SANDIA REPORT

SAND2004-6383
Unlimited Release
Printed January 2005

Model-Building Codes for Membrane Proteins

W. Michael Brown, Jean-Loup Faulon, Genetha A. Gray, Thomas W. Hunt, Joseph S. Schoeniger, David Shirley, Alex Slepoy, Malin M. Young and Kenneth L. Sale

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND 2004-6383
Unlimited Release

Model-Building Codes for Membrane Proteins

W. Michael Brown[†], Jean-Loup Faulon[†], Genetha A. Gray[‡], Thomas W. Hunt[#], Joseph S. Schoeniger[§], David Shirley[¥], Alex Slepoy[#], Malin M. Young[§] and Kenneth L. Sale[§*]

[§]Biosystems Research Department
Sandia National Laboratories
Livermore, CA 94551

[‡]Computational Sciences and Mathematics Research Department
Sandia National Laboratories
Livermore, CA 94551

[†]Computational Biology
Sandia National Laboratories
Albuquerque, NM 87123

[#]Computational Materials and Molecular Biology
Sandia National Laboratories
Albuquerque, NM 87123

[¥]Scientific Computing Systems
Sandia National Laboratories
Albuquerque, NM 87123

Abstract

We have developed a novel approach to modeling the transmembrane spanning helical bundles of integral membrane proteins using only a sparse set of distance constraints, such as those derived from MS3-D, dipolar-EPR and FRET experiments. Algorithms have been written for searching the conformational space of membrane protein folds matching the set of distance constraints, which provides initial structures for local conformational searches. Local conformation search is achieved by optimizing these candidates against a custom penalty function that incorporates both measures derived from statistical analysis of solved membrane protein structures and distance constraints obtained from experiments. This results in refined helical bundles to which the interhelical loops and amino acid side-chains are added. Using a set of only 27 distance constraints extracted from the literature, our methods successfully recover the structure of dark-adapted rhodopsin to within 3.2 Å of the crystal structure.

This page intentionally left blank

Table of Contents

General Introduction

Integral membrane proteins are essential components of the cell membrane that participate in many important cellular processes such as cell intoxication and pathogenesis, energy transduction, cell signaling, mediation of senses and immune recognition. Their significance is emphasized by the fact that approximately one third of the proteins encoded for by a typical genome are membrane proteins, and approximately 70 percent of current pharmaceuticals are thought to act on membrane proteins. Despite their obvious importance, in contrast to over 27,000 soluble proteins structures, the structures of fewer than 75 integral membrane proteins have been solved. Given the difficulties, such as the instability of membrane proteins in environments lacking phospholipids, their tendency to aggregate and precipitate, and protein abundance, expression and purification issues, it is unlikely that generating high-resolution structural data from traditional methods such as X-ray crystallography and NMR will yield a significant increase in the number of solved membrane protein structures in the near future.

The Interfacial Biosciences Grand Challenge (IBIG) supported the development of algorithms for modeling the geometry of transmembrane helical bundles, as well as new theoretical approaches to protein side-chain packing and loop building. In this work we built on the successes demonstrated under IBIG by developing and validating an integrated set of software tools for membrane protein modeling. Our goal was to develop methods to model transmembrane proteins using a set of sparse distance constraints, thus leveraging the many recent advances in techniques for measuring distances within protein in their native environment, including chemical cross-linking combined with mass spectrometry (MS-3D), which was developed for membrane proteins under IBIG, site directed spin labeling combined with electron paramagnetic resonance (SDSL-EPR) and fluorescence resonance energy transfer (RET). We combined this low-to-moderate resolution structural data with constraints derived from analysis of existing membrane protein structures, such as structural rules derived from helix - helix interactions in known structures, to determine the structure of the transmembrane spanning domain. Lastly, the loop domains and side-chains are added to the structure.

This document presents the results of this work in the form of four chapters representing the following four stages of our membrane protein modeling method:

1. Complete enumeration of the membrane protein folds satisfying a set of distance constraints (Chapter 1)
2. Ranking and refining these structures using an empirical scoring function based on solved membrane protein structures (Chapter 2)
3. Optimization of the empirical scoring function (Chapter 3)
4. Addition of amino acid-side chains the backbone level model structures (Chapter 4).

Chapter 1: A deterministic algorithm for constrained enumeration of transmembrane protein folds

W. Michael Brown, Jean-Loup Faulon, Ken Sale, Joseph S. Schoeniger, Malin M. Young

Abstract

A deterministic algorithm for enumeration of transmembrane protein folds is presented. Using a set of sparse pairwise atomic distance constraints (such as those obtained from chemical cross-linking, FRET, or dipolar EPR experiments), the algorithm performs an exhaustive search of secondary structure element packing conformations distributed throughout the entire conformational space. The end result is a set of distinct protein conformations, which can be scored and refined as part of a process designed for computational elucidation of transmembrane protein structures.

Introduction

Integral membrane proteins compose roughly 20% of the total proteins encoded by the human genome[1] and play essential roles in energy transduction, cell signaling, mediation of senses, and immune recognition. Their obvious importance in human physiology and great potential as targets for pharmaceutical therapies has made structure elucidation of membrane proteins highly attractive. However, due to limited solubility, low protein abundance and expression, and sample purity issues, the efficacy of traditional structure determination by X-ray crystallography or NMR spectroscopy has been limited such that only a handful of protein structures have been solved. Therefore, attention has been given to the development of computational strategies for membrane protein structure determination.

In light of the limited number of structures available for homology modeling, Bowie proposed a four stage computational approach for structure determination based on the Popot & Engelman model[2] for helix-bundle membrane protein folding:

1. Prediction of transmembrane regions within the primary sequence
2. Construction and optimization of individual helices
3. Assembly of the helix bundle
4. Addition of interhelical loops and side-chains

Success has been achieved in the accurate prediction of transmembrane-spanning secondary structure elements[3-5] (stages 1-2), and our laboratory has demonstrated success in the development of algorithms for the optimal bundling of transmembrane helices[6-8] (stage 3). The latter effort has focused on a two-step approach which makes use of sparse pairwise distance constraints that can be determined experimentally from experiments such as NMR, chemical cross-linking, dipolar EPR, or FRET. The first step involves a search of the conformational space of membrane protein bundles to find those matching the experimental distance constraints. Second, the top-scoring helical bundles are refined using a Monte Carlo simulated annealing protocol designed for local minimization of a custom penalty function.[8] Finally, for stage 4, the addition of loops can be performed

using commercially available software such as WHATIF,[9] SCWRL,[10] and Jackal.[11]

In our previous work on optimal bundling, step 1 was performed by mapping membrane helices onto a library of helix-bundle templates calculated to represent the possible protein folds for a given number of helices.[7] For each template, every possible mapping was considered using 1˚ increments for helical axis spins. For each mapping, the distance restraints and associated experimental errors were checked and the conformation was thrown out if any were violated. The template library was calculated as described by Bowie[12] based on statistics from 45 transmembrane helices and 88 helix packing interactions. While this approach can be successful, it suffers from several important drawbacks:

1. The library generation is stochastic and reproducible results are not guaranteed.
2. The distributions from which the helix center of mass distances and packing angles are drawn are not uniform. Therefore, helix bundles with helix center of mass distances lying outside the mean will be poorly represented, for example.
3. The approach suffers from an inherent error that is impossible to assess.

The last drawback can be seen when no experimental errors are added to the distance restraints. Because the template library is a discrete set of transmembrane helix bundles, it is highly unlikely that mapping onto one of these templates will produce a protein conformation where the interatomic distances match exactly. In this case, no protein conformations will be found. Because the library is generated in a stochastic process, it is impossible to determine how much tolerance should be added to a distance restraint to account for the discrete spacing between conformations. Even when experimental error is added to the distance restraints, there is a chance that the helix bundle representing the minimum RMSD from the desired structure will be thrown out due to violation of a single distance restraint.

We have therefore developed a deterministic algorithm for enumeration of potential transmembrane helix bundles. The program is called ESSEB (Enumeration of Secondary Structure Element Bundles). Given a set of individual helices and corresponding distance restraints, the program will output a set of distinct protein conformations, which can be scored and refined in order to predict the structure of integral membrane proteins.

The ESSEB Algorithm

Overview

The ESSEB algorithm works by dividing the conformational space of each secondary structure element (SSE) into a set of cells. For each cell there is a representative conformation and for each atom in the SSE for which a distance restraint is available, there is an associated internal error. The internal error for a distance restraint is the maximum distance that the atom, when positioned in any conformation within a cell, can be from the atom in the representative conformation. The algorithm works recursively by positioning one representative conformation of an SSE. All distance restraints are checked with a tolerance that includes both the experimental and internal error. If all restraints are satisfied, every representative conformation of the next SSE is

checked, otherwise, the program moves on to the next representative conformation of the current SSE.

In addition to the distance restraints, other constraints on protein conformation can be enforced. These include the distance of closest approach between SSE axes, a restraint which prevents the cross-over of loops connecting adjacent SSEs, and a restriction on the minimum and maximum distances between axis end-points. Any protein conformation satisfying all of the restraints is enumerated for later scoring and possible refinement. Additionally, in order to make run-times feasible, a divide-and-conquer approach is used in which the cells of each SSE in an accepted protein conformation can be further divided such that the internal errors are reduced and the new representative conformations can be evaluated.

Dividing the Conformational Space

A protein conformation is considered as an arrangement of SSE axes. The SSE axis vector is calculated as described previously[13] as the eigen vector corresponding to the minimum eigen value for the inertia matrix calculated for all C-α atoms in the SSE. The endpoints of the SSE axis are the projection of the N-terminal nitrogen atom onto the axis vector and the projection of the C-terminal carbon atom onto the axis vector. For integral membrane proteins, the axis centers are restricted to coordinates in the xy-plane (the plane is intended to represent the center of the membrane bilayer). Therefore, a protein conformation consisting of n SSEs can be described by a set of n quaternions and n-1 axis centers (the first SSE is always positioned at the origin). The axis-center space for one SSE is divided by considering its position relative to another SSE. Defining d_{min} as the minimum distance from another SSE and d_{max} as the maximum distance, the axis-center space consists of the area in the bilayer plane between the concentric circles with radii d_{min} and d_{max} (Figure 1). An axis center within this area is defined by a polar angle (θ) and a distance (d). The angular space for an SSE is defined by the Euler angles ϕ, θ, and ψ where θ represents the angle with the z-axis (normal to the bilayer plane), ϕ represents the axis-angle spin about the z-axis, and ψ represents the spin about the SSE axis itself. Using these definitions, the conformation of an SSE is defined by (d, θ, ϕ, θ, ψ) where d and θ describe the axis center and ϕ, θ, and ψ describe the quaternion. The conformational space is divided into SSE cells. The range of conformational space contained within a SSE cell is given by (Δd, $\Delta\theta$, $\Delta\phi$, $\Delta\theta$, $\Delta\psi$) and each cell is represented by a SSE axis conformation lying in the “middle” of this space (see Figures 1 and 2a).

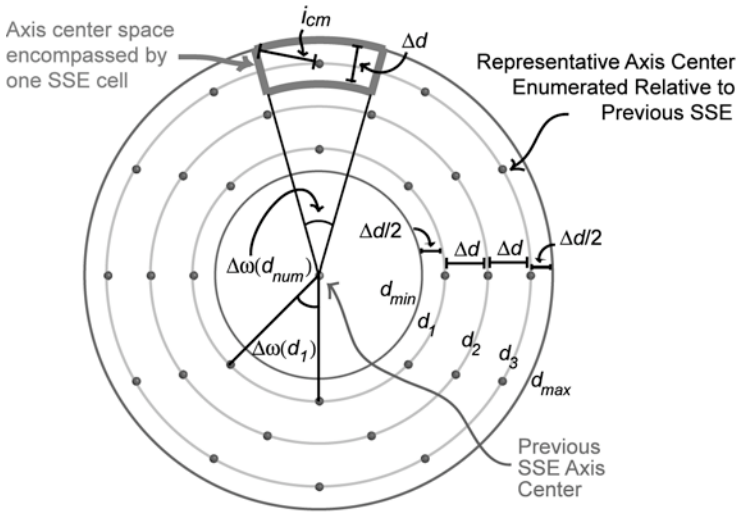

Figure 1. Division of the axis center conformaional space for one SSE using d_{num}=3 and $\_{num}$=12

Internal Error

Due to the fact that the SSE constraints are evaluated on representative conformations, which are intended to represent a range of SSE conformations within a cell, the tolerance on each distance restraint is calculated such that if any conformation within the cell satisfies the constraints, the representative conformation will be accepted. We describe this tolerance as internal error. There are two internal error values, which are useful in describing the internal error for an atom — i_{cm} and $i\_$. The internal error for axis center displacement (i_{cm}) is given by the distance from the axis center at (d,□) to the axis center at (d+□d/2,□+□□/2):

$$i_{cm} = \left(2d^2 \left(1 - \cos\left(\frac{\Delta\theta}{2}\right)\right) d^2 \left(1 - \cos\left(\frac{\Delta\theta}{2}\right)\right) \cdot d \cdot \Delta d + d \cdot \Delta d + \frac{\Delta d^2}{4}\right)^{1/2} \quad (1)$$

The internal error for axis end point displacement ($i_{\square}$), assuming a fixed axis center, is given by the distance from the end point at (□,□) to the end point at (□+□□/2,□+□□/2) using an axis length of l:

$$i_{\square} = \frac{l\sqrt{2}}{2}\left(1 - \sin\theta\sin\left(\theta + \frac{\Delta\theta}{2}\right)\cos\left(\frac{\Delta\theta}{2}\right) - \cos\theta\cos\left(\theta + \frac{\Delta\theta}{2}\right)\right)^{1/2} \quad (2)$$

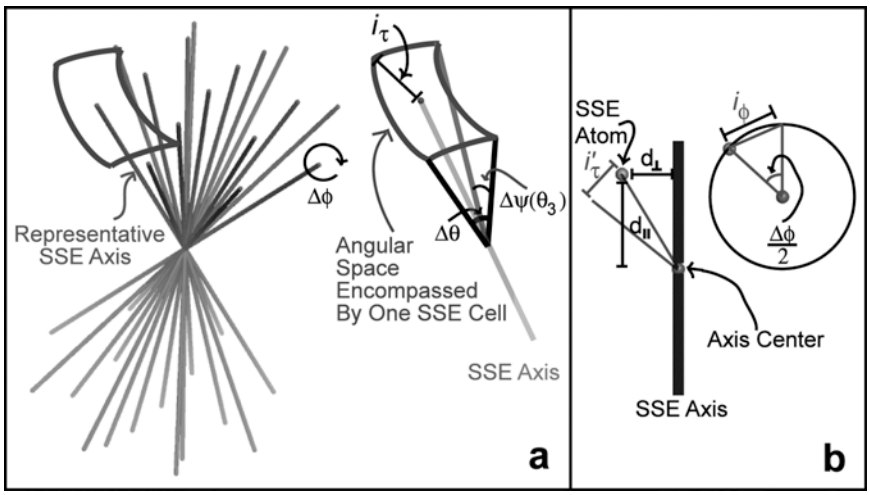

Figure 2: (a) Division of the angular conformational space for one SSE using _num=3 and _num=8 (b) Internal error components for an SSE atom

The internal error for an atom within the SSE (i_{atom}) can be estimated as the sum of the maximum individual displacements arising from axis center movement (i_{cm}), axis angle movement of the atom ($i_{\square}$'), and axis spin movement of the atom ($i_{\square}$). This number represents a slight over-approximation because it assumes maximum displacement in the same direction for all three components, and the axis angle displacement vector cannot have the exact same direction as the vectors measured by i_{cm} and $i_{\square}$ unless $\square\square$ is equal to zero. Let $d_{||}$ represent the distance from the atom to the axis center projected onto the axis, and let $d_{\_}$ represent the distance from the atom to the axis. The internal error for the atom can then be calculated as (see also Figure 2b):

$$\begin{aligned} i_{atom} &= i_{cm} + i'_{\square} + i_{\square} \\ &= i_{cm} + \frac{d_{\|}}{l} i_{\square} + 2d_{\square} \sin\left(\frac{\square\square}{4}\right) \end{aligned} \tag{3}$$

RMSD Upper Bounds

If every representative conformation within the conformational space of the protein is enumerated, what is the maximum RMSD that an arbitrary conformation can have from an enumerated conformation? We answer this question by calculating two upper bounds, the all-atom RMSD and the $RMSD_{HEL}$. The $RMSD_{HEL}$ was introduced previously[12] in order to study the deviation between helical axes and is useful in that it is independent of the atomic content of the SSEs within a protein. It is calculated simply as the RMSD of the axis centers and both sets of endpoints. Assuming that the conformational space for an SSE is divided such that $i_{\_}$ and i_{cm} remain constant for each of the cells, a tight upper bound to the $RMSD_{HEL}$ can be calculated.

Let **a** represent the vector of maximum displacement of an axis endpoint with fixed center within a cell and $\mathbf{a}_x$, $\mathbf{a}_y$, and $\mathbf{a}_z$ represent the Cartesian component vectors of **a**. Likewise, let **b** represent the vector of maximum displacement for the other axis end point. Because the axis is rotating about its center, we have:

$$i_{\theta} = |\mathbf{a}| = |\mathbf{b}|$$
$$\mathbf{a} = \theta\mathbf{b} \tag{4}$$

The maximum deviation for the endpoint moving by **a** is:

$$d_a = \left(\left(i_{cm} + |\mathbf{a}_x|\right)^2 + \mathbf{a}_y^2 + \mathbf{a}_z^2\right)^{1/2}$$
$$= \left(i_{cm}^2 + 2|\mathbf{a}_x| i_{cm} + i_{\theta}^2\right)^{1/2} \tag{5}$$

Likewise, by making the substitution in equation 4, the maximum deviation for the endpoint moving by **b** is:

$$d_b = \left(i_{cm}^2 - 2|\mathbf{a}_x| i_{cm} + i_{\theta}^2\right)^{1/2} \tag{6}$$

The maximum deviation for the axis center is simply i_{cm}, and therefore the upper bound to the RMSD_{HEL} for a single SSE can be calculated as:

$$\text{RMSD}_{\text{HEL}} = \sqrt{i_{cm}^2 + \frac{2}{3} i_{\theta}^2} \tag{7}$$

The upper bound for the entire protein conformation can be calculated by including the internal errors for each of the SSEs within the summation, noting that the i_{cm} for the first SSE is always 0 and that the i_{cm} for the second SSE is also reduced because the axis-center space covers only the x-axis.

For the all-atom RMSD, the calculation is less straightforward. Simply summing the squares of the internal errors for each atom results in a significant over-approximation because it neglects the fact that the axis is rotating about its' center. We therefore calculate the all-atom RMSD upper bound by assuming that the axis halves are symmetric. This assumes that for every atom that moves on one side of the axis, an atom on the other half moves in an opposite direction. Then, analogous to the RMSD_{HEL} calculation, the all-atom RMSD can be calculated as:

$$\text{RMSD}_{\text{UPPER}} \leq \left(\frac{1}{n} \sum_{j=1}^{n} \left(\left(i_{cm} + i_{\theta}(j)\right)^2 + i_{\theta}(j)^2\right)\right)^{1/2} \tag{8}$$

In this equation, the summation is over each atom *j* in the SSE with *i(j)* representing an internal error component for atom *j*. It is also important to note that neither the RMSD_{HEL} upper bounds nor the all-atom RMSD upper bounds represent optimal (fitted) upper bounds.

Local SSE Enumeration

A local enumeration of SSE cells is implemented by enumerating each SSE relative to a previous SSE (and thus absolute axis center positions are not considered). The local enumeration for each SSE is specified by a division count (d_{num}, θ_{num}, ϕ_{num}, θ_{num}, ϕ_{num}). The axis centers are enumerated at distances d_j between d_{min} and d_{max}, indexed at *j* from 1 to d_{num} (see Figure 1). For each cell:

$$d_j = d_{\min} + \frac{\delta d(2j - 1)}{2} \quad , \quad \delta d = \frac{d_{\max} - d_{\min}}{d_{num}} \tag{9}$$

Using a constant $\Delta\theta$ would lead to a decrease in the axis-center space covered by each cell with decreasing distance. Therefore, $\Delta\theta$ is calculated at each distance d_i to maintain an approximately constant i_{cm}. θ_{num} is used only to calculate $\Delta\theta$ at d_{d_num} and the resulting value for i_{cm} is used to calculated $\Delta\theta$ at each other distance (as derived from equation 1):

$$\Delta\theta(d_j) = 2\theta \bullet \left\lfloor \frac{2\theta}{2\cos^{-1}\left(\frac{8d_j^2 + 4d_j\Delta d + \Delta d^2 - 4i_{cm}^2}{8d_j^2 + 4d_j\Delta d}\right)} + 1 \right\rfloor^{-1} \tag{10}$$

The angular orientations for the axes are enumerated at θ_j between 0 and θ_{max} indexed at j between 1 and θ_{num}:

$$\theta_j = \frac{\Delta\theta(2j-1)}{2} \quad , \quad \Delta\theta = \frac{\theta_{\max}}{\theta_{num}} \tag{11}$$

The value for $\Delta\theta$ is calculated at each θ_j in order to maintain an approximately constant i_θ (as derived from equation 2):

$$\Delta\theta(\theta_i) = 2\theta \bullet \left\lfloor \frac{2\theta}{2\cos^{-1}\left(\frac{1 - \cos\theta\cos(\theta + \Delta\theta/2) - 2i_\theta^2}{l^2\sin\theta\sin(\theta + \Delta\theta/2)}\right)} + 1 \right\rfloor^{-1} \tag{12}$$

The axis spin division is straightforward:

$$\Delta\theta = \frac{2\theta}{\theta_{num}} \tag{13}$$

The equations allow the conformational space for an SSE to be divided with approximately constant values for i_θ and i_{cm} such that the RMSD upper bounds and internal errors for each atom remain consistent. The axis-center space for the first SSE consists of only the origin, and thus the i_{cm} for the first SSE is zero. The axis-center space for the second SSE consists of only the x-axis between d_{min} and d_{max}, and therefore the i_{cm} for the second SSE is calculated with $\Delta\theta$ equal to zero. It is also important to note that d_{min} and d_{max} are adjusted for the third and higher SSEs based on the i_{cm} for the previous SSE. This is to account for internal error in the positioning of the previous SSE. The "global" enumeration will enumerate any combination of representative axis conformations that satisfy all of the constraints as described in the Overview. The software performing the global enumeration takes as input an $RMSD_{HEL}$ or all-atom RMSD upper bound and calculates, based on this value, the division counts for each SSE needed in order to satisfy the upper bound. This is performed in an iterative manner, starting with division counts of 1. At each iteration, the division count, which results in the greatest decrease in internal error, is increased until the upper bound is satisfied.

Divide-and-Conquer

In order to decrease the run-times required for enumeration, a divide-and-conquer strategy is implemented. In this case, two RMSD upper bounds are supplied — an initial and final upper bound. The conformational space is originally divided according to the initial upper bound. If a representative conformation satisfies the constraints, its' cell is split, the internal errors are recalculated, and the new set of representative conformations

are evaluated. This division continues until a representative conformation is accepted for a cell whose RMSD upper bound is below the final value. The software allows two options for enumeration during divide-and-conquer. In the first, every accepted conformation per original cell is accepted. This option allows the final RMSD upper bound to be satisfied. In the second option, only the first accepted SSE-bundle within an original cell is accepted. This option prevents the enumeration of the large number of very similar protein conformations needed to satisify a small RMSD upper bound, while still allowing a low internal error condition for acceptance of protein conformations.

The options for dividing a conformational cell include splitting the axis-center space (Δd or $\Delta\theta$), the angular space ($\Delta\theta$ or $\Delta\theta$), or the axis-spin space ($\Delta\theta$). The decision is made by assessing which split will result in the largest decrease in internal error for a cell based on the values for the i_{cm}, the i_{θ}, and the i_{θ} for the atom farthest from the SSE axis. For the axis-center space, the cell can be split by dividing Δd, $\Delta\theta$, or both. The decision is made by assessing the decrease in i_{cm} for each method. Unless the i_{cm} decreases by at least twice the amount that would be obtained by splitting Δd or $\Delta\theta$, the split into fourths is not performed and the better of the other two methods is taken. When splitting $\Delta\theta$, the $\Delta\theta$ for each of the new cells is set to maintain an approximately equal i_{cm}. The split of the angular space is made in a manner analogous to that of the axis-center space.

Parameterization

The parameters that govern the conformational space for an integral membrane helical bundle are taken from the ranges reported in a survey by Bowie.[14] The maximum angle of a helical axis with the bilayer plane (θ_{max}) is set at 40º. The minimum distance between axis centers (d_{min}) is set to 6 Å and the d_{max} between helical axes adjacent in terms of primary sequence is set to 13.4 Å. The distance of closest approach between two SSE axes is set to lie in the same range as the axis center distances.

Results and Discussion

As described above, the $RMSD_{HEL}$ is useful because it only represents the deviation of the axis center and endpoints, and is therefore independent of specific protein structures. The total number of representative conformations that exist for a given $RMSD_{HEL}$ upper bound are plotted in Figure 3a for a 7-helix bundle containing axes 30 Å in length. There is an exponential increase in the number of representative conformations that must be evaluated with decreasing $RMSD_{HEL}$. This increase is relevant in that the $RMSD_{HEL}$ is intimately related to the internal error for each distance restraint. This presents an interesting problem in view of the objective to determine the percentage of total conformational space in which a helix bundle satisfies experimental distance restraints. By decreasing the $RMSD_{HEL}$ upper bound, the internal error becomes reduced. However, there is an exponential increase in the number of potential conformations that must be evaluated. Likewise, there is an exponential increase in the number of conformations that lie within a fixed percentage of the conformational space and must later be considered for scoring and refinement. On the other hand, by increasing the $RMSD_{HEL}$, the constraints become relaxed such that the percent of conformational space in which the helix bundle can exist increases.

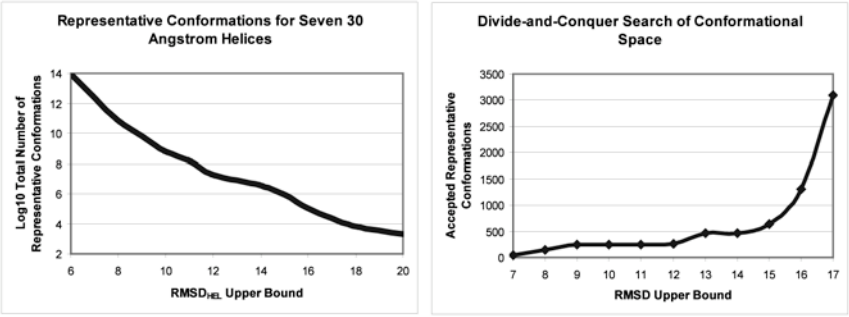

Figure 3: (a) Total number of possible representative conformations as a function of the $RMSD_{HEL}$ Upper Bound. The numbers were calculated using seven 30 angstrom SSE axes. (b) Divide-and-conquer search of the conformational space for bovine rhodopsin using an initial RMSD upper bound of 17 angstroms

Our solution to this problem is a divide-and-conquer approach. The algorithm prevents the enumeration of a large number of very similar conformations, while at the same time allowing for a small internal error in distance restraint evaluation. We have tested the efficacy of this approach using the crystal structure of bovine rhodopsin (1F88.pdb). Thirty-eight distance restraints were calculated between atoms in amino acid pairs that could potentially be cross-linked (K-K, K-D, K-E, K-C, and C-C) as reported in our previous work.[7] The "experimental" errors were set to ±2 angstroms. When using as input an all-atom RMSD upper bound of 17Å, 3088 out of a total of 2.6_10^{9} representative conformations were enumerated, indicating that they satisfied both the distance restraints and other helical-bundle constraints listed in the overview. In terms of conformational cells, this represents less than 1.2_10^{-6} percent of the total conformation space.

By implementing the divide-and-conquer approach, this number can be reduced further. Figure 3b illustrates the effect of the final divide-and-conquer RMSD upper bound on the number of accepted conformations. Using a final all-atom RMSD upper bound of 7Å, the number of accepted conformations is reduced to 48. Out of the enumerated conformations, the minimum all-atom RMSD from the crystal structure is found to be 4.3Å. The superimposition of the structures is shown in Figure 5.

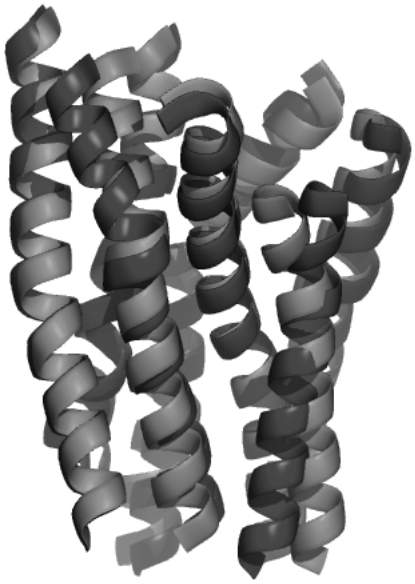

Figure 4. Superimposition of the helix bundle from the crystal structure of bovine rhodopsin with a representative conformation selected from enumeration.

Summary

We have described a deterministic algorithm for enumeration of transmembrane protein folds, which does not suffer from the drawbacks of the stochastic approach we have used previously. Additionally, we have demonstrated the efficacy of the approach via the constrained enumeration of potential helix bundles for bovine rhodopsin using sparse distance constraints. The results show that an exhaustive search of the potential conformational space of a transmembrane helix bundle is possible, based only a set of sparse experimental distance restraints. The source code for ESSEB is available upon request from the authors.

Acknowledgments

Funding for this work was provided by Sandia National Laboratories under contract DE-AC04-94AL85000. Sandia is a multiprogram laboratory operated by Sandia Corp., a Lockheed Martin Company, for the U.S. Department of Energy (DOE)'s National Nuclear Security Administration.

References

1. D. Boyd, C. Schierle and J. Beckwith, *Protein Sci.* **7**, 201 (1998).
2. J. L. Popot and D. M. Engelman, *Biochemistry* **29**, 4031 (1990).
3. G. D. Rose, *Nature* **272**, 586-90 (1978).
4. S. Jayasinghe, K. Hristova and S. H. White, *Protein Sci.* **10**, 455 (2001).
5. S. Jayasinghe, K. Hristova and S. H. White, *J. Mol. Biol.* **312**, 927 (2001).
6. J. L. Faulon, M. D. Rintoul and M. M. Young, *Journal Phys. A* **35**, 1 (2002).
7. J. L. Faulon, K. Sale and M. Young, *Protein Sci.* **12**, 1750 (2003).
8. K. Sale, J. L. Faulon, G. Gray, J. Schoeniger and M. Young, *Protein Sci.* **13,** 2613 (2004).

9. G. Vriend, *J. Mol. Graphics* **8**, 52 (1990).
10. M. J. Bower, F. E. Cohen and R. L. Dunbrack, Jr., *J. Mol. Biol.* **267**, 1268 (1997).
11. Z. Xiang, C. S. Soto and B. Honig, *Proc. Natl. Acad. Sci. USA* **99**, 7432 (2002).
12. J. U. Bowie, *Protein Sci.* **8**, 2711 (1999).
13. C. Chothia, M. Levitt and D. Richardson, *J. Mol. Biol.* **145**, 215 (1981).
14. J. U. Bowie, *J. Mol. Biol.* **272**, 780 (1997).

Chapter 2: Optimal bundling of transmembrane helices using sparse distance constraints

Ken Sale, Jean-Loup Faulon, Genetha A. Gray, Joseph S. Schoeniger and Malin M. Young

Abstract

We present a two-step approach to modeling the transmembrane spanning helical bundles of integral membrane proteins using only sparse distance constraints, such as those derived from chemical cross-linking, dipolar EPR and FRET experiments. In step one, using an algorithm we developed (Faulon et al. 2003), the conformational space of membrane protein folds matching a set of distance constraints is explored to provide initial structures for local conformational searches. In step two, these structures refined against a custom penalty function that incorporates both measures derived from statistical analysis of solved membrane protein structures and distance constraints obtained from experiments. We begin by describing the statistical analysis of the solved membrane protein structures from which the theoretical portion of the penalty function was derived. We then describe the penalty function, and, using a set of six test cases, demonstrate that it is capable of distinguishing helical bundles that are close to the native bundle from those that are far from the native bundle. Finally, using a set of only 27 distance constraints extracted from the literature, we show that our method successfully recovers the structure of dark-adapted rhodopsin to within 3.2 Å of the crystal structure.

Introduction

Integral membrane proteins are essential components of the cell membrane that participate in many important cellular processes such as energy transduction, cell signaling, mediation of senses such as vision, cell intoxication and pathogenesis, and immune recognition. Their significance is emphasized by the fact that approximately one-third of the proteins encoded for by a typical genome are membrane proteins (Buchan et al. 2002). Furthermore, at least 70 percent of current pharmaceuticals are thought to act on membrane proteins (Wilson and Bergsma 2000). Despite their obvious importance, to date, the structures of fewer than 75 integral membrane proteins have been solved (see (White 2003) and references therein), and this number includes redundant structures across species. This is a vast contrast to the over 25,000 soluble proteins whose structures have been solved using X-ray crystallography and NMR. Reasons for the slow progress in the structural analysis of membrane proteins include the instability of membrane proteins in environments lacking phospholipids, their tendency to aggregate and precipitate, and protein abundance, expression and purification issues. These characteristics highlight why the application of standard structure determination methods to membrane proteins is non-trivial.

Given the nature of the difficulties in generating high-resolution structural data from methods such as X-ray crystallography and NMR, it is unlikely that these experimental techniques will yield a significant increase in the number of solved membrane protein structures in the near future. As an alternative approach, the focus here

is on modeling transmembrane proteins using a set of sparse distance constraints, thus leveraging the many recent advances in techniques for measuring distances within a protein. Such methods include chemical cross-linking combined with mass spectrometry (Bennett et al. 2000; Rappsilber et al. 2000; Young et al. 2000; Back et al. 2002; Taverner et al. 2002; Dihazi and Sinz 2003; Kruppa et al. 2003; Novak 2003; Schilling et al. 2003), site directed spin labeling combined with electron paramagnetic resonance (SDSL-EPR) (Rabenstein and Shin 1995; Farrens et al. 1996; Hustedt et al. 1997; McHaourab et al. 1997; Steinhoff et al. 1997; Hustedt and Beth 1999; Altenbach et al. 2001; Borbat et al. 2001; Liu et al. 2001; Persson et al. 2001; Radzwill et al. 2001; Brown et al. 2002; Perozo et al. 2002; Hubbell et al. 2003), disulfide bond formation mapping (Cai et al. 1999; Yu et al. 1999; Cai et al. 2001) and fluorescence resonance energy transfer (FRET) (Matyus 1992; Hillisch et al. 2001; Klostermeier and Millar 2001; Parkhurst et al. 2001; Rye 2001; Szollosi et al. 2002; Sekar and Periasamy 2003). These methods produce low-to-moderate resolution structural data that can be used in conjunction with computational predictions, such as structural rules derived from helix-helix interactions in known structures (Bowie 1997; 1999), to determine a transmembrane protein structure to moderate resolution.

The modeling challenge of constructing a transmembrane helical bundle that is consistent with a set of low-to-moderate resolution experimental constraints can be simplified by considering some of the relative characteristics of a transmembrane protein. The low dielectric environment of a lipid bilayer favors the formation of regular secondary structural elements (SSE), such as helices and beta sheets, by increasing the strength of hydrogen bonds (White and Wimley 1999; Kim and Cross 2002). The thermodynamic disadvantages of transferring non-hydrogen bonded peptides from a water to a lipid environment (+5 kcal/mol per H-bond, (Engelman et al. 1986)) imply that transmembrane proteins fold and assemble in a multi-stage process (Jacobs and White 1989; Popot and Engelman 1990). We assume the two-stage model (Popot and Engelman 1990) and describe the construction of transmembrane protein models as two separate tasks: (1) defining the transmembrane SSEs and (2) determining their relative orientations or packing.

While not a solved problem, transmembrane spanning SSEs can be accurately predicted from sequence information using widely accepted methods such as sliding-window hydrophobicity analysis (Rose 1978; Jayasinghe et al. 2001a; b). However, subsequent prediction of the association of these helices into the final transmembrane protein fold is not well established. Structural constraints imposed by the lipid bilayer on transmembrane SSEs do limit the number of possible membrane protein folds (White and Wimley 1998), and several ab-initio and potential based computational approaches for predicting interhelical packing have been proposed (Bowie 1997; 1999; Nikiforovich et al. 2001; Dobbs et al. 2002; Fleishman and Ben-Tal 2002; Vaidehi et al. 2002; Kim et al. 2003).

Several of these approaches incorporate experimental data into their models. For example, Nikiforovich et al.(Nikiforovich et al. 2001) use the similarity between the X-ray structures of bacteriorhodopsin and rhodopsin to estimate helix packing in the membrane plane. Specifically, the intersections between the helical axes and the membrane plane are fixed at values derived from the two X-ray structures. Vaidehi et al.

orient each helical axis of the helical bundle according to the 7.5 Å electron density map of rhodopsin (Vaidehi et al. 2002). Herzyk and Hubbard developed an automated approach to modeling seven helix transmembrane receptors using a combination of data from electron microscopy, neutron diffraction, mutagenesis, chemical cross-linking, site-directed spin labeling, disulfide mapping, FTIR difference spectroscopy, solid state ^{13}C NMR, semi-empirical calculations on ligand-protein interaction, multiple sequence alignment and hydrophobicity (Herzyk and Hubbard 1995). Using a potential function designed to constrain model structures to satisfying these data, they built a model structure of bacteriorhodopsin that was within 1.87 Å RMSD of the structure determined by electron microscopy. By combining several types of data, they have laid the groundwork for developing scoring functions that constrain helical bundles using experimental data. In this work, we take a similar approach; however, rather than using data taken from a variety of experiments, we develop a function based solely on distance constraints and data mined from structures in the PDB.

In this paper, we describe a two-step approach to modeling the transmembrane spanning, helical bundles of integral membrane proteins using sparse distance constraints. Since many of the known membrane protein structures are all alpha-helical, we limit our discussion to modeling helical bundles. The method is as follows: step 1) search the conformational space of membrane protein folds to find those matching a given set of distance constraints (Faulon et al. 2003); step 2) refine the helical bundles from step one using a Monte Carlo simulated annealing protocol designed for local minimization of a custom penalty function referred to as Bundler. The Bundler function scores a helical bundle based on its consistency with the structural features of known transmembrane bundles as well as with distance constraints from experimental methods such as chemical cross-linking, NMR, FRET and EPR. In the following sections the Bundler penalty function is described in detail and validated across a set of six known transmembrane protein structures to show that it is capable of distinguishing between structures close to and far from the native structure. We also demonstrate that our two-step approach can recover the transmembrane helical bundle of the dark-adapted rhodopsin structure (1f88) to within 3.2 Å RMSD of the native structure using only 27 experimental distance constraints gathered from the literature.

Results

We begin this section by presenting a statistical analysis of a set of non-redundant helical transmembrane proteins. This is followed by a description of the penalty function, referred to as Bundler, and validation of the penalty function as a tool to differentiate near native helical bundles from those far from the native bundle is then described. Using a set of six membrane proteins crystal structures the penalty function is validated by showing that helical bundles with lower RMSD from the X-ray structure score lower than those with higher RMSD. Lastly, we demonstrate the method on the structure of dark-adapted rhodopsin using a set of distance constraints taken from the literature.

Statistical analysis of membrane protein structures

The set of 14 membrane proteins listed in Table 1, with all-alpha helical

transmembrane domains, was examined to extract statistical information about their helix packing distances, angles and number of nearest neighbors. Since the structure of the individual helices comprising the helical bundle is not likely to be known, we assumed that in most cases the bundle will initially be modeled using idealized helices and concluded that collecting statistics on an idealized set of the 14 transmembrane proteins would result in the most useful statistical parameters for the scoring function. Idealized representations of the 14 proteins were constructed by superimposing perfect alpha-helical structures of the appropriate lengths onto the helices in the transmembrane domains. The Cα level RMSD between the individual idealized helices and their corresponding helices from the PDB structure ranged from 0.56 Å (1PRC, 17 aa) to 4.07 Å (1QLAC, 35 aa), while across all helices of the transmembrane domain, the Cα level RMSDs ranged from 1.15 Å (1FQY, 136 aa) to 2.37 Å (1QLAC, 160 aa).

Table 1: Structures used for statistical characterization of transmembrane protein bundles

PDB ID	*Name*	*Number of AAs*
1BL8	KcsA Potassium Channel	388
1C3W	Bacteriorhodopsin	222
1E12	Halorhodopsin	239
1EHK	Ba3 Cytochrome C Oxygenase	743
1EUL	Calcium ATPase	994
1EZVC	Cytochrome bc1 Complex	385
1F88	Rhodopsin	338
1FQY	AQP1 –Aquaporin Water Channel	226
1FX8	GlpF-Glycerol Facilitator Channel	254
1JGJ	Sensory Rhodpsin II	217
1MSL	McsL Mechanosensitive Channel	545
1OCC	aa3 Cytochrome C Oxidase	1780
1PRC	Photosynthetic Reaction Center	605
1QLAC	Fumerate Reductase Complex	254

Statistics collected on the 14 idealized representative structures are listed in Table 2. Means and standard deviations were calculated for the distances between the centers of mass for consecutive helices ($\Delta_{COM,cons}$), distances between the centers of mass for all helical pairs (Δ_{COM}), the minimum approach distance of the helical axes for consecutive helices ($\Delta_{min,cons}$), the minimum approach distance of all helix axial pairs (Δ_{min}), the packing angle of helical axes (Δ_{pack}), and the number of helical neighbors (n_{neigh}) with a minimum pairwise approach distance less than 15 Å. Note that in Table 2, *N* indicates the sample size.

Table 2: Statistics describing transmembrane protein helical bundles[1]

Statistic	μ	σ	*N*
$\Delta_{COM,cons}$	12.8 Å	5.3 Å	86
Δ_{COM}	18.6 Å	7.32 Å	336
$\Delta_{min,cons}$	10.7 Å	5.2 Å	86
Δ_{min}	16.3 Å	7.4 Å	336
Δ_{pack}	30.9°	16.4°	336
n_{neigh}	3.4	1.4	102
Δ_{pack}	37.1	2.5	16

[1] All statistics were calculated on the set of proteins listed in Table 1 with the exception of the packing density, Δ_{pack}, which was calculated on the proteins listed in Table 3.

Fleishman and Ben-Tal have suggested that short loops, less than 20 amino acids, play an important role in determining the packing of helices in membrane protein structures (Fleishman and Ben-Tal 2002). Hence, in addition to experimentally

determined distances, we include distances generated by correlating loop lengths to helix-end to helix-end distances. Using our set of 14 helical membrane proteins, we correlated the helix-end to helix-end distances with the number of amino acids in the loop connecting the two helices (Figure 5). Across the span of loop lengths, this correlation is quite low ($R^2 = 0.4$). However, dividing this sample into a group with loops containing seven or fewer amino acids ($R^2 = 0.8$) and loops with eight or more amino acids ($R^2 = 0.2$) allowed us to develop a set of guidelines for deriving helix-end to helix-end distance constraints given the number of amino acids in the loop. The least squares line through the points with seven or fewer amino acids is $D = 1.2x(\pm 0.2) + 4.9(\pm 1.0)$, where D is the helix-end to helix-end distances and x is the number of amino acids. Using a 95% confidence interval around this least squares line and the minimum and maximum distances for loops with 8 or more amino acids, we obtain the following upper (*UB*) and lower (*LB*) bounds for distance constraints between helix ends:

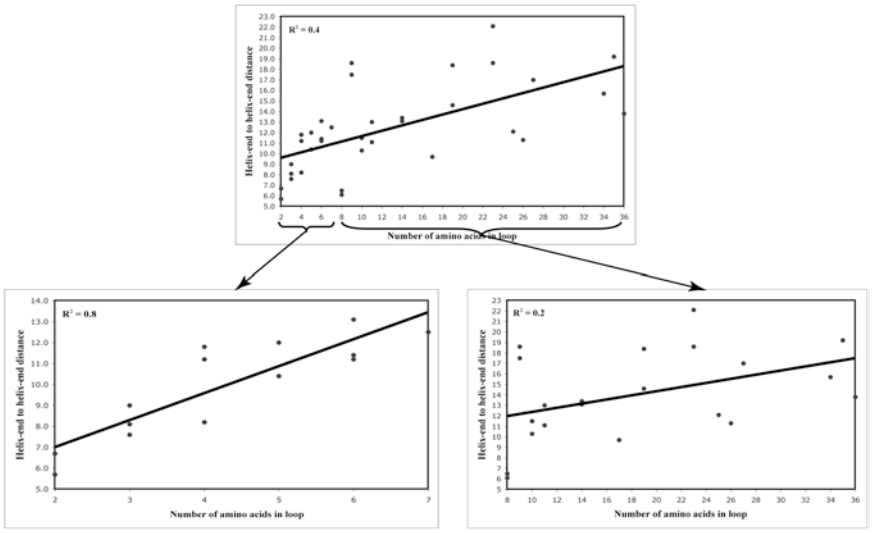

Figure 5: Correlation of helix-end to helix-end distance with number of amino acids in the loop. Statistics are for the 36 helix-end to helix-end distances extracted from the set of 14 non-redundant structures given in Error! Reference source not found.

$$\begin{array}{ll} \#AA \leq 7 & \begin{cases} LB = 0.7x + 2.9 \\ UB = 1.6x + 6.9 \end{cases} \\ \#AA \geq 8 & \begin{cases} LB = 5 \\ UB = 25 \end{cases} \end{array} . \qquad (1)$$

For loops ranging from 4 to 8 residues the upper bounds are 13.5 Å, 15.2 Å, 16.8 Å, 18.1 Å and 20.1 Å, respectively, which compare well to the values of 14.7 Å, 15.7 Å, 18.2 Å, 18.2 Å and 20.7 Å reported by Hertzig and Hubbard (Herzyk and Hubbard 1995).

Penalty function

The Bundler penalty function incorporates distance constraints determined via experimental methods such as chemical crosslinking, dipolar EPR, FRET and NMR. Bundler assesses a possible helical bundle and assigns it a score reflecting, in part, its degree of consistency with a set of experimental distance constraints. Given a large enough experimental distance constraint set, such a function would require no additional

considerations; however, measuring distances in membrane proteins is difficult, so it is likely that only a sparse number of distance constraints will be available. Moreover, it is expected that the available distances will not be error free. Therefore, to improve its viability, Bundler also includes penalties for violating a set of helix packing parameters determined by the analysis of a set of membrane protein structures from the PDB. Note that while after the first step of the overall modeling procedure only helical bundles satisfying the distance constraints remain, it is still necessary to include a distance constraints penalty to avoid allowing the bundle to deviate far from experimental results in favor of the structure survey-based constraints. The total penalty, *P*, is thus the sum of a distance constraint penalty and the structure-based penalties:

$$P = P_{\text{distance constraints}} + P_{\text{structure}} \tag{2}$$

Distance Constraints Penalty (P_{dist})

Distance constraints provide moderate resolution structural information and are a crucial component in our modeling of helical membrane proteins (Faulon et al. 2003). Bundler penalizes structures that violate distance constraints according to a "soft" square well potential defined as

$$P_{\text{dist}} = k_{\text{dist}} \begin{cases} \left(d_{ij} - l_{ij}\right)^2, & d_{ij} < l_{ij} \\ 0, & l_{ij} \le d_{ij} \le u_{ij}, \\ \left(u_{ij} - d_{ij}\right)^2, & d_{ij} > u_{ij} \end{cases} \tag{3}$$

where l_{ij} and u_{ij} are the lower and upper limits on the distance between atoms *i* and *j*, respectively; d_{ij} is the distance between atoms *i* and *j* in the current bundle; and k_{dist} is a force constant and was set to 500.

Structure based penalties

The structure based piece of the scoring function consists of penalties for helical bundles with packing angles, packing distances, and/or packing densities outside the ranges determined from analysis of a non-redundant set of helical transmembrane protein structures. It also incorporates a van der Waals repulsive potential, a "compactness" penalty for having too few neighboring helices, and a penalty for unlikely side-chain interactions. Summing these terms gives the total structure based penalty

$$P_{\text{structure}} = P_{\text{packing distance}} + P_{\text{packing angle}} + P_{\text{packing density}} + P_{\text{vdw}} + P_{\text{contacts}} + P_{\text{side-chain preference}}. \tag{4}$$

Below, we describe each of the terms of (4) in detail.

Packing Distance Penalty (P_{pdist})

The mean distance between the centers of mass of consecutive helices, as derived from the set of 14 non-redundant helical transmembrane protein structures Table 1, is 12.8 ± 5.3 Å, while the mean distance between consecutive helical line segments is 10.7 ± 5.2 Å. A packing distance penalty is applied if either the centers of mass of the consecutive helices or the minimum distance between the two helical axes falls outside 1.5 standard deviations of their respective mean. The packing distance penalty is defined

as a soft square well potential,

$$P_{\Box} = k_{\Box} \begin{cases} (\Box_{ij} - \Box_l)^2, & \Box_{ij} < \Box_l \\ 0, & \Box_l \le \Box_{ij} \le \Box_u, \\ (\Box_u - \Box_{ij})^2, & \Box_{ij} > \Box_u \end{cases} \quad \Box_l = \bar{\Box} - 1.5 s_{\Box} \text{ and } \Box_u = \bar{\Box} + 1.5 s_{\Box}, \tag{5}$$

where $\bar{\Box}$ and $s_{\Box}$ are the mean and standard deviation of the interhelical distance, respectively; $\Box_{ij}$ is the distance between the centers of mass of helix *i* and helix *j* in the current structure; and $k_{\Box}$ is a force constant, which we set at 50. The packing distance term is summed over the set of distinct helical pairs.

Packing Density Penalty (P_{pdens})

Packing density is defined as the ratio of atomic volume to solvent accessible volume (Richards 1974). Since average protein packing density does not vary significantly with secondary structure class (Chothia 1975), we increased our sample size

Table 3: Packing density statistics

PDB ID	*Number of AAs*	*Name*	*TM Class*	*Packing Density*
1BL8	388	KcsA Potassium Channel	□	37.0
1BXW	172	OmpA	□	37.0
1C3W	222	Bacteriorhodopsin	□	38.0
1E12	239	Halorhodopsin	□	38.0
1EHK	743	ba3 Cytochrome C Oxygenase	□	38.0
1EK9	423	TolC Outer Membrane Protein	□	37.0
1EUL	994	Calcium ATPase	□	37.0
1EZVC	385	Cytochrome bc1 Complex	□	37.0
1F88	338	Rhodopsin	□	37.0
1FEP	669	FepA	□	37.0
1FQY	226	AQP1-Aquaporin Water Channel	□	36.0
1FX8	254	GlpF-Glycerol Facilitator Channel	□	38.0
1JGJ	217	Sensory Rhodopsin	□	38.0
1LGH	198	Light Harvesting Complex	□	37.0
1MAL	421	Maltoporin	□	37.0
1MSL	545	MscL Mechanosensitive Channel	□	35.0
1OCC	1780	aa3 Cytochrome C Oxidase	□	37.0
1PHO	330	PhoE	□	37.0
1PRC	605	Photosynthetic Reaction Center	□	36.0
1QD5	257	OMPLA	□	37.0
1QJ8	148	OmpX	□	39.0
1QLAC	254	Fumerate Reductase Complex	□	37.0
2FCP	705	FhuA	□	37.0
2MPR	421	Maltoporin	□	37.0
2OMF	340	OmpF	□	38.0
2POR	301	Porin	□	38.0
3LKF	292	LukF	□	37.0
7AHL	293	Alpha-hemolysin	□	36.0

for calculating packing density statistics by analyzing a non-redundant set of 28 alpha-helical and/or beta strand-containing membrane proteins Table 3from which the mean backbone packing density was 37.1 ± 2.5. Structures with a packing density greater than

1.5 standard deviations away from the mean are penalized using a soft square well potential

$$P_{\rho} = k_{\rho}\begin{cases}(\rho - \rho_l)^2, & \rho < \rho_l \\ 0, & \rho_l \le \rho \le \rho_u, \\ (\rho_u - \rho)^2, & \rho > \rho_u\end{cases} \quad \text{where } \rho_l = \bar{\rho} - 1.5 s_{\rho} \text{ and } \rho_u = \bar{\rho} + 1.5 s_{\rho}, \tag{6}$$

where $\bar{\rho}$ and s_{ρ} are the mean and standard deviation of the packing density, respectively; and k_{ρ} is a force constant, which we set at 500.

Packing Angle Penalty (P_{angle})

The helix packing angle score penalizes structures in which the angle between the helical axes of consecutive pairs of helices is outside 1.5 standard deviations of the average angle. The mean packing angle between consecutive pairs of helices, calculated over the non-redundant set of 14 helical transmembrane proteins in Table 3, is 30.9 ± 16.3 Å. Packing angle violations are penalized according to a soft square well potential,

$$P_{\Omega} = k_{\Omega}\begin{cases}(\Omega_{ij} - \Omega_l)^2, & \Omega_{ij} < \Omega_l \\ 0, & \Omega_l \le \Omega_{ij} \le \Omega_u, \\ (\Omega_u - \Omega_{ij})^2, & \Omega_{ij} > \Omega_u\end{cases} \quad \text{where } \Omega_l = \bar{\Omega} - 1.5 s_{\Omega} \text{ and } \Omega_u = \bar{\Omega} + 1.5 s_{\Omega}, \tag{7}$$

where $\bar{\Omega}$ and s_{Ω} are the mean and standard deviation of the packing angles, respectively; and Ω_{ij} is the angle between helix *i* and helix *j*. The force constant is k_{Ω} = 5. The packing angle penalty is summed over the set of consecutive helical pairs.

van der Waals Repulsion (P_{vdw})

In order to avoid overlapping helices, we include a van der Waals potential. Since our helix bundling is done at the Cα level of atomic detail, we use only the van der Waals repulsive function (Brünger et al. 1998),

$$P_{vdw} = k_{vdw}\begin{cases}0, & r_{ij} \ge sR_{ij} \\ \left(s^2 R_{ij}^2 - r_{ij}^2\right)^2, & r_{ij} < sR_{ij}\end{cases}, \tag{8}$$

to prevent interhelical clashes. Here, *s* is a predetermined van der Waals scaling factor and was set to 1; r_{ij} is the distance between Cα atoms *i* and *j*; R_{ij} is the distance at which atoms i and j begin to repel each other; and k_{vdw} is a weighting constant and is set at 5. This piece of the penalty function is summed over the set of all pairs of Cα atoms, and for computing efficiency, we consider only Cα – Cα clashes.

Contact Penalty (P_{contact})

Our analysis of the 14 membrane proteins listed in Table 1 revealed that the helices are usually in contact with at least two neighbor helices. To guarantee that this is

the case in our candidate helical bundles, we apply a simple linear penalty to any structure containing a helix that is not in contact with at least two neighbors and define a contact penalty as

$$P_{contact} = k_{contact}(2 - c), \tag{9}$$

Here, $c < 2$ is the number of helices with a center of mass that is less than $\sigma_{(COM)} - 1.5\sigma_{(COM)}$ of the center of mass of the specified helix and $k_{contact}$= 500. A contact penalty score is calculated for each helix in the bundle.

Side-Chain Interaction Preference Penalty (P_{sc})

The amino acids in membrane proteins show a preference for which amino acids they interact with on neighboring helices (Adamian and Liang 2001; Nikiforovich et al. 2001; Adamian et al. 2003). To evaluate this characteristic in our candidate helical bundles, we incorporate the membrane helical interfacial pairwise (MHIP) amino acid interaction propensity matrix of Adamian and Liang (Adamian and Liang 2001) into our penalty function. The entries of this matrix have been adjusted to reflect penalties for low propensity pair interactions rather than bonuses for favored pair interactions by subtracting the propensity score for each amino acid pair from the value of the highest propensity pair (Table 4). Note that the penalty for the strongest interacting pairs, such as CYS – GLN, which have an MHIP = 6.0, is now 0.0, while the penalty on the weakest interacting pairs, such as ARG – SER with an MHIP = 0.0, is now 6.0, the largest value in Table 4. The side-chain propensity penalty is simply the sum of the pair wise propensity over all side-chain pairs, for which the C_α atoms are within 4.9 Å of each other,

Table 4: Helical interfacial side-chain packing penalties[1]

	ALA	CYS	ASP	GLU	PHE	GLY	HIS	ILE	LYS	LEU	MET	ASN	PRO	GLN	ARG	SER	THR	VAL	TRP	TYR
ALA	4.7	4.3	4.8	5.2	4.9	4.9	4.7	5.0	5.3	5.1	4.3	4.9	3.9	5.0	5.5	5.1	5.0	5.2	4.9	5.2
CYS	4.3	5.2	6.0	5.2	4.2	3.6	4.7	4.9	6.0	5.0	4.5	5.2	5.4	6.0	5.6	3.8	4.8	5.7	5.6	5.7
ASP	4.8	6.0	6.0	5.6	5.7	5.9	5.4	5.0	3.8	5.3	5.5	1.2	4.2	6.0	2.3	4.8	5.0	5.9	5.6	3.2
GLU	5.2	5.2	5.6	4.4	5.5	5.3	5.0	5.6	4.3	5.5	5.0	4.7	4.1	5.6	4.8	5.0	5.0	5.3	5.9	5.3
PHE	4.9	4.2	5.7	5.5	4.3	4.7	4.9	5.2	5.6	4.9	4.6	5.5	5.4	5.0	5.6	5.0	5.3	5.1	4.6	5.2
GLY	4.9	3.6	5.9	5.3	4.7	3.0	2.9	5.4	5.6	5.0	4.7	4.4	5.4	4.6	5.4	5.0	5.4	5.0	4.6	4.4
HIS	4.7	4.7	5.4	5.0	4.9	2.9	2.1	5.3	5.5	5.3	5.0	5.8	5.7	3.5	5.7	4.7	3.7	5.5	4.1	4.8
ILE	5.0	4.9	5.0	5.6	5.2	5.4	5.3	4.7	5.5	5.0	4.9	4.9	4.8	5.0	5.8	5.4	5.1	5.2	5.0	5.5
LYS	5.3	6.0	3.8	4.3	5.6	5.6	5.5	5.5	6.0	5.3	3.8	3.2	5.0	4.4	5.2	4.9	5.8	5.6	5.4	3.5
LEU	5.1	5.0	5.3	5.5	4.9	5.0	5.3	5.0	5.3	4.9	5.0	5.1	5.3	5.2	5.4	4.9	5.4	5.0	5.0	5.0
MET	4.3	4.5	5.5	5.0	4.6	4.7	5.0	4.9	3.8	5.0	4.5	5.2	4.6	5.0	4.7	4.1	5.3	5.1	4.8	5.4
ASN	4.9	5.2	1.2	4.7	5.5	4.4	5.8	4.9	3.2	5.1	5.2	0.0	4.8	3.6	5.3	4.6	5.2	5.1	5.3	4.5
PRO	3.9	5.4	4.2	4.1	5.4	5.4	5.7	4.8	5.0	5.3	4.6	4.8	4.2	5.1	5.4	4.8	4.7	5.4	4.8	4.0
GLN	5.0	6.0	6.0	5.6	5.0	4.6	3.5	5.0	4.4	5.2	5.0	3.6	5.1	6.0	3.2	3.5	4.6	5.3	4.7	3.7
ARG	5.5	5.6	2.3	4.8	5.6	5.4	5.7	5.8	5.2	5.4	4.7	5.3	5.4	3.2	6.0	0.0	5.0	5.0	1.0	5.0
SER	5.1	3.8	4.8	5.0	5.0	5.0	4.7	5.4	4.9	4.9	4.1	4.6	4.8	3.5	0.0	1.6	4.5	5.2	4.9	5.0
THR	5.0	4.8	5.0	5.0	5.3	5.4	3.7	5.1	5.8	5.4	5.3	5.2	4.7	4.6	5.0	4.5	4.9	4.9	4.9	4.8
VAL	5.2	5.7	5.9	5.3	5.1	5.0	5.5	5.2	5.6	5.0	5.1	5.1	5.4	5.3	5.0	5.2	4.9	5.0	5.1	5.5
TRP	4.9	5.6	5.6	5.9	4.6	4.6	4.1	5.0	5.4	5.0	4.8	5.3	4.8	4.7	1.0	4.9	4.9	5.1	5.2	5.1
TYR	5.2	5.7	3.2	5.3	5.2	4.4	4.8	5.5	3.5	5.0	5.4	4.5	4.0	3.7	5.0	5.0	4.8	5.1	5.1	5.4

[1] Penalties are the membrane helical interfacial pairwise contact propensities from Adamian and Liang (Adamian and Liang 2001) for which each propensity has been subtracted from the highest propensity to yield a penalty for preferred side-chains not interacting.

$$P_{sc} = \sum_{ij} P_{ij}, \quad d_{ij} \le 4.9 \text{ Å} \tag{10}$$

where P_{ij} is the interaction penalty of amino acids i and j and d_{ij} is the distance between the two C_α atoms.

Total Score

The total score is the sum of the individual components, which are summed over the appropriate set of pairwise interactions. Let m be the number of helices, n the number of amino acids, $\square$ the set of amino acids among which distances have been measured, $\square$ the set of $m(m-1)/2$ distinct helical pairs and $\square$ the set of $n(n-1)/2$ distinct C$\square$ pairs. Then, the Bundler penalty can be written as

$$P = \sum_{(i,j)\square\square} P_{\exp} + \sum_{(i,j)\square\square} P_{angle} + \sum_{(i,j)\square\square} P_{dist} + P_{density} + \sum_{(i,j)\square\square} P_{vdw} + \sum_{(i,j)\square\square} P_{sc} + \sum_{i\square\square} P_{contacts}. \quad (11)$$

Scoring Function Validation

Given the small sample size of transmembrane helical bundles from which to draw a picture of the "average" transmembrane helical bundle, we did not necessarily expect Bundler to identify the native structure as the least penalized bundle. Rather, we expected to be able to coarsely group bundles in such a way that their penalty would identify how near or far a given model bundle is from the native bundle and that these groupings would be dependent on the class of membrane protein from which a helical bundle is a member. This is a reasonable expectation when one considers that the minimum score structure represents the average bundle across a diverse set of transmembrane helices. As a result, we placed only modest demands on the Bundler penalty function. Our principal requirement is that it can be calibrated in such a way that the score of near-native structures clearly differentiates them from structures that are not likely to be native bundles.

To determine whether or not Bundler is capable of distinguishing the known helical bundle from a set of helical bundles close to the PDB structure, we analyzed the helical bundles of six known membrane proteins. Helical bundles were extracted as is (i.e. any distortions from ideality were maintained) from the PDB files, and only those portions of the transmembrane helices completely embedded in the membrane were considered. For example, the two short helices, 76 – 86 and 192 – 202, of Aquaporin (1fqy.pdb) only partially insert into the membrane and thus were excluded. For each structure, we derived a set of C$\square$ to C$\square$ distances corresponding to pairs of amino acids (K-K, K-D, K-E, K-C and C-C) that could potentially be obtained via chemical cross-linking using commercially available chemical cross-linkers and then added a 4 Å error to each distance. Five hundred bundles were generated for each test case by running a Monte Carlo simulated annealing algorithm at 500 °K, a temperature high enough to generate a set of structures with an RMSD spectrum of several angstroms. Specifically, we considered the following six helical bundles (PDB identifier, number of helices and number of distance constraints, respectively, are given in parentheses): bacteriorhodopsin (1c3w, 7, 60), halorhodopsin (1e12, 7, 9), rhodopsin (1f88, 7, 38), aquaporin-1 (1fqy, 6, 17), sensory rhodopsin (1jgj, 7, 18), and a subunit of fumarate reductase flavoprotein (1qlaC, 5, 58).

Figure 6 displays the results for all six test cases as plots of the Bundler function value versus distance from the known structure measured using the RMSD across the C atoms (C-RMSD). The scatter plots show the results for a representative case of 500 structures generated as outlined above for each of the test proteins. In all cases, the helical bundle from the PDB file has the lowest Bundler penalty. Moreover, the general trend is that bundles closer in C-RMSD to the known structure have lower Bundler penalty scores than those farther from the known structure. In the case of aquaporin,

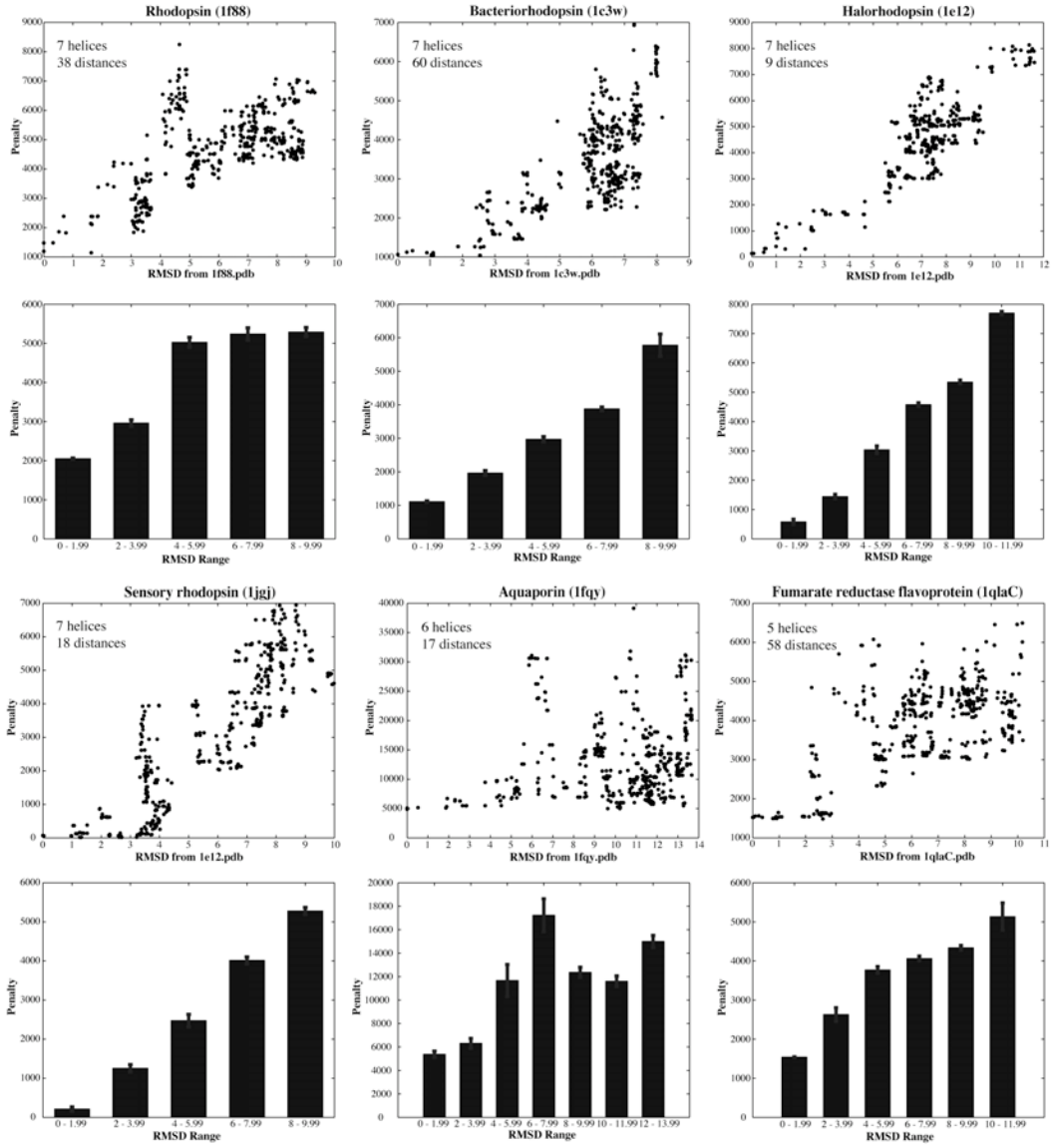

Figure 6: Bundler penalty as a function of root mean square deviation from the x-ray structure for six integral membrane proteins. Sets of 500 structures were generated using a Monte Carlo simulated annealing algorithm at a single high temperature as described in the text. Scatter plots show the results for a typical single set of 500 structures. Bar charts show the mean and standard error of 10 sets of 500 structures each generated with different random number streams. The number of helices and number of distances are provided in the inset of each scatter plot.

while the known structure had the lowest penalty, the correlation between distance from the known structure and penalty was not as strong. This lack of correlation for aquaporin

may be due to the fact that we are including only the transmembrane helices that span the membrane and omitting two short helices that are only partially inserted into the membrane, which most impacts the contacts penalty portion of the Bundler score: omission of the two short helices removes neighbors within the cut-off distance from several helices which increases the contacts penalty.

To further test the robustness of Bundler at predicting native like helical bundles, we generated 10 sets, using different random number streams, of 500 structures for each of the six test proteins. These structures were then grouped into 2 Å bins and the mean and standard deviation of the penalty was calculated within each bin (Figure 6). Overall, the structures with lower Bundler scores correspond to structures closer to the target or native structure. Thus, it is reasonable to expect that the models with the lowest Bundler scores represent structures within a few Angstroms of their corresponding native bundle. The variation in penalty within each group is small, suggesting that the trend is not due to the presence of a few very low penalty structures and a few very high penalty structures. We can thus be confident that a bundle with a higher Bundler score is not close to the native-like bundle and the bundles with the lowest Bundler penalty represent the most native-like bundles amongst the set of possible models. Excluding aquaporin, these results also provide sufficient evidence that an upper bound on the Bundler penalty can be set and used to pick a subset of models for further refinement. For example, model bundles with a Bundler penalty of less than 2000, or more conservatively 3000, are good candidates for further refinement by penalty function minimization.

Two-step approach to modeling transmembrane helical bundles using sparse distance constraints to build the rhodopsin helical bundle

The overall goal of this work was to develop a technique for building the transmembrane helical bundles of integral membrane proteins given a sparse set of distance constraints. In this section, we demonstrate a two-step approach to modeling transmembrane helical bundles. This method combines our previous work on searching the conformational space of membrane protein bundles satisfying a set of distance constraints (Faulon et al. 2003) with Monte Carlo simulated annealing (MCSA) of the empirical scoring function described in the previous sections. The method is designed to provide a computationally efficient means of searching the conformational space of the helical bundle by first searching the global space of all possible helical bundles to find those satisfying a given set of distance constraints and then searching the local conformational space of each of these candidate models. Each step is detailed in the Methods section.

The method is demonstrated using the seven transmembrane helices from the rhodopsin crystal structure 1f88.pdb, and a set of 27 distances constraints compiled from various experiments, reported in the literature, and summarized by Yeagle et al., (Yeagle et al. 2001). These included dipolar EPR distances (Farrens et al. 1996; Yang et al. 1996; Albert et al. 1997; Galasco et al. 2000), disulfide mapping distances (Yu et al. 1995; Sheikh et al. 1996; Cai et al. 1997; Cai et al. 1999; Yu et al. 1999) and distances from electron cryo-microscopy (Unger and Schertler 1995; Yeagle et al. 2001). These distance constraints are given in Table 5 and have an average error of ± 3.75 Å.

Since the published EPR dipolar distances are between nitroxide spin labels, they do not directly correspond to distances between helical axes. To better represent these distances, we determined the error associated with interpreting spin-spin distances as Cβ-Cβ distances by comparing the two measures in proteins for which distances have been measured by EPR and a crystal structure is also available. We used a total of sixteen measures for this analysis including six from rhodopsin (1F88) (Farrens et al. 1996; Yang et al. 1996; Palcewski et al. 2000), four from human carbonic anhydrase II (Hakansson et al. 1992; Persson et al. 2001), four from T4-lysozyme (3LZM) (Matsumura et al. 1989;

Table 5: Experimental distances used for modeling the rhodopsin helical bundle[1]

Helix1	*Helix2*	*Residue1*	*Residue2*	*Minimum Distance*	*Maximum Distance*	*Experimental Method*	*Reference*
C	F	139	248	6	20	Dipolar SDSL-EPR[2]	(Farrens et al. 1996)
C	F	139	249	9	26	″	″
C	F	139	250	9	26	″	″
C	F	139	251	6	20	″	″
C	F	139	252	9	26	″	″
A	G	65	316	4	19	″	(Yang et al. 1996)
E	F	204	276	4	8	Disulfide Mapping[3]	(Yu et al. 1995)
C	E	140	222	4	8	″	(Yu et al. 1999)
C	E	140	225	4	8	″	″
C	F	135	250	4	8	″	″
C	E	136	222	4	8	″	(Cai et al. 1997)
C	E	136	225	4	8	″	″
B	C	71	134	9	13	Electron Diffraction[4]	(Unger et al. 1995; Yeagle et al. 2001)
B	C	90	116	5	10	″	″
B	D	71	153	5	10	″	″
B	D	86	172	15	20	″	″
C	E	136	226	6	9	″	″
C	E	125	215	6	9	″	″
D	E	152	225	18	22	″	″
E	F	216	258	9	13	″	″
F	G	253	305	6	8	″	″
F	G	264	298	6	8	″	″
A	G	39	286	9	14	″	″
C	F	114	268	14	18	″	″
D	F	171	268	17	20	″	″
B	F	73	250	10	15	″	″
A	F	62	250	16	20	″	″
A	F	47	264	16	19	″	″

[1] Helices A, B, C, D, E, F, G correspond to residues 33-65, 70-101, 105-140, 149-173, 199-226, 245-278 and 284-309, respectively.

[2] Reported distance ranges were adjusted to account for the error involved in using spin-spin distances as Cβ – Cβ distances as described in the text.

[3] Cβ – Cβ distances from disulfide mapping were set to 5.68 Å ± (reported error) as described in the text.

[4] Cβ – Cβ distances correspond to distances measured from the top, middle and bottom of consecutive helices as described by Yeagle *et al*. (Yeagle et al. 2001).

McHaourab et al. 1997), and one each from maltose-binding protein liganded form (1MDP) (Sharff et al. 1995; Hall et al. 1997) and maltose-binding protein un-liganded form (1DMB) (Sharff et al. 1993). From this analysis, we determined the difference between spin-spin distances and Cα-Cα distances to be 4.3 ± 1.8 Å. We used this distance to adjust the lower and upper limits of the reported distances to better represent the inter-nitroxide distances as helix backbone distances. We use the reported distance plus 6 Å as an upper bound and either the minimum of the reported distance minus 6 Å and 4 Å as a lower bound. For the disulfide mapping distances, we use a Cα to Cα distance of 5.68 Å, which corresponds to two Cα to Sγ bonds (1.82 Å) and one Sγ to Sγ bond (2.04 Å), plus or minus the reported error.

In a recent paper (Faulon et al. 2003), we described a method for searching the conformation space of a set of transmembrane helices for bundles matching a given set of distance constraints. Applying this method to the seven rhodopsin helices using the 27 distance constraints given in Table 5 reduced the approximately 7.0×10^{11} possible seven-helix configurations to only 87 helical bundles with Cα – RMSDs ranging from 4.3 to 9.5 Å (Faulon et al. 2003). Thus, given only 27 distance constraints from a variety of experimental methods with differing levels of error, we were able to extract a reasonable number of structures suitable for further refinement from an overwhelmingly large dataset of possible helix bundles.

We refined each of these 87 structures using the Monte Carlo simulated annealing (MCSA) protocol described in the Methods section. The local conformation space of each helical bundle was searched for the structure with the minimum Bundler penalty function value. Since our goal is only to search the local conformational space of each bundle in a way that allows uphill moves over small barriers within a larger penalty function minima, we use a starting temperature of 30 and a geometric cooling schedule with the cooling constant set at 0.9 (i.e., $T_i = 0.9T_{i-1}$). A temperature cycle was terminated after either a total of 1000 structures were generated or after 100 structures were accepted, whichever occurred first. The MCSA simulations were run for 34 temperature steps.

Figure 7: Comparison of the predicted helical bundle (black) to the X-ray structure (1F88.pdb) helical bundle (gray). The Cα – RMSD between the two structures is 3.2 Å. As is clearly visible the helices are correctly arranged and most of the deviation is due to differences in helical tilt angles.

The least penalized structure in this cluster has a penalty of 3.3 and a Cα – RMSD from the known structure of 4.1 Å. Compared to the scores of the decoy structures tabulated in Figure 6 the Bundler penalty on this structure is much lower than those of the lowest RMSD helix assemblies. This indicates that models with Bundler penalties in the

range of 1000 to 2000 should have properties most similar to those of an “average” membrane protein helical bundle, while satisfying a set of experimental distance constraints. Among the 87 refined bundles, several have minimized penalties around 1000. The least penalized bundle among these has a Bundler score of 1003.3 and a Cα – RMSD of 3.2 Å (Figure 7). This result again provides evidence that simply minimizing an empirical structure based penalty function may not produce the ultimate best structure. Minimization drives the structure toward an “average” structure, which is not the most native-like structure for a particular protein. It is therefore essential to calibrate the function to a particular family of structures. Our results show that for seven helix bundles, the most native-like structures have Bundler penalties between approximately 1000 and 2000, which provides a better stopping criterion for our MCSA refinement protocol. For example, we could anneal the structure using a faster cooling schedule, until reaching a penalty of 2000 and then slow the cooling to more thoroughly sample conformations with Bundler scores between 1000 and 2000. The search will ultimately be stopped when the Bundler penalty drops below 1000.

Discussion

Due to the difficulties of using the standard structure determination methods for structural modeling of transmembrane proteins, it is important to develop methods using more easily obtainable, but lower resolution, data. With this in mind, we have developed a method for using sparse distance constraints to model the transmembrane spanning domain. Development of such a method is particularly timely and important given the progress in using methods such as chemical cross-linking, dipolar EPR and FRET for providing distance constraints.

We have presented a two-step approach to modeling transmembrane helical bundles and demonstrated its effectiveness by accurately modeling the transmembrane helical bundle of dark-adapted rhodopsin. In the first step, the set of all possible helical bundles is generated and filtered to find the set of bundles that satisfy the set of distance constraints using a previously reported algorithm (Faulon et al. 2003). In step two, the structures from step one are refined using a Monte Carlo simulated annealing protocol to minimize a scoring function that penalizes helical arrangements that violate distance constraints and that violate constraints derived from a statistical analysis of solved membrane protein structures from the PDB. Using a set of 27 experimental distance constraints extracted from the literature, we modeled the helical bundle of dark-adapted bovine rhodopsin to within 3.2 Å of the X-ray structure.

A major component of this work was the development and validation of a penalty function designed to discriminate near native helical bundles from those far from the native structure and thus build transmembrane helical bundles that are consistent with both experimental distance constraints and other helical bundles from known structures. Because the majority of known transmembrane protein structures are seven helix bundles, it is not surprising that the Bundler penalty function works very well for this class of membrane proteins (Figure 6). However, we have also illustrated that Bundler can be useful for modeling other classes of helical bundles (e.g. aquaporin, fumarate reductase flavoprotein).

In the case of aquaporin, the correlation between the RMSD from the crystal

structure and the Bundler score is less pronounced than for the other validation cases. Inspection of Bundler's components revealed that the relatively higher scores are due to larger contacts penalties resulting from a reduction in the number of neighboring helices within the cut-off distance, presumably caused by removal of the two partially inserted helices. Moreover, there is a high side-chain interaction preference penalty and a high helix packing angle penalty for some of the lower RMSD bundles. This again is likely due to the removal of the partially inserted helices, which in this case is likely to have removed favorable side-chain interactions and reduced the overall helix packing, allowing non-typical helix tilt angles.

Clearly a structure based penalty function for helical membrane bundles is a work in progress that will continually be updated as more structures become available. In addition to refinements of the penalty as the database of solved membrane protein structures grows, we are also investigating the value of increasing the level of molecular detail by either representing each side-chain atom explicitly or using a reduced side-chain representation such as that described by Herzyk and Hubbard (Herzyk and Hubbard 1993). Additionally, the penalty function force constants are based on the assumption that the variance of a component scales with its importance as a predictor and as such are somewhat arbitrary. Refinement of these parameters against, for example, our databases of decoy structures may also improve the penalty function. We are also exploring ways to include, either explicitly or implicitly, ligands such as retinal. Increased structural detail will impact the packing parameters of the helical bundle by enhancing the level of detail of side-chain van der Waals interactions and by increasing the accuracy of packing density calculations. The inclusion of ligands such as retinal may be necessary in order to more accurately predict helix-helix interactions that are unlike those of an average bundle. For example, in helix bundles containing a ligand, additional van der Waals interactions between helix atoms and the ligand may be necessary to force the associated helices of the bundle outside the range of allowed distances or angles derived from idealized versions of solved structures without ligands.

Moreover, our results in using this method to recover the structure of rhodopsin prompt questions as to whether similar results could be obtained using fewer distances and how the accuracy of a helical bundle generally varies with number of distance. In response, we note that the determination of accuracy solely as a function of the number of distances is non-trivial. Previously, we showed that the number possible helical bundles simultaneously satisfying a set of distance constraints varies with the number of distances, the error on these distances, and the radius of the associated distance graph or, in other words, the way in which the distances are distributed among and connect the helices (Faulon et al. 2003). This result likely carries over to the accuracy of modeling membrane proteins using Bundler, however, we have not yet carried out the extensive analysis required to confirm this assumption. For now, it suffices to say that only a modest number of distances are needed to build accurate models of transmembrane helical bundles using the approach outlined here.

It remains to be seen whether or not a truly general function useful for refining helix bundles with a range of secondary structural elements can be developed. While it is likely that the form of the penalty function presented in this paper utilizes many necessary structural components, the determination of a broader range of structures with

a varying number of transmembrane secondary structural elements may result in separate sets of statistical parameters that depend on the number of these elements. Regardless of such future findings, the approach proposed here is general, and Bundler is easily adaptable to new statistics based parameterization.

Methods

Representation of the helical bundle

For the test cases used in this study, the helices were obtained using the helix definitions provided in the PDB file. All side chain atoms beyond the Cβ were removed (i.e. we represent the helix in its native form at the Cβ level of detail). Helices are treated as rigid bodies with the helical axis defined as the line segment between the unweighted centers of mass of the last four residues of the C and N termini.

Assembly of membrane protein dataset

The membrane proteins used in this work were selected from the list of solved structures kindly provided by Professor Stephen H. White at the University of California, Irvine (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html). Proteins without definable backbone atom positions were not used (eg. 2PPS, 1FE1). Monomers, if they form a compact folding unit, were used. An exception was made for small monomers that pack together to form a helical bundle; in those cases, the entire bundle was used (eg. 1BL8). If the structure of a single protein was solved more than once, we selected the structure of the highest resolution, and if the structure was solved for multiple species, the structure for the species with the highest resolution was chosen. Heteromultimeric complexes were parsed to remove all but the transmembrane bundle subunits (eg. 1EZVC). Helices that only partially span the membrane were removed from the final bundle structures (eg. 1FQY).

Determination of force constants

The variance in the measured properties of transmembrane protein bundles is a good indicator of the importance of a given property in predicting the fold of a helical bundle. We use the variance from our analysis of a set of non-redundant structures to guide our choices of force constants in the Bundler penalty function. Those measures having the smallest variances as a percentage of the mean were assumed to be better descriptors of a helical bundle and were assigned a force constant of 500. The largest variance measure, the packing angle, is assigned a force constant of 5, and the remaining force constants were given intermediate values.

We have recently shown the importance of distance constraints in exploring the conformational space of helical bundles and in reducing the number of candidate structures for local conformational search to a reasonable number (Faulon et al. 2003). To accurately represent this importance in Bundler, we set the force constant for experimental distance constraints to the highest value of 500.

Conformational Search under a set of distance constraints

Details of our procedure for exploring the conformational space of membrane

protein folds matching distance constraints are provided in (Faulon et al. 2003) and are summarized in the methods section. Briefly, the procedure generates an exhaustive set of helix bundles within a specified RMSD by positioning the helices such that distance constraints are satisfied. The data required by step 1 is a set of individual helices in PDB format that we assume has been modeled and optimized and a set of distances. Step 1 results in a set of all possible helical bundles matching the distances such that the bundles in the set differ from one another by some user defined RMSD. These helical arrangements are described at an atomistic level suitable for further refinement by local conformational search (step 2).

Monte Carlo Simulated Annealing

In step 2 of our procedure for building an optimized helical bundle, we refine a subset of the structures from the conformational search step 1 using the Bundler penalty function developed in this paper and a Monte Carlo simulated annealing (MCSA) protocol to search the local conformational space of the bundle.

Helical Bundle

A helical bundle is defined as any arrangement of the helices in Cartesian coordinate space. The helix z-axis (z' in Figure 8) is defined as the line segment connecting the average coordinates of the N and C termini for each helix (Figure 4). Each helix has six degrees of freedom consisting of translations in the global (x, y, z) axis system and rotations in the (x', y', z') axis system (Figure 8), giving a system wide total of 6*n* degrees of freedom, where *n* is the number of helices.

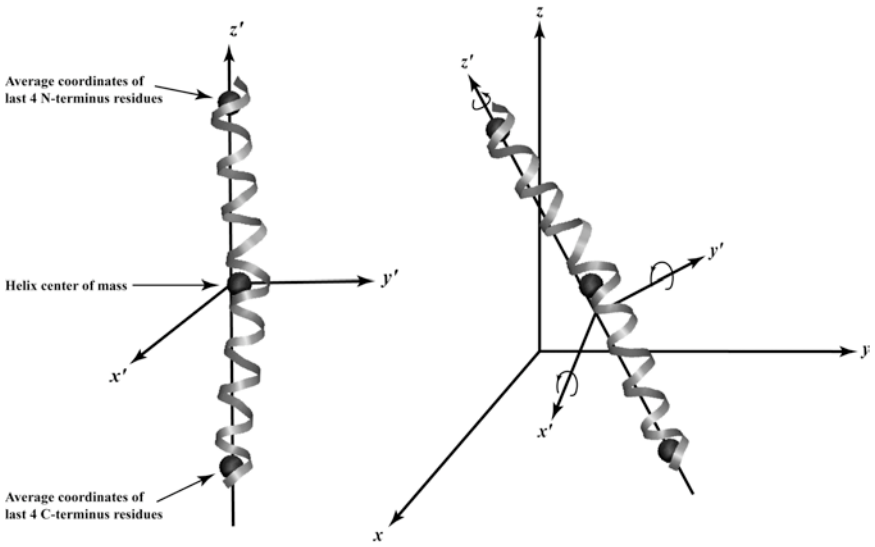

Figure 8: Definition of helix axis system (left) and helix degrees of freedom (right). The helix z-axis is defined as the vector connecting the average coordinates of the last four residues of the helix N and C termini. Helix degrees of freedom include translations in the global (x, y, z) axis system and x', y' and z' rotations around the helix axes.

Monte Carlo Sampling

Starting from the last accepted arrangement, a new helical bundle is generated by randomly selecting one of the secondary structural elements (SSEs) and randomizing its position by either translation in the global axis system (x, y, z) or rotation in the local axis system (x', y', z') (Figure 8). Similar to those used by Hertzyk and Hubbard (Herzyk and Hubbard 1995), four moves are possible (Figure 8): (1) translation along the z, (2) two consecutive translations along the x and y, (3) rotation around z' or (4) two consecutive rotations around x' and y'.

The size of this move is chosen randomly within some user defined limits. If the Bundler penalty of the new structure is lower than that of the current lowest scoring structure, then that structure is accepted as the current structure. Otherwise, the Boltzmann probability factor, *p*, is calculated as $e^{-\Delta P/T}$, where ΔP is the difference in total penalty between the least penalized structure and the newly generated structure and *T* is the temperature, which in this case is simply a parameter for controlling the probability of a given helical bundle (Kirkpatrick et al. 1983). Then *p* is compared to a random number, *r*, from a uniform [0,1] distribution. If $p < r$, the new configuration is accepted as the new best structure; otherwise, the new bundle is rejected (Metropolis et al. 1958).

Cooling Schedule

The cooling schedule used for the refinements of step 2 started at $T = 30$ and was reduced at each new temperature cycle according to a geometric temperature schedule with the temperature reduction factor set to 0.95 (i.e., $T_i = 0.95T_{i-1}$). Thirty-four temperature cycles were completed, and each temperature cycle terminated after either 1000 Monte Carlo steps were completed or after 100 candidate structures were accepted.

Structural analysis and data processing

Membrane protein statistics were calculated using in-house software. Root mean square deviation calculations and various manipulations of PDB files were performed using the Multiscale Modeling Tools in Structural Biology, MMTSB, tool set (Feig et al. 2001). Molecular visualization and renderings were obtained using VMD (Humphrey et al. 1996). All analysis of the penalty data was done using programs written in MATLAB 6.5 (The Math Works Inc., Natick, MA).

References

Adamian, L., Jackups, R., Jr., Binkowski, T.A., and Liang, J. 2003. Higher-order interhelical spatial interactions in membrane proteins. *J Mol Biol* **327:** 251-272.

Adamian, L., and Liang, J. 2001. Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. *J Mol Biol* **311:** 891-907.

Albert, A.D., Watts, A., Spooner, P., Grobner, G., Young, J., and Yeagle, P.L. 1997. A solid state NMR characterization of the substrate binding specificity and dynamics for the L-fucose-H+ membrane transport protein of E. coli. *Biochim. Biophys. Acta* **1328:** 74 - 82.

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

Back, J.W., Sanz, M.A., De Jong, L., De Koning, L.J., Nijtmans, L.G., De Koster, C.G., Grivell, L.A., Van Der Spek, H., and Muijsers, A.O. 2002. A structure for the yeast prohibitin complex: Structure prediction and evidence from chemical crosslinking and mass spectrometry. *Protein Sci* **11:** 2471-2478.

Bennett, K.L., Kussmann, M., Bjork, P., Godzwon, M., Mikkelsen, M., Sorensen, P., and Roepstorff, P. 2000. Chemical cross-linking with thiol-cleavable reagents combined with differential mass spectrometric peptide mapping--a novel approach to assess intermolecular protein contacts. *Protein Sci* **9:** 1503-1518.

Borbat, P.P., Costa-Filho, A.J., Earle, K.A., Moscicki, J.K., and Freed, J.H. 2001. Electron spin resonance in studies of membranes and proteins. *Science* **291:** 266-269.

Bowie, J.U. 1997. Helix packing in membrane proteins. *J. Mol. Biol* **272:** 780-789.

Bowie, J.U. 1999. Helix-bundle membrane protein fold templates. *Protein Science* **8:** 2711-2719.

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

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. 1998. Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Cryst. D54***:** 905-921.

Buchan, D., Shephard, D., Lee, D., Peral, F., Rison, S., Thorton, J., and Orengo, C. 2002. Structural assignment for whole genes and genomes using the CATH domain structure database. *Genome Research* **12:** 503-514.

Cai, K., Klein-Seetharaman, J., Altenbach, C., Hubbell, W.L., and Khorana, H.G. 2001. Probing the dark state tertiary structure in the cytoplasmic domain of rhodopsin: proximities between amino acids deduced from spontaneous disulfide bond formation between cysteine pairs engineered in cytoplasmic loops 1, 3, and 4. *Biochemistry* **40:** 12479-12485.

Cai, K., Klein-Seetharaman, J., Hwa, J., Hubbell, W.L., and Khorana, H.G. 1999. Structure and Function in Rhodopsin. Effects of Disulfide Cross-Links in the Cytoplasmic Face of Rhodopsin on Transducin Activation and Phosphorylation by Rhodopsin Kinase. *Biochemistry* **38:** 12893 - 11898.

Cai, K., Langen, R., Hubbell, W.L., and Khorana, H.G. 1997. Structure and function in rhodopsin: topology of the C-terminal polypeptide chain in relation to the cytoplasmic loops. *Proc. Natl. Acad. Sci. USA* **94:** 14267 - 14272.

Chothia, C. 1975. Structural invariants in protein folding. *Nature* **254:** 304-308.

Dihazi, G.H., and Sinz, A. 2003. Mapping low-resolution three-dimensional protein structures using chemical cross-linking and Fourier transform ion-cyclotron resonance mass spectrometry. *Rapid Commun Mass Spectrom* **17:** 2005-2014.

Dobbs, H., Orlandini, E., Bonaccini, R., and Seno, F. 2002. Optimal potentials for predicting inter-helical packing in transmembrane proteins. *Proteins* **49:** 342-349.

Engelman, D.M., Steitz, T.A., and Goldman, A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem* **15:** 321-353.

Farrens, D.L., Altenbach, C., Yang, K., Hubbell, W.L., and Khorana, H.G. 1996. Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin. *Science* **274:** 768 - 770.

Faulon, J.-L., Sale, K., and Young, M. 2003. Exploring the conformational space of membrane protein folds matching distance constraints. *Protein Science* **12:** 1750-1761.

Feig, M., Karanicolas, J., and Brooks, C.L.I. 2001. MMTSB Tool Set. MMTSB NIH Research Resource, The Scripps Research Institute.

Fleishman, S.J., and Ben-Tal, N. 2002. A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. *J Mol Biol* **321:** 363-378.

Galasco, A., Crouch, R.K., and Knapp, D.R. 2000. Intrahelic arrangement in the integral membrane protein rhodopsin investigated by site-specific chemical cleavage and mass spectrometry. *Biochemistry* **39:** 4907 - 4914.

Hakansson, K., Carlsson, M., Svensson, L.A., and Liljas, A. 1992. Structure of native and apo carbonic anhydrase II and structure of some of its anion-ligand complexes. *Journal of Molecular Biology* **227:** 1192.

Hall, J.A., Thorgeirsson, T.E., Liu, J., Shin, Y.K., and Nikaido, H. 1997. Two modes of ligand binding in maltose-binding protein of Escherichia coli. Electron paramagnetic resonance study of ligand-induced global conformational changes by site-directed spin labeling. *J Biol Chem* **272:** 17610-17614.

Herzyk, P., and Hubbard, R.E. 1993. A reduced representation of proteins for use in restraint satisfaction calculations. *Proteins* **17:** 310-324.

Herzyk, P., and Hubbard, R.E. 1995. Automated method for modeling seven-helix transmembrane receptors from experimental data. *Biophys J* **69:** 2419-2442.

Hillisch, A., Lorenz, M., and Diekmann, S. 2001. Recent advances in FRET: distance determination in protein-DNA complexes. *Curr Opin Struct Biol* **11:** 201-207.

Hubbell, W.L., Altenbach, C., Hubbell, C.M., and Khorana, H.G. 2003. Rhodopsin structure, dynamics, and activation: a perspective from crystallography, site-directed spin labeling, sulfhydryl reactivity, and disulfide cross-linking. *Adv Protein Chem* **63:** 243-290.

Humphrey, W., Dalke, A., and Schulten, K. 1996. VMD - Visual Molecular Dynamics. *Journal of Molecular Graphics* **14:** 33-38.

Hustedt, E.J., and Beth, A.H. 1999. Nitroxide spin-spin interactions: applications to protein structure and dynamics. *Annu Rev Biophys Biomol Struct* **28:** 129-153.

Hustedt, E.J., Smirnov, A.I., Laub, C.F., Cobb, C.E., and Beth, A.H. 1997. Molecular distances from dipolar coupled spin-labels: the global analysis of multifrequency continuous wave electron paramagnetic resonance data. *Biophys J* **72:** 1861-1877.

Jacobs, R.E., and White, S.H. 1989. The nature of the hydrophobic binding of small peptides at the bilayer interface: implications for the insertion of transbilayer helices. *Biochemistry* **28:** 3421-3437.

Jayasinghe, S., Hristova, K., and White, S.H. 2001a. Energetics, stability, and prediction of transmembrane helices. *J Mol Biol* **312:** 927-934.

Jayasinghe, S., Hristova, K., and White, S.H. 2001b. MPtopo: A database of membrane protein topology. *Protein Sci* **10:** 455-458.

Kim, S., Chamberlain, A.K., and Bowie, J.U. 2003. A simple method for modeling transmembrane helix oligomers. *Journal of Molecular Biology* **329:** 831-840.

Kim, S., and Cross, T.A. 2002. Uniformity, ideality, and hydrogen bonds in transmembrane alpha-helices. *Biophys J* **83:** 2084-2095.

Kirkpatrick, S., Gerlatt, C.J., and Vecchi, M. 1983. Optimization by simulated annealing. *Science* **220:** 671-680.

Klostermeier, D., and Millar, D.P. 2001. Time-resolved fluorescence resonance energy transfer: a versatile tool for the analysis of nucleic acids. *Biopolymers* **61:** 159-179.

Kruppa, G.H., Schoeniger, J., and Young, M.M. 2003. A top down approach to protein structural studies using chemical cross-linking and Fourier transform mass spectrometry. *Rapid Commun Mass Spectrom* **17:** 155-162.

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

Matsumura, M., Wozniak, J.A., Sun, D.P., and Matthews, B.W. 1989. Structural studies of mutants of T4 lysozyme that alter hydrophobic stabilization. *J Biol Chem* **264:** 16059-16066.

Matyus, L. 1992. Fluorescence resonance energy transfer measurements on cell surfaces. A spectroscopic tool for determining protein interactions. *J Photochem Photobiol B* **12:** 323-337.

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

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. 1958. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21:** 1087 - 1092.

Nikiforovich, G.V., Galaktionov, S., Balodis, J., and Marshall, G.R. 2001. Novel approach to computer modeling of seven-helical transmembrane proteins: Current progress in the test case of bacteriorhodopsin. *Acta Biochimica Polinica* **48:** 53-64.

Novak, P., Young, M. M., Schoeniger, J., Kruppa, G. H. 2003. A top down approach to protein structure studies using chemical crosslinking and fourier transform mass spectrometry. *European Journal of Mass Spectrometry.* **9:** 623-631.

Palcewski, K., Kumasaka, T., Hori, T., Behnke, C.A., Motoshima, H., Fox, B.A., Le Trong, I., Teller, D.C., Okada, T., Stenkamp, R.E., et al. 2000. Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science* **289:** 739-745.

Parkhurst, L.J., Parkhurst, K.M., Powell, R., Wu, J., and Williams, S. 2001. Time-resolved fluorescence resonance energy transfer studies of DNA bending in double-stranded oligonucleotides and in DNA-protein complexes. *Biopolymers* **61:** 180-200.

Perozo, E., Cuello, L.G., Cortes, D.M., Liu, Y.S., and Sompornpisut, P. 2002. EPR approaches to ion channel structure and function. *Novartis Found Symp* **245:** 146-158; discussion 158-164, 165-148.

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

Popot, J.L., and Engelman, D.M. 1990. Membrane protein folding and oligomerization: the two-stage model. *Biochemistry* **29:** 4031-4037.

Rabenstein, M.D., and Shin, Y.K. 1995. Determination of the distance between two spin labels attached to a macromolecule. *Proc Natl Acad Sci U S A* **92:** 8239-8243.

Radzwill, N., Gerwert, K., and Steinhoff, H.J. 2001. Time-resolved detection of transient movement of helices F and G in doubly spin-labeled bacteriorhodopsin. *Biophys J* **80:** 2856-2866.

Rappsilber, J., Siniossoglou, S., Hurt, E.C., and Mann, M. 2000. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal Chem* **72:** 267-275.

Richards, F. 1974. The interpretation of protein structures: total volume, group volume distributions and packing density. *Journal of Molecular Biology* **82:** 1-14.

Rose, G.D. 1978. Prediction of chain turns in globular proteins on a hydrophobic basis. *Nature* **272:** 586-590.

Rye, H.S. 2001. Application of fluorescence resonance energy transfer to the GroEL-GroES chaperonin reaction. *Methods* **24:** 278-288.

Schilling, B., Row, R.H., Gibson, B.W., Guo, X., and Young, M.M. 2003. MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. *J Am Soc Mass Spectrom* **14:** 834-850.

Sekar, R.B., and Periasamy, A. 2003. Fluorescence resonance energy transfer (FRET) microscopy imaging of live cell protein localizations. *J Cell Biol* **160:** 629-633.

Sharff, A.J., Rodseth, L.E., and Quiocho, F.A. 1993. Refined 1.8-A structure reveals the mode of binding of beta- cyclodextrin to the maltodextrin binding protein. *Biochemistry* **32:** 10553.

Sharff, A.J., Rodseth, L.E., Szmelcman, S., Hofnung, M., and Quiocho, F.A. 1995. Refined structures of two insertion/deletion mutants probe function of the maltodextrin binding protein. *Journal of Molecular Biology* **246:** 8.

Sheikh, S.P., Zvyaga, T.A., Lichtarge, O., Sakmar, T.P., and Bourne, H.R. 1996. Rhodopsin activation blocked by metal-ion-binding sites linking transmembrane helices C and F. *Nature* **383:** 347 - 350.

Steinhoff, H.J., Radzwill, N., Thevis, W., Lenz, V., Brandenburg, D., Antson, A., Dodson, G., and Wollmer, A. 1997. Determination of interspin distances between spin labels attached to insulin: comparison of electron paramagnetic resonance data with the X-ray structure. *Biophys J* **73:** 3287-3298.

Szollosi, J., Nagy, P., Sebestyen, Z., Damjanovicha, S., Park, J.W., and Matyus, L. 2002. Applications of fluorescence resonance energy transfer for mapping biological membranes. *J Biotechnol* **82:** 251-266.

Taverner, T., Hall, N.E., O'Hair, R.A., and Simpson, R.J. 2002. Characterization of an antagonist interleukin-6 dimer by stable isotope labeling, cross-linking, and mass spectrometry. *J Biol Chem* **277:** 46487-46492.

Unger, V.M., and Schertler, G.F. 1995. Low resolution structure of bovine rhodopsin determined by electron cryo-microscopy. *Biophysical Journal* **68:** 1776-1786.

Vaidehi, N., Floriano, W.B., Trabanino, R., Hall, S.E., Freddolino, P., Choi, E.J., Zamanakos, G., and Goddard III, W.A. 2002. Prediction of structure and function of G protrein-couple receptors. *PNAS* **99:** 12622-12627.

White, S.H. 2003. Membrane proteins of known 3D structure. *http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html*.

White, S.H., and Wimley, W.C. 1998. Hydrophobic interactions of peptides with membrane interfaces. *Biochim Biophys Acta* **1376:** 339-352.

White, S.H., and Wimley, W.C. 1999. Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct* **28:** 319-365.

Wilson, S., and Bergsma, D. 2000. Orphan G-protein coupled receptors: novel drug targets for the pharmaceutical industry. *Drug Dessign and Discovery* **17:** 105-114.

Yang, K., Farrens, D.L., Hubbell, W.L., and Khorana, H.G. 1996. Structure and function in rhodopsin. Single cysteine substitution mutants in the cytoplasmic interhelical E-F loop region show position-specific effects in transducin activation. *Biochemistry* **35:** 14040 - 14046.

Yeagle, P.L., Choi, G., and Albert, A.D. 2001. Studies on the structure of the G-protein-coupled receptor rhodopsin including the putative G-protein binding site in unactivated and activated forms. *Biochemistry* **40:** 11932-11937.

Young, M., Tang, N., Hempel, J., Oshiro, C., Taylor, E., Kuntz, I., Gibson, B., and Dollinger, G. 2000. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **97:** 5802-5806.

Yu, H., Kono, M., McKee, T.D., and Oprian, D.D. 1995. A general method for mapping tertiary contacts between amino acid residues in membrane embedded proteins. *Biochemistry* **34:** 14963 - 14969.

Yu, H., Kono, M., and Oprian, D.D. 1999. State-dependent Disulfide Cross-linking in Rhodopsin. Biochemistry 38: 12028 - 12032.

Chapter 3: Optimizing an Empirical Scoring Function for Transmembrane Protein Structure Determination

Genetha Anne Gray and Tamara G. Kolda, Ken Sale and Malin M. Young

Abstract

We examine the problem of transmembrane protein structure determination. Like many questions that arise in biological research, this problem cannot be addressed generally by traditional laboratory experimentation alone. Instead, an approach that integrates experiment and computation is required. We formulate the transmembrane protein structure determination problem as a bound-constrained optimization problem using a special empirical scoring function, called Bundler, as the objective function. In this paper, we describe the optimization problem and its mathematical properties, and we examine results obtained using two different derivative-free optimization algorithms.

Introduction

In this study, we formulate the transmembrane protein structure determination problem as a bound-constrained nonlinear optimization problem,

$$\begin{aligned} &\min f(x) \\ &\text{s.t. } l \le x \le u, \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a nonlinear function; $x, l, u \in \mathbb{R}^n$; and l and u are given lower and upper bounds on x, respectively. In this application, the objective function f is an empirical scoring function designed to rate the validity of proposed transmembrane protein structures. The variable $x \in \mathbb{R}^n$ represents the spatial positions of certain components of the transmembrane protein, and the bounds l and u are derived using observed properties of these components.

There is a wide variety of optimization methods available for finding a solution to (1). However, the effectiveness and efficiency of these algorithms can be application-specific. Hence, answering the question of which to use is not easy. In this paper, we examine the transmembrane protein structure identification problem and its model formulation. We consider two different optimization algorithms that are appropriate for this application. We discuss why we chose these two methods and compare and contrast numerical results for a transmembrane protein of known structure.

This paper is organized as follows: in the Biological Background section, we discuss the biological significance of transmembrane proteins and the importance of determining their structures. Then, in the Transmembrane Protein Structure Determination section, we describe the mathematical formulation of the transmembrane protein structure determination problem and give some details of the scoring function. In the Optimization Methods section we review the basic characteristics of the optimization methods that we chose to apply to the problem, motivate their use, and give the details of their implementations. The results of our numerical study are presented in the Numerical

Results section. Finally, we summarize our work and draw conclusions.

Biological Background

Approximately one third of the proteins encoded for by a typical genome are transmembrane proteins (Buchan et al., 2002), and they participate in many important cellular processes. Some transmembrane proteins form a channel through which certain ions and molecules can enter or leave the cell. Others act as signal transduction receptors or play roles in cell recognition, senses mediation, or cell-to-cell communication. Many diseases are the result of transmembrane protein malfunction, absence, or mutation. Hence, these proteins are an important target of drug design. In fact, a large percentage of the current pharmaceuticals act on transmembrane proteins (Wilson and Bergsma, 2000). Additional information about the structure and function of transmembrane proteins can be found in texts such as (Brandon and Tooze, 1999; Banaszak, 2000; and Creighton, 1992), and references therein.

Like all proteins, a transmembrane protein is a macromolecule consisting of a chain of amino acids. The defining characteristic of a transmembrane protein is that this chain traverses the cell membrane one or more times. For example, a G-protein-coupled receptor, one type of transmembrane protein involved in signal transduction, spans the cell membrane seven times. The portion of the transmembrane protein within the cell membrane consists primarily of hydrophobic amino acids, while the portion outside the cell membrane consists mainly of hydrophilic amino acids. These characteristics, in conjunction with the makeup of the cell membrane, dictate the overall structure of transmembrane proteins. In particular, due to the chemical environment of the membrane interior, the amino acids that are inside the cell membrane form stable secondary structures including α-helices and β-sheets. To date, two major structural classes of transmembrane domains have been observed: all α-helical and all β-stranded. We will limit the subsequent discussion to the all α-helical case and note that 20 % to 30 % of a genome's proteins are likely to have a transmembrane helical domain (Wallin and von Heigne, 1998; Krough et al., 2001). In this study, we consider a transmembrane protein to consist of a bundle of connected α-helices. Figure 1 contains an illustration of the transmembrane protein rhodopsin contained in a retina cell membrane. In this cartoon, the seven linked cylinders, labeled A through G, represent the seven α-helices that traverse the cell membrane. (Note that this cartoon was obtained from the G-protein-coupled receptor website (Hulsen and Lutje-Hulsik, 2001).)

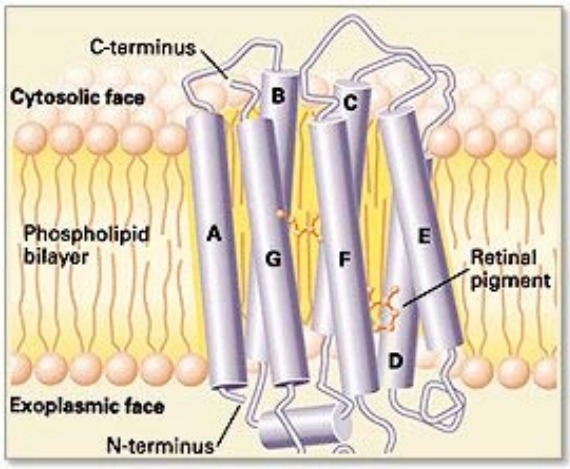

Figure 1: Illustration of the transmembrane protein rhodopsin.

As of May 2004, the protein data bank (PDB) contains over 25,000 structures, and

its size is increasing exponentially (Berman et al., 2000). However, the majority of the proteins found in the PDB are soluble proteins. In contrast, the structures of only about 80 transmembrane proteins have been determined (see (White, 1998) and references therein). This is due to the fact that experimental structure determination methods such as X-ray crystallography and nuclear magnetic resonance (NMR) have been difficult to apply to transmembrane proteins. Furthermore, since so few transmembrane protein structures have been determined, few suitable templates exist for homology modeling (Herzyk and Hubbard, 1998). Therefore, the development of an integrated computational/experimental model to address transmembrane protein structure and function questions is an important challenge in the field of structural biology. The modeling of transmembrane proteins can be broken up into the separate tasks of defining the transmembrane helices and determining the relative orientation of these helices. A process known as *sliding-window hydrophobicity* is an accurate and well established method of predicting transmembrane helices given their amino acid sequences (Rose, 1978; Jayasinghe et al., 2001a; Jayasinghe et al., 2002). No widely accepted method has yet emerged to subsequently ascertain the spatial locations of these helices. Because the cell membrane imposes certain structural constraints on the positions of the helices and thus limits the number of possible structures, several ab-initio computational approaches have been proposed (Bowie, 1999; Nikiforovich et al., 2001; Vaidehi et al., 2002). One such procedure is based on the fact that the conformational space of membrane proteins can be effectively sampled and enumerates all the possible helical bundles (Bowie, 1999).

However, this method neglects the orientations of the individual helices around their respective axes. Several other promising methods have been specifically designed for G-protein coupled receptors but have yet to be validated for other transmembrane proteins (Nikiforovich et al., 2001; Vaidehi et al., 2002; Dobbs et al., 2002).

Transmembrane Protein Structure Determination

In (Faulon et al., 2003; Sale et al., 2004), a novel two-step approach for determining the spatial location of the transmembrane protein helices is proposed. The first step, described in detail in (Faulon et al., 2003), involves searching the conformational space of transmembrane proteins to find a set of candidate helical bundles matching some given experimental distance constraints. The second step refines and reduces this set of bundles via optimization techniques. Using the structures obtained in step one as starting points, solutions to problem (1) are sought, where the objective function f assigns a score to each candidate helical arrangement that indicates how similar it is to the true structure. The minimization problem of step two is the focus of this paper.

Mathematical Description of the Problem

In this study, determining the structure of the transmembrane protein focuses on describing the relative orientation of the helices, or how they bundle. Each helix is assumed to be a rigid body, so we describe its position in space using its center of mass and a line segment defined by the two points centered in the terminal turns of the helical ends. We define a three-dimensional reference space for each helix using its initial center of mass and initial helix axial line segment. In other words, the position of each helix is defined in terms of its original location. Then, the variables in (1) are merely the x, y, and z translations from the original centers of mass of each helix and the x, y, and z rotations

about the initial helix axial line segment for each helix. This is illustrated in Figure 2. Hence, a transmembrane protein with *m* helices has 6*m* variables. At this time, we do not consider the loops that connect the helices as part of the structure determination but note that they can be added via existing techniques after the helical positions have been established (Vriend, 1990; Xiang and Honig, 2001).

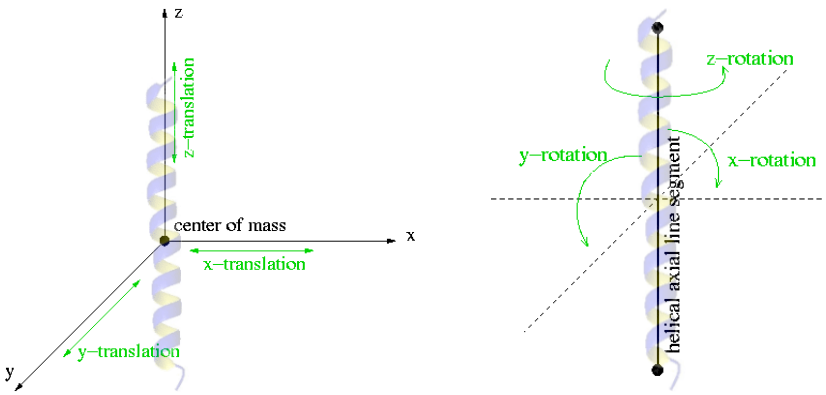

Figure 2: Depiction of the six positional variables associated with each helix.

Most of the 6*m* variables have simple bounds that derive from the fact that transmembrane proteins reside in the cell membrane. The restrictions on the *x* and *y* rotations of each helix are based on the survey of helix tilt angles given in (Bowie, 1997). The *z* rotational variables have no such limitations and are allowed to vary in the entire *z* rotational space. Both the *x* and the *y* translations are confined to a space that is approximately one third of the total radius of the membrane protein. These constraints are based on the study of helix packing behavior presented in (Bowie, 1997). The *z* translation variables have the tightest bounds to restrict the helical portions of the transmembrane protein to the interior of the cell membrane.

We now need a way to compare possible structures and decide which one best approximates the transmembrane protein in question. If the structure were known, such comparisons could be made simply using root mean square deviation (RMSD), a way of comparing two protein structures by calculating the sum of the distances of comparable atoms (see, for example, (Leach, 2001)). However, the overall goal of this work is to identify unknown transmembrane protein structures, so we must develop another technique. We use a penalty scoring function known as *Bundler* to rate each structure (Sale et el., 2004). Bundler measures how well a structure conforms to specific criteria based on experimental data and helix bundling features described in the literature, and it does not require any a priori knowledge of the location of the helices. The Bundler score is smallest for those structures that most closely meet the specified criteria. Thus, we define an objective function *f* for problem (1) using Bundler to give this structure a score. Therefore, minimizing *f* is the computational tool for determining the structure of a transmembrane protein.

The Scoring Function: Bundler

As previously stated, the Bundler scoring function combines experimental data and topological models created from a survey of known transmembrane helix packing interactions. For each structure, the score is calculated as the sum

$$P = P_E + P_I, \tag{2}$$

where P_E quantifies the structure's violation of a set of experimental distance constraints and P_I quantifies how well the structure satisfies some helix packing parameters determined by analyzing a set of 16 nonredundant membrane proteins (Sale et al., 2004).

It has been shown that distance constraints are important in determining transmembrane protein structure. In fact, the number of possible structures decreases exponentially with the number of distance constraints available and increases exponentially with the error on the distance measures (Faulon et al., 2003). Bundler incorporates experimental distance constraints in the term

$$P_{\mathrm{E}} = \sum_{(a,b)\in\Omega} K_{\mathrm{E}} * \begin{cases} (d_{ab} - l_{ab})^2, & d_{ab} < l_{ab} \\ 0, & l_{ab} \le d_{ab} \le u_{ab} \\ (u_{ab} - d_{ab})^2, & d_{ab} > u_{iab} \end{cases} \quad (3)$$

where l_{ab} and u_{ab} are predetermined upper and lower bounds on the distance between atoms a and b, respectively; d_{ab} is the distance between atoms a and b in the current structure; Ω is a subset of atom pairs; and K_E is a force constant. The distance constraints where l_{ab} and u_{ab} are obtained from experimental methods such as chemical cross-linking, dipolar electron paramagnetic resonance (dipolar EPR) (Berliner et al., 2001), fluorescence resonance energy transfer (FRET), or NMR for assembling transmembrane helical proteins (Krishna and Berliner, 1999). Note that these constraints are not procurable for every pair of atoms in the structure. Instead, experimental distance constraints are only available for a small subset, Ω, of all atom pairs.

Obtaining enough distance constraints to determine a structure uniquely is difficult, particularly for transmembrane proteins (Faulon et al., 2002; Faulon et al., 2003). Furthermore, these distances are not error-free. Therefore, to better identify desirable structures, Bundler also includes a term that measures correspondence to observed helix packing properties (determined from an analysis of known structures). This term, P_I is actually a sum of six different terms:

$$P_I = P_{\Delta} + P_{\Delta} + P_{\Delta} + P_{sc} + P_{vdw} + P_c. \quad (4)$$

Each term checks a different helical bundling property.

The packing distance score, P_{Δ}, and packing angle score, P_{Δ}, consider all the helical pairs in the bundle and penalize them if they are too far apart or too close together. More specifically, the packing distance score gauges how far apart two helices are in terms of their centers of mass, and the packing angle score examines the angle between two helices in terms of their axial line segments. Let Δ denote the set of $m(m-1)/2$ distinct helical pairs (i, j). Then, the packing distance score is defined as

$$P_{\Box} = \sum_{(i,j)\Box\Box} K_{\Box} * \begin{cases} (\Box_{ij} - \Box_l)^2, & \Box_{ij} < \Box_l \\ 0, & \Box_l \Box \Box_{ij} \Box \Box_u \\ (\Box_u - \Box_{ij})^2, & \Box_{ij} > \Box_u \end{cases}. \quad (5)$$

Here, $\Box_l = \bar{\Box} - 1.5 s_{\Box}$ and $\Box_u = \bar{\Box} + 1.5 s_{\Box}$, where $\bar{\Box}$ and $s_{\Box}$ are the mean and standard deviation of the inter-helical distances, respectively, which are calculated using a set of 16 known structures; $\Box_{ij}$ is the distance between the centers of mass of helices *i* and *j* in the current structure; and $K_{\Box}$ is a given force constant. Similarly, the packing angle score is defined as

$$\sum_{(i,j)\Box\Box} K_{\Box} * \begin{cases} (\Box_{ij} - \Box_l)^2, & \Box_{ij} < \Box_l \\ 0, & \Box_l \Box \Box_{ij} \Box \Box_u \\ (\Box_u - \Box_{ij})^2, & \Box_{ij} > \Box_u \end{cases}, \quad (6)$$

where $\Box_l = \bar{\Box} - 1.5 s_{\Box}$ and $\Box_u = \bar{\Box} + 1.5 s_{\Box}$, and $\bar{\Box}$ and $s_{\Box}$ are the mean and standard deviation of the inter-helical packing angles; $\Box_{ij}$ is the inter-helical packing angle between helices *i* and *j* in the current structure; and $K_{\Box}$ is a given force constant.

The packing density is defined as the ratio of atomic volume to solvent accessible volume (Richards, 1974). It gauges how efficiently a protein folds together or, equivalently, how much interior space is left unused. The packing density score, $P_{\Box}$, is defined as

$$P_{\Box} = K_{\Box} * \begin{cases} (\Box - \Box_l)^2, & \Box < \Box_l \\ 0, & \Box_l \Box \Box \Box \Box_u \\ (\Box_u - \Box)^2, & \Box > \Box_u \end{cases}, \quad (7)$$

where $\Box_l = \bar{\Box} - 1.5 s_{\Box}$ and $\Box_u = \bar{\Box} + 1.5 s_{\Box}$, and $\bar{\Box}$ and $s_{\Box}$ are the mean and standard deviation of the observed packing density; $\Box$ is the packing density of the current structure; and $K_{\Box}$ is a given force constant. It penalizes structures that are packed too tightly or too loosely.

In transmembrane proteins, it has been observed that amino acids have a preference for which amino acids they interact with on neighboring helices (Adiman and Liang, 2001; Nikiforovich et al., 2001; Adamian et al., 2003). The side-chain interaction propensity score, P_{sc}, incorporates this into Bundler. It is based on the membrane helical inter-facial pairwise (MHIP) amino acid interaction propensity table in (Adiman and Liang, 2001), and it penalizes structures containing amino acid pairs that are in contact contrary to their normal observed behavior. Let $\Box_i$ be the set of C$\Box$ atoms (see, for example, (Brandon and Tooze, 1999)) in helix *i* and $\circ$ be the set of *m* consecutive helical pairs. (Note that two helices are a consecutive pair if they are directly connected by an

outer loop.) Then, the side-chain propensity score is defined as

$$P_{sc} = \sum_{(i,j)\in\Omega} \left(\sum_{a\in A_i, b\in A_j} K_{sc} * (p - p_{ab}) \right) \tag{8}$$

where p is the maximum propensity score in the MHIP table, p_{ab} is the MHIP propensity value of atoms a and b, and K_{sc} is a constant.

To prevent inter-helical clashes, Bundler includes the van der Waals repulsive function (Brünger, 1992)

$$P_{vdw} = \sum_{(a,b)\in\Psi} K_{vdw} * \begin{cases} 0, & r_{ab} \geq sR_{ab} \\ \left(s^2 R_{ab}^2 - r_{ab}^2\right)^2, & r_{ab} < sR_{ab} \end{cases}, \tag{9}$$

Here, Ψ is the set of all distinct pairs of Cα atoms, r_{ab} is the distance between Cα atoms a and b in the current structure, R_{ab} is the observed distance at which atoms a and b interact or repulse, s is a predetermined van der Waals scaling factor, and K_{vdw} is a given constant.

Finally, to ensure that each helix has at least two neighboring helices, Bundler includes a contact score. This piece of the scoring function guarantees that the helices are packed tightly and prevents any one helix from being excluded from the bundle. It is defined as

$$P_c = \sum_{i\in\Phi} K_c * \begin{cases} 0, & c_i \geq 2 \\ (2 - c_i), & c_i < 2 \end{cases} \tag{10}$$

where Φ is the set of helices; c_i is the number of helices that helix i is in contact with; and K_c is a given constant. Two helices are defined to be in contact if their centers of mass are within a given distance of one another. This distance bound is calculated using the analysis of the 16 known structures.

Observe that both the side-chain interaction propensity score, P_{sc}, and the contact score, P_c, introduce discontinuities in the Bundler scoring function. Moreover, P_E, P_Ω, P_Ψ, and P_Φ contain points at which the derivative is undefined. These properties of Bundler are worth noting as they affect our choice of optimization method. We also note that all the pieces of the Bundler scoring function contain at a least one constant as well as some predetermined bounds. Setting these parameters is an important component of the transmembrane protein structure determination problem but does not effect the optimization of Bundler and is thus not addressed in this paper.

Optimizing Bundler

In this paper, we are interested in the details of optimizing the Bundler scoring function, and so we have included only a basic description of Bundler. Further details and more specific explanations of the function's development and validation are not critical to our numerical study and can be found in (Sale et al., 2004). However, we wish to make some comments and observations about Bundler in terms of our optimization

goals and expectations.

First, we reiterate the fact that the Bundler scoring function incorporates real data obtained via laboratory experimentation. Hence, there is a certain amount of noise in our objective function. At present, there is no regularization term in the Bundler scoring function to prevent fitting this noise, and thus it is not productive to demand that an optimization algorithm yield a structure with a Bundler score of zero. Moreover, we have observed that small variations in Bundler scores result in only noise-level differences in the structures (Sale et al., 2004), and so we do not require a high level of accuracy from the optimization method.

Secondly, we remind the reader that optimizing the Bundler scoring function is the second step of a method for determining the spatial locations of the helices of a transmembrane protein. In the first step, the discrete conformational space is reduced to hundreds or even thousands of candidate helical bundles to be used as the starting points in the second step, minimizing (1). In order to attain a small number of final candidates for further study, we require a fast and efficient optimization method capable of further refining the results of step one.

Finally, it should be noted that the Bundler scoring function incorporates helix packing parameters defined using a very small sample (16 nonredundant proteins) of transmembrane helical bundles. Until this set can be dramatically increased, we do not necessarily expect Bundler to identify the true (or native) structure as the structure with the absolute lowest score. Instead, we have designed the Bundler scoring function to serve as an empirical measure for differentiating between groups of bundles that are far from the native structure from those that are near. It is still unclear to us how low the Bundler score of a structure must be in order for that structure to be of use in our process of protein structure determination. We believe that the threshold of useful scores will vary from protein to protein and thus must be determined empirically.

Optimization Methods

Since the Bundler scoring function is non-smooth and contains discontinuities, we have chosen to apply derivative-free methods to obtain a solution to (1). Although we focus on two particular methods here, there are many other derivative-free methods (see, for example, (Powell, 1998; Kolda et al., 2003) and references therein). Moreover, finite differencing could be used to approximate the gradient so that we could use derivative-based methods. However, because Bundler is discontinuous and directly incorporates noisy experimental distance constraints, such approximations may contain too much error to be useful (Hough and Meza, 2002). In this paper, we present results using simulated annealing and parallel pattern search, described below.

Simulated Annealing

Simulated annealing (SA) is arguably the most widely used optimization method for molecular conformation problems. For just a few of the many examples of the use of simulated annealing in computational biology, (Ghosh et al., 2002; Perkins and Dean, 1993; Campbell et al., 1998; Goodsell and Olson 1990; Brünger et al., 1997). The SA algorithm is a computational analogue to the industrial annealing process in which metal alloys are slowly cooled to obtain an optimal molecular configuration. This controlled

cooling process is very important since a less stable configuration is obtained when the alloy is cooled too quickly. Computationally, annealing is implemented by allowing optimization steps that do not necessarily reduce the objective function. The idea is that a few bad steps can be accepted in order to get on the best path to the solution.

The SA algorithm is based on the Metropolis method (Metropolis et al., 1958) of obtaining the equilibrium configuration of a group of atoms at a given temperature. A connection between the Metropolis method and Monte Carlo simulation was first described in (Pincus, 1970). The simulated annealing optimization technique that is used today was proposed in (Kirkpatrick et al., 1983). It begins with a Metropolis Monte Carlo simulation at a high temperature. After a sufficient number of Monte Carlo steps have been taken, the temperature is reduced and the Metropolis Monte Carlo is continued. This process is repeated until a specified final temperature is reached. At high temperatures, a relatively large number of random steps will be accepted, and, as the temperature decreases, fewer steps are accepted.

The main advantage of SA over other optimization methods is that it is global. In theory, the algorithm can avoid becoming trapped in bad local minima regardless of its starting point. Furthermore, SA is easy to implement. Unfortunately, SA also has many well-documented disadvantages. It requires extensive computational work (van Laarhoven and Aarts, 1987; Moret and Shapiro, 1991; Aarts et al., 1997; Elmohamed et al., 1998), and it is sensitive to the choices of its many parameters (Elmohamed et al., 1998, Piccioni, 1987, Stiles, 1994; Aarts et al., 1997; Randelman and Grest 1986; van Laarhoven and Aarts, 1987). For example, there are at least a dozen different temperature cooling schedules from which to choose (Kirkpatrick et al., 1983; Geman and Geman, 1984; van Laarhoven and Aarts, 1987; Salamon et al., 2002). Finally, because the steps in SA are taken randomly, the algorithm does not employ any knowledge gained in previous iterations (Ali and Storey, 1997).

Since SA is the method of choice in the computational biology community and since it is also easy to implement, it was the first optimizer that we tried. In our implementation of SA, we use the geometric annealing schedule,

$$T_{new} = \alpha T_{old} \tag{11}$$

where $\alpha = 0.95$. We selected this schedule on the basis of numerical experiments, and our selection is supported by (Johnson et al., 1989; Johnson et al., 1991). The initial temperature and number of temperature cycles were chosen independently for each of the numerical tests presented in this paper. Each temperature cycle is terminated after either 1000 structures are generated or 100 structures are accepted. New structures are generated as follows: first, one of the helices is randomly selected. Then, starting from the arrangement of the last accepted structure, the position of the selected helix is randomized either by translating it or by rotating it around the *x* and *y* or the *z*-axis. The type and amount of randomization are randomly chosen within a user-defined maximum.

Our SA algorithm is implemented in C and uses the PDB Record I/O Libraries to read and write Brookhaven PDB formatted files (Couch et al., 1995). Our implementation of SA is serial. Although some parallelized versions exist (Kliewer and Tschöke, 1998; Stiles et al., 1989; Lee, 1995), none are compatible with MPI libraries such as MPICH-1.2.4 (Gropp et al., 1996; Gropp and Lusk, 1996). We chose to use a

basic implementation of SA but note that there are many sophisticated variations. For example, re-annealing has been shown to be effective by adapting to changes in parameter sensitivities when the search becomes trapped (Ingber, 1989; Ingber, 1993). Other adaptive and multi-start modifications of SA have also been shown to be successful (Piccioni, 1987; Stiles, 1994; Ali and Storey, 1997; Salamon et al., 2002).

Asynchronous Parallel Pattern Search

To contrast SA, we opted to apply an algorithm from a completely different class of optimization methods---pattern searches. Because this class of methods is generally overlooked in computational biology, we were particularly interested in examining its applicability and performance.

Pattern search methods are practical for solving problems such as (1) when the derivative of the objective function is unavailable and approximations are unreliable. They use a predetermined pattern of points to sample the given function domain. When certain requirements on the form of the points in the pattern are followed, it can be shown that if the objective function is smooth, global convergence to a stationary point is guaranteed (Dolan et al., 2000; Lewis and Torczon, 1996; Torczon, 1997). Bundler, our objective function, is not smooth, but further analysis reveals that pattern search may still find a minimum even for non-smooth functions (Kolda et al., 2003). We also note that pattern search methods are typically used for optimization problems with fewer than 100 variables (Kolda et al., 2003). Most transmembrane proteins have fewer than 13 helices, and we are interested in proteins that have 12 or fewer. Hence, the transmembrane protein structure determination problem that we consider contains at most 72 variables, and pattern search is a reasonable choice.

The majority of the computational cost of pattern search methods is in the function evaluations, so parallel pattern search (PPS) techniques have been derived to reduce the overall computational time. Specifically, PPS exploits the fact that, once the points in the search pattern have been defined, the function values at these points can be computed simultaneously (Dennis and Torczon, 1991; Torczon, 1992). The particular implementation of PPS that we use is asynchronous. Asynchronous parallel pattern search (APPS) (Kolda and Gray, 2004) retains the positive features of PPS, but it does not assume that the amount of time required for an objective function evaluation is constant or that the processors are homogeneous. It does not have any required synchronizations and thus requires less total time than PPS to return results (Hough et al., 2001). Furthermore, it has been shown that APPS is globally convergent under the standard assumptions for PPS (Kolda and Torczon, 2004). Finally, there is an existing open source version of APPS, called APPSPACK, which is easy to install and use (Kolda and Gray, 2004).

APPSPACK is available in MPI, PVM, and serial modes. To make use of the parallel machines at our disposal, we opted to use the MPI mode of APPSPACK version 3.0. This mode and version requires a minimum of three processors: one master agent to coordinate the search, one cache agent to save and look up points at which the function has already been evaluated, and at least one worker to perform function evaluations. For our problem, the use of cache is particularly advantageous as it should reduce the required number of new function evaluations and increase algorithm efficiency. The

default MPI version of APPSPACK requires that the function evaluations be run as separate executables and communicates with the worker tasks via file input and output. In our case, we customized APPSPACK to avoid the overhead of system calls and file I/O and improve overall efficiency.

We found that APPS required almost no tuning. We used the default values for all the parameters except the convergence tolerance, which was set to be 0.01. The default tolerance of 10^{-4} is small with respect to the variable sensitivity in our application, and thus we increased it in order to reduce the number of function evaluations required to obtain convergence. We also note that our implementation uses the coordinate direction search pattern.

Numerical Results

In this section, we present numerical results obtained using experimental distance constraints for rhodopsin. Rhodopsin is a transmembrane protein that is located in the retinal rods of the eye, and it plays a role in vision. It is a G-protein-coupled receptor made up of seven transmembrane helices and thus has 42 variables in its structure determination problem. The 3-D structure of the dark-adapted form of rhodopsin is known, having been determined using x-ray crystallography (Palczewski et al., 2000). Moreover, a set of experimental distance constraints for dark-adapted rhodopsin has been compiled in (Yeagle et al., 2001) making it an appropriate test case for our numerical experiments. In this paper, we use the structure of rhodopsin determined in (Palczewski et al., 2000), PDB entry 1F88, as the true structure. Because we are using a known structure, we can compute the difference between the true structure and any other structure using RMSD computed across the Cα atoms. Although we cannot use RMSD when trying to ascertain structures that have not yet been determined, we use it in our study to add clarity to the comparisons.

Motivation

We begin by presenting a simple example that motivated this study. Here, we use only one starting point, which was obtained by randomizing the true structure of

Figure 3: Comparison of the true and calculated locations of the helices of dark-adapted rhodopsin.

rhodopsin. The subsequent starting structure has an initial Bundler score of 11,342 and an RMSD of 15.0. We first tried optimizing this structure using our SA algorithm. After extensive tuning, the best structure we were able to produce resulted from using a starting temperature of 500 and 290 temperature cycles. This structure has a final Bundler score of 377 and an RMSD of 4.5. Next, we applied the APPS algorithm.

On our first try, we were able to produce a structure with a score of 122 and an RMSD 3.4. Figure 3 contains two pictures that illustrate the spatial positions of the helices relative to the known structure. In both pictures, the light gray cylinders represent the α-helices of dark-adapted rhodopsin. In picture (a), on the left, the dark cylinders depict the locations of the helices found using simulated annealing. Picture (b), on the right, shows the helices' locations determined by APPS as the dark cylinders. Note that APPS determines the orientation of all seven helices relatively well. In contrast, two of the helices determined by SA are a poor match.

We also examined the computational efficiency of each method. As previously discussed, SA often requires extensive computational work. This example was no exception. The SA algorithm required 81,800 function evaluations and 61 hours of run time on a single processor. In comparison, the APPS algorithm required only 32,458 function evaluations and 17 minutes of run time on 86 processors. It is difficult to compare the two algorithms directly since SA is serial and APPS is parallel. However, we can note that APPS required fewer total function evaluations than SA. Moreover, if SA were parallelized in the most efficient manner possible and run on 86 parallel processors, it would still take almost 45 minutes to obtain a solution.

This result led us to pursue a more thorough evaluation. We now present this study and its results.

Numerical Study

For our numerical study, 87 starting structures were selected from 7.0×10^{13} possible candidates using the procedure described in detail in (Faulon et al., 2003) and a set of 27 distance constraints, D_1 obtained from (Yeagle et al., 2001). This procedure resulted in structures that have no experimental distance penalty, i.e., $P_E = 0$ where $P_E = 0$ is as defined in (3), for each of the 87 structures with respect to D_1.

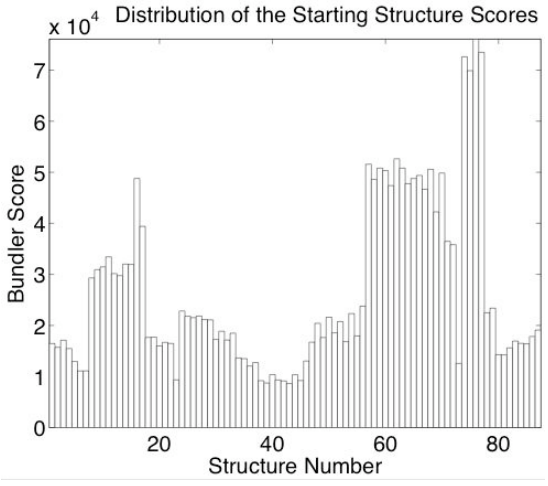

Figure 4: Distribution of the intial Bundler scores.

To fully test the capabilities of the optimization methods, we use a different set of distance constraints, D_2. The set D_2 contains upper and lower bounds for the same 27 pairs of atoms as D_1 but the range of these bounds is tighter as detailed in (Yeagle et al., 2001; Sale et al., 2004). The average Bundler score of the starting structures

is 26,555 with a maximum of 76,080 and a minimum of 8,608. The distribution of these scores is shown in Figure 4.

To optimize the 87 structures, we first applied our SA algorithm with a starting temperature of 300 and 125 temperature cycles. Next, we applied APPS to the same set of 87 starting structures. The results of this test set are displayed in Figure 5 and summarized in Figure 6. Note that APPS produces a much wider range of final scores than SA, and it appears that with APPS, some of starting structures get stuck in bad local minima. In contrast, the majority of the scores achieved by SA are below 200 and in fact, 40 of the 87 structures differ by less than six Bundler points.

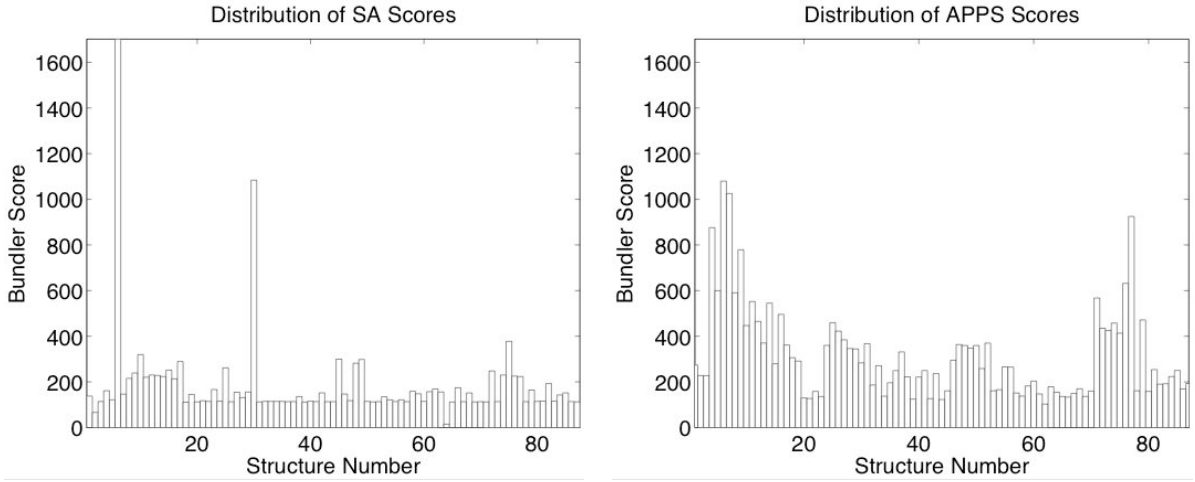

Figure 5: Final Bundler scoring distribution for SA (left) and APPS (right).

We can conclude that overall, this implementation of SA more *effectively* reduces the Bundler score. However, recall the aim of this particular project: to produce at least one structure with an empirically low score as efficiently as possible. In considering this goal, the fact that SA produced more structures with low scores does not necessarily give it an edge over APPS. The APPS algorithm does yield some structures with low scores.

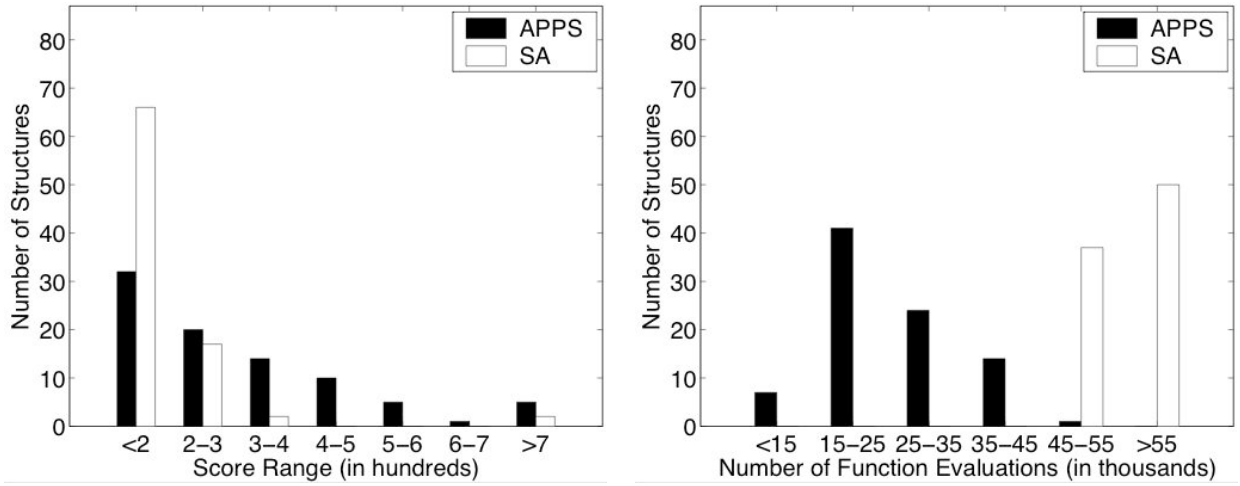

Figure 6: Summary of the final APPS and SA Bundler values (left) and required number of function evaluations (right).

Further study showed that for each SA structure with a score of less than 200, there is at least one APPS structure with a score of less than 200 such that its RMSD with respect to the SA structure is less than three. Given the errors in the distance constraints, we can therefore conclude that SA and APPS are finding the same minima. Furthermore, as Figure 6 shows, the computational cost of the success of SA is quite high. Each structure requires a minimum of 49,500 function evaluations to produce a solution. In comparison, the maximum number of function evaluations needed by APPS is 48,812, and 24 of the runs required fewer than 20,000 evaluations. Therefore, we conclude that APPS more *efficiently* reduces the Bundler scoring function. In fact, APPS is better suited than this implementation of SA for our transmembrane protein structure determination problem.

Next, we decided to more closely examine our implementation of SA to see if there was a simple way of improving the efficiency of SA without sacrificing too much of its effectiveness. One way of reducing the number of SA function evaluations is

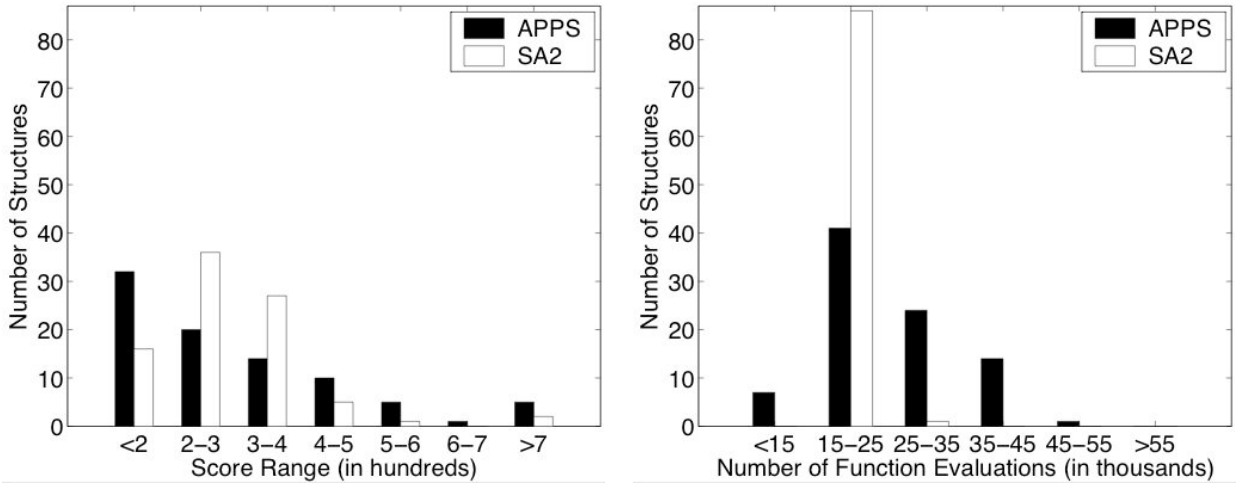

Figure 7: Summary of the final APPS and SA2 Bundler values (left) and required number of function evaluations (right).

reducing the number of temperature cycles. We use SA2 to denote the SA algorithm terminated after only 60 temperature cycles, or approximately one third of the number of function evaluations of the previous implementation. The results of this comparison are shown in Figure 7. The APPS and SA2 algorithms obtain solutions to problem (1) for the 87 different starting points in a similar number of function evaluations. An RMSD comparison of the structures with scores less than 200 showed that SA2 and APPS find the same minima. Hence, for our problem, both SA2 and APPS achieve our goals. The SA2 algorithm is less effective than SA at reducing the scoring function, but it still produces structures with a low Bundler score, and it is more efficient than SA.

The explicit distribution of the final SA2 Bundler scores is shown in Figure 8. The average final Bundler score is 306 with a maximum of 1883 and a minimum of 132. The results of this test allowed us to conclude that APPS is indeed a practical choice for our problem and that it is competitive with simulated annealing, the method of choice in the computational biology community.

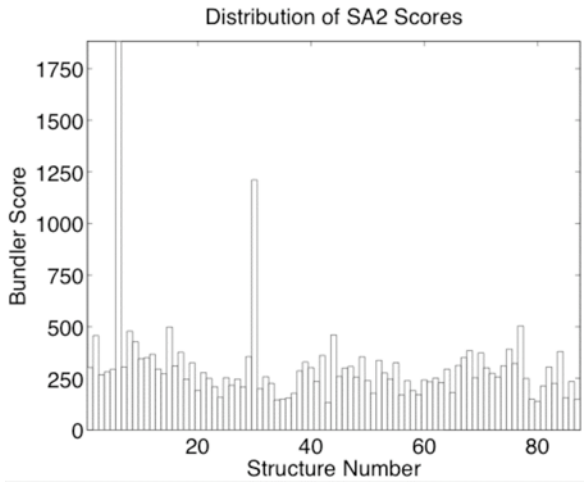

Figure 8: Distribution of the final SA2 Bundler scores.

For our final test, we tried to reduce the number of SA temperature cycles by using a lower starting temperature. Here, we use an initial temperature of 30 and do 75 temperature cycles. By beginning with a lower temperature, we will not accept as many randomized steps and thus we are, in effect, doing a more localized search. We use SA3 to denote this algorithm and summarize its results in Figure 9. Although SA3 still requires fewer function evaluations than SA, it does not successfully produce structures with low Bundler scores. None of the SA3 final scores are below 275, and, in fact, only two structures have scores below 300. The final Bundler score distribution for SA3 is given in Figure 10, and the average final Bundler score is 561 with a maximum of 2386 and a minimum of 274. In addition, SA3 is overall less efficient than both the SA2 and the APPS algorithms. Therefore, we can conclude that the simulated annealing algorithm using these particular parameters, a low initial temperature and a small number of temperature cycles, is not a viable alternative for solving our problem.

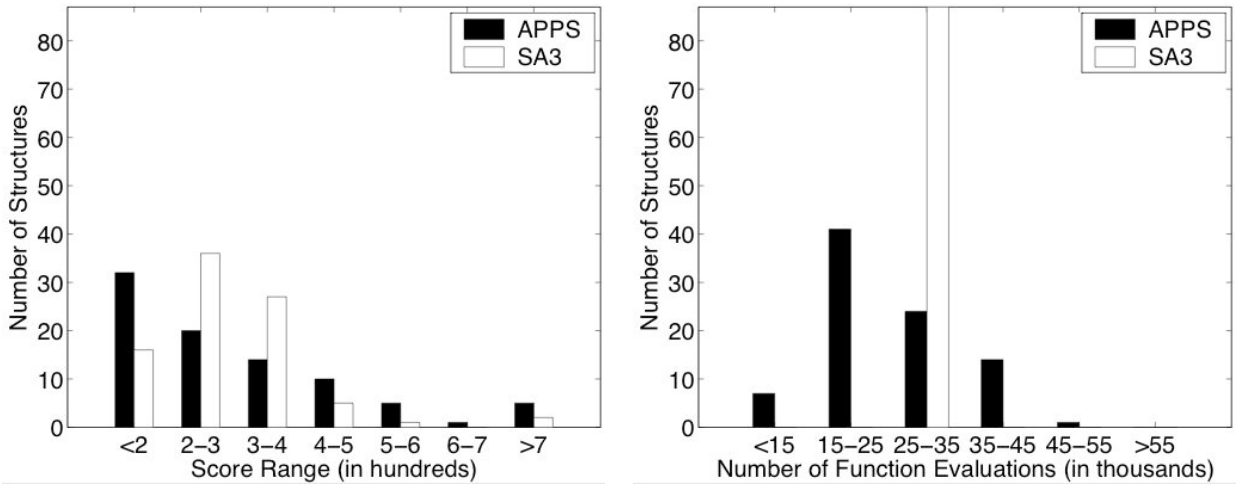

Figure 9: Sumary of the final APPS and SA3 Bundler values (left) and required number of function evaluations (right).

Conclusions

In this paper, we discuss the transmembrane protein structure identification problem. In particular, we focus on the second step of an innovative two-step method that combines laboratory and computational techniques (Faulon et al., 2003; Sale et al., 2004) to determine the spatial locations of the transmembrane helices. This second step refines a large set of possible helical bundles, generated in step one, by optimizing an empirical scoring function known as Bundler. The optimization of Bundler raises the question of finding an appropriate minimization algorithm.

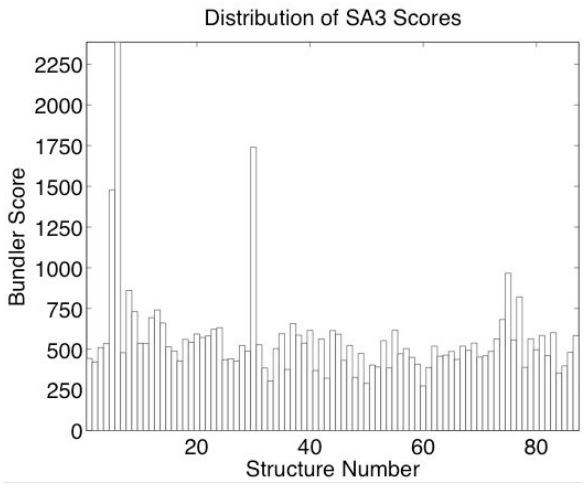

Figure 10: Distribution of the final SA3 scores.

Because Bundler is a discontinuous function that incorporates noisy experimental data, we opted to use a derivative-free method. We considered two very different algorithms: SA and APPS. The SA algorithm imitates an industrial cooling process and uses Metropolis Monte Carlo to generate new points. In contrast, APPS is a pattern search method that uses a predetermined pattern of points to sample a given function domain. In testing these methods, we had to consider the goal of our project: identifying at least one structure with a Bundler score low enough to warrant further study. Since Bundler was designed using only a small set of transmembrane proteins and inexact laboratory data, a high level of accuracy is neither expected nor desired from the optimization. However, efficiency is important as hundreds or even thousands of structures must be optimized. Therefore, we were looking for an optimization method that is both efficient and sufficiently accurate.

Given the numerous variations of the SA algorithm and the number of documented successes using SA, we are confident that we could eventually find a suitable version of the SA algorithm to solve our transmembrane protein structure determination problem. In fact, we have demonstrated that the SA2 implementation is sufficient for identifying the helical placement of rhodopsin. However, it is unclear whether or not this algorithm would be sufficient for a general transmembrane protein or if its parameters would be biased for certain proteins. We must also consider the fact that the Bundler scoring function will likely undergo a series of minor improvements, and we do not want any of these small changes to require that the optimization algorithm be re-tuned.

To our knowledge, APPS has not been previously applied to a problem of protein structure determination. We did not encounter any difficulties in applying it to our transmembrane protein problem. In fact, on our first try, we were able to produce desirable results. Moreover, in light of our efficiency and accuracy specifications, APPS was superior to both our SA and the SA3 algorithms and was comparable to our SA2 algorithm.

Despite their similar performance, APPS still has two implicit advantages over our SA2. First, APPS is easy to fine tune. Note that SA2 is exactly the same as our original SA algorithm with the exception of one small change to one parameter. However, the results obtained from these two algorithms are significantly different. Choosing the total number of temperature cycles is difficult. We must complete enough cycles to sufficiently reduce the Bundler score but not so many as to ignore our efficiency requirements. This is of concern to us because we will be using different sets of distance constraints and making minor changes to the Bundler scoring function. Second, because it is parallel, APPS will require less wall-clock time for problems with a large number of starting structures. Therefore, we have chosen to use APPS as the optimizer for this problem.

Acknowledgments

We gratefully acknowledge the Mathematical, Information, and Computational Sciences Program of the United States Department of Energy and the Laboratory Directed Research and Development program at Sandia National Laboratories for their support of this research. Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

References

Aarts, E. H., Korst, J., and van Laarhoven, P. J. (1997). Local Search in Combinatorial Optimization, chapter 4, Simulated Annealing. John Wiley and Sons, New York.

Adamian, L., Jackups, Jr., R., Binkowski, T. A., and Liang, J. (2003). Higher-order inter-helical spatial interactions in membrane proteins. J. Mol. Biol., 327:251-272.

Adiman, L. and Liang, J. (2001). Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. J. Mol. Biol., 311:891-907.

Ali, M. M. and Storey, C. (1997). Aspiration based simulated annealing algorithm. J. Glob. Opt., 11:181-191.

Banaszak, L. (2000). Foundations of Structural Biology. Academic Press, San Diego, CA.

Berliner, L. J., Eaton, S. S., and Eaton, G. R. (2001). Distance Measurements in Biological Systems by EPR, volume 19 of Biological Magnetic Resonance. Kluwer Academic Publishersn Plenum Publishing Corp., New York.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Research, 28:235-242. http://www.rcsb.org/pdb.

Bowie, J. U. (1997). Helix packing in membrane proteins. J. Mol. Biol., 272:780-789.

Bowie, J. U. (1999). Helix-bundle membrane protein fold templates. Protein Sci., 8:2711-2719.

Brandon, C. and Tooze, J. (1999). Introduction to Protein Structure. Garland Publishing

Inc., New York, 2nd edition.

Brunger, A. T. (1992). X-PLOR: A System for X-ray Crystallography and NMR, Version 3.1. Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT. http://www.ocms.ox.ac.uk/mirrored/xplor/manual/htmlman/htmlman.html.

Brunger, A. T., Adams, P. D., and Rice, L. M. (1997). New applications of simulated annealing in X-ray crystallography and solution NMR. Structure, 5:325-336.

Buchan, D., Shepherd, D., Lee, D., Pearl, F., Rison, S., Thorton, J., and Orengo, C. (2002). Gene3D: Structural assignment for whole genes and genomes using the CATH domain structure database. Genome Res., 12:503{514.

Campbell, B. J., Bellussi, G., Carluccio, L., Perego, G., Cheetham, A. K., Cox, D. E., and Millin, R. (1998). The synthesis of the new zeolite, ERS-7, and the determination of its structure by simulated annealing and synchrotron X-ray powder di®raction. J. Chem.Soc. Chem. Commun., 16:1725-1726.

Couch, G. S., Pettersen, E. F., Huang, C. C., and Ferrin, T. E. (1995). Annotating PDB les with scene information. J. Molec. Graphics, 13:153-158.

Creighton, T. E. (1992). Proteins: Structures and Molecular Properties. W.H. Freeman & Co., New York, 2nd edition.

Dennis, Jr., J. E. and Torczon, V. (1991). Direct search methods on parallel machines. SIAM J. Opt., 1:448-474.

Dobbs, H., Orlandini, E., Bonaccini, R., and Seno, F. (2002). Optimal potentials for predicting inter-helical packing in transmembrane proteins. Proteins, 49:342-349.

Dolan, E. D., Lewis, R. M., and Torczon, V. (2000). On the local convergence properties of parallel pattern search. Technical Report 2000-36, NASA Langley Research Center, Institute for Computer Applications in Science and Engineering, Hampton, VA.

Elmohamed, S., Fox, G., and Coddington, P. (1998). A comparison of annealing techniques for academic course scheduling. In 2nd International Conference on the Practice and Theory of Automated Timetabling, pages 146-166, Syracuse, NY.

Faulon, J.-L., Rintoul, M. D., and Young, M. M. (2002). Constrained walks and self-avoiding walks: implications for protein structure determination. J. Phys. A.: Math. Gen., 35:1-19.

Faulon, J.-L., Sale, K., and Young, M. M. (2003). Exploring the conformational space of membrane protein folds matching distance constraints. Protein Sci., 12:1750-1761.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Patt. Anal. Mach. Intel., 6:721-741.

Ghosh, A., Elber, R., and Scheraga, H. A. (2002). An atomically detailed study of the folding pathways of protein A with the stochastic di®erence equation. Proc. Natl.

Acad. Sci., 99:10394-10398.

Goodsell, D. S. and Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. Proteins: Str. Func. and Genet., 8:195{202.

Gropp, W., Lusk, E., Doss, N., and Skjellum, A. (1996). A high-performance, portable implementation of the MPI message passing interface standard. Parallel Comp., 22:789-828.

Gropp, W. D. and Lusk, E. (1996). User's guide for mpich, a portable implementation of MPI. Technical Report ANL-96/6, Mathematics and Computer Science Division, Argonne National Laboratory.

Herzyk, P. and Hubbard, R. E. (1998). Using experimental information to produce a model of the transmembrane domain of the ion channel phospholamban. Biophys. J., 74:1203-1214.

Hough, P. D., Kolda, T. G., and Torczon, V. (2001). Asynchrounous parallel pattern search for nonlinear optimization. SIAM J. Sci. Comput., 23:134{156.

Hough, P. D. and Meza, J. C. (2002). A class of trust-region methods for parallel optimiza- tion. SIAM J. Optim., 13:264-282.

Hulsen, T. and Lutje-Hulsik, D. (2001). GPCR Websire: Pictures about GPCRs (G protein-coupled receptors). University of Nijmegen, The Netherlands. http://www.cmbi.kun.nl/~dlutjehu/pictures.html.

Ingber, L. (1989). Very fast simulated re-annealing. Math. Comp. Mod., 12:967-973.

Ingber, L. (1993). Simulated annealing: Practice versus theory. Math. Comp. Mod., 18:29-57.

Jayasinghe, S., Hristova, K., and White, S. H. (2001a). Energetics, stability, and prediction of transmembrane helices. J. Mol. Biol., 312:927-934.

Jayasinghe, S., Hristova, K., and White, S. H. (2001b). MPtopo: A database of membrane protein topology. Protein Sci., 10:455-458.

Johnson, D. S., Aragon, C. R., McGeoch, L. A. M., and Schevon, C. (1989). Optimization by simulated annealing: An experimental evaluation; part I, graph partitioning. Oper. Res., 37:865-892.

Johnson, D. S., Aragon, C. R., McGeoch, L. A. M., and Schevon, C. (1991). Optimization by simulated annealing: an experimental evaluation; part II, graph coloring and number partitioning. Oper. Res., 39:378-406.

Kirkpatrick, S., Gerlatt, Jr., C. D., and Vecchi, M. P. (1983). Optimization by simulated annealing. Science, 220:671-680.

Kliewer, G. and TschÄoke, S. (1998). A general parallel simulated annealing library (parSA) and its applications in industry. PAREO 1998: First meeting of the PAREO working group on Parallel Processing in Operations Research, Versailles, France. http://www.uni-paderborn.de/fachbereich/AG/monien/SOFTWARE/PARSA/.

Kolda, T. G. and Gray, G. A. (2004). APPSPACK 4.0: asynchronous parallel pattern

search for derivative-free optimization. in preparation for ACM Trans. Math. Software. Software and documentation available at http://software.sandia.gov/appspack/.

Kolda, T. G., Lewis, R. M., and Torczon, V. (2003). Optimization by direct search: New perspectives on some classical and modern methods. SIAM Rewiew, 45:385-482.

Kolda, T. G. and Torczon, V. (2004). On the convergence of asynchronous parallel pattern search. SIAM J. Opt., 14:939-964.

Krishna, N. R. and Berliner, L. J. (1999). Structural Computation and Dynamics in Protein, volume 17 of Biological Magnetic Resonance. Kluwer Academic Publishersn Plenum Publishing Corp., New York.

Krough, A., Larsson, B., von Heijne, G., and Sonnhammer, E. (2001). Predicting transmem-brane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Bio., 305:567-580.

Leach, A. R. (2001). Molecular Modeling. Principles and Applications. Prentice Hall, New York, 2nd edition.

Lee, F. H. A. (1995). Parallel simulated annealing on a message-passing multi-computer. PhD thesis, Utah State University, Department of Electrical Engineering, Logan, UT.

Lewis, R. M. and Torczon, V. (1996). Rank ordering and positive basis in pattern search algorithms. Technical Report 96-71, NASA Langley Research Center, Institute for Computer Applications in Science and Engineering, Hampton, VA.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1958). Equations of state calculations by fast computing machines. J. Chem. Phys., 21:1087-1092.

Moret, B. M. E. and Shapiro, H. D. (1991). Algorithms from P to NP, volume I. Benjamin/Cummings Publishing Company, Redwood City, CA.

Nikiforovich, G. V., Galaktionov, S., Balodis, J., and Marshall, G. R. (2001). Novel approach to computer modeling of seven-helical transmembrane proteins: current progress in the test case of bacteriorhodopsin. Acta Biochimica Polinica, 48:53-64.

Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Trong, I. L., Teller, D. C., Okada, T., Stenkamp, R. E., Yamamoto, M., and Miyano, M. (2000). Crystal structure of rhodopsin: a G protein-coupled receptor. Science, 289:739-745.

Perkins, T. D. J. and Dean, P. M. (1993). An exploration of a novel strategy for superposing several °exible molecules. J. Comp.Aided Mol. Design, 7:155-172.

Piccioni, M. (1987). A combined multistart-annealing algorithm for continuous global optimization. Technical Report 87-45, The University of Maryland, Systems and Research Center, College Park MD.

Pincus, M. (1970). Monte Carlo method for the approximate solution of certain types of

constrained optimization problems. Oper. Res, 18:1225-1228.

Powell, M. J. D. (1998). Direct search algorithms for optimization calculations. Acta Numer., 7:287-336.

Randelman, R. E. and Grest, G. S. (1986). N-city traveling salesman problem|optimization by simulated annealings. J. of Stat. Phys., 45:885-890.

Richards, F. M. (1974). The interpretation of protein structures: total volume, group volume, distributions and packing density. J. Mol. Biol., 82:1-14.

Rose, G. D. (1978). Prediction of chain turns in globular proteins on a hydrophobic basis. Nature, 272:586-590.

Salamon, P., Sibani, P., and Frost, R. (2002). Facts, Conjectures, and Improvements for Simulated Annealing. Monographs on Mathematical Modeling and Computation 7. SIAM, Philadelphia, PA.

Sale, K. L., Gray, G. A., Faulon, J.-L., Schoeniger, J., and Young, M. M. (2004). Optimal bundling of transmembrane helices of integral membrane proteins using sparse distance constraints. Prot. Sci., 13:2613-2627.

Stiles, G. S. (1994). The e®ect of numerical precision upon simulated annealing. Phys. Lett. A., 185:253-261.

Stiles, G. S., Bosworth, K. W., Morgan, T. W., Lee, F. H., and Pennington, R. J. (1989). Parallel optimization of distributed database networks. In Proc. First Int'l Conf. Applications of Transputers, Amsterdam, The Netherlands. IOS Press.

Torczon, V. (1992). PDS: Direct search methods for unconstrained optimization on either sequential or parallel machines. Technical Report TR92-09, Rice University, Department of Computational & Applied Math, Houston, TX.

Torczon, V. (1997). On the convergence of pattern search algorithms. SIAM J. Opt., 7:1-25.

Vaidehi, N., Floriano, W. B., Trabanino, R., Hall, S. E., Freddolino, P., Choi, E. J., Zamanakos, G., and Goddard, III, W. A. (2002). Prediction of structure and function of G protein-couple receptor. Proc. Natl. Acad. Sci., 99:12622-12627.

van Laarhoven, P. M. J. and Aarts, E. H. L. (1987). Simulated Annealing: Theory and Applications. Dordrecht Reidel Publishing Company, Dordrecht, The Netherlands. Republished in 1989 by Kluwer Academic, Boston, MA.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. J. Mol. Graphics, 8:52-56.

Wallin, E. and von Heijne, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaen, and eukaryotic organisms. Prot. Sci., 283:489-506.

White, S. (1998). Membrane Proteins of Known 3D Structure. Stephen White Laboratory at UC Irvine. http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html. Wilson, S. and Bergsma, D. (2000). Orphan G-protein coupled receptors: Novel drug targets for the pharmaceutical industry. Drug Des. Discov., 17:105-114.

Xiang, Z. and Honig, B. (2001). Extending the accuracy limits of prediction for side chain conformations. J. Mol. Biol., 311:421-430.

Yeagle, P. L., Choi, G., and Albert, A. D. (2001). Studies on the structure of the G-protein coupled receptor rhodopsin including the putative G-protein binding site in unactivated and activated forms. Biochemistry, 40:11932-11937.

Chapter 4: Using a Detailed Atomistic Potential to Place Side-Chains onto Poorly-Folded Backbones

Alex Slepoy, Thomas W. Hunt and David Shirley

Abstract

We compare the performance of a simplified energy function with a full atomistic one with implicit solvation in a side-chain placement algorithm. We find that, while their performance is comparable for well-folded structures, the full atomistic energy function use often results in significantly better sidechain conformations for backbones deformed from the well-folded structure.

Introduction

Sidechain placement plays a critical part in computational protein structure determination, homology modeling, and flexible ligand docking. Often cited difficulties arise from a need to explore a large conformational space and inaccuracies in the energy function approximations. The extent of the conformational space that placement methods need to search has been reduced by discovery of strong localization propensities in side-chain conformations. This discovery led to a development of *rotamer* libraries [1], where the allowed conformations for a given sidechain are limited to a small discrete set. Such an approximation permits the search to be conducted in a reduced integer representation, leading to an exact ground state solution in computationally feasible time.

It is possible to guarantee an exact ground state given a particular energy function and a set of discrete rotamers through use of Branch-And-Bound [5] type of algorithms (BAB), that avoid performing an exhaustive enumeration by use of bounding functions. Preconditioners [4], like Dead End Elimination (DEE), are often able to reduce the combinatorial space so that enumeration becomes possible. Widely-used methods [1] use hybrid combinations of the above methods and other preconditioners to improve execution speed. It appears that the hurdle of the conformational search has been overcome.

The debate continues over the required accuracy of the energy function [7]. Less-detailed energy functions lead to a faster performance, which is still an important consideration, especially in the protein design problems. However, the simplicity of the reduced energy function leads to a limited range of applicability and a potential for considerable error. Recent research suggests that including solvent effects and electrostatics may be critical to an accurate sidechain placement [6,8].

A convincing case has been made that the quality of the placement prediction also highly depends on the relative proximity to the native fold of the backbone conformation [9]. We wish to investigate a proposition that the accuracy of the energy function becomes more important when the backbone is not well-folded. Our investigation involves generating a set of backbone structures that are progressively deformed from the

well-folded conformation, placing sidechains on these using a simplified partially information-based energy function and a full atomistic model that includes an implicit solvent, and evaluating the quality of placement. We chose the latest release of SCWRL [1] for the method that uses a simplified energy function. This is a widely-used method that finds the exact ground state of its given energy function. For an atomistic function, we chose CHARMM19 [3] with an implicit solvation mode EEF1 [2], parameterized for this force field.

To achieve sidechain placement using the CHARMM energy function with the EEF1 solvent model, we have developed a software application called CENTIPEDE that includes a set of combinatorial methods employed in SCWRL. We closely follow the SCWRL algorithm in the ground state search to avoid introducing simulation artifacts. However, certain alterations to the algorithm were unavoidable, as described below [in Section 3].

Comparison Scheme

Since the backbones we populate are deformed from the native structure, for which the sidechain conformations are available in the database, we need an alternate method to evaluate the quality of the placement. We choose the CHARMM software to evaluate the energy of the resulting structures from both SCWRL and CENTIPEDE using the CHARMM19 energy function with the EEF1 solvent. Though this may seem a recursive strategy on our part, it is useful to mention here that, for well-folded structures SCWRL and CENTIPEDE performed similarly. The energies of the resulting structures were not found to favor either energy function, except in a few cases occasioned by the SCWRL's use of smaller steric constraint parameters. The comparison is also mixed for the perturbed structures [in Section 4].

CENTIPEDE

CENTIPEDE is a deterministic sidechain confirmation prediction code that finds the lowest energy combination of rotamers from a user specified rotamer library. CENTIPEDE employs a two-step process to find the lowest energy combination of rotamers. First, rotamers are eliminated with a Dead End Elimination algorithm (DEE). After the DEE step, a Branch And Bound (BAB) algorithm picks out the lowest energy combination from the reduced set of rotamers. We did not follow the graph component decomposition strategy, which makes SCWRL runtime orders of magnitude shorter. CHARMM-based residue interaction graph includes electrostatic and solvent effects, which prevent such a decomposition. The graph contains a single giant highly connected component that has no low order articulation points. Thus, CENTIPEDE run times are orders of magnitude slower then SCWRL even for the relatively small protein fragments.

Energy Function

CENTIPEDE's energy function is based on the CHARMM19 force field and the EEF1 implicit solvation model [2]. We chose EEF1 as the solvation model for its pair-wise decomposition properties, which allowed us to calculate the solvent contribution to the sidechain interaction energies.

Dead End Elimination

CENTIPEDE's dead end elimination algorithm uses several variants to reduce the number of possible rotamer combinations. The simplest DEE variant that CENTIPEDE uses is Goldstein DEE, which is the cheapest DEE scheme in terms of computational cost, but usually the least effective. After the Goldstein DEE algorithm has exhausted all the rotamers that it can eliminate, CENTIPEDE employs Split DEE, which is computationally costlier than Goldstein DEE, but potentially more effective at reducing the set of rotamers.

Branch and Bound

After the set of rotamer combinations has been reduced by DEE, the number of rotamer combinations is typically still quite large, so CENTIPEDE uses a branch and bound algorithm to determine the lowest energy combination of rotamers without necessarily calculating the energy of every rotamer combination. Branch and bound can skip over certain rotamer combinations by determining if an incomplete combination of rotamers cannot possibly be in the full combination of rotamers that minimizes the energy of the structure. Here are two ways that CENTIPEDE can avoid calculating the energy of any rotamer combination that contains a given fixed subset of rotamers. A lower bound on the energy of any rotamer combination containing the rotamer subset can be computed, and if this lower bound is greater than the energy of a previously computed rotamer combination, then it is unnecessary to compute the energy of any combination of rotamers that contains the subset of rotamers. During Split DEE, pairs and triples of rotamers, known as dead end pairs and triples are often discovered. Any rotamer combination that contains a dead end pair or triple cannot possibly be the lowest energy combination of rotamers, so the energy of such combinations do not have to be computed.

Test Set of Protein Fragments

Several sets of perturbed structures were generated to compare CENTIPEDE and SCWRL rotamer placement quality.

1BDC-Immunoglobulin-Binding Protein

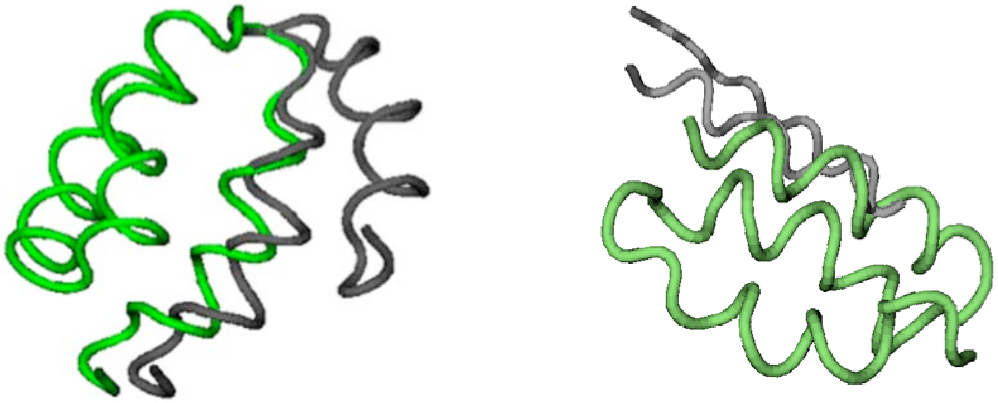

Figure 1: Two sets of structures were generated from the protein 1BDC. First, three helices were extracted from the structure. One set of test structures was produced by progressively rotating a single helix out of its native orientation with respect to the other helices. Another set of structures was produced progressively altering the ϕ and ψ angles of one of three helices so that it eventually flattened out.

1A7F-Insulin Mutant

Figure 2: A test set of 1A7F based structures was generated by progressively altering the ϕ and ψ angles of the original structure's helix so that the helix flattened into a planar confirmation.

1B03-Viral Protein

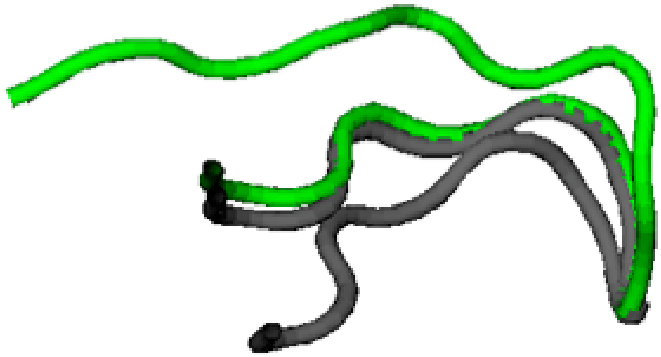

Figure 3: **A test set was derived from the initial 1B03 structure by progressively altering the ϕ angle in a beta hairpin so that the beta hairpin was perturbed from its native planar confirmation.**

1UBQ-Chromosomal Protein

Figure 4: **Two test sets were derived from the initial 1UBQ structure. The first 46 residues of ubiquitin were extracted, and the test structures were derived from this extracted structure. One test set was created by successively perturbing the tertiary structure of 1UBQ. This was done by altering a backbone ϕ angle successively. The second set was produced by altering a ϕ angle in a beta hairpin in the same fashion as the 1B03 derived test set.**

Results and Discussion

We have performed two basic types of deformations on our test set. The first type of deformation moves secondary structure domains to new relative positions with respect to other domains without distorting their secondary structure. The second type of the deformation explicitly distorts secondary structure by unfolding helices or prying apart beta-sheets. We find that, in a large number of cases, CENTIPEDE is able to identify a lower energy sidechain solution then SCWRL for the backbones that are distorted from the native fold. This is particularly true for the second type of the deformation. Secondary structure distortion leads to the unexplored territory for the knowledge-based portion of the SCWRL energy function, whose training set consists entirely of the native structures with intact secondary structure.

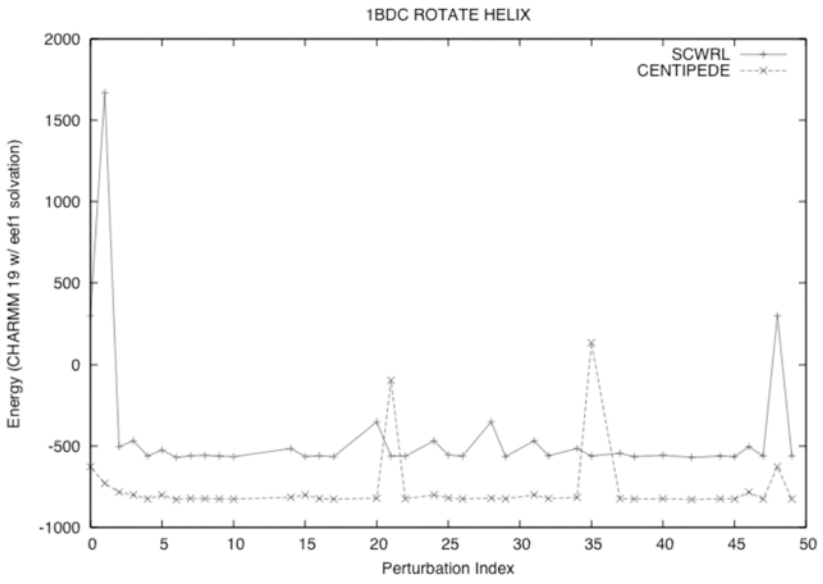

Figure 5: Perturbation index reflects a single axis tick per 3 degrees of angle rotation. CENTIPEDE energies are consistently lower except for two perturbed structures. SCWRL has a large conflict for unperturbed structure, rooted in the use of small steric clash parameter.

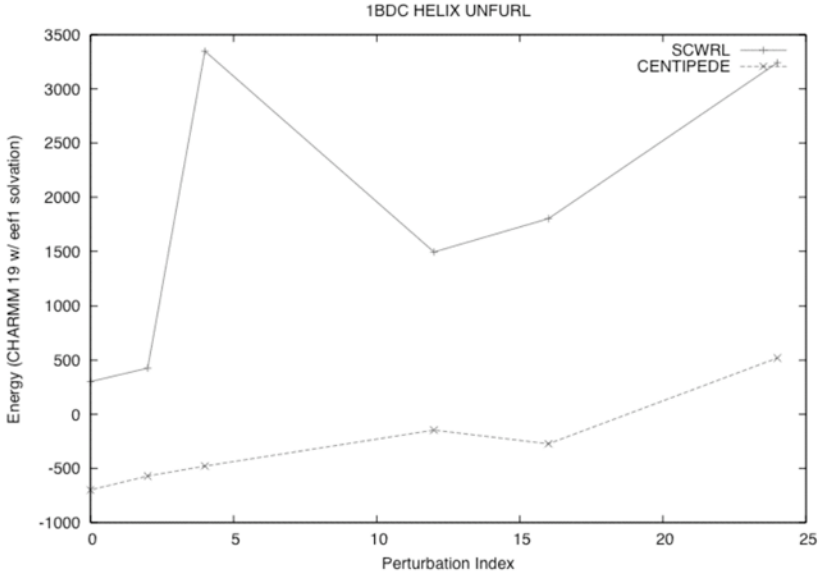

Figure 6: Here perturbation coordinate is proportional to the initial angle value, progressively taking all angles toward zero, which corresponds to a planar structure. Again, SCWRL is unable to find low energy conformations, producing many extreme conflicts. CENTIPEDE states appear significantly more controlled, always finding a much lower energy value.

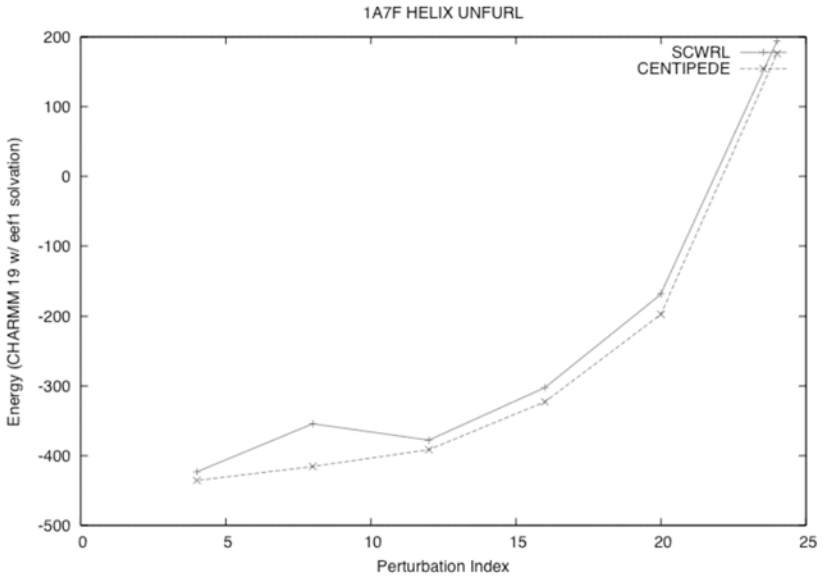

Figure 7: Here the two methods are in closer agreement. Centipede however always finds a lower energy structure.

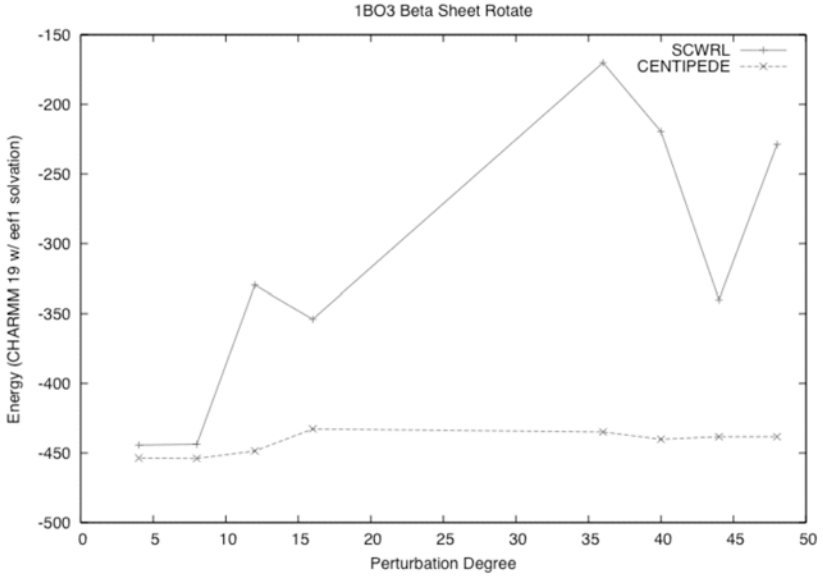

Figure 8: For a distorted beta sheet, SCWRL manages to find reasonable solutions at the low perturbation range, but performs less efficiently at the larger perturbation scale.

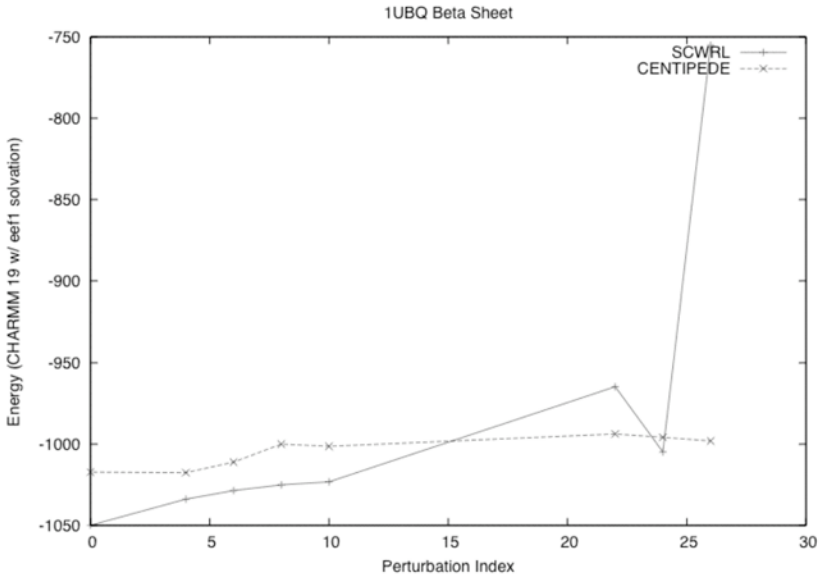

Figure 9: For Ubiquitin, we get anomalous results. SCWRL outperforms CENTIPEDE in the small perturbation regime, but has difficulty for the larger ones.

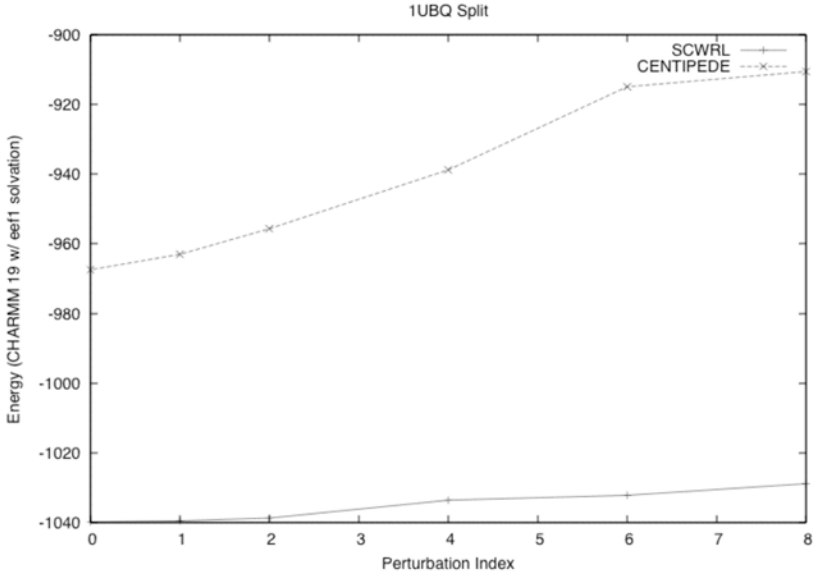

Figure 10: For this single dihedral rotation, SCWRL outperforms CENTIPEDE all the way. However, this perturbation represents no secondary structure distortion.

This preliminary study indicates that a more detailed energy function may be appropriate under certain circumstances. It appears to perform better whenever a significant secondary structure distortion occurs. This makes it useful in the homology modeling, when large portions of the backbone are unknown. It is also valuable in the structure determination methods where the native conformation of the backbone is not available. We intend to continue studying this effect by looking closely into the individual contributions of the different fractions of the full energy function to the accuracy of the sidechain placement.

References

[1] A. A. Canutescu, A. A. Shelenkov, R. L. Dunbrack, Jr. A graph theory algorithm for protein side-chain prediction. *Protein Science* 12, 2001-2014 (2003)

[2] Lazaridis T, Karplus M. Effective energy function for proteins in solution. *Proteins*, 1999 May 1;35(2):133-52

[3] Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem*. 4:187-217

[4] N.A. Pierce, J. A. Spriet, J. Desmet, S.L. Mayo Conformational Splitting: A More Powerful Criterion for Dead-End Elimination *J Comput Chem*, 21(11):999-1009, 2000

[5] James R. Bitner , Edward M. Reingold, Backtrack Programming Techniques, Communications of the ACM, Volume 18 , Issue 11:651-656

[6] Eyal E, Najmanovich R, McConkey BJ, Edelman M, Sobolev V., Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. ,*J Comput Chem*. 2004 Apr 15;25(5):712-24.

[7] Petrella, R.J., Lazaridis, T., and Karplus, M. 1998. Protein sidechain conformer prediction: A test of the energy function. Fold. Des. 3: 353-377.

[8] Ronald W. Peterson, P. Leslie Dutton and A. Joshua Wand, Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library, *Protein Science* (2004), 13:735-751

[9] E.S. Huang, P. Koehl, M. Levitt, R.V. Pappu, J.W. Ponder, *Proteins: Structure, Function, and Genetics.*, 33:204-217 (1998)

Distribution

1	MS9004	Rick Stulen, 8100
1	MS9951	Len Nepolitano, 8140
1	MS9951	Malin Young, 8141
1	MS9951	Joe Schoeniger, 8141
1	MS9951	Jean-Loup Faulon, 9212
1	MS1110	Alex Slepoy, 9235
1	MS0310	William Michael Brown, 9212
1	MS0807	David Shirley, 9328
1	MS1110	Thomas W. Hunt, 9235
1	MS9159	Genetha A. Gray, 8962
5	MS9951	Kenneth Sale, 8141
3	MS9018	Central Technical Files, 8945-1
1	MS0899	Technical Library, 9616
1	MS0123	D. Chavez, LDRD Office, 1011