits at position that would block the motion predicted by the original model. Additional experimental and computational effort is needed to fully determine the conformation of β and γ .

EPR distance and accessibility data was incorporated with membrane protein structure prediction using the BioChemical Library. Twenty membrane proteins of up to 312 residues were predicted to an RMSD100 accuracy of better than 8 Å. Three homo multimer proteins were also benchmarked and sampled under 6 Å. The multimer proteins contained up to 595 total residues. Previously membrane protein fold was done with Rosetta without restraints (Yarov-Yarovoy, Schonbrun et al. 2006). It was able to predict predict up to 145 residues to within 4 Å RMSD accuracy. However, this depended on the topology and number of transmembrane spanning regions of the protein. For example, the 278 residues of rhodopsin, which has seven transmembrane spanning helices, were predicted to an accuracy of 9.2 Å RMSD accuracy. Membrane protein folding in Rosetta was demonstrated using one or two restraints to constraint a pair of transmembrane helices together (Barth, Wallner et al. 2009). This introduced a special protocol that allowed the contact to be made without regard for the connectivity of the backbone between the helices. The remaining helices are then sequentially introduced to fold the rest of the protein. By predicting structures in using the BCL algorithm, sequentially distance contacts can be rapidly sampled as secondary structure elements are free to move without hindrance from being connected by an intact protein backbone. Using EPR accessibility restraints in addition to distance restraints improved membrane protein structure prediction trajectories. This is in contrast to the result of using accessibility restraints during soluble protein folding, where accessibility restraints did not improve the accuracy of models generated. Future work could be done to introduce accessibility measurements as predictors of secondary structure. This would be based on work that has shown the periodicity of a series accessibility measurements can indicate the secondary structure type on which the measurements are being made (Zou and McHaourab 2009; Lietzow and Hubbell 2004).

APPENDIX

BCL::CLUSTER: A METHOD FOR CLUSTERING BIOLGOICAL MOLECULES COUPLED WITH VISUALIZATION IN THE PYMOL MOLECULAR GRAPHICS SYSTEM

This section provides a guide to clustering procedures used in the dissertation. It is based off (Alexander, Woetzel et al.).

Abstract

Clustering algorithms are used as data analysis tools in a wide variety of applications in Biology. Clustering has become especially important in protein structure prediction and virtual high throughput screening methods. In protein structure prediction, clustering is used to structure the conformational space of thousands of protein models. In virtual high throughput screening, databases with millions of drug-like molecules are organized by structural similarity, e.g. common scaffolds. The tree-like dendrogram structure obtained from hierarchical clustering can provide a qualitative overview of the results, which is important for focusing detailed analysis. However, in practice it is difficult to relate specific components of the dendrogram directly back to the objects of which it is comprised and to display all desired information within the two dimensions of the dendrogram. The current work presents a hierarchical agglomerative clustering method termed bcl::Cluster. bcl::Cluster utilizes the Pymol Molecular Graphics System to graphically depict dendrograms in three dimensions. This allows simultaneous display of relevant biological molecules as well as additional information about the clusters and the members comprising them.

Introduction

Hierarchical clustering is the procedure of iteratively grouping similar objects together, and a cluster is constituted by this group of similar objects (Johnson 1967). A distance measure is necessary to calculate the similarity between two objects. The purpose of clustering is to facilitate the identification of data patterns or classification of objects. Clustering methods are used in a wide variety of scientific applications and several different clustering algorithms can be applied to a dataset (for recent reviews of clustering methods see (Xu and Wunsch 2005; Omran, Engelbrecht et al. 2007)).

In particular, clustering is utilized in *de novo* protein structure prediction in order to aid in the selection of native-like models (Betancourt and Skolnick 2001; Skolnick, Kolinski et al. 2001). Theoretically, the native protein structure resides in the global energy minimum and can be identified unambiguously as the point of lowest free energy in the conformational space. However, vastness of the conformational space requires evaluation of millions of protein models to generate some that are “native-like”, i.e. reasonably similar in structure to the native conformation; typically a root mean square distance (RMSD) of backbone atoms smaller 7.5 Å. Such models have a score significantly higher than the native conformation. Further, the scoring functions used in protein structure prediction to estimate protein free energy are designed for fast evaluation of models. Stabilizing interactions within the protein model are not evaluated at atomic detail which reduces accuracy. In result, some non-native conformations will achieve scores similar to the native-like conformations.

Clustering is used to overcome this limitation (Shortle, Simons et al. 1998). Although the depth of the energy minimum in which the native conformation resides is reduced, the width of the energy funnel is less affected. Therefore, upon clustering predicted models according to structural similarity as measured by the RMSD, large clusters have

an increased likelihood to contain native-like conformations. Global Distance Test (GDT) (Zemla 2003) and distance matrices (Lesk 1997) are alternative distance measures used in the process.

Clustering is also used in the analysis of libraries of small, often drug-like, molecules. Often millions of such molecules are included in (virtual) high-throughput screening or generated by structure generators (Leach, Gillet et al. ; Yongye, Bender et al. ; Priestle 2009). Clustering structures the chemical space and identifies, for example, sets of similar compounds that share a common biological activity (Mueller, Rodriguez et al. ; Willett, Winterman et al. 1986). Similarity measures compare the configuration of small molecules either based on the largest common substructure (Barnard 1993) or based on a vector of descriptors, so-called fingerprints (Duan, Dixon et al. ; Sastry, Lowrie et al. ; Willett 2006). The Tanimoto (Godden, Xue et al. 2000) coefficient is a popular similarity measure (for review see (Willett, Barnard et al. 1998)).

The focus of the current work is to introduce a hierarchical agglomerative clustering method (bcl::Cluster). The goal of bcl::Cluster is to facilitate the clustering and analysis of biological molecules such as proteins and ligands by allowing visualization of the molecules within the context of the dendrogram. bcl::Cluster uses the Pymol Molecular Graphics System (Pymol) to display the dendrogram and the biomolecules.

Methods

bcl::Cluster is implemented as a part of the BioChemical Library, an in-house developed, object oriented, C++ programming library. The code has been developed with flexibility and extensibility as a priority. Key aspects of the method are elaborated on below.

Input

bcl::Cluster relies upon pre-calculated pair-wise distances between objects in order to perform clustering. As input formats, bcl::Cluster reads data in the format of a distance matrix or a pair-wise list of distances, where the objects to be clustered are represented by an identifier. Both input formats are independent of the actual type of object that is being clustered. Therefore, although the graphical output of the method is tailored to biological molecules, bcl::Cluster is generally applicable. The separation of the calculation of distances between individual objects and the clustering algorithm allows bcl::Cluster the flexibility to work with any numerical distance measure for any type of object. The bcl library is used to compute a variety of similarity measures such as GDT (Zemla 2003), longest continuous segment (Zemla 2003), MaxSub (Siew, Elofsson et al. 2000), average distance matrix error (Lesk 1997), RMSD (Rao and Rossmann 1973), RMSD100 (Carugo and Pongor 2001), largest common substructure (Krissinel and Henrick 2004), and the Tanimoto coefficient (Godden, Xue et al. 2000).

Distance Measures

bcl::Cluster allows the use of similarity or dissimilarity distance measures for clustering. In the case of a similarity distance measure, objects with a greater distance value are more similar. An example of such a measure would be the Tanimoto coefficient frequently used to calculate the similarity of small molecules (Godden, Xue et al. 2000). A dissimilarity distance measure is one where objects with a smaller distance value are more similar. The RMSD value between two proteins is an example of a dissimilarity distance measure (Maiorov and Crippen 1994).

Clustering Algorithm

bcl::Cluster uses a hierarchical agglomerative clustering algorithm (Serna 1996). Each individual object starts out in a cluster containing only that object. The method continues to iteratively combine the most similar cluster pairs until only a single cluster remains.

The similarity, or linkage, between two clusters can be calculated in several ways in bcl::Cluster. Average linkage between two clusters is calculated as the average pair-wise distance between all objects in two clusters. Single linkage between two clusters is calculated as the distance of the most similar pair of objects between the two clusters. Complete linkage between two clusters is calculated as value of the most dissimilar pair of objects between the two clusters. Lastly, total linkage is calculated similarly to average linkage but also considers pair-wise distances within the two clusters when calculating the average distance. This differs from average linkage which only considers pair-wise distances between clusters.

Clustering Cutoff

For practical applications, it is typically not necessary to compute the entire hierarchy of cluster agglomerations. For example, in the case of clustering protein models, clustering can be stopped once linkage values are reached where combining two clusters would produce a cluster encompassing proteins of different topology, i.e. at a RMSD of approximately 7.5 Å. By allowing the user to limit the extent of clustering, the time and memory requirements of bcl::Cluster can be reduced.

Pre-clustering

As mentioned in the description of the clustering algorithm, a hierarchy of clusters is obtained by iteratively combining pairs of clusters until only a single cluster remains that contains all previous clusters. Reducing the number of iterations that are needed until all clusters are combined will reduce the number of linkage values that need to be calculated and increase the speed of the clustering algorithm. To this end, bcl::Cluster offers the ability to perform a “pre-clustering” step before the hierarchical clustering takes place. The pre-clustering step consists of a single pass through all objects where objects that are within a defined similarity are automatically combined to form a cluster. As the clusters are formed during the single iteration through all objects, an object will be added

to a cluster if it is within the predefined similarity cutoff of any object within the cluster. In this manner, the pre-clustering step is using single linkage. After pre-clustering, agglomerative clustering proceeds as normal albeit some initial clusters will already contain multiple objects.

Pymol Visualization

The Python programming language can be used to interface with Pymol in a scriptable manner. Python scripts can be written which perform calculations based on data extracted directly from Pymol and perform functions within Pymol. In addition, Pymol allows simple shapes such as spheres and cylinders as well as text to be generated. These generated objects are termed compiled graphics objects, CGOs. bcl::Cluster takes advantage of these features. After clustering is complete, bcl::Cluster generates a Python script which will create the dendrogram and load any molecules for display in Pymol.

Results

A set of protein models and a set of small molecules with distance matrices are used to demonstrate bcl::Cluster. Up to 1000 protein models are used, with an RMSD matrix containing values ranging from 0.0 Å to 18.8 Å. Five small molecules are used with a randomly filled distance matrix assumed to be a similarity measure. The values range from 0.2 to 1.0.

Pymol Dendrogram Output

In Pymol, the dendrogram is displayed in conjunction with additional text information. The scale of linkages is shown on the right side of the dendrogram (Figure 46A). In addition, information about each cluster can be displayed in front of the dendrogram (Figure 46B). The information contains in order from top to bottom along the cluster (Figure 46A): a.) the identifier for the object which is the center of the cluster, where the

center object is calculated as the object with the smallest average distance to all other objects in the cluster; b.) a unique identification number for the cluster which can be used as a guide to find the cluster in text files created by bcl::Cluster; c.) the size of the cluster interms of the number of objects that are contained within the cluster; d.) the linkage of the cluster.

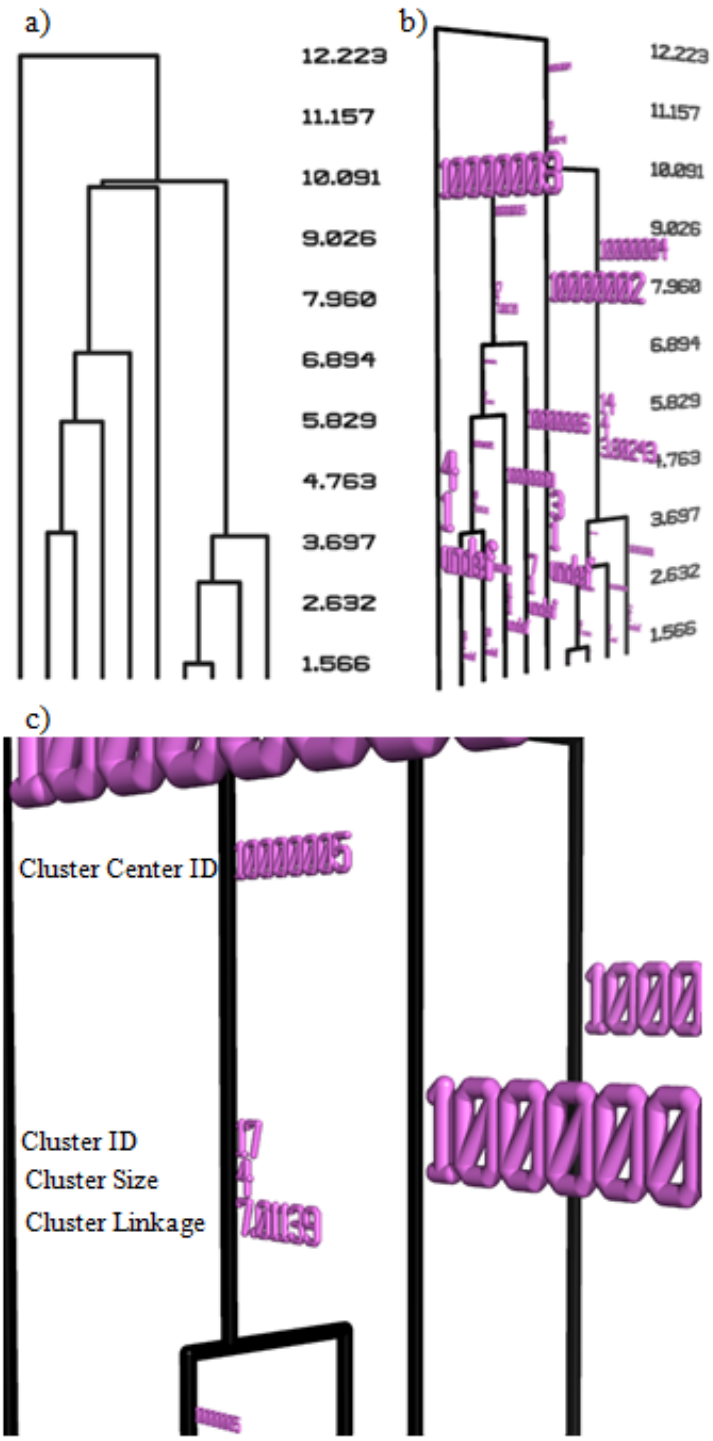


Figure 46 Dendrograms and cluster information generated using Pymol from the output of bcl::Cluster.
(a) Simple display of a dendrogram. The numbers at right denote linkage levels of clusters. (b) Clusters within dendrograms can be labeled with information about each cluster. Displaying the dendrogram in Pymol allows the user to dynamically adjust the view. (c) A zoomed-in view of a specific cluster with the information about the cluster labeled.

Cluster Color Gradient

Visualization of the dendrogram in Pymol provides additional opportunities to aid in the analysis of clustering beyond directly viewing the biological molecules. Pymol allows the colors of CGOs to be specified. In bcl::Cluster, the individual clusters in the dendrogram can be colored Visualization of the dendrogram in Pymol provides additional opportunities to aid in the analysis of clustering beyond directly viewing the biological molecules. Pymol allows the colors of CGOs to be specified. In bcl::Cluster, the individual clusters in the dendrogram can be colored according to a gradient indicative of some numerical descriptor. For example, the color of a cluster can indicate how similar the members of the cluster are to the native protein structure (Figure 47A).

Cluster radius

When defining the cylinder CGOs that comprise the dendrogram in Pymol, the desired radius is specified. bcl::Cluster can vary the radius of the cylinders according to the number of objects that are within the cluster corresponding to a cylinder (Figure 47B). Scaling the visual size of a cluster with the number of members allows the user to quickly determine which clusters in the dendrogram contain the largest number of members.

Pre-Clustering Procedure

The pre-clustering procedure allows similar objects to be grouped into a cluster prior to hierarchical clustering (Figure 48). Selecting an appropriate value for the distance threshold for combining objects allows pre-clustering to take place without affecting the upper regions of the dendrogram. In a test case using 1000 proteins with a pre-clustering threshold set so that the effect is similar to that seen in Figure 48, clustering is finished 20% faster.

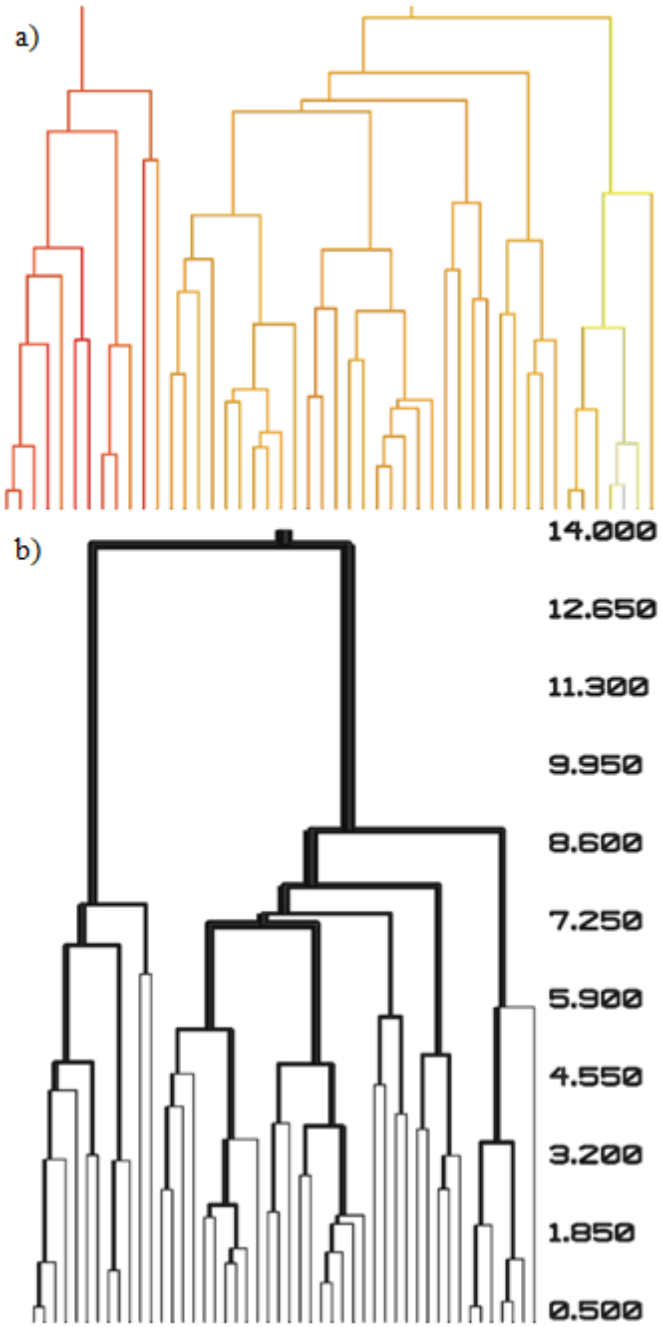


Figure 47 The flexibility in generating dendrograms in Pymol allows the dendrogram itself to contain more information than just the cluster hierarchy.
(a) Clusters of the dendrogram are color coded according to the average RMSD to an experimental structure of cluster members. The color scheme goes from red (very similar to experimental structure) to yellow to white (less similar to experimental strucdture). (b) Clusters in the dendrogram are scaled in size according to the number of members contained within the cluster. Clusters are scaled by 3.0*sqrt(number of members - 1).

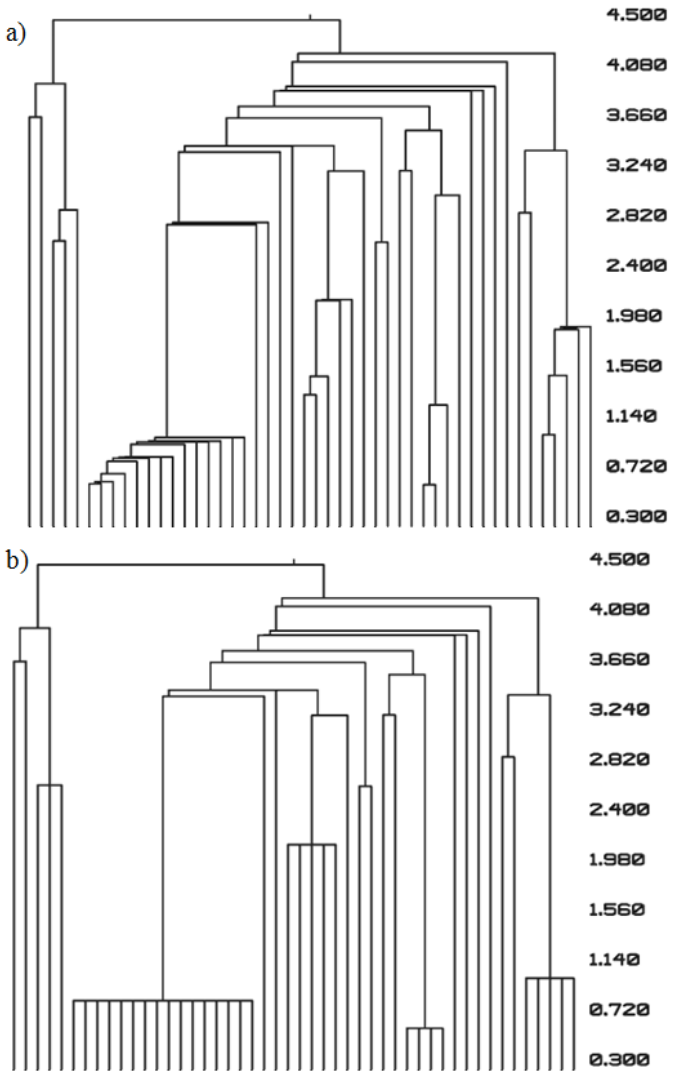


Figure 48 Comparison of the clustering results (a) without pre-clustering and (b) with pre-clustering.

In a set of 50 protein models, a pre-clustering threshold of 3.0 Å RMDS was used to create clusters of the most similar models before hierarchical clustering was performed. The dendrogram that is obtained with the added pre-clustering step shows several models were initially clustered together. As hierarchical clustering progresses, the differences between (a) and (b) diminish. Pre-clustering is performed in a single pass through all the objects being clustered, and it therefore reduces the number of iterations that must take place during the hierarchical clustering step.

Display of Biological Molecules

For every cluster, the biological molecule which is the center of the cluster is displayed as a representative of that cluster (Fig. 4(a) and Fig. 5). The additional cluster information previously described can be shown along with the biological molecules (Fig. 4(b)) but is easily hidden in the Pymol environment if desired. In Fig. 5, the small molecule distance measurement is assumed to be a similarity measure, so larger distance values indicate a higher similarity between objects. As a result, the dendrogram is inverted compared to when a similarity measurement is used, as in the case of the protein model dendrogram (Fig. 4).

Text Output

In addition to the Python script for displaying the dendrogram in Pymol, bcl::Cluster outputs information about the dendrogram in text format to facilitate quantitative analysis. Every member of every cluster is listed on a separate line with additional information (Fig. 6).

Discussion

This work describes the bcl::Cluster clustering method which has been developed to allow straight forward analysis of clustering of biological molecules. Pymol provides the graphical interface which displays the dendrogram resulting from the hierarchical agglomerative algorithm of bcl::Cluster. Using Pymol allows other information to be displayed to the user in addition to the dendrogram such as the actual molecular structures of the objects being clustered, cluster sizes, and color coding according to some other numerical descriptor. The user can then quickly focus on the areas of interest in the dendrogram.

One of the advantages of using Pymol is that the display of the clustering results is dynamic. The user can perform any function of Pymol while viewing the results such as

zooming, translating, and hiding certain objects. When looking at large, complex dendrograms, this functionality makes it easier to view the results as compared to if the dendrogram was displayed as a static picture. However, one limitation of bcl::Cluster is the computational power needed to display a complex dendrogram and many proteins or ligands in real time in Pymol. This limitation can be partially overcome by hiding objects within the Pymol environment, but with very large datasets the dendrogram alone will grow to be the limiting factor in what can be displayed. However, the bcl::Cluster text output provides the information needed to analyze clustering results for datasets too large to view in Pymol.

The object oriented nature of the bcl::Cluster code allows additional functionality to be easily added in the future. One extension would be to add other clustering algorithms. Additional formats for inputting distance values or outputting results can also be added. The application is available from the bcl::Commons website (http://bclcommons.vueinnovations.com/).

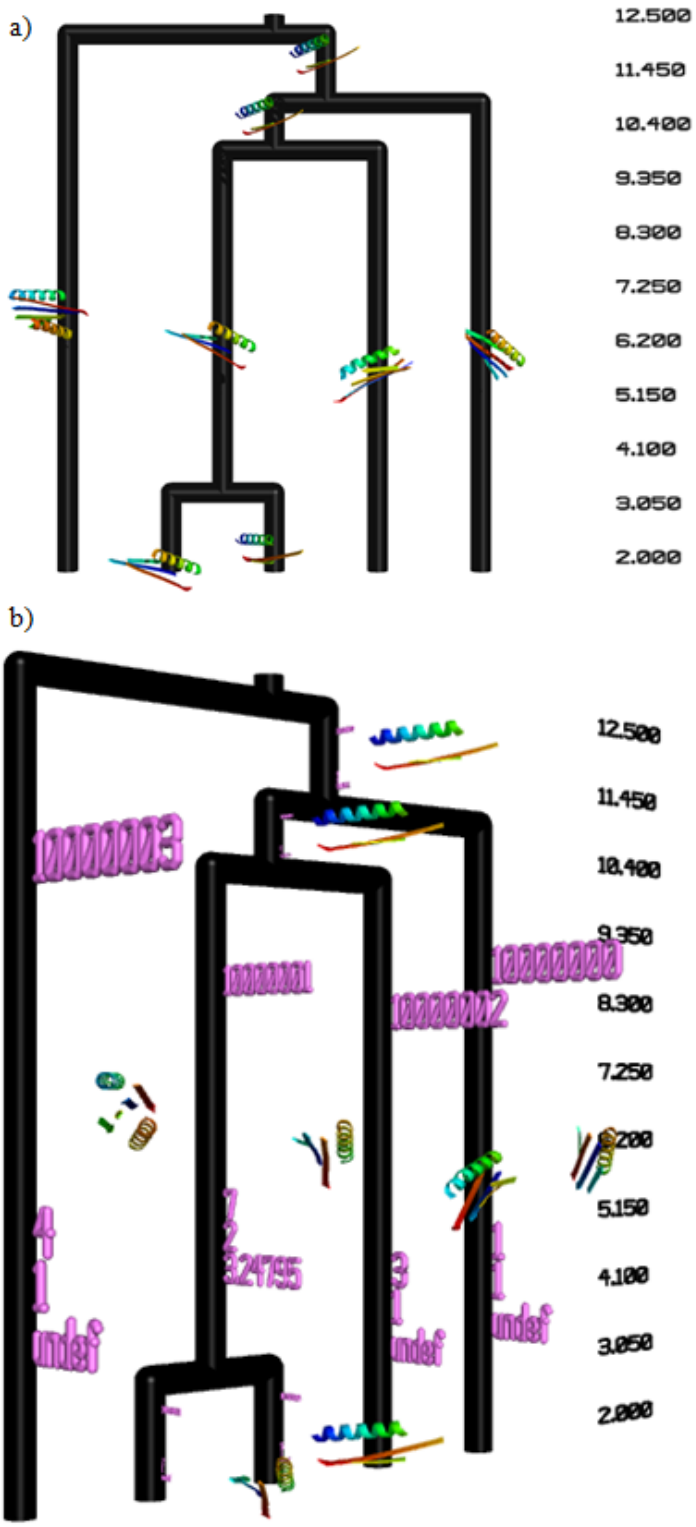


Figure 49 Display of clustered proteins directly within the context of the dendrogram. The protein which is the center of the cluster is displayed as the representative. (a) Simple view of the dendrogram with cluster center protein structures displayed. (b) Additional information about each cluster can also be displayed in conjunction with the protein structures. The cluster center id (Figure 46) indicates the coordinate file from which the structure is created.

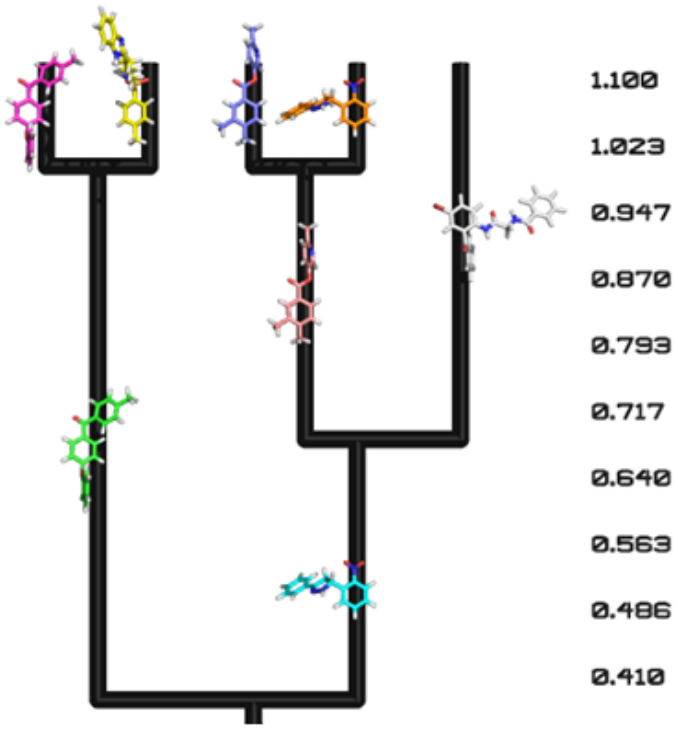


Figure 50 Clustered small molecules displayed within the resulting dendrogram. Here the distances used were similarity measures such as the Tanimoto coefficient.

NODE 4 : Member : a : Size : 3 : Leaf : 0 : Linkage : 0.243243

NODE 4 : Member : b : Size : 3 : Leaf : 0 : Linkage : 0.243243

NODE 4 : Member : c : Size : 3 : Leaf : 0 : Linkage : 0.243243

NODE 3 : Member : c : Size : 1 : Leaf : 1 : Linkage : nan

NODE 5 : Member : a : Size : 2 : Leaf : 0 : Linkage : 0.393939

NODE 5 : Member : b : Size : 2 : Leaf : 0 : Linkage : 0.393939

NODE 1 : Member : a : Size : 1 : Leaf : 1 : Linkage : nan

NODE 2 : Member : b : Size : 1 : Leaf : 1 : Linkage : nan

Figure 51 Sample text output from a dendrogram created from three objects.

For objects (a, b, and c), each member of each cluster is listed on a separate line. The cluster identification and linkage is given for each member. Also, whether or not the node is at the base of the dendrogram (Leaf) is indicated by a boolean (one for true, zero for false). Linkages for clusters of only one member are undefined

Supplemental Information

Here specific command line options are described and examples provided.

input_format

TableLowerTriangle uses table format with lower triangle filled in :

bcl::storage::Table<double>	1000_0000	1000_0001	1000_0002	1000_0003	1000_0004
1000_0000	0	0	0	0	0
1000_0001	13.5371	0	0	0	0
1000_0002	11.9716	14.9337	0	0	0
1000_0003	11.7247	8.8339	14.01	0	0
1000_0004	6.89769	3.24795	4.82608	13.9589	0

TableUpperTriangle uses table format with lower triangle filled in

bcl::storage::Table<double>	1000_0000	1000_0001	1000_0002	1000_0003	1000_0004
1000_0000	0	13.5371	11.9716	11.7247	6.89769
1000_0001	0	0	14.9337	8.8339	3.24795
1000_0002	0	0	0	14.01	4.82608
1000_0003	0	0	0	0	13.9589
1000_0004	0	0	0	0	0

Linkage

Complete – the most different inter-cluster distance between any two members in two clusters (does not take into account intra-cluster distances in the two clusters).

Single – the most similar inter-cluster distance between any two members in two clusters (does not take into account intra-cluster distances in the two clusters).

Average – For two nodes *a* and *b* with members *i* and *j*, it is calculated as $\frac{\sum_{i,j} distance^{i,j}}{size_a \cdot size_b}$

Total - For two nodes *a* and *b* with members *i* and *j*, it is calculated as

$$\frac{\sum_{i,j} distance^{i,j} + \sum_{i,i} distance^{i,i} + \sum_{j,j} distance^{j,j}}{size_a \cdot size_b + \frac{size_a * (size_a - 1)}{2} + \frac{size_b * (size_b - 1)}{2}}$$

This is similar to average linkage but it also takes into account the intra-node members.

output_file

Using the TableLowerTriangle input file (in a file named distances.txt) from above with the following command line will result in the output files shown below named cluster_output.Rows.txt and cluster_output.Centers.txt:

cluster.exe Cluster –distance_input_file distances.txt -input_format TableLowerTriangle –output_format Rows Centers –linkage Average –output_file cluster_output

Rows – outputs information about the dendrogram with one member of a cluster per line. The numberical identifier for the cluster is given in the second column. Every member will have the identifier of the cluster to which it belongs. The name of the member is given in the 6th column, and the size of the cluster to which the member belongs is given in the 10th cluster. If the member is part of a cluster that is at the base of the dendrogram, the 14th column will have a “1” to indicate the member is part of a cluster that is a leaf. The linkage of the cluster in which the member resides is given in the 18th column. The linkage of a cluster that has only one member is undefined and indicated as “nan”. In this example, node 9 has four members and a linkage of 10.8021.

```
NODE 6 : Member : 1000_0000 : Size : 5 : Leaf : 0 : Linkage : 12.1319
NODE 6 : Member : 1000_0001 : Size : 5 : Leaf : 0 : Linkage : 12.1319
NODE 6 : Member : 1000_0002 : Size : 5 : Leaf : 0 : Linkage : 12.1319
NODE 6 : Member : 1000_0003 : Size : 5 : Leaf : 0 : Linkage : 12.1319
NODE 6 : Member : 1000_0004 : Size : 5 : Leaf : 0 : Linkage : 12.1319
NODE 4 : Member : 1000_0003 : Size : 1 : Leaf : 1 : Linkage : nan
NODE 9 : Member : 1000_0001 : Size : 4 : Leaf : 0 : Linkage : 10.8021
NODE 9 : Member : 1000_0004 : Size : 4 : Leaf : 0 : Linkage : 10.8021
NODE 9 : Member : 1000_0002 : Size : 4 : Leaf : 0 : Linkage : 10.8021
NODE 9 : Member : 1000_0000 : Size : 4 : Leaf : 0 : Linkage : 10.8021
NODE 8 : Member : 1000_0001 : Size : 3 : Leaf : 0 : Linkage : 9.87989
NODE 8 : Member : 1000_0004 : Size : 3 : Leaf : 0 : Linkage : 9.87989
NODE 8 : Member : 1000_0002 : Size : 3 : Leaf : 0 : Linkage : 9.87989
NODE 7 : Member : 1000_0001 : Size : 2 : Leaf : 0 : Linkage : 3.24795
NODE 7 : Member : 1000_0004 : Size : 2 : Leaf : 0 : Linkage : 3.24795
NODE 2 : Member : 1000_0001 : Size : 1 : Leaf : 1 : Linkage : nan
NODE 5 : Member : 1000_0004 : Size : 1 : Leaf : 1 : Linkage : nan
NODE 3 : Member : 1000_0002 : Size : 1 : Leaf : 1 : Linkage : nan
NODE 1 : Member : 1000_0000 : Size : 1 : Leaf : 1 : Linkage : nan
```

Centers – analogous to the output from “Rows” above, except that only the center member of each cluster is output. The cluster center member is calculated as the member that is most similar to all other members in the cluster. For each member the distance to all other members is summed, and the member with the sum indicating it is most similar to other members is the center.

```
    NODE 6 : Member : 1000_0004 : Size : 5 : Leaf : 0 : Linkage : 12.1319
NODE 4 : Member : 1000_0003 : Size : 1 : Leaf : 1 : Linkage : nan
NODE 9 : Member : 1000_0004 : Size : 4 : Leaf : 0 : Linkage : 10.8021
NODE 8 : Member : 1000_0004 : Size : 3 : Leaf : 0 : Linkage : 9.87989
NODE 7 : Member : 1000_0001 : Size : 2 : Leaf : 0 : Linkage : 3.24795
NODE 2 : Member : 1000_0001 : Size : 1 : Leaf : 1 : Linkage : nan
NODE 5 : Member : 1000_0004 : Size : 1 : Leaf : 1 : Linkage : nan
NODE 3 : Member : 1000_0002 : Size : 1 : Leaf : 1 : Linkage : nan
NODE 1 : Member : 1000_0000 : Size : 1 : Leaf : 1 : Linkage : nan
```

remove_nodes_below_size

Using the command line below will remove any nodes that have a size below 2.

cluster.exe Cluster –distance_input_file distances.txt -input_format TableLowerTriangle –output_format Rows Centers –linkage Average –output_file cluster_output -remove_nodes_below_size 2

remove_internally_similar_nodes

After clustering is finished, clusters within clusters that have a linkage less than the given value will be removed from the hierarchy. This ensures that all members are represented in visual output.

cluster.exe Cluster –distance_input_file distances.txt -input_format TableLowerTriangle –output_format Rows Centers –linkage Average –output_file cluster_output -remove_internally_similar_nodes 10

height_cutoff

The clustering will be stopped when a cluster is formed that has a linkage greater than the supplied cutoff.

cluster.exe Cluster –distance_input_file distances.txt -input_format TableLowerTriangle –output_format Rows Centers –linkage Average –output_file cluster_output -height_cutoff 9

output_pymol

Enables creation of a python script that can be run in Pymol to create a dendrogram. In Pymol go to “File” then “Run” then select the appropriate file. Each cluster has text indicating from top to bottom : a) the name of the cluster center member b) the cluster identification c) the size of the cluster d) the linkage of the cluster. In the example below the python script is being output to a file named dendrogram.py.

```
cluster.exe Cluster -distance_input_file distances.txt -input_format TableLowerTriangle -
linkage Average -output_format Rows Centers -output_file cluster_output -output_pymol
100 25 50 25 10 dendrogram.py
```

The unit length, width, and spacing of the cylinders can be adjusted to change the dimensions of the dendrogram.

```
cluster.exe Cluster -distance_input_file distances.txt -input_format TableLowerTriangle -
linkage Average -output_format Rows Centers -output_file cluster_output -output_pymol
100 5 10 25 10 dendrogram.py
```

The dendrogram will be affected by other flags such as remove_nodes below size.

```
cluster.exe Cluster -distance_input_file distances.txt -input_format TableLowerTriangle -
linkage Average -output_format Rows Centers -output_file cluster_output -output_pymol
100 25 50 25 10 dendrogram.py -remove_nodes_below_size 2
```

The dendrogram will also be affected by the remove_internally_similar_nodes flag :

```
cluster.exe Cluster -distance_input_file distances.txt -input_format TableLowerTriangle -
linkage Average -output_format Rows Centers -output_file cluster_output -output_pymol
100 25 50 25 10 dendrogram.py -remove_internally_similar_nodes 10
```

The dendrogram will also be affected by the height_cutoff flag :

```
cluster.exe Cluster -distance_input_file distances.txt -input_format TableLowerTriangle -
linkage Average -output_format Rows Centers -output_file cluster_output -output_pymol
100 5 10 25 10 dendrogram.py -height_cutoff 9
```

pymol_label_output_string

This is the default method for labeling the dendrogram as seen above.

pymol_label_output_protein_model_from_string

The dendrogram will be labeled with the actual protein models. The protein model PDB file names are created from the member names and the prefix and postfix parameters passed to this flag. In this example, the four objects clustered are named `1000_0000`, `1000_0001`, `1000_0002`, `1000_0003`, `1000_0004`. They correspond to Protein Data Base formatted files (PDBs) that are located in "/home/user/pdbs/" and the files start with "model". The PDB files end in "_final.pdb". So the PDB for `1000_0001` is "/home/user/pdbs/model_1000_0001_final.pdb".

```
cluster.exe Cluster -distance_input_file distances.txt -input_format TableLowerTriangle -
linkage Average -output_format Rows Centers -output_file cluster_output -output_pymol
50 5 50 25 10 dendrogram_a.py -pymol_label_output_protein_model_from_string
/home/user/pdbs/model _final.pdb
```

The unit values for the length, radius, and separation of the cylinders in the dendrogram should be adjusted according to the size of the protein.

pymol_label_output_small_molecule

The dendrogram will be labeled with molecules taken from an SDF formatted file. In this case, clustering is being done on a similarity measure, so clusters have linkages indicated high similarity are at the top of the dendrogram. The input file assumes the objects numbered from 0 to N and the number of the object corresponds to its position within the SDF file. See below:

```
    bcl::storage::Table<double> 0 1 2 3 4
0 1.000000 1.000000 0.393939 0.243243 0.222222
1        0 1.000000 1.000000 0.255814 0.181818
2        0        0 1.000000 1.000000 0.368421
3        0        0        0 1.000000 1.000000
4        0        0        0        0 1.000000
```

An example command line is below. The SDF file is provided as "/home/user/molecules.sdf".

```
cluster.exe Cluster -output_file cluster_results.txt -distance_input_file distances.txt -
input_format TableUpperTriangle -output_format Centers -linkage Average -
distance_definition greater -output_pymol 100 1 10 20 10 dendrogram.py -
pymol_label_output_small_molecule /home/user/molecules.sdf
```

pymol_set_min_max_girth

The minimum and maximum that the dendrogram reaches to can be set with this flag.

```
cluster.exe Cluster -output_file cluster_output.txt -distance_input_file distances.txt -
input_format TableUpperTriangle -output_format Centers -linkage Average -
distance_definition greater -output_pymol 100 1 10 20 10 dendrogram.py -
pymol_label_output_small_molecule /home/user/molecules.sdf -
pymol_set_min_max_girth 0.0 1.2
```

pymol_scale_node_with_size

The radius of the cylinder representing a cluster is increases as the number of members in the clusters increases

```
cluster.exe Cluster -distance_input_file distances.txt -input_format TableLowerTriangle -
linkage Average -output_format Rows Centers -output_file cluster_output -output_pymol
50 5 50 25 10 dendrogram_a.py -pymol_label_output_protein_model_from_string
/home/user/pdbs/model _final.pdb -pymol_set_min_max_girth 0.0 13.0 -
pymol_scale_node_with_size
```

pymol_color_nodes_by_description

Colors clusters in a gradient based on some numerical descriptor. Each cluster is colored according to the average of the numerical descriptors for all its members. The gradient goes from Red (small average numerical descriptor) to White (large average numerical descriptor). The numerical descriptor could be some score, or RMSD to a native structure, etc. An example numerical descriptor file is given below. The member names in the descriptor file must match the member names in the distance input file.

```
1000_0000 16.4436
1000_0001 10.4263
1000_0002 17.3385
1000_0003 9.7106
1000_0004 24.9397
```

The minimum and maximum descriptor values for red and white, respectively, are set so that extreme values will not affect the gradient.

cluster.exe Cluster -distance_input_file distances.txt -input_format TableLowerTriangle -linkage Average -output_format Rows Centers -output_file cluster_output -output_pymol 50 5 50 25 10 dendrogram_a.py -pymol_label_output_protein_model_from_string /home/user/pdbs/model _final.pdb -pymol_set_min_max_girth 0.0 13.0 -pymol_color_nodes_by_description descriptions.ls 9 25

precluster

Before clustering begins, makes a single pass through all objects and uses single linkage to populate clusters. So, during the single pass through all objects, two objects will be combined into a cluster if they have a distance that meets the provided cutoff. Additional objects will be added to a cluster if the distance from the current object to any of the objects already in the cluster meets the cutoff threshold. This is used to speed up clustering since having prepopulated clusters will reduce the number of iterations

needed to develop the whole hierarchy. However, if the threshold is too generous the results in the important parts of the dendrogram could be affected.

```
cluster.exe Cluster -distance_input_file distances.txt -input_format TableLowerTriangle -
linkage Average -output_format Rows Centers -output_file cluster_output -output_pymol
50 5 50 25 10 dendrogram_a.py -pymol_label_output_protein_model_from_string
/home/user/pdbs/model _final.pdb -pymol_set_min_max_girth 0.0 13.0 -precluster 5
```

output_node_members

Prints a file for every cluster. Each file lists the objects that are contained within that cluster. The files are named “dendrogram_node_*X*.ls”, where *X* is a cluster identifier. The clusters are taken from the dendrogram after it has been filtered by remove_internally_similar_nodes and remove_nodes_below_size.

```
cluster.exe Cluster -distance_input_file distances.txt -input_format TableLowerTriangle -
linkage Average -output_format Rows Centers -output_file cluster_results.txt -
output_node_members
```

This work was performed in the directory *cluster.*

GUIDE TO SAMPLING AND FITTING OF MODEL ENSEMBLES TO EPR DISTANCE PROBABILITY DISTRIBUTIONS

This section presents the protocol and command lines for creating an ensemble of structurally perturbed structures and then finding the subset of the ensemble which can reproduce experimentally measured EPR distance probability distributions.

Generating an ensemble of models

The BCL can be used to perturb specific domains of a protein. The protocol is based on the standard BCL folding method and one of the available protocols under the name “Dock”. The special aspect of Dock is that conformations user-targeted regions of the protein can be selectively sampled. The domains consist of secondary structure elements, and regions of flexibility can be removed from the model before sampling. The specified domains are then perturbed as a rigid body. Below is an example of a file for specifying the domain of a protein that should be perturbed.

```
DomainSpecifier
translate min = 0.0 max = 10.0
rotate min = 0.0 max = 1.048
REMOVE 'A'   58    62
HELIX  'A'   63    90
COIL   'A'   91    97
HELIX  'A'   98   110
 REMOVE 'A'  111   120
DomainSpecifierEnd
```

This allows the amount of maximum translation and rotation per Monte Carlo move to be specified. If the “REMOVE” tag is used, this secondary structure element will be

removed from the structure before sampling begins. The character identifier of the chain within which the SSEs are located is also given, along with the first and last residue number of the SSEs of interest. Only the “REMOVE” string is a recognized identifier, a description for each line is necessary and allows the flexibility for additional key identifiers in the future.

The domain specification file is used in conjunction with the other standard flags of the BCL Fold application. An example command line is provided below.

```
bcl.exe Fold -protocols Default Dock -mutate_protocols Default Dock -score_protocols Default Dock -prefix decoys//m_354 -nmodels 2 -native m_12_bound_gdp_0001.pdb -start_model m_12_bound_gdp_0001.pdb -use_native_pool -mc_number_iterations 750 500 -min_sse_size 0 0 0 -aaclass AABackBone -mc_temperature_fraction 0.5 0.2 -quality RMSD -domain_specify helical_bound_23.domain -random_seed $seed_number
```

Fitting of sampled models to EPR distance probability distributions

Given an ensemble of protein structures, the BCL can be used to find the subset which closely reproduces an arbitrary number of EPR distance distributions. Two inputs are needed for this. The first is a file listing the protein models that will be used for fitting. The second is a file listing the files containing each EPR distance distribution. The distance distribution files should have two columns, one with the distance and the other with the probability. The baseline of the probabilities should be at zero. A command line that can be used to perform fitting is given below.

```
bcl.exe FitEPRDistribution -exp_hist_list epr_distributions_trimmed.ls -exp_hist_data_columns 0 1 -model_list relaxed_best_renum_full.ls 0 -num_fits 1000 -start_size_range 5 20 -prefix fit_01/ -terminate_criteria 0.1 2500 -use_pdbid_numbering -message_level Standard -random_seed 2011102609
```

GUIDE TO EPR RESTRAINT BASED MEMBRANE PROTEIN FOLDING IN THE BCL

This section presents the protocol and specific command line options for membrane protein structure prediction using EPR distance and accessibility restraints in the BCL. Additional information about the individual flags can be obtained by using the “-help” flag on the appropriate application.

Secondary Structure Element Pool Generation

The BioChemical Library (BCL) can be used to generate secondary structure element (SSE) pools. To generate SSE pools using only Octopus (Viklund and Elofsson 2008):

```
create_sse_pool.exe CreateSSEPool -ssmethods OCTOPUS -pool_min_sse_lengths 5 3
-sse_threshold 0.0 0.0 0.0 -prefix
/home/alexanns/bclepr/membrane/multimer/data_benchmark/" $1 " -join_separate -
evaluate_pool -pdb /home/alexanns/bclepr/membrane/multimer/data_benchmark/" $1
".pdb -factory SSPredHighest -chain_id
```

To generate SSE pools using Octopus, Jufo, and PsiPred (Jones 1999):

```
create_sse_pool.exe CreateSSEPool -ssmethods OCTOPUS JUFO PSIPRED -
pool_min_sse_lengths 5 3 -sse_threshold 0.0 0.0 0.0 -prefix
/home/alexanns/bclepr/membrane/multimer/data_benchmark/" $1 " -join_separate -
evaluate_pool -pdb /home/alexanns/bclepr/membrane/multimer/data_benchmark/" $1
".pdb -factory SSPredHighest -chain_id
```

Obtaining Simulated EPR Distance Restraints

The command line that can be used for selecting restraints using the application of the BCL is shown below below:

bcl-all-static.exe OptimizeDataSetPairwise -fasta $fasta_files -pool_min_sse_lengths 0 0 -pool $pool_filename -distance_min_max 20 45 -nc_limit 10 -ensembles $ensemble_file -mc_number_iterations 1000000 1000000 -prefix $output_prefix.$build_number -nmodels $nstruct -read_scores_optimization $score_weights -read_mutates_optimization $mutate_weights -message_level Standard -pymol_output -data_set_size_fraction_of_sse_resis 0.1 -random_seed $seed

The inputted score table format is, for example :

bcl::storage::Table<double> seq_sep data_set_size sse_connection distance_range_0 exposure_0

Weights 1 1 1 1 1

The weight table for mutates is similar to :

bcl::storage::Table<double>	add_single	swap
weights	1	100

Secondary structure element pools containing predictions from Octopus, Jufo, and Psipred can be provided to allow selection of restraints between secondary structure elements. The pools should be modified as needed to ensure that any selected restraints would have coordinates available in the experimental structure, if their purpose is for benchmarking with known structures.

After the set of restraints is selected, the experimental structure is used to determine the restraint distances. In addition, an amount can be added or subtracted to mimic the uncertainty in EPR distance measurements. This is accomplished using BCL application and command line below :

bcl-all-static.exe SimulateDistanceRestraints -pdb $input_pdb_filename -simulate_distance_restraints -output_file $output_cst_file -min_sse_size 0 0 0 -add_distance_uncertainty /blue/meilerlab/home/alexanns/workspace_bcl/bcl/histogram/sl-cb_distances.histograms -restraint_list $dataset_prefix$x$dataset_postfix 0 1 5 6 -random_seed $seed

Obtaining Simulated EPR Accessibility Restraints

The BCL application and command line below is used to generate accessibility restraints for use during structure prediction :

```
bcl-all-static.exe SimulateAccessibilityRestraints -pdb $pdb_filename -output_file $output_filename -accessibility_environments Oxygen -min_sse_size 0 0 999
```

Protein Structure Prediction Trajectories

The command line for predicting a number of structures for a monomeric protein using an SSE pool containing all SSE predictions using EPR distance and accessibility restraints is :

```
bcl-all-static.exe Fold -native 1IWGA.pdb -pool_separate 1 -pool_min_sse_lengths 5 3 -quality RMSD GDT_TS -superimpose RMSD -message_level Critical -function_cache -sspred JUFO PSIPRED -sspred_path_prefix 1IWG -stages_read stages.txt -pool 1IWGA.SSPredHighest_PSIPRED_JUFO_OCTOPUS.pool -nmodels 20 -prefix 1IWGAbuild_01_0_0_ -protein_storage pdbs/ Overwrite  -membrane -restraint_types DistanceEPR AccessibilityEPR -restraint_prefix 1IWGA/cst/1IWGA_sim.0.050_true -loop_closure_threshold 0.1 -loop_rama_mutate_prob 0.0 -ccd_fraction '[0.5,1.0]' -random_seed $seed_number
```

The command line for predicting multimeric proteins additionally has the flags specifying the symmetry and native for comparison :

```
-native_multimer 1BL8A.pdb -symmetry C4
```

For structure prediction runs using the Octopus generated pool for the first two stages of assembly, a pool prefix flag must be used to specify how to find the appropriate pool files, and the standard pool flag can also be used to specify a starting pool :

```
-pool_prefix 1BL8A -pool 1BL8A.SSPredHighest_OCTOPUS.pool
```

After protein structure prediction, the BCL application and command line below is used to rank models by score and RMSD100 value, allowing the models to be analyzed further:

```
bcl-all-static.exe FoldAnalysis -output_table score_sorted_eprsum.tbl -protein_storage
pdbs/ -message_level Standard  -sort epr_distance epr_upper_penalty
epr_lower_penalty
```

MODELING THE CONFORMATION OF RECEPTOR BOUND VISUAL ARRESTIN

The information in this section presents the modeling methods in the submitted manuscript entitled “The conformation of receptor-bound visual arrestin” (Kim, Vishnivetskiy et al. submitted).

RosettaEPR Protein Modeling Based on DEER Distance Restraints

A crystal structure of free arrestin (PDB ID 1CF1 (Hirsch, Schubert et al. 1999)) was used as the template for comparative modeling. The crystal structure contains four copies of the protein in the asymmetric unit (chains A-D). The four copies display structural plasticity in loop regions involving residues 67-79, 132-143, 152-169, and 335-345. The average per-residue-RMSD values for these four regions are 6.0 Å, 0.2 Å, 2.3 Å, and 5.6 Å, respectively. To calculate the average per-residue-RMSD values, the structures are superimposed using GDT (Zemla 2003) with a 2 Å cutoff. Next, the Cα coordinates for each residue are collected from the four structures. The RMSD for a residue is then calculated : $RMSD_{residue} = \sqrt{\frac{\sum_{m_1}^{N}\sum_{m_2}^{N} d_{m_1 m_2}^2}{0.5 * (N*N-1)}}$, where *N* is the number of structures; *m1* and *m2* are each one of the structures; *d* is the distance between the Cα coordinates for the current residue of interest of *m1* and *m2*.The average per residue RMSD over all residues is 1.1 Å. Although residues 132-143 show low per residue RMSD within the experimental structures, the EPR distance measurements show large changes upon binding, indicating flexibility in these residues (Table 33).

Table 33 Twenty-five distance measurements made in the free (*Free (d1)*) and receptor bound (*+P-Rh* (d2)*) state of visual arrestin, which were used for modeling. *d2-d1* indicates the change in distance between the bound and free state.

	Arrestin Mutant	Median distance (Å)		
		Free (d1)	+ P-Rh* (d2)	d2-d1
1	32/356	15.5	16.5	1.0
2	72/173	19.0	21.5	2.5
3	72/348	39.0	37.0	-2.0
4	74/60	22.5	27.0	4.5
5	74/139	27.0	23.0	-4.0
6	74/157	34.0	40.0	6.0
7	74/173	21.5	24.0	2.5
8	74/240	36.0	37.0	1.0
9	74/344	50.0	47.0	-3.0
10	85/244	34.5	35.5	1.0
11	139/60	39.0	31.0	-8.0
12	139/173	28.0	18.0	-10.0
13	139/197	43.0	55.0	12.0
14	139/227	45.5	49.0	3.5
15	139/244	23.0	35.0	12.0
16	139/251	16.0	34.0	18.0
17	139/267	41.0	45.0	4.0
18	139/344	33.0	44.0	11.0
19	157/173	30.0	34.5	4.5
20	173/240	36.0	33.0	-3.0
21	197/267	27.5	27.5	0.0
22	197/344	22.0	21.0	-1.0
23	244/272	37.5	37.0	-0.5
24	244/344	32.0	25.0	-7.0
25	267/344	21.0	24.5	3.5

Modeling the Unbound State of Arrestin

The experimentally determined structure of unbound arrestin was subjected to the Rosetta relaxation protocol (Bradley, Misura et al. 2005; Misura and Baker 2005). Chains A, C, and D were used as starting point for modeling free arrestin. Prior to relaxation, residues with missing density in the experimental structure were constructed to ensure a complete, continuous structure. Reconstruction included residues 1-9, 362-373, and 385-404 chains A, C, and D. Chain B was excluded from modeling unbound arrestin, since it has missing density for residues 70 -77, a critical loop region for which eight EPR distances have been measured for this study. Twenty five EPR distance restraints obtained for free arrestin (Table 33) guided the relaxation trajectories using a knowledge-based EPR distance potential (Hirst, Alexander et al. 2011). The command line flags used are given below :

relax.linuxgccrelease -fa_input -database rosetta_database/ -in:file:fullatom -out::overwrite -out::file::fullatom -constraints::cst_fa_file distances.rosetta_cst -constraints::cst_fa_weight 4 -constraints::epr_distance -in:file:s start_model.pdb -out:prefix m_5_ -nstruct 10 -use_input_sc

During relaxation the local interactions of all atoms are optimized within the Rosetta energy potential. A total of 2853 relaxations were conducted. The best structure after relaxation according to EPR distance restraint agreement is derived from chain D of the crystal structure. This model did not fulfill restraints involving residues 335-348 and 195-202, and the restraint between residue 72 and residue 173.

In order to obtain a structural model that also fulfills restraints involving these residues, larger conformational changes were applied to the best structure after relaxation according to EPR distance restraint agreement. Residues 335-348 and 195-202 were reconstructed using the Rosetta loop building protocol (Qian, Raman et al.

2007; Wang, Bradley et al. 2007) in 1000 independent trials. The command line flag used are given below :

loopmodel.linuxgccrelease -fa_input –database rosetta_database/ -loops::loop_file 1cf1.loops -loops::frag_sizes 9 3 1 -loops::frag_files aa1cf1A09_05.200_v1_3 aa1cf1A03_05.200_v1_3 none -loops::build_initial -loops::remodel quick_ccd -loops::refine -loops::relax -out::overwrite -out::file::fullatom -constraints::cst_file distances.rosetta_cst -constraints::epr_distance -loops::input_pdb m_90_start_model_0006.pdb -out:prefix m_12_ -nstruct 20

The best model according to the knowledge-based EPR distance potential score was then used as the basis for modeling residues 67-79. During loop construction, all twenty five EPR distances were used to restrain the generated conformations. The command line used was the same as above, with only the loop definition file being changed appropriately. The best model according to EPR restraints agrees better with the EPR distance data than any of the four crystallographic conformations. The average restraint score is -0.91 for the best model, whereas the average restraint scores for the experimental structure chains A, C, and D are -0.84, -0.83, and -0.88, respectively. The best score would be -1.0 and 0.0 the worst. Chain B has an average restraint score of -0.90, not taking into account restraints involving missing density. For the best unbound model, only distances 139-197, 139-227, and 139-244 score worse than -0.85. These are still within tolerance, having scores of -0.59, -0.69, and -0.57, respectively. The restraints involving the rebuilt loop regions show an improved score compared to the model that resulted from relaxation. Specifically, distance 72-173 improves from 0.00 to -1.0; 74-344 improves from -0.03 to -0.98; 139-197 improves from 0.00 to -0.59; 139-344 improves from 0.00 to -0.91; 197-344 improves from 0.00 to -0.85; and 244-344 improves from 0.00 to -0.90.

Modeling the P-Rh* Bound State Arrestin

A total of 4037 independent loop building trajectories were conducted using Rosetta starting from chains A, B, C, and D of the experimental structure of arrestin. Twenty five distances measured by EPR of arrestin in the R*-bound state (Table 33) were used during loop construction to bias the structures towards the bound conformation. The residues that were re-constructed included 1-9, 67-79, 132-143, 152-169, 247-254, 264-272, 335-345, 362-373, and 385-404. The command line used is given below :

> loopmodel.linuxgccrelease -fa_input –database rosetta_database/ -loops::loop_file 1cf1.loops -loops::frag_sizes 9 3 1 -loops::frag_files aa1cf1A09_05.200_v1_3 aa1cf1A03_05.200_v1_3 none -loops::build_initial -loops::remodel quick_ccd -loops::refine no -loops::relax no -out::overwrite -out::file::fullatom -constraints::cst_file distances.rosetta_cst -constraints::epr_distance -loops::input_pdb 1cf1A.pdb -out:prefix m_0_ -nstruct 100

The top ten models by EPR restraint score have better agreement with the bound state EPR data than any of the four starting crystallographic conformations. The average restraint score is better than or equal to -0.90 for each of the ten models. The average restraint score for the starting conformations is between -0.65 and -0.68. Across the ten models, each distance restraint has an average score of better than -0.80, except for distances 139-244, 139-251, and 197-344. These three distances have average scores of -0.67, -0.67, and -0.78, respectively, across the ten models.

This work was performed in the directory *arrestin*.

BIBLIOGRAPHY

(2005). Mathematica Champaign, Illinois, Wolfram Research, Inc.

Abramson, J., I. Smirnova, et al. (2003). "Structure and Mechanism of the Lactose Permease of Escherichia coli." Science **301**(5633): 610-615.

Alexander, N., A. Al-Mestarihi, et al. (2008). "De Novo High-Resolution Protein Structure Determination from Sparse Spin-Labeling EPR Data." Structure **16**(2): 181-195.

Alexander, N., N. Woetzel, et al. Bcl::Cluster: A method for clustering biological molecules coupled with visualization in the Pymol Molecular Graphics System. Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on.

Altenbach, C., W. Froncisz, et al. (2005). "Accessibility of nitroxide side chains: absolute Heisenberg exchange rates from power saturation EPR." Biophys J **89**(3): 2103-2112.

Altenbach, C., D. A. Greenhalgh, et al. (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 of the United States of America **91**(5): 1667-1671.

Altenbach, C., A. K. Kusnetzow, et al. (2008). "High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation." Proceedings of the National Academy of Sciences.

Altenbach, C., T. Marti, et al. (1990). "Transmembrane protein structure: spin labeling of bacteriorhodopsin mutants." Science **248**(4959): 1088-1092.

Altenbach, C., K. J. Oh, et al. (2001). "Estimation of inter-residue distances in spin labeled proteins at physiological temperatures: experimental strategies and practical limitations." Biochemistry **40**(51): 15471-15482.

Altenbach, C., K. Yang, et al. (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**(38): 12470-12478.

Baker, D. (2000). "A surprising simplicity to protein folding." Nature **405**(6782): 39-42.

Barnard, J. M. (1993). "SUBSTRUCTURE SEARCHING METHODS - OLD AND NEW." Journal of Chemical Information and Computer Sciences **33**(4): 532-538.

Barth, P., J. Schonbrun, et al. (2007). "Toward high-resolution prediction and design of transmembrane helical protein structures." Proceedings of the National Academy of Sciences of the United States of America **104**(40): 15682-15687.

Barth, P., B. Wallner, et al. (2009). "Prediction of membrane protein structures with complex topologies using limited constraints." Proceedings of the National Academy of Sciences **106**(5): 1409-1414.

Berg, B. v. d., W. M. Clemons, et al. (2004). "X-ray structure of a protein-conducting channel." Nature **427**(6969): 36-44.

Berman, H. M., T. Battistuz, et al. (2002). "The Protein Data Bank." Acta Crystallogr D Biol Crystallogr **58**(Pt 6 No 1): 899-907.

Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucleic Acids Research **28**(1): 235-242.

Betancourt, M. R. and J. Skolnick (2001). "Finding the needle in a haystack: Educing native folds from ambiguous ab initio protein structure predictions." Journal of Computational Chemistry **22**(3): 339-353.

Bill, R. M., P. J. F. Henderson, et al. (2011). "Overcoming barriers to membrane protein structure determination." Nat Biotech **29**(4): 335-340.

Bokoch, M. P., Y. Zou, et al. (2010). "Ligand-specific regulation of the extracellular surface of a G-protein-coupled receptor." Nature **463**(7277): 108-112.

Bonneau, R., I. Ruczinski, et al. (2002). "Contact order and ab initio protein structure prediction." Protein Sci **11**: 1937-1944.

Bonneau, R., C. E. M. Strauss, et al. (2001). "Improving the Performance of Rosetta Using Multiple Sequence Alignment Information and Global Measures of Hydrophobic Core Formation." Proteins: Struct., Funct., Genet. **43**: 1-11.

Bonneau, R., J. Tsai, et al. (2001). "Rosetta in CASP4: Progress in ab initio protein structure prediction." Proteins **45 Suppl**(5): 119-126.

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

Borbat, P. P., H. S. McHaourab, et al. (2002). "Protein structure determination using long-distance constraints from double-quantum coherence ESR: study of T4 lysozyme." J Am Chem Soc **124**(19): 5304-5314.

Borbat, P. P., K. Surendhran, et al. (2007). "Conformational motion of the ABC transporter MsbA induced by ATP hydrolysis." Plos Biology **5**(10): 2211-2219.

Bower, M. J., F. E. Cohen, et al. (1997). "Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool." J Mol Biol **267**(5): 1268-1282.

Bowers, P. M., C. E. M. Strauss, et al. (2000). "Denovo protein structure determination using sparse NMR data." J. Biomol. NMR **18**: 311-318.

Bradley, P., D. Chivian, et al. (2003). "Rosetta in CASP5: Progress in ab initio protein structure prediction." Proteins: Struct., Funct., Genet. **53**(Suppl 6): 457-468.

Bradley, P., L. Malmstrom, et al. (2005). "Free modeling with Rosetta in CASP6." Proteins **61 Suppl 7**: 128-134.

Bradley, P., K. M. Misura, et al. (2005). "Toward high-resolution de novo structure prediction for small proteins." Science **309**(5742): 1868-1871.

Brocchieri, L. and S. Karlin "Protein length in eukaryotic and prokaryotic proteomes." Nucleic Acids Research **33**(10): 3390-3400.

Brown, L. J., K. L. Sale, et al. (2002). "Structure of the inhibitory region of troponin by site directed spin labeling electron paramagnetic resonance." Proc Natl Acad Sci U S A **99**(20): 12765-12770.

Bryson, K., L. J. McGuffin, et al. (2005). "Protein structure prediction servers at University College London." Nucleic Acids Res **33**(Web Server issue): W36-38.

Call, M. E., K. W. Wucherpfennig, et al. (2010). "The structural basis for intramembrane assembly of an activating immunoreceptor complex." Nat Immunol **11**(11): 1023-1029.

Canutescu, A. A. and R. L. Dunbrack (2003). "Cyclic coordinate descent: A robotics algorithm for protein loop closure." Protein Science **12**(5): 963-972.

Carugo, O. and S. Pongor (2001). "A normalized root-mean-square distance for comparing protein three-dimensional structures." Protein Sci **10**(7): 1470-1473.

Cavalli, A., X. Salvatella, et al. (2007). "Protein structure determination from NMR chemical shifts." Proceedings of the National Academy of Sciences of the United States of America **104**(23): 9615-9620.

Chakrapani, S., L. G. Cuello, et al. (2008). "Structural dynamics of an isolated voltage-sensor domain in a lipid bilayer." Structure **16**(3): 398-409.

Chakrapani, S., P. Sompornpisut, et al. "The activated state of a sodium channel voltage sensor in a membrane environment." Proceedings of the National Academy of Sciences of the United States of America **107**(12): 5435-5440.

Chiang, Y. W., P. P. Borbat, et al. (2005). "The determination of pair distance distributions by pulsed ESR using Tikhonov regularization." J Magn Reson **172**(2): 279-295.

Chiang, Y. W., P. P. Borbat, et al. (2005). The determination of pair distance distributions by pulsed ESR using Tikhonov regularization, Academic Press Inc Elsevier Science.

Choe, H. W., Y. J. Kim, et al. (2011). "Crystal structure of metarhodopsin II." Nature **471**(7340): 651-655.

Claxton, D. P., M. Quick, et al. "Ion/substrate-dependent conformational dynamics of a bacterial homolog of neurotransmitter:sodium symporters." Nat Struct Mol Biol **17**(7): 822-829.

Coleman, D. E., A. M. Berghuis, et al. (1994). "Structures of active conformations of Gi alpha 1 and the mechanism of GTP hydrolysis." Science **265**(5177): 1405-1412.

Coleman, D. E. and S. R. Sprang (1998). "Crystal structures of the G protein Gi alpha 1 complexed with GDP and Mg2+: a crystallographic titration experiment." Biochemistry **37**(41): 14376-14385.

Columbus, L., T. Kalai, et al. (2001). "Molecular motion of spin labeled side chains in alpha-helices: Analysis by variation of side chain structure." Biochemistry **40**(13): 3828-3846.

Czogalla, A., A. Pieciul, et al. (2007). "Attaching a spin to a protein - site-directed spin labeling in structural biology." Acta Biochimica Polonica **54**(2): 235-244.

Dalmas, O., L. G. Cuello, et al. "Structural Dynamics of the Magnesium-Bound Conformation of CorA in a Lipid Bilayer." Structure **18**(7): 868-878.

Daura, X. (2006). "Molecular dynamics simulation of peptide folding." Theoretical Chemistry Accounts **116**(1-3): 297-306.

Dinner, A. R. (2000). "Local deformations of polymers with nonplanar rigid main-chain internal coordinates." Journal of Computational Chemistry **21**(13): 1132-1144.

Dong, J., G. Yang, et al. (2005). "Structural basis of energy transduction in the transport cycle of MsbA." Science **308**(5724): 1023-1028.

Doyle, D. A., J. M. Cabral, et al. (1998). "The Structure of the Potassium Channel: Molecular Basis of K+ Conduction and Selectivity." Science **280**(5360): 69-77.

Dratz, E. A., J. E. Furstenau, et al. (1993). "NMR structure of a receptor-bound G-protein peptide." Nature **363**(6426): 276-281.

Duan, J. X., S. L. Dixon, et al. "Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods." Journal of Molecular Graphics & Modelling **29**(2): 157-170.

Dunbrack, R. L. (2002). "Rotamer libraries in the 21(st) century." Current Opinion in Structural Biology **12**(4): 431-440.

Dunbrack, R. L., Jr. and M. Karplus (1993). "Backbone-dependent rotamer library for proteins. Application to side-chain prediction." J Mol Biol **230**(2): 543-574.

Durham, E., B. Dorr, et al. (2009). "Solvent accessible surface area approximations for rapid and accurate protein structure prediction." J Mol Model.

Dutzler, R., E. B. Campbell, et al. (2002). "X-ray structure of a ClC chloride channel at 3.0[thinsp][angst] reveals the molecular basis of anion selectivity." Nature **415**(6869): 287-294.

Eisenberg, D., R. M. Weiss, et al. (1984). "The Hydrophobic Moment Detects Periodicity in Protein Hydrophobicity." Proceedings of the National Academy of Sciences of the United States of America **81**(1): 140-144.

Eshaghi, S., D. Niegowski, et al. (2006). "Crystal Structure of a Divalent Metal Ion Transporter CorA at 2.9 Angstrom Resolution." Science **313**(5785): 354-357.

Faham, S., D. Yang, et al. (2004). "Side-chain Contributions to Membrane Protein Structure and Stability." Journal of Molecular Biology **335**(1): 297-305.

Fajer, M. I., H. Z. Li, et al. (2007). "Mapping electron paramagnetic resonance spin label conformations by the simulated scaling method." Journal of the American Chemical Society **129**(45): 13840-13846.

Fanucci, G. E. and D. S. Cafiso (2006). "Recent advances and applications of site-directed spin labeling." Curr Opin Struct Biol **16**(5): 644-653.

Farahbakhsh, Z. T., C. Altenbach, et al. (1992). "Spin labeled cysteines as sensors for protein-lipid interaction and conformation in rhodopsin." Photochem Photobiol **56**(6): 1019-1033.

Fenwick, R. B., S. Esteban-Martin, et al. (2011). "Understanding biomolecular motion, recognition, and allostery by use of conformational ensembles." European Biophysics Journal with Biophysics Letters **40**(12): 1339-1355.

Finn, R. D., J. Mistry, et al. (2006). "Pfam: clans, web tools and services." Nucleic Acids Res **34**(Database issue): D247-251.

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

Fleishman, S. J., S. E. Harrington, et al. (2006). "Quasi-symmetry in the cryo-EM structure of EmrE provides the key to modeling its transmembrane domain." J Mol Biol **364**(1): 54-67.

Fleissner, M. R., D. Cascio, et al. (2009). "Structural origin of weakly ordered nitroxide motion in spin-labeled proteins." Protein Science **18**(5): 893-908.

Francis, D. M., B. Różycki, et al. (2011). "Structural basis of p38α regulation by hematopoietic tyrosine phosphatase." Nat Chem Biol **7**(12): 916-924.

Frishman, D. and P. Barth (2010). Prediction of three-dimensional transmembrane helical protein structures. Structural Bioinformatics of Membrane Proteins, Springer Vienna**:** 231-249.

Ganguly, S., Brian E. Weiner, et al. "Membrane Protein Structure Determination using Paramagnetic Tags." Structure **19**(4): 441-443.

Gautier, A., H. R. Mott, et al. (2010). "Structure determination of the seven-helix transmembrane receptor sensory rhodopsin II by solution NMR spectroscopy." Nat Struct Mol Biol **17**(6): 768-774.

Ginalski, K., A. Elofsson, et al. (2003). "3D-Jury: a simple approach to improve protein structure predictions." Bioinformatics **19**(8): 1015-1018.

Ginalski, K., M. von Grotthuss, et al. (2004). "Detecting distant homology with Meta-BASIC." Nucleic Acids Res **32**(Web Server issue): W576-581.

Godden, J. W., L. Xue, et al. (2000). "Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients." Journal of Chemical Information and Computer Sciences **40**(1): 163-166.

Gray, J. J., S. Moughon, et al. (2003). "Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations." Journal of Molecular Biology **331**(1): 281-299.

Grossfield, A. (2011). "Recent progress in the study of G protein-coupled receptors with molecular dynamics computer simulations." Biochimica et Biophysica Acta (BBA) - Biomembranes **1808**(7): 1868-1878.

Guo, Z. F., D. Cascio, et al. (2008). "Structural determinants of nitroxide motion in spin-labeled proteins: Solvent-exposed sites in helix B of T4 lysozyme." Protein Science **17**(2): 228-239.

Guo, Z. F., D. Cascio, et al. (2007). "Structural determinants of nitroxide motion in spin-labeled proteins: Tertiary contact and solvent-inaccessible sites in helix G of T4 lysozyme." Protein Science **16**(6): 1069-1086.

Haley, D. A., M. P. Bova, et al. (2000). "Small heat-shock protein structures reveal a continuum from symmetric to variable assemblies." J Mol Biol **298**(2): 261-272.

Hamm, H. E. (2001). "How activated receptors couple to G proteins." Proc Natl Acad Sci U S A **98**(9): 4819-4821.

Hamm, H. E., D. Deretic, et al. (1988). "Site of G protein binding to rhodopsin mapped with synthetic peptides from the alpha subunit." Science **241**(4867): 832-835.

Harrison, S. C. (2004). "Whither structural biology?" Nat Struct Mol Biol **11**(1): 12-15.

Herrmann, R., M. Heck, et al. (2006). "Signal transfer from GPCRs to G proteins: role of the G alpha N-terminal region in rhodopsin-transducin coupling." J Biol Chem **281**(40): 30234-30241.

Higashijima, T., K. M. Ferguson, et al. (1987). "Effects of Mg2+ and the beta gamma-subunit complex on the interactions of guanine nucleotides with G proteins." J Biol Chem **262**(2): 762-766.

Hilger, D., Y. Polyhach, et al. (2009). "Backbone Structure of Transmembrane Domain IX of the Na+/Proline Transporter PutP of Escherichia coli." Biophysical Journal **96**(1): 217-225.

Hirsch, J. A., C. Schubert, et al. (1999). "The 2.8 A crystal structure of visual arrestin: a model for arrestin's regulation." Cell **97**(2): 257-269.

Hirst, S. J., N. Alexander, et al. (2011). "RosettaEPR: an integrated tool for protein structure determination from sparse EPR data." J Struct Biol **173**(3): 506-514.

Hofmann, K. P., P. Scheerer, et al. (2009). "A G protein-coupled receptor at work: the rhodopsin model." Trends in Biochemical Sciences **34**(11): 540-552.

Horwitz, J. (1992). "Alpha-crystallin can function as a molecular chaperone." Proc Natl Acad Sci U S A **89**(21): 10449-10453.

Horwitz, J. (1993). "Proctor Lecture. The function of alpha-crystallin." Invest Ophthalmol Vis Sci **34**(1): 10-22.

Hubbell, W. L. and C. Altenbach (1994). "INVESTIGATION OF STRUCTURE AND DYNAMICS IN MEMBRANE-PROTEINS USING SITE-DIRECTED SPIN-LABELING." Current Opinion in Structural Biology **4**(4): 566-573.

Hubbell, W. L., H. S. McHaourab, et al. (1996). "Watching proteins move using site-directed spin labeling." Structure **4**(7): 779-783.

Isas, J. M., R. Langen, et al. (2004). "Structure and dynamics of a helical hairpin that mediates calcium-dependent membrane binding of annexin B12." Journal of Biological Chemistry **279**(31): 32492-32498.

Jeschke, G. (2002). "Distance measurements in the nanometer range by pulse EPR." Chemphyschem **3**(11): 927-932.

Jeschke, G., A. Bender, et al. (2004). "Sensitivity enhancement in pulse EPR distance measurements." Journal of Magnetic Resonance **169**(1): 1-12.

Jeschke, G. and Y. Polyhach (2007). "Distance measurements on spin-labelled biomacromolecules by pulsed electron paramagnetic resonance." Physical Chemistry Chemical Physics **9**(16): 1895-1910.

Johnson, S. C. (1967). "HIERARCHICAL CLUSTERING SCHEMES." Psychometrika **32**(3): 241-254.

Jones, D. T. (1999). "Protein secondary structure prediction based on position-specific scoring matrices." J Mol Biol **292**(2): 195-202.

Kahsay, R. Y., G. Gao, et al. (2005). "An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes." Bioinformatics **21**(9): 1853-1858.

Kamarainen, J.-K., V. Kyrki, et al. (2003). "Improving similarity measures of histograms using smoothing projections." Pattern Recogn. Lett. **24**(12): 2009-2019.

Kang, C. and Q. Li (2011). "Solution NMR study of integral membrane proteins." Current Opinion in Chemical Biology **15**(4): 560-569.

Karakas, M., N. Woetzel, et al. (2012). "BCL::Fold - De novo prediction of complex and large protein topologies by assembly of secondary structure elements." PLoS Biol (submitted).

Karplus, K., R. Karchin, et al. (2003). "Combining local-structure, fold-recognition, and new fold methods for protein structure prediction." Proteins **53 Suppl 6**: 491-496.

Karplus, K., K. Sjolander, et al. (1997). "Predicting protein structure using hidden Markov models." Proteins **Suppl 1**: 134-139.

Karplus, M. and J. A. McCammon (2002). "Molecular dynamics simulations of biomolecules." Nature Structural Biology **9**(9): 646-652.

Kaufmann, K. W., G. H. Lemmon, et al. (2010). "Practically useful: what the Rosetta protein modeling suite can do for you." Biochemistry **49**(14): 2987-2998.

Kazmier, K., N. S. Alexander, et al. "Algorithm for selection of optimized EPR distance restraints for de novo protein structure determination." Journal of Structural Biology **173**(3): 549-557.

Kelley, L. A., R. M. MacCallum, et al. (2000). "Enhanced genome annotation using structural profiles in the program 3D-PSSM." J Mol Biol **299**(2): 499-520.

Kim, H. J., S. C. Howell, et al. (2009). "Recent advances in the application of solution NMR spectroscopy to multi-span integral membrane proteins." Progress in Nuclear Magnetic Resonance Spectroscopy **55**(4): 335-360.

Kim, K. K., R. Kim, et al. (1998). "Crystal structure of a small heat-shock protein." Nature **394**(6693): 595-599.

Kim, M., S. A. Vishnivetskiy, et al. (submitted). "The conformation of receptor-bound visual arrestin." Proc. Natl. Acad. Sci. U.S.A.

Kisselev, O. G., J. Kao, et al. (1998). "Light-activated rhodopsin induces structural binding motif in G protein alpha subunit." Proc Natl Acad Sci U S A **95**(8): 4270-4275.

Kortemme, T., D. E. Kim, et al. (2004). "Computational alanine scanning of protein-protein interfaces." Sci STKE **2004**(219): pl2.

Koteiche, H. A., A. R. Berengian, et al. (1998). "Identification of protein folding patterns using site-directed spin labeling. Structural characterization of a beta-sheet and putative substrate binding regions in the conserved domain of alpha A-crystallin." Biochemistry **37**(37): 12681-12688.

Koteiche, H. A. and H. S. McHaourab (1999). "Folding pattern of the alpha-crystallin domain in alphaA-crystallin determined by site-directed spin labeling." J Mol Biol **294**(2): 561-577.

Krissinel, E. B. and K. Henrick (2004). "Common subgraph isomorphism detection by backtracking search." Software-Practice & Experience **34**(6): 591-607.

Krivov, G. G., M. V. Shapovalov, et al. (2009). "Improved prediction of protein side-chain conformations with SCWRL4." Proteins: Structure, Function, and Bioinformatics **77**(4): 778-795.

Kroncke, B. M., P. S. Horanyi, et al. (2010). "Structural Origins of Nitroxide Side Chain Dynamics on Membrane Protein Î±-Helical Sites." Biochemistry **49**(47): 10045-10060.

Kuhlman, B. and D. Baker (2000). "Native protein sequences are close to optimal for their structures." Proc. Natl. Acad. Sci. U. S. A. **97**(19): 10383-10388.

Kuhlman, B. and D. Baker (2000). "Native protein sequences are close to optimal for their structures." Proceedings of the National Academy of Sciences of the United States of America **97**(19): 10383-10388.

Kuhlman, B., G. Dantas, et al. (2003). "Design of a novel globular protein fold with atomic-level accuracy." Science **302**(5649): 1364-1368.

Lambright, D. G., J. P. Noel, et al. (1994). "STRUCTURAL DETERMINANTS FOR ACTIVATION OF THE ALPHA-SUBUNIT OF A HETEROTRIMERIC G-PROTEIN." Nature **369**(6482): 621-628.

Lambright, D. G., J. Sondek, et al. (1996). "The 2.0 angstrom crystal structure of a heterotrimeric G protein." Nature **379**(6563): 311-319.

Langen, R., K. J. Oh, et al. (2000). "Crystal structures of spin labeled T4 lysozyme mutants: implications for the interpretation of EPR spectra in terms of structure." Biochemistry **39**(29): 8396-8405.

Leach, A. R., V. J. Gillet, et al. "Three-Dimensional Pharmacophore Methods in Drug Discovery." Journal of Medicinal Chemistry **53**(2): 539-558.

Lesk, A. M. (1997). "Extraction of well-fitting substructures: Root-mean-square deviation and the difference distance matrix." Folding & Design **2**(3): S12-S14.

Lesley, S. A., P. Kuhn, et al. (2002). "Structural genomics of the Thermotoga maritima proteome implemented in a high-throughput structure determination pipeline." Proc Natl Acad Sci U S A **99**(18): 11664-11669.

Li, J., P. C. Edwards, et al. (2004). "Structure of Bovine Rhodopsin in a Trigonal Crystal Form." Journal of Molecular Biology **343**(5): 1409-1438.

Lietzow, M. A. and W. L. Hubbell (2004). "Motion of Spin Label Side Chains in Cellular Retinol-Binding Protein:â€‰ Correlation with Structure and Nearest-Neighbor Interactions in an Antiparallel Î²-Sheetâ€ " Biochemistry **43**(11): 3137-3151.

Linder, M. E., I. H. Pang, et al. (1991). "Lipid modifications of G protein subunits. Myristoylation of Go alpha increases its affinity for beta gamma." J Biol Chem **266**(7): 4654-4659.

Lindert, S., N. Alexander, et al. (2012). "EM-Fold: De Novo Atomic-Detail Protein Structure Determination from Medium-Resolution Density Maps." Structure **20**(3): 464-478.

Lindert, S., R. Staritzbichler, et al. (2009). "EM-Fold: De Novo Folding of α-Helical Proteins Guided by Intermediate-Resolution Electron Microscopy Density Maps." Structure **17**(7): 990-1003.

Lindert, S., P. L. Stewart, et al. (2009). "Hybrid approaches: applying computational methods in cryo-electron microscopy." Current Opinion in Structural Biology **19**(2): 218-225.

Liu, Y. S., P. Sompornpisut, et al. (2001). "Structure of the KcsA channel intracellular gate in the open state." Nat Struct Biol **8**(10): 883-887.

Lovell, S. C., J. M. Word, et al. (2000). "The penultimate rotamer library." Proteins-Structure Function and Genetics **40**(3): 389-408.

Madej, M. G., H. R. Nasiri, et al. (2006). "Evidence for transmembrane proton transfer in a dihaem-containing membrane protein complex." Embo Journal **25**(20): 4963-4970.

Maiorov, V. N. and G. M. Crippen (1994). "SIGNIFICANCE OF ROOT-MEAN-SQUARE DEVIATION IN COMPARING 3-DIMENSIONAL STRUCTURES OF GLOBULAR-PROTEINS." Journal of Molecular Biology **235**(2): 625-634.

Mandell, D. J., E. A. Coutsias, et al. (2009). "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling." Nat Methods **6**(8): 551-552.

Marin, E. P., A. G. Krishna, et al. (2002). "Disruption of the alpha5 helix of transducin impairs rhodopsin-catalyzed nucleotide exchange." Biochemistry **41**(22): 6988-6994.

Martin, E. L., S. Rens-Domiano, et al. (1996). "Potent peptide analogues of a G protein receptor-binding region obtained with a combinatorial library." J Biol Chem **271**(1): 361-366.

Maslennikov, I., C. Klammt, et al. (2010). "Membrane domain structures of three classes of histidine kinase receptors by cell-free expression and rapid NMR analysis." Proceedings of the National Academy of Sciences.

Matsumura, M. and B. W. Matthews (1989). "CONTROL OF ENZYME-ACTIVITY BY AN ENGINEERED DISULFIDE BOND." Science **243**(4892): 792-794.

McGuffin, L. J. and D. T. Jones (2003). "Improvement of the GenTHREADER method for genomic fold recognition." Bioinformatics **19**(7): 874-881.

McHaourab, H. S., A. R. Berengian, et al. (1997). "Site-directed spin-labeling study of the structure and subunit interactions along a conserved sequence in the alpha-crystallin domain of heat-shock protein 27. Evidence of a conserved subunit interface." Biochemistry **36**(48): 14627-14634.

McHaourab, H. S., M. A. Lietzow, et al. (1996). "Motion of spin-labeled side chains in T4 lysozyme. Correlation with protein structure and dynamics." Biochemistry **35**(24): 7692-7704.

Medkova, M., A. M. Preininger, et al. (2002). "Conformational changes in the amino-terminal helix of the G protein alpha(i1) following dissociation from Gbetagamma subunit and activation." Biochemistry **41**(31): 9962-9972.

Meiler, J. (2003). "JUFO3D: Coupled Prediction of Protein Secondary and Tertiary Structure (server)." www.meilerlab.org.

Meiler, J. and D. Baker (2003). "Coupled Prediction of Protein Secondary and Tertiary Structure." PNAS **100**(21): 12105-12110.

Meiler, J. and D. Baker (2003). "Rapid protein fold determination using unassigned NMR data." Proceedings of the National Academy of Sciences of the United States of America **100**(26): 15404-15409.

Meiler, J., P. Bradley, et al. (2003). ROSETTA in CASP5: Progress in de novo protein structure prediction (poster). Molecular Modeling Workshop. Erlangen, Bavaria, Germany.

Meiler, J., M. Müller, et al. (2001). "Generation and Evaluation of Dimension Reduced Amino Acid Parameter Representations by Artificial Neural Networks." J. Mol. Model. **7**(9): 360-369.

Metropolis, N. R., A.; Rosenbluth, M.; Teller A. (1953). "Equations of state calculations by fast computing machines." Journal of Chemical Physics **21**: 1087 - 1091.

Millar, R. P. and C. L. Newton "The Year In G Protein-Coupled Receptor Research." Molecular Endocrinology **24**(1): 261-274.

Millhauser, G. L., W. R. Fiori, et al. (1995). [24] Electron spin labels. Methods in Enzymology. S. Kenneth, Academic Press. **Volume 246:** 589-610.

Misura, K. M. and D. Baker (2005). "Progress and challenges in high-resolution refinement of protein structure models." Proteins **59**(1): 15-29.

Misura, K. M., D. Chivian, et al. (2006). "Physically realistic homology models built with ROSETTA can be more accurate than their templates." Proc Natl Acad Sci U S A **103**(14): 5361-5366.

MOE MOE (Molecular Operating Environment). Montreal, Quebec, Canada, Chemical Computing Group Inc.

Mueller, R., A. L. Rodriguez, et al. "Identification of Metabotropic Glutamate Receptor Subtype 5 Potentiators Using Virtual High-Throughput Screening." Acs Chemical Neuroscience **1**(4): 288-305.

Murakami, S., R. Nakashima, et al. (2002). "Crystal structure of bacterial multidrug efflux transporter AcrB." Nature **419**(6907): 587-593.

Murata, T., I. Yamato, et al. (2005). "Structure of the Rotor of the V-Type Na+-ATPase from Enterococcus hirae." Science **308**(5722): 654-659.

Nakamura, T., S. Mine, et al. (2008). "Tertiary structure and carbohydrate recognition by the chitin-binding domain of a hyperthermophilic chitinase from Pyrococcus furiosus." Journal of Molecular Biology **381**(3): 670-680.

Natochin, M., M. Moussaif, et al. (2001). "Probing the mechanism of rhodopsin-catalyzed transducin activation." J Neurochem **77**(1): 202-210.

Nederveen, A. J., J. F. Doreleijers, et al. (2005). "RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank." Proteins **59**(4): 662-672.

Noel, J. P., H. E. Hamm, et al. (1993). "THE 2.2-ANGSTROM CRYSTAL-STRUCTURE OF TRANSDUCIN-ALPHA COMPLEXED WITH GTP-GAMMA-S." Nature **366**(6456): 654-663.

Okada, T., M. Sugihara, et al. (2004). "The Retinal Conformation and its Environment in Rhodopsin in Light of a New 2.2Å Crystal Structure." Journal of Molecular Biology **342**(2): 571-583.

Oldham, W. M. and H. E. Hamm (2007). How do receptors activate g proteins? Mechanisms and Pathways of Heterotrimeric G Protein Signaling. San Diego, Elsevier Academic Press Inc. **74:** 67-93.

Oldham, W. M., N. Van Eps, et al. (2006). "Mechanism of the receptor-catalyzed activation of heterotrimeric G proteins." Nat Struct Mol Biol **13**(9): 772-777.

Oldham, W. M., N. Van Eps, et al. (2007). "Mapping allosteric connections from the receptor to the nucleotide-binding pocket of heterotrimeric G proteins." Proc Natl Acad Sci U S A **104**(19): 7927-7932.

Omran, M. G. H., A. P. Engelbrecht, et al. (2007). "An overview of clustering methods." Intelligent Data Analysis **11**(6): 583-605.

Overington, J. P., B. Al-Lazikani, et al. (2006). "How many drug targets are there?" Nat Rev Drug Discov **5**(12): 993-996.

Palczewski, K., T. Kumasaka, et al. (2000). "Crystal structure of rhodopsin: A G protein-coupled receptor." Science **289**(5480): 739-745.

Pannier, M., S. Veit, et al. (2000). "Dead-time free measurement of dipole-dipole interactions between electron spins." Journal of Magnetic Resonance **142**(2): 331-340.

Park, J. H., P. Scheerer, et al. (2008). "Crystal structure of the ligand-free G-protein-coupled receptor opsin." Nature **454**(7201): 183-187.

Perozo, E., D. M. Cortes, et al. (1998). "Three-dimensional architecture and gating mechanism of a K+ channel studied by EPR spectroscopy." Nature Structural Biology **5**(6): 459-469.

Perozo, E., D. M. Cortes, et al. (1999). "Structural rearrangements underlying K+-channel activation gating." Science **285**(5424): 73-78.

Polyhach, Y., E. Bordignon, et al. "Rotamer libraries of spin labelled cysteines for protein studies." Physical Chemistry Chemical Physics **13**(6): 2356-2366.

Preininger, A. M., M. A. Funk, et al. (2009). "Helix dipole movement and conformational variability contribute to allosteric GDP release in Galphai subunits." Biochemistry **48**(12): 2630-2642.

Preininger, A. M., J. Parello, et al. (2008). "Receptor-mediated changes at the myristoylated amino terminus of Galpha(il) proteins." Biochemistry **47**(39): 10281-10293.

Priestle, J. P. (2009). "3-D clustering: a tool for high throughput docking." Journal of Molecular Modeling **15**(5): 551-560.

PyMOL The PyMOL Molecular Graphics System, Version 1.2r1, Schrödinger, LLC.

Qian, B., S. Raman, et al. (2007). "High-resolution structure prediction and the crystallographic phase problem." Nature **450**(7167): 259-264.

Rabenstein, M. D. and Y. K. Shin (1995). "Determination of the distance between two spin labels attached to a macromolecule." Proc Natl Acad Sci U S A **92**(18): 8239-8243.

Raman, S., O. F. Lange, et al. "NMR Structure Determination for Larger Proteins Using Backbone-Only Data." Science: science.1183649.

Raman, S., R. Vernon, et al. (2009). "Structure prediction for CASP8 with all-atom refinement using Rosetta." Proteins: Structure, Function, and Bioinformatics **77**(S9): 89-99.

Rao, S. T. and M. G. Rossmann (1973). "COMPARISON OF SUPER-SECONDARY STRUCTURES IN PROTEINS." Journal of Molecular Biology **76**(2): 241-&.

Rasmussen, S. G., B. T. Devree, et al. (2011). "Crystal structure of the beta(2) adrenergic receptor-Gs protein complex." Nature.

Resh, M. D. (1999). "Fatty acylation of proteins: new insights into membrane targeting of myristoylated and palmitoylated proteins." Biochim Biophys Acta **1451**(1): 1-16.

Rohl, C. A. and D. Baker (2002). "De novo determination of protein backbone structure from residual dipolar couplings using rosetta." Journal of the American Chemical Society **124**(11): 2723-2729.

Rohl, C. A., C. E. Strauss, et al. (2004). "Protein structure prediction using Rosetta." Methods Enzymol **383**: 66-93.

Rost, B. (2001). "Review: protein secondary structure prediction continues to rise." J Struct Biol **134**(2-3): 204-218.

Rost, B. and C. Sander (1993). "Improved prediction of protein secondary structure by use of sequence profiles and neural networks." Proc. Natl. Acad. Sci. USA **90**: 7558-7562.

Rost, B., G. Yachdav, et al. (2004). "The PredictProtein server." Nucleic Acids Res **32** (Web Server issue): W321-326.

Sale, K., C. Sar, et al. (2002). "Structural determination of spin label immobilization and orientation: A Monte Carlo minimization approach." Journal of Magnetic Resonance **156**(1): 104-112.

Sale, K., L. Song, et al. (2005). "Explicit treatment of spin labels in modeling of distance constraints from dipolar EPR and DEER." J Am Chem Soc **127**(26): 9334-9335.

Sali, A. (1998). "100,000 protein structures for the biologist." Nature structural biology **5**(12)): 1029-1032.

Salwinski, L. and W. L. Hubbell (1999). "Structure in the channel forming domain of colicin E1 bound to membranes: The 402-424 sequence." Protein Science **8**(3): 562-572.

Sastry, M., J. F. Lowrie, et al. "Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments." Journal of Chemical Information and Modeling **50**(5): 771-784.

Scheerer, P., M. Heck, et al. (2009). "Structural and kinetic modeling of an activating helix switch in the rhodopsin-transducin interface." Proceedings of the National Academy of Sciences of the United States of America **106**(26): 10660-10665.

Scheerer, P., J. H. Park, et al. (2008). "Crystal structure of opsin in its G-protein-interacting conformation." Nature **455**(7212): 497-502.

Serna, A. (1996). "Implementation of hierarchical clustering methods." Journal of Computational Physics **129**(1): 30-40.

Sgourakis, N. G. and A. E. Garcia (2010). "The Membrane Complex between Transducin and Dark-State Rhodopsin Exhibits Large-Amplitude Interface Dynamics on the Sub-Microsecond Timescale: Insights from All-Atom MD Simulations." Journal of Molecular Biology **398**(1): 161-173.

Shapovalov, Maxim V. and Roland L. Dunbrack (2011). "A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions." Structure (London, England : 1993) **19**(6): 844-858.

Shi, J., T. L. Blundell, et al. (2001). "FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties." J Mol Biol **310**(1): 243-257.

Shinitzky, M. and B. Rivnay (1977). "Degree of exposure of membrane proteins determined by fluorescence quenching." Biochemistry **16**(5): 982-986.

Shortle, D., K. T. Simons, et al. (1998). "Clustering of low-energy conformations near the native structures of small proteins." Proceedings of the National Academy of Sciences of the United States of America **95**(19): 11158-11162.

Siew, N., A. Elofsson, et al. (2000). "MaxSub: an automated measure for the assessment of protein structure prediction quality." Bioinformatics **16**(9): 776-785.

Simons, K. T., C. Kooperberg, et al. (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., I. Ruczinski, et al. (1999). "Improved Recognition of Native-Like Protein Structures Using a Combination of Sequence-Dependent and Sequence-Independent Features of Proteins." Proteins: Structure, Function, and Genetics **34**: 82-95.

Sippl, M. J. (1995). "Knowledge-based potentials for proteins." Curr Opin Struct Biol **5**(2): 229-235.

Skolnick, J., A. Kolinski, et al. (2001). "Ab initio protein structure prediction via a combination of threading, lattice folding, clustering, and structure refinement." Proteins-Structure Function and Genetics: 149-156.

Soding, J., A. Biegert, et al. (2005). "The HHpred interactive server for protein homology detection and structure prediction." Nucleic Acids Res **33**(Web Server issue): W244-248.

Sompornpisut, P., H. Mchaourab, et al. (2002). "Spectroscopically- determined solvent accessibilities as constraints for global fold discrimination in proteins." Biophysical Journal **82**(1): 474.

Stamler, R., G. Kappe, et al. (2005). "Wrapping the alpha-crystallin domain fold in a chaperone assembly." J Mol Biol **353**(1): 68-79.

Steinbacher, S., R. Bass, et al. (2007). Structures of the Prokaryotic Mechanosensitive Channels MscL and MscS. Current Topics in Membranes. P. H. Owen, Academic Press. **Volume 58:** 1-24.

Stevens, R. C., S. Yokoyama, et al. (2001). "Global efforts in structural genomics." Science **294**(5540): 89-92.

Sui, H., B.-G. Han, et al. (2001). "Structural basis of water-specific transport through the AQP1 water channel." Nature **414**(6866): 872-878.

Tate, C. G. (2010). Practical Considerations of Membrane Protein Instability during Purification and Crystallisation. Heterologous Expression of Membrane Proteins: Methods and Protocols. I. MusVeteau, Humana Press Inc, 999 Riverview Dr, Ste 208, Totowa, Nj 07512-1165 USA. **601:** 187-203.

Tate, C. G. and G. F. X. Schertler (2009). "Engineering G protein-coupled receptors to facilitate their structure determination." Current Opinion in Structural Biology **19**(4): 386-395.

Tesmer, J. J. G. (2010). "The quest to understand heterotrimeric G protein signaling." Nature Structural & Molecular Biology **17**(6): 650-652.

Tombolato, F., A. Ferrarini, et al. (2006). "Dynamics of the nitroxide side chain in spin-labeled proteins." Journal of Physical Chemistry B **110**(51): 26248-26259.

Tombolato, F., A. Ferrarini, et al. (2006). "Modeling the effects of structure and dynamics of the nitroxide side chain on the ESR spectra of spin-labeled proteins." Journal of Physical Chemistry B **110**(51): 26260-26271.

Tsai, J., R. Bonneau, et al. (2003). "An improved protein decoy set for testing energy functions for protein structure prediction." Proteins-Structure Function and Genetics **53**(1): 76-87.

Tsukihara, T., H. Aoyama, et al. (1996). "The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 angstrom." Science **272**(5265): 1136-1144.

Tusnady, G. E., Z. Dosztanyi, et al. (2004). "Transmembrane proteins in the Protein Data Bank: identification and classification." Bioinformatics **20**(17): 2964-2972.

Ulmschneider, J. P. and W. L. Jorgensen (2003). "Monte Carlo backbone sampling for polypeptides with variable bond angles and dihedral angles using concerted rotations and a Gaussian bias." Journal of Chemical Physics **118**(9): 4261-4271.

Unwin, N. (2005). "Refined Structure of the Nicotinic Acetylcholine Receptor at 4 Å Resolution." Journal of Molecular Biology **346**(4): 967-989.

Van Eps, N., L. L. Anderson, et al. (2010). "Electron paramagnetic resonance studies of functionally active, nitroxide spin-labeled peptide analogues of the C-terminus of a G-protein alpha subunit." Biochemistry **49**(32): 6877-6886.

Van Eps, N., W. M. Oldham, et al. (2006). "Structural and dynamical changes in an alpha-subunit of a heterotrimeric G protein along the activation pathway." Proc Natl Acad Sci U S A **103**(44): 16194-16199.

Van Eps, N., A. M. Preininger, et al. "Interaction of a G protein with an activated receptor opens the interdomain interface in the alpha subunit." Proceedings of the National Academy of Sciences.

van Montfort, R. L., E. Basha, et al. (2001). "Crystal structure and assembly of a eukaryotic small heat shock protein." Nat Struct Biol **8**(12): 1025-1030.

Vásquez, V., M. Sotomayor, et al. (2008). "Three-Dimensional Architecture of Membrane-Embedded MscS in the Closed Conformation." Journal of Molecular Biology **378**(1): 55-70.

Viklund, H. and A. Elofsson (2008). "OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar." Bioinformatics **24**(15): 1662-1668.

Wall, M. A., D. E. Coleman, et al. (1995). "The structure of the G protein heterotrimer Gi alpha 1 beta 1 gamma 2." Cell **83**(6): 1047-1058.

Wang, C., P. Bradley, et al. (2007). "Protein-protein docking with backbone flexibility." Journal of Molecular Biology **373**(2): 503-519.

Wang, C., O. Schueler-Furman, et al. (2005). "Improved side-chain modeling for protein–protein docking." Protein Science **14**(5): 1328-1339.

Wang, G. and R. L. Dunbrack, Jr. (2003). "PISCES: a protein sequence culling server." Bioinformatics **19**(12): 1589-1591.

Wang, Y., Y. Huang, et al. (2009). "Structure of the formate transporter FocA reveals a pentameric aquaporin-like channel." Nature **462**(7272): 467-472.

Wang, Y., Y. Zhang, et al. (2006). "Crystal structure of a rhomboid family intramembrane protease." Nature **444**(7116): 179-180.

Ward, A., C. L. Reyes, et al. (2007). "Flexibility in the ABC transporter MsbA: Alternating access with a twist." Proc Natl Acad Sci U S A **104**(48): 19005-19010.

Weaver, L. H. and B. W. Matthews (1987). "Structure of bacteriophage T4 lysozyme refined at 1.7 A resolution." J Mol Biol **193**(1): 189-199.

Westbrook, J., Z. Feng, et al. (2003). "The Protein Data Bank and structural genomics." Nucleic Acids Res **31**(1): 489-491.

Westfield, G. H., S. G. Rasmussen, et al. (2011). "Structural flexibility of the G alpha s alpha-helical domain in the beta2-adrenoceptor Gs complex." Proc Natl Acad Sci U S A **108**(38): 16086-16091.

Willett, P. (2006). "Similarity-based virtual screening using 2D fingerprints." Drug Discovery Today **11**(23-24): 1046-1053.

Willett, P., J. M. Barnard, et al. (1998). "Chemical similarity searching." Journal of Chemical Information and Computer Sciences **38**(6): 983-996.

Willett, P., V. Winterman, et al. (1986). "IMPLEMENTATION OF NON-HIERARCHICAL CLUSTER-ANALYSIS METHODS IN CHEMICAL INFORMATION-SYSTEMS - SELECTION OF COMPOUNDS FOR BIOLOGICAL TESTING AND CLUSTERING OF SUBSTRUCTURE SEARCH OUTPUT." Journal of Chemical Information and Computer Sciences **26**(3): 109-118.

Woetzel, N., M. Karakaş, et al. (2012). "BCL::Score - Knowledge based energy potentials for ranking protein models represented by idealized secondary structure elements " PLoS Biol(submitted).

Wüthrich, K. (1986). NMR of Proteins and Nucleic Acids (1H-NMR shifts of amino acids). New York, Chichester, Brisbane, Toronto, Singapore, John Wiley & Sons.

Xu, R. and D. Wunsch (2005). "Survey of clustering algorithms." Ieee Transactions on Neural Networks **16**(3): 645-678.

Yang, Y., T. A. Ramelot, et al. (2010). "Combining NMR and EPR Methods for Homodimer Protein Structure Determination." Journal of the American Chemical Society **132**(34): 11910-11913.

Yarov-Yarovoy, V., J. Schonbrun, et al. (2006). "Multipass membrane protein structure prediction using Rosetta." Proteins-Structure Function and Bioinformatics **62**(4): 1010-1025.

Yongye, A. B., A. Bender, et al. "Dynamic clustering threshold reduces conformer ensemble size while maintaining a biologically relevant ensemble." Journal of Computer-Aided Molecular Design **24**(8): 675-686.

Zemla, A. (2003). "LGA: a method for finding 3D similarities in protein structures." Nucleic Acids Research **31**(13): 3370-3374.

Zhang, P., J. Wang, et al. (2010). "Structure and mechanism of the S component of a bacterial ECF transporter." Nature **468**(7324): 717-720.

Zhang, Y. (2008). "Progress and challenges in protein structure prediction." Current Opinion in Structural Biology **18**(3): 342-348.

Zhang, Y., M. E. DeVries, et al. (2006). "Structure Modeling of All Identified G Protein–Coupled Receptors in the Human Genome." PLoS Comput Biol **2**(2): e13.

Zhou, Y., T. Cierpicki, et al. (2008). "NMR Solution Structure of the Integral Membrane Enzyme DsbB: Functional Insights into DsbB-Catalyzed Disulfide Bond Formation." Molecular Cell **31**(6): 896-908.

Zou, P., M. Bortolus, et al. (2009). "Conformational Cycle of the ABC Transporter MsbA in Liposomes: Detailed Analysis Using Double Electron-Electron Resonance Spectroscopy." Journal of Molecular Biology **393**(3): 586-597.

Zou, P. and H. S. McHaourab (2009). "Alternating Access of the Putative Substrate-Binding Chamber in the ABC Transporter MsbA." Journal of Molecular Biology **393**(3): 574-585.

Zou, P., K. Surendhran, et al. (2007). "Distance Measurements by Fluorescence Energy Homotransfer: Evaluation in T4 Lysozyme and Correlation with Dipolar Coupling between Spin Labels." Biophysical Journal **92**(4): L27-L29.

Zwanzig, R., A. Szabo, et al. (1992). "Levinthal's Paradox." Proceedings of the National Academy of Sciences of the United States of America **89**(1): 20-22.