SANDIA REPORT

SAND2003-8795
Unlimited Release
Printed December 2003

Mapping Membrane Protein Interactions in Cell Signaling Systems

Marites Ayson, Joohee Hong, Yooli Light, Masood Hadi, Pam Lane, Nichole Wood, Rick Jacobsen, Joe Schoeniger, Malin Young

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-AC04-94-AL85000.

Sandia National Laboratories

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.doe.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 ordering:	http://www.ntis.gov/ordering.htm

SAND2003-8795
Unlimited Release
Printed December 2003

Mapping Membrane Proteins in Cell Signaling Systems

Marites Ayson, Joohee Hong, Yooli Light, Masood Hadi, Pam Lane, Nichole Wood, Rick Jacobsen, Joe S. Schoeniger, and Malin M. Young

Biosystems Research Department
Sandia National Laboratories
P.O. Box 969
Livermore, CA 94551-9951

ABSTRACT

We proposed to apply a chemical cross-linking, mass spectrometry and modeling method called MS3D to the structure determination of the rhodopsin-transducin membrane protein complex (RTC). Herein we describe experimental progress made to adapt the MS3D approach for characterizing membrane protein systems, and computational progress in experimental design, data analysis and protein structure modeling. Over the past three years, we have developed tailored experimental methods for all steps in the MS3D method for rhodopsin, including protein purification, a functional assay, cross-linking, proteolysis and mass spectrometry. In support of the experimental effort, we have put a data analysis pipeline in place that automatically selects the mono-isotopic peaks in a mass spectrometric spectrum, assigns them and stores the results in a database. Theoretical calculations using 24 experimentally-derived distance constraints have resulted in a backbone-level model of the activated form of rhodopsin, which is a critical first step towards building a model of the RTC. Cross-linked rhodopsin-transducin complexes have been isolated via gel electrophoresis and further mass spectrometric characterization of the cross-links is underway.

Intentionally Left Blank

CONTENTS

Introduction 8

Methods Development For Bacteriorhodopsin 10

Experimental progress (BR) 10

Separation of the BR monomeric fraction 10

Solubilization, enzymatic digestion and reverse phase chromatography 11

ESI MS of BR peptides 11

Computational progress 12

Experimental design 12

MS spectrum assignment 12

MS/MS spectrum assignment 12

On- and off-lattice protein structure modeling 12

BR structure modeling 13

Rhodopsin Cross-linking 14

Experimental Methods 14

Cysteine-lysine and lysine-lysine cross-linking 14

Lysine labeling studies 14

Cysteine labeling studies 15

Data analysis 15

Rhodopsin Cross-linking : Results and Discussion 15

Rhodopsin-Transducin Cross-linking 17

Protein purification 17

Functional assay of the RTC 17

Protein Expression 17

Cross-linking the RTC 18

GTP-eluted Cross-linked RTC 19

Cross-linking at Lower Concentrations 19

Cross-linking at Lower Concentrations with GTP Elution 20

Membrane Protein Structural Modeling 21

Methods 22

Representation of the helical bundle 22

Selection of membrane protein representative set 22

Determination of force constants 22

Conformational search under a set of distance constraints 23

Monte Carlo Simulated Annealing 23

Helix bundle definition 23

Monte Carlo sampling 24

Cooling schedule 26

Structural analysis and data processing 26
Results 26
Statistical analysis of membrane protein structures 27
Penalty function 28
Distance Constraints Penalty (P_{dist}) 29
Structure based penalties 30
Packing Distance Penalty (P_{pdist}) 30
Packing Density Penalty (P_{pdens}) 31
Packing Angle Penalty (P_{angle}) 31
van der Waals Repulsion (P_{vdw}) 32
Contact Penalty ($P_{contact}$) 32
Side-Chain Interaction Preference Penalty ($P_{contact}$) 32
Total Score 32
Scoring Function Validation 33
Two-step approach to modeling transmembrane helical bundles using sparse distance constraints to build the rhodopsin helical bundle 34
Discussion of BUNDLER Results 38
Light-adapted rhodopsin model 39

Mass Spectrometry Data Reduction and Analysis 40
Data reduction 40
MS spectrum assignment 40
MS/MS spectrum assignment 41

References 42

Distribution 47

FIGURES

Figure 1. Representative BR peptide ESI FT-MS spectrum. 10
Figure 2. Sequence coverage of bacteriorhodopsin. Peptides observed in the CAN:H2O fraction are blue, and peptides in the CMW fraction are red. Helical BR regions are shown in green. 11
Figure 3. Structural model (calculation 2, 6.34 Å RMSD, using 15 simulated experimental constraints). 13
Figure 4. CNBr peptide sequence coverage of rhodopsin as determined by LCMS. 14
Figure 5. Schematic of the 9 observed crosslinks mapped on to the rhodopsin tertiary structure. Unambiguous crosslinks are shown with solid lines; ambiguous crosslinks with dashed lines. Lysine-lysine crosslinks are in cyan and lysine-cysteine crosslinks are orange. 16
Figure 6. Purified rhodopsin and transducin. 17
Figure 7. Transferred Gel 18
Figure 8. Western Blot 18

Figure 9. Transferred gel stained with Coomassie blue of cross-linked and uncross-linked GTP-eluted RTC 19
Figure 10. Western blot of cross-linked and uncross-linked GTP-eluted RTC. 19
Figure 11. Transferred gel of cross-linked and uncross-linked RTC at lower crosslinker concentrations. 20
Figure 12. Western blot of cross-linked and uncross-linked RTC at lower crosslinker concentrations. 20
Figure 13. Transferred gel of cross-linked and uncross-linked RTC at lower crosslinker concentrations. 20
Figure 14. Western blot of cross-linked and uncross-linked RTC at lower crosslinker concentrations 20
Figure 15. 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. 26
Figure 16. Correlation of helix-end to helix-end distance and number of amino acids in the loop. 30
Figure 17. 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. 35
Figure 18. Comparison of predicted helical bundle (black) to the native bundle (gray). The C□ – 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. 36
Figure 19. Proposed model of rhodopsin light activation. Left: Ribbon diagram of the rhodopsin crystal structure helical bundle (1f88). The arrows indicate the predicted helical movements. Right: Optimized light-adapted rhodopsin model generated by a distance geometry calculation using 24 literature-derived experimental constraints. 39

TABLES

Table. 1. Summary of distance geometry calculations using 4 sets of simulated cross-linking constraints. 13
Table. 2. Cross-linking results for the dark-adapted and light-activated conformational states of rhodopsin. 15
Table. 3. Cysteine-lysine cross-linking results for dark-adapted and light-activated rhodopsin. 15
Table. 4. Structures used to derive statistical characterization of membrane protein bundles 27
Table. 5. Statistics describing membrane protein bundles. 28
Table. 6. Experimental distances used for the Rhodopsin structure[1] 37

Mapping Membrane Proteins in Cell Signaling Systems

Introduction

We proposed to develop and apply a novel high-throughput experimental and computational technique to map the self-organization of membrane protein complexes. The cell membrane is the interface between the external world and the machinery of life. The thirty percent of proteins that are membrane proteins (MPs) regulate the flow of information into and out of the cell through signaling cascades that stimulate cellular responses. MPs thus have critically important roles acting as receptors for drugs and bio-regulators, transporting substances into the cell, adhering pathogens to cells, and mediating the immune response. They carry out these roles by associating with other MPs and cytoplasmic proteins. MP interactions are thus central to processes of infection and cell signaling that are central to the action of infectious agents and toxins (including known biological and chemical warfare agents).

Signaling cascades depend on the molecular details of specific membrane protein-protein interactions. Little is known about these critically important complexes because standard structure determination techniques (X-ray crystallography and NMR) can rarely be applied to characterizing protein-protein interactions within the membrane. The development of a technique to analyze MP interactions at the molecular level thus has major implications for understanding basic life processes, pharmaceutical development, and the development of CBW agent detectors and countermeasures. In a broader context, this project will address a major challenge in the post-genomic era, which is to understand how the tens of thousands of gene products (proteins) encoded by the genome interact with each other to perform essential cellular functions. Success in this project will thus position Sandia to participate in the anticipated future growth in funding related to the Human Genome Project.

To develop and validate our approach, we chose as a model cell receptor complex the well-studied visual proteins rhodopsin (RO) and transducin (TR). RO is a member of the largest vertebrate gene superfamily, the G protein–coupled receptors (GPCRs). A canonical signaling system includes a G protein (guanine nucleotide–binding regulatory protein) and 7 transmembrane helix-containing GPCR. The G-protein-coupled photoreceptor rhodopsin undergoes a conformational change upon light-activation that initiates the vertebrate visual signaling cascade. Light-activated rhodopsin, a GPCR, catalyzes guanine nucleotide exchange by transducin, a G protein, which ultimately leads to a change in membrane cation conductance and a neural signal. Although the 11-cis to all-trans isomerization of the retinal cofactor is well-understood, less is known about the protein structural changes that are induced by the absorption of light.

The rhodopsin-transducin complex (RTC) is a particularly suitable test case for structure-function studies, as large quantities of protein and several high-resolution crystal structures of transducin and rhodopsin are available [1-4]. **However, the structure of the RTC is still unknown, and its solution would represent a major scientific advance.**

Recently, a technique called MS3D was validated for low-resolution structure determination of soluble proteins [5]. Application of MS3D to soluble proteins showed that, with the aid of sensitive mass spectrometry instrumentation and new computational

tools, the distance information derived from chemical cross-linking could be increased by at least an order of magnitude.

Chemical cross-linking has an extensive history.[6-8] Until the advent of scanning mutagenesis and Electro-Spray Ionization (ESI) and Matrix Assisted Laser Desorption Ionization (MALDI) mass spectrometry, the general experience with the method might be characterized as "nasty and brutish" but definitely not "short". The procedure was extremely laborious and fraught with numerous complications.[6] The questions asked largely concerned simple proximity issues. What has changed is the prospect of using the great advances in mass spectrometry of proteins and peptides to assign significant numbers (10-100) of intra-molecular cross-links in proteins in a few automated experiments, to obtain distance constraints between the cross-linked residues. Some years ago, it was shown that distance constraint information from various experiments could be combined to produce three-dimensional structures whose resolutions were determined by the amount and type of information.[9] The work of Gordon Crippen and Timothy Havel was instrumental in developing a mathematical technique called distance geometry (DG) for converting distance constraints into molecular structures.[10] It remains an important method for deriving structures from NMR studies in solution and for homology modeling.[11-15] Alternative approaches, such as constrained molecular dynamics can also be utilized.[16, 17]

The major structural genomics consortia have had many successes in solving protein structures, but there is significant debate about the rate at which protein structures can be solved using existing approaches employing NMR and X-Ray crystallography.[18, 19] Given the major initiatives already underway in high throughput protein structure determination[20], we propose to develop the MS3D technology as an initial automated experiment to assist crystallographic and NMR studies through the identification of potential novel folds and the discovery of distant homologs of known proteins, and to provide distance constraints that can be used to derive preliminary, low-resolution structural models. We can also add value to genomic level homology modeling, by providing experimental distance constraints to test existing and newly developed homology models.[21] This area is of great interest to the proteomics community and to basic scientists in biochemistry, biophysics, chemical biology, and to drug discovery scientists. It is clear that a new method, complementary to existing methods employing NMR and X-Ray methods would be a very significant development and fill an important niche.

With this goal in mind, we adapted the MS3D technique to probe the structures of membrane protein complexes, specifically the RTC. Our plan was to experimentally derive distance constraints from this complex using molecular cross-linking, proteolysis, and mass spectrometry (MS). We would then use the distance constraints, in conjunction with genomic, proteomic, and structural information derived for rhodopsin and transducin, to solve the structure of the complex.

Methods Development For Bacteriorhodopsin

Significant progress was made during the first year. The accomplishments are hereafter divided into experimental progress, which has primarily been focused on the adaptation of the MS3D technique to membrane proteins, and computational progress, in the areas of experimental design, data analysis and protein structure modeling. Both the experimental and computational efforts were initially directed towards validation of our methods on a well-understood proof-of-concept system, bacteriorhodopsin (BR). BR is an appropriate test system for our purpose as it is structurally homologous to rhodopsin, has had its crystal structure solved to high resolution in multiple intermediate states, should react with multiple cross-linking reagents, and is commercially available (Sigma). Once our experimental and computational methods were validated for BR, we proceeded with our planned cross-linking experiments on rhodopsin and the rhodopsin-transducin complex (RTC).

Experimental progress (BR)

We successfully completed a series of control experiments with uncross-linked BR. These experiments demonstrated that all steps in the MS3D protocol (except the initial cross-linking step) could be performed with BR – namely the separation, solubilization and digestion of the monomeric fraction, and the acquisition of mass spectrometry data on the proteolytic digest.

Separation of the BR monomeric fraction

We employed preparative gel electrophoresis to separate monomeric BR from higher-order cross-linked species. We used a gradient (4-20% tris-glycine) preparative 2D gel for electrophoresis. The monomer band was subjected to electroelution following Coomassie staining using a Bio-Rad electroeluter. The eluted protein was then precipitated using 2:5:2 chloroform:methanol:water (CMW) to remove residual SDS.

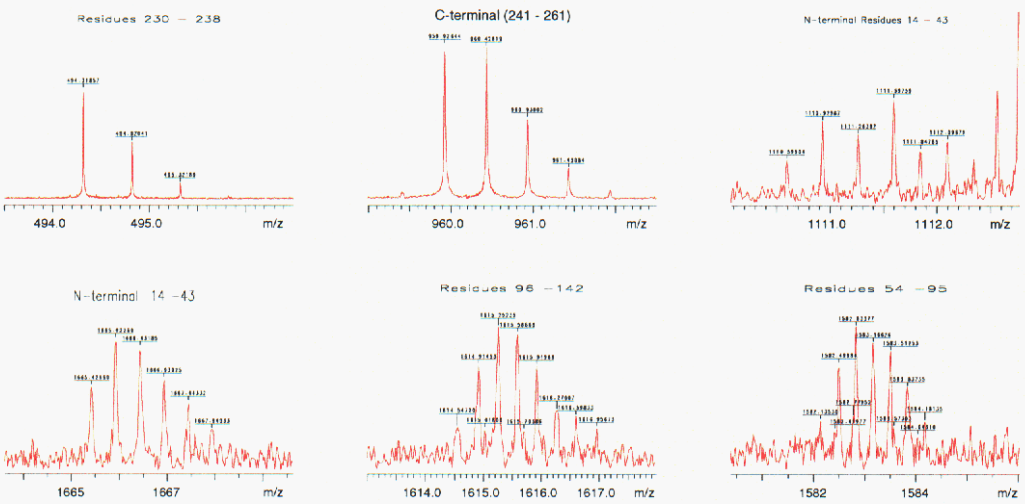


Figure 1. Representative BR peptide ESI FT-MS spectrum.

Solubilization, enzymatic digestion and reverse phase chromatography

We developed a protocol for the solubilization and digestion of BR using two ionic detergents. BR solubilized in either 0.1% SDS or 0.1% CTAB in 50mM ammonium bicarbonate was digested using sequencing grade modified trypsin in a ratio of 1:25. Multiple bands on SDS-PAGE confirmed the success of digestion under these conditions.

ESI MS of BR peptides

In order to determine the identity and number of tryptic products, as well as conditions for detergent-free solubilization and ESI, we performed the following experiment: Purple membrane (Sigma) was delipidated using CMW extraction and the protein pellet was suspended by brief sonication in 40% ACN:60% 50mM ammonium bicarbonate pH7.9 at a concentration of 1mg/ml. Sequencing-grade modified trypsin (Promega) was added at 50:1 BR:trypsin and the mixture was incubated overnight at 37 C with vigorous shaking. The resulting suspension was centrifuged at 10000g for 5 min, and the supernatant (containing more polar peptides) was decanted and acidified to 2% acetic acid. The pellet was dissolved in 4:4:1 chloroform:methanol:water, 2% acetic acid and centrifuged to remove any undissolved material. The samples were then directly electrosprayed into the FTICR with an infusion pump, and representative spectra are shown in Figure 1 (512 scans, ~40 microliters of sample total).

The top row shows peptides found in the ACN:H2O phase, and the bottom in the CMW phase. The high mass resolution of the FTICR (~1 ppm) allowed peptides to be unambiguously assigned. The sequence is shown with representative ACN:H2O peptides in blue, and CMW peptides in red. Helical regions of BR are indicated with green boxes (Figure 2).

Although products from all cleavage sites were not seen, complete sequence coverage was reproducibly achieved. The chloroform phase was able to solubilize hydrophobic peptides effectively, which suggests that ternary solvents using chloroform would allow LC/MS or direct infusion ESI for analysis of these peptides. LC/MS, unless needed to compensate for ionization suppression, may be supplanted by direct ESI/FTICR analysis of these digest mixtures.

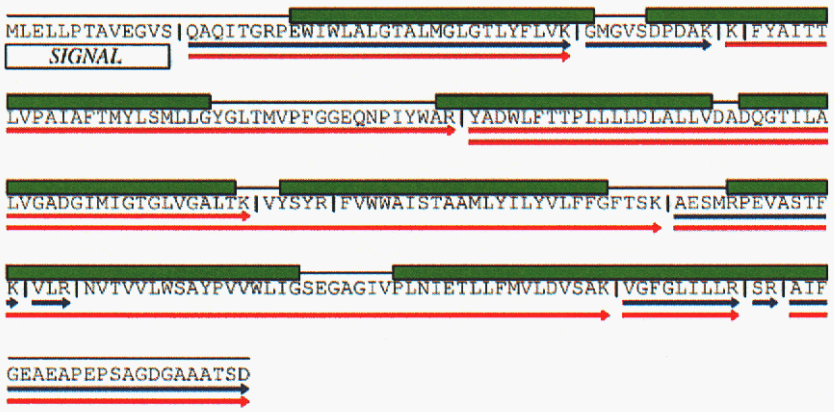


Figure 2. Sequence coverage of bacteriorhodopsin. Peptides observed in the CAN:H2O fraction are blue, and peptides in the CMW fraction are red. Helical BR regions are shown in green.

Computational progress

The computational progress during the first year was in the areas of experimental design, data analysis and protein structure modeling.

Experimental design

Assignment of cross-linked peptides in MS spectra can be complicated when there are multiple assignments for a m/z peak within a given mass error. Resolution of this issue usually involves a MS/MS experiment to disambiguate the assignments, which can be time consuming. What would be desirable is an experimental design strategy to reduce the frequency of ambiguous assignments. Careful selection of a protease could lower the number of overlapping assignments, and thus reduce the ambiguity problem. A prototype tool, MSDesign, was developed to assist experimentalists with protease selection. It calculates the theoretical libraries of peptides and cross-linked peptides for a user-defined set of proteases, and produces statistics on the total number of overlapping species, the number of overlapping cross-linked species, and the peptide coverage of the resolvable cross-linked species over a given ppm error range.

MS spectrum assignment

A new version of the spectrum assignment program, **ASAP**, was developed supporting flexible amino acid modification and protease specificity definitions, fully combinatoric theoretical library generation, control spectrum subtraction, and N^{14} and N^{15} peptide assignments.

MS/MS spectrum assignment

As mentioned above, MS/MS experiments are performed to identify a molecular ion. Software currently exists to assign MS/MS spectra of peptides (e.g. MS-Product in ProteinProspector, http://prospector.ucsf.edu). However, no tools are available to assign the more complicated MS/MS spectra of singly-labeled and cross-linked peptides. We developed a software tool, **MS2Assign**, to assign the MS/MS spectra of singly-labeled and cross-linked peptides that is available now for the analysis of in-house MS/MS spectra. A paper we published in collaboration with Professor Bradford Gibson in 2003 (Buck Institute for Aging) discussed the successful application of MS2Assign to the interpretation of MS/MS data [22].

On- and off-lattice protein structure modeling

We performed theoretical calculations both on- and off-lattice to enumerate the number of constraints required to uniquely determine a protein structure and to map the size of the solution space as a function of the number of provided constraints. Results from these calculations show that: 1) lattice protein structures can be retrieved in linear time when O(n) distances are provided, 2) real space protein structures can be retrieved in linear time using O(n) precise ($\varepsilon < 0.1$ Å) distances, 3) when precision decreases the number of distances required increases but the solution space size decreases, 4) short-range topological distances are more discriminating than short-range geometrical distances, and 5) the solution space is of manageable size when about half or more of the short-range topological and/or geometrical distances are provided. These results were published in the *Journal of Physics A* in 2002 [23].

BR structure modeling

The calculations described above are for systems in which the constraints are known to a precision of $\leq$ 0.1 Å. To investigate the feasibility of model building with looser constraints, we performed four distance geometry (DG) calculations with simulated cross-linking data to build structural models of BR. In calculations 1 and 2, rigid helices from the BR crystal structure 1c3w were assembled using theoretical constraints (derived from [24]) and simulated constraints from BR cross-linking with BS^3 (amine-amine, Cα-Cα distance=5-24 Å) and BS^3+EDC (amine-carboxyl, Cα-Cα distance=5-12.5 Å). In the third calculation, ideal alpha helices were positioned using the theoretical and simulated BS^3+EDC constraint set. The fourth calculation was a control, in which only the theoretical constraints were used to generate a model. The results are summarized in Table 1. Shown in Figure 3 is the most accurate structural model generated to date.

Figure 3. Structural model (calculation 2, 6.34 Å RMSD, using 15 simulated experimental constraints).

Table. 1. Summary of distance geometry calculations using 4 sets of simulated cross-linking constraints.

Cross-linker(s)	Helices	Exptl Constr	Tors < Constr	Packing Constr	Loop Constr	Mean RMSD
BS^3	Crystal	9	6	45	37	7.26 Å
BS^3+EDC	Crystal	15	6	45	37	6.55 Å
BS^3+EDC	Ideal	15	6	45	37	6.84 Å
None	Crystal	0	6	45	37	8.10 Å

Rhodopsin Cross-linking

After validation of our methods on bacteriorhodopsin, we undertook a program to obtain structural information on the mechanism of rhodopsin activation using a combination of chemical cross-linking, protein fragmentation and high-resolution mass spectrometry on dark-adapted and light-activated rhodopsin. We hypothesized that differential cross-linking patterns will yield information about the structural differences between the ground (dark-adapted) and the meta II (light-activated) states of rhodopsin. To date, we have identified 9 cross-links in the ground and meta II states of rhodopsin. The cross-linking data correlates well with the results of cysteine and lysine accessibility studies we performed using the chemical labels maleimide (cys) and NHS-acetate (lys). In this chapter, we will discuss the structural implications of the accessibility and cross-linking results.

Experimental Methods

Cysteine-lysine and lysine-lysine cross-linking

The rhodopsin-rich rod outer segment membrane (ROS) was purified from bovine retinas, as described by Palczewski et al [1]. Native rhodopsin was cross-linked in ROS with a library of 10 commercially available crosslinkers (K-K: DST, DSG, DSS, BS3, EGS; K-C: GMBS, EMCS, SMCC, LC-SMCC, SIA from Pierce). In the lysine-lysine cross-linking reactions, cysteines were reduced with TCEP (Pierce) and alkylated with 4-vinylpyridine before subsequent purification steps. The monomeric protein was separated from intermolecularly cross-linked species and contaminating proteins by preparative SDS-PAGE using a column gel (5 cm length, 1 cm diameter, 11 % acrylamide resolving gel; 2 cm length, 4 % acrylamide stacking gel). The eluted monomer was precipitated (and SDS removed) by chloroform:methanol:water (CMW) extraction and then chemically digested by overnight incubation with 4 M CNBr in 70% TFA.

Lysine labeling studies

Rhodopsin was incubated with either 50:1 or 1000:1 NHS-acetate:Rhodopsin for 30 min. at 37C, pH 7.5. The reaction was quenched with Tris and prepared for LC-MS using preparative SDS-PAGE, CMW extraction and CNBr digestion as described above.

Figure 4. CNBr peptide sequence coverage of rhodopsin as determined by LCMS.

Cysteine labeling studies

Rhodopsin was incubated with 1000:1 maleimide:rhodopsin for 30 min. at 37C, pH 7. Unreacted maleimide was removed by washing membrane (ROS) or buffer exchange (solubilized rhodopsin). The sample was then reduced with TCEP, alkylated using 4-VP, and prepared for LC-MS analysis in the same manner as above.

Data analysis

The resulting mixtures of peptide fragments was analyzed by LC/MS with a 7 Tesla FTICR mass spectrometer (Bruker). Peptide masses were identified from MS spectra using the Sandia MS2Pro software suite integrated into the Xmass package (Bruker).

Table. 2. Cross-linking results for the dark-adapted and light-activated conformational states of rhodopsin.

Lysine-Lysine			Seen in Meta-II?
Linker	Length (max)	Crosslinks	
DST	6.3 Å	K66/67xK325/339	Yes
		K311xK66/67	Yes
		K311xK325/339	? (trace)
		K325xK339	? (trace)
		K66xK67	Yes
DSG	7.4 Å	K66/67xK325/339	Yes
		K311xK66/67	Yes
		K311xK325/339	Yes
		K325xK339	Yes
		K66xK67	Yes
DSS	11.3 Å	K66/67xK325/339	Yes
		K311xK66/67	Yes
		K311xK325/339	Yes
		K325xK339	Yes
		K66xK67	Yes
EGS	14.5 Å	K66/67xK325/339	n.d.

Table. 3. Cysteine-lysine cross-linking results for dark-adapted and light-activated rhodopsin.

Cysteine-Lysine			Seen in Meta-II?
Linker	Length (max)	Dark crosslinks	
SIA	4 Å	C316xK66/67	Yes
		C316xK325/339	Yes
sGMBS	9.7 Å	C316xK66/67	Yes
		C316xK325/339	Yes
		C140xK141	Yes
EMCS	11 Å	C316xK66/67	Yes
		C316xK325/339	Yes
		C316xK311	Yes
		C140xK141	Yes
LC-SMCC	15 Å	C316xK66/67	Yes
		C316xK325/339	Yes
		C316xK311	Yes
		C140xK141	Yes

Rhodopsin Cross-linking : Results and Discussion

The 10 crosslinkers tested all reacted well with rhodopsin in ROS down to a protein:crosslinker ratio of 1:1.5, based on the formation of multimeric protein complexes as seen following analysis by SDS-PAGE. Monomeric protein was purified with up to 90% recovery using preparative gel electrophoresis. Complete MS coverage was routinely seen in control (uncross-linked) rhodopsin CNBr digestion experiments (Figure 6).

Once a protocol for identifying CNBr digest products was worked out for control protein, the same was used for cross-linked and labeled rhodopsin. Results of our cross-linking experiments on both dark-adapted and light-activated rhodopins are summarized in Tables 2 and 3.

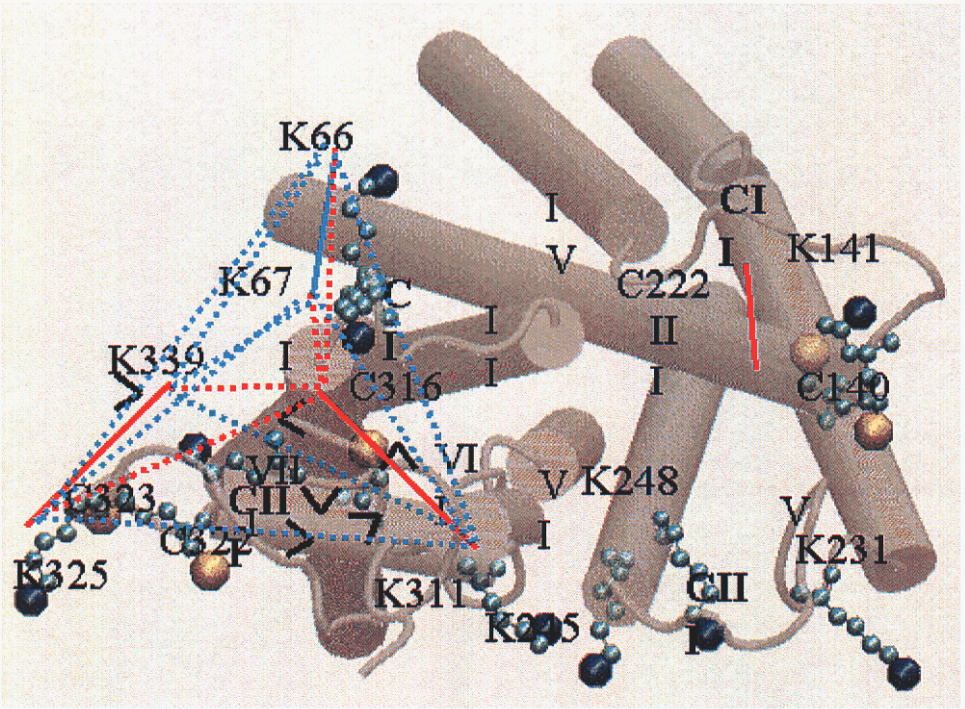


Figure 5. Schematic of the 9 observed crosslinks mapped on to the rhodopsin tertiary structure. Unambiguous crosslinks are shown with solid lines; ambiguous crosslinks with dashed lines. Lysine-lysine crosslinks are in cyan and lysine-cysteine crosslinks are orange.

We have thus far identified 9 peptide pairs that are cross-linked by one or more of the crosslinkers in our library. When mapped to the rhodopsin crystal structure, these cross-links are localized to helices I, VII and VIII (Figure 5). We observe few, if any, differences in the cross-linking pattern between the dark-adapted and the light-activated structures of rhodopsin indicating that there is little movement in this region of the structure. This result is in agreement with low-resolution structure information in the literature [25].

The remainder of the structure contains theoretically cross-linkable residue pairs, but they are not observed in our experiments. Possible reasons why are: the peptide fragments containing them are too large/non-ionizable to be observed, the residues are inaccessible and/or unreactive, or there may be structural considerations preventing the formation of an internal cross-link (steric hindrance, dimerization, etc.). We investigated the reactivity of the lysines and cysteines in rhodopsin by performing labeling experiments with NHS acetate and maleimide, respectively. We observed profound differences in lysine and cysteine reactivities that correlate well with our cross-linking results and with theoretical predictions of solvent accessibility (data not shown).

Rhodopsin-Transducin Cross-linking

Herein we describe experimental progress made to develop expression systems for rhodopsin, alpha-, beta- and gamma-transducins and tailored experimental methods to probe the formation and product types generated by cross-linking the rod outer segment membrane (ROS) with cys-cys (Bis-Maleimidohexane,BMH), cys-lys (N-(e-Maleimidocaproyloxy) succinimide ester, EMCS), and lys-lys (Disuccinimidyl suberate, DSS) crosslinkers. We carried out a set of experiments that investigated the effect of crosslinker concentration on the amount and types of RTC cross-linking products formed. Gel electrophoresis and Western blots of the intact cross-linked ROS and GTP-eluted fractions showed that alpha-transducin is cross-linked to rhodopsin by all three cross-linking reagents.

Protein purification

We successfully purified all components of the RTC from bovine retinas. The rod outer segment membrane (ROS) was isolated from the retina by homogenization followed by separation on a sucrose density gradient. Rhodopsin represents >95% of ROS protein after a series of washes performed in the dark remove soluble and membrane-associated proteins. The analytical protein gel in Figure 6 (left gel) shows a fraction of solubilized ROS following membrane washes. Transducin, which consists of three subunits (α, β, γ), was purified by light-induced binding to rhodopsin in the ROS followed by elution with GTP, which causes the rhodopsin-transducin complex to dissociate. The individual subunits were separated into α and βγ fractions by column chromatography, as shown in Figure 6.

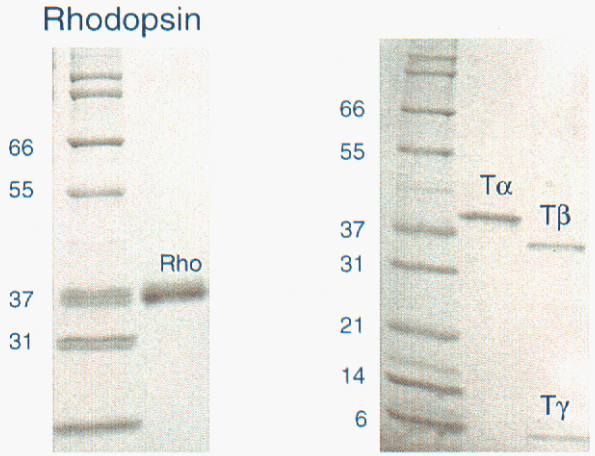


Figure 6. Purified rhodopsin and transducin.

Functional assay of the RTC

We prepared an assay to determine whether we have functional RTC. The assay consists of 1) mixing purified transducin with light-exposed ROS, 2) washing off the unbound transducin, 3) eluting the bound transducin with GTP and 4) running a gel on the eluted transducin. We have determined that purified transducin binds rhodopsin in a light-dependent manner and can be GTP-eluted, indicating that the purified proteins are functional.

Protein Expression

Obtaining quantities of transducin sufficient for cross-linking studies from bovine retinal tissue proved to be extremely laborious and time-consuming. We therefore shifted focus to constructing expression systems for both rhodopsin and transducin. As a first step, we obtained the cDNAs for the mouse and human genes encoding rhodopsin, alpha, beta and gamma transducin.

We made entry vectors from all of the cDNAs and confirmed them by sequencing. The entry vectors were converted into expression vectors containing N-terminally histidine (HIS) tagged rhodopsin and glutathione S-transferase (GST) tagged transducin for expression in *E.coli* and Baculovirus.

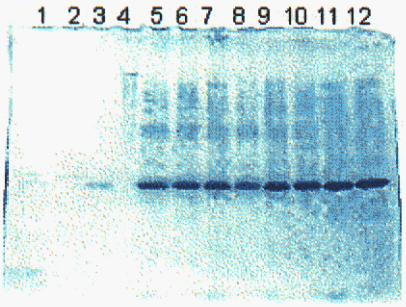
The mouse and human set were characterized in *in vitro* extracts and we found cross-reacting bands at the correct molecular weights based on HIS- and GST-specific western blots.

We further characterized the human GST and HIS tagged set in *E. coli* and found that approximately 30% of the expressed GST tagged proteins are in the soluble fraction. Most of the HIS tagged proteins are in the insoluble fraction. These results are consistent with expectations, as transducin is a soluble protein and rhodopsin is primarily insoluble.

We also created a Baculovirus system for rhodopsin expression in SF9 and sf21 cells. The virus was amplified and titered, the MOI determined and a small-scale expression run performed. The yield of rhodopsin was about >1 ug/L. We are in the process of optimizing this yield.

We obtained the N-myristoyl transferase genes NMT1 and NMT2 for post-translational processing (addition of a myristoyl group) of the N-terminal glycine of alpha-transducin. Alpha transducin was co-expressed alpha transducin in *E. coli* with NMT1 and NMT2 to obtain functional product. We can detect the expression of unprocessed and processed alpha transducin by Western blots and we are currently working out optimum expression conditions prior to scaleup.

We are also in the process of co-expressing alpha, beta and gamma transducin in SF9 and SF21 cells. We have the primary virus and are in the process of going through virus purification.



Well #	Content
1	Transducin complex(control)
2	Alpha Transducin (control)
3	Beta and Gamma Transducin (control)
4	See Blue Standard
5	5xBMH in ROS/Transducin
6	50x BMH in ROS/Transducin
7	5X DSS in ROS/Transducin
8	50x DSS in ROS/Transducin
9	5xEMCS in ROS/Transducin
10	50x EMCS in ROS/Transducin
11	control, no crosslinker but GTP eluted
12	control, no crosslinker and none GTP elution

Figure 7. Transferred Gel

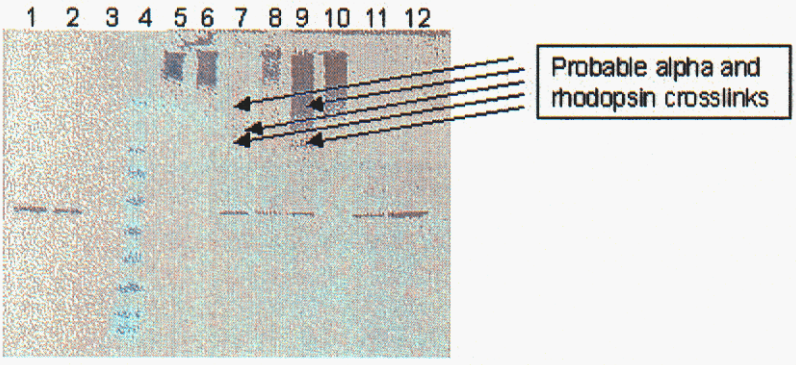


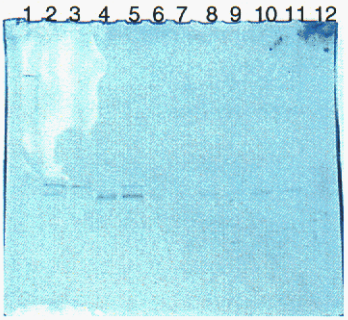
Figure 8. Western Blot

Cross-linking the RTC

The purified ROS/transducin complex was treated with cys-cys (Bis-Maleimidohexane,BMH), cys-lys (N-(e-Maleimidocaproyloxy) succinimide ester, EMCS), and lys-lys (Disuccinimidyl suberate, DSS) crosslinkers. The results of the cross-linking experiments are shown in Figure 7 and Figure 8.

The top figufre shows a transferred gel stained with Coomassie blue. As the concentration of crosslinker:protein increased (lanes 5-10), higher molecular weight bands appeared on the gel. These bands most likely correspond to the formation of cross-linked protein complexes.

To probe these higher molecular weight species, we did a Western blot with anti-alpha transducin antibody. Whereas no higher molecular weight species are observed in the transducin control lanes 1 and 2, we see them appear as transducin-containing bands in lanes 5-10 for each of the 3 cross-linking reactions. Lanes 11 and 12 are ROS controls that contain residual transducin even after GTP elution in the absence of crosslinker.



Well #	Content
1	See Blue standard
2	Transducin complex (control)
3	Alpha Transducin (control)
4	Beta and Gamma Transducin (control)
5	5x BMH in ROS/Transducin, GTP eluted
6	50x BMH in ROS/Transducin, GTP eluted
7	5X DSS in ROS/Transducin, GTP eluted
8	50x DSS in ROS/Transducin, GTP eluted
9	5x EMCS in ROS/Transducin, GTP eluted
10	50x EMCS in ROS/Transducin, GTP eluted
11	control, no crosslinker but GTP eluted
12	control, no crosslinker and none GTP elution

Figure 9. Transferred gel stained with Coomassie blue of cross-linked and uncross-linked GTP-eluted RTC

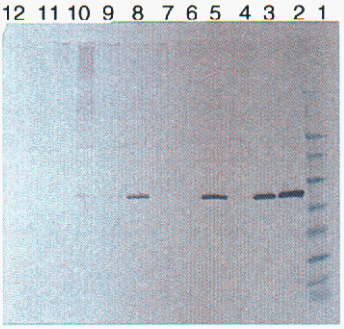


Figure 10. Western blot of cross-linked and uncross-linked GTP-eluted RTC.

GTP-eluted Cross-linked RTC

Once we confirmed that we could observe cross-linked RTC complex formation, we were interested in probing what proteins were principally involved in the cross-linking reactions. We hoped to accomplish this by GTP eluting the transducin that is not cross-linked with rhodopsin from the ROS after the cross-linking reaction. The presence of higher molecular weight bands on a gel or a Western blot for a GTP-eluted cross-linking reaction would indicate that alpha transducin crosslinks the beta and/or gamma subunits. Absence of such bands would indicate that alpha transducin crosslinks rhodopsin preferentially and is therefore not elutable from the ROS by the addition of GTP.

Figures 9 and 10 show the results of cross-linking with BMH, DSS and EMCS followed by a GTP wash. The GTP wash fractions were gel assayed and the gels were blotted with anti-alpha transducin antibody. No higher molecular mass species were observed in lanes 5-10 of the Western blot, indicating that alpha transducin crosslinks principally with rhodopsin and not with the beta and gamma subunits.

Cross-linking at Lower Concentrations

The appearance of higher molecular weight "smearing" on Figures 7 and 8 indicate that a complex mixture of cross-linked complexes are formed during the cross-linking reaction. To reduce production of higher order cross-linking products, we repeated the cross-linking experiments but at lower crosslinker concentrations. The results of this experiment are shown in Figures 11 and 12.

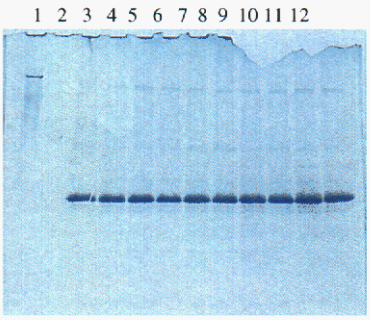
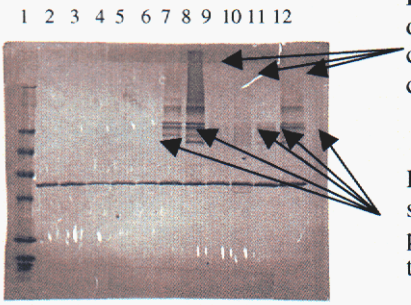


Figure 11. Transferred gel of cross-linked and uncross-linked RTC at lower crosslinker concentrations.



Less smearing effect due to lower concentration of crosslinker used.

Higher molecular mass species indicates probable rhodopsin and transducin crosslinks.

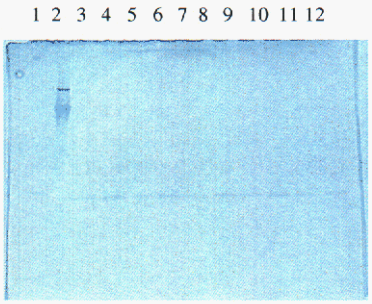
Figure 12. Western blot of cross-linked and uncross-linked RTC at lower crosslinker concentrations.

Well #	Content
1	See Blue Plus
2	Alpha Transducin (control)
3	ROS/transducin, no TCEP, no GTP elution
4	ROS/transducin, no TCEP, with GTP elution
5	ROS/transducin, with TCEP, no GTP elution
6	ROS/transducin, with TCEP, with GTP elutio
7	0.2 X BMH, ROS pellet
8	0.5 X BMH, ROS pellet
9	2X DSS, ROS pellet
10	5X DSS, ROS pellet
11	0.5X EMCS, ROS pellet
12	1.5X EMCS, ROS pellet

Figure 12 shows a distinct ladder of higher molecular weight species that is in sharp contrast to the "smearing" seen in Figure 8. This indicates that reducing the crosslinker concentration lowered the amount of non-specific protein-protein cross-linking.

Cross-linking at Lower Concentrations with GTP Elution

For completeness, we investigated the effect of lowered crosslinker concentrations on the products eluted with GTP from the ROS. The results are shown in Figures 13 and 14.



Well #	Content
1	Alpha Transducin Control
2	See Blue Plus
3	ROS/transducin, no TCEP, no GTP elution
4	ROS/transducin, no TCEP, with GTP elution
5	ROS/transducin, with TCEP, no GTP elution
6	ROS/transducin, with TCEP, with GTP elutio
7	0.2 X BMH, GTP wash supernatant
8	0.5 X BMH, GTP wash supernatant
9	2X DSS, GTP wash supernatant
10	5X DSS, GTP wash supernatant
11	0.5X EMCS, GTP wash supernatant
12	1.5X EMCS, GTP wash supernatant

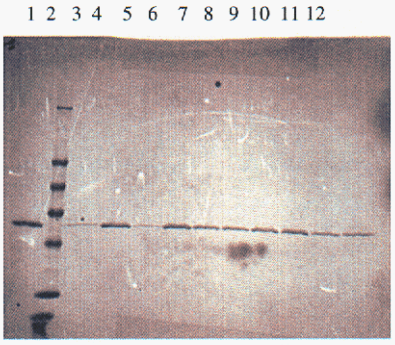


Figure 13. Transferred gel of cross-linked and uncross-linked RTC at lower crosslinker concentrations.

Figure 14. Western blot of cross-linked and uncross-linked RTC at lower crosslinker concentrations

As in Figures 9 and 10, we observe no higher molecular weight species in the GTP elution fractions at lower crosslinker concentrations, indicating that the alpha transducin subunit does not form cross-linked complexes with either the beta or gamma subunit. The higher order species we observe in the samples derived directly from ROS must

therefore contain alpha-transducin/rhodopsin at the minimum (as shown by Western blotting) and potentially rhodopsin/rhodopsin cross-linked species as well.

The presence of clearly distinguishable bands on the ROS Western blot makes it possible to isolate and characterize each of the cross-linking products represented on the gel. We are currently in the process of cutting these bands out of the gel and characterizing them using proteolysis and mass spectrometry.

Membrane Protein Structural Modeling

The modeling challenge with constructing a transmembrane helical bundle that is consistent with a set of low-to-moderate resolution experimental constraints is, in some ways, more straight-forward than for soluble proteins. 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 [26, 27]. Due to the thermodynamic disadvantages of transferring non-hydrogen bonded peptides from a water to a lipid environment (+5 kcal/mol per H-bond, [28]), transmembrane proteins fold and assemble in a multi-stage process [29, 30]. We assume the two-stage model [30] and consider the building of transmembrane proteins as the separate tasks of defining the transmembrane SSEs and determining their relative orientations or packing.

While not a solved problem, transmembrane spanning SSEs can often be accurately predicted from sequence information using widely accepted methods such as sliding-window hydrophobicity analysis [31-33]. 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 limit the number of possible membrane protein folds [34], and thus several *ab initio* and potential based computational approaches for predicting interhelical packing have been developed [35-41].

Several of these approaches incorporate experimental data into their models. For example, Nikiforovich et al.[36] use the similarity between the X-ray structures of bacteriorhodopsin and rhodopsin to estimate the 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. [37] orient each helical axis of the helix bundle according to the 7.5 Å electron density map of rhodopsin. 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, semiempirical calculations on ligand-protein interaction, multiple sequence alignment and hydrophobicity [42]. Although these methods use energetic calculations and molecular simulations to further refine the helical arrangements, they are potentially biased toward the structures of bacteriorhodopsin and rhodopsin and have yet to be validated for other membrane proteins.

Here, we describe our two-step approach for using sparse distance constraints to model the transmembrane spanning bundles of integral membrane proteins. As many of the known membrane protein structures are all alpha-helical, we will limit the discussion to modeling helical bundles in this work. The two-step method is as follows: First, we searched the conformational space of membrane protein bundles to find those matching a given set of distance constraints [43]; Second, we refined the top-scoring helical bundles with a Monte Carlo simulated annealing protocol designed for local minimization of a custom penalty function/ The penalty function scores a helical bundle

based on its consistency with the structural features of known transmembrane bundles and with distance constraints from experimental methods such as chemical cross-linking, NMR, FRET and EPR.

In this chapter, we describe our penalty function and validate it across a set of known transmembrane protein structures to show that it is capable of distinguishing structures close to the native structure from those far from the native structure. Herein we show that we can construct accurate transmembrane helical bundle models for six disparate transmembrane proteins using simulated cross-linking data sets. We also demonstrate that our two-step approach can recover the transmembrane helical bundle of dark-adapted rhodopsin structure (1F88) to within 4 Å using only 27 experimental distance constraints gathered from the literature.

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 C□ were removed (i.e. we represent the helix in its native form at the C□ 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.

Selection of membrane protein representative set

Membrane proteins 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. 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 penalty function. Those measures having the smallest variances as a percentage of the mean are 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 [43]. Therefore, to accurately represent this importance in our penalty function we set the force constant for experimental distance constraints to the highest value of 500.

Conformational search under a set of distance constraints

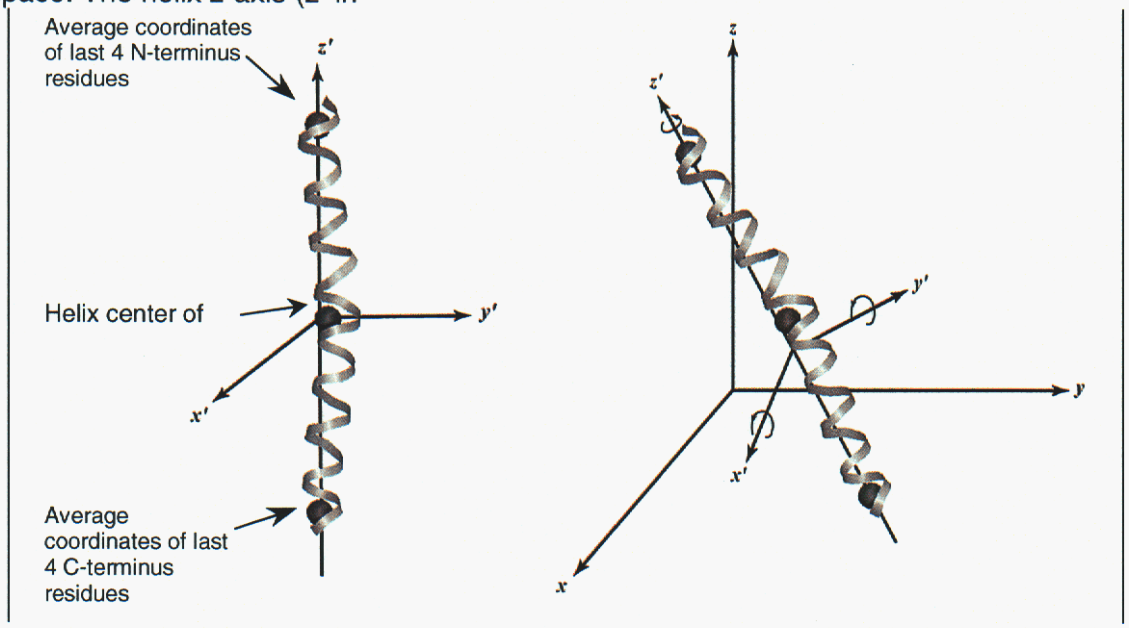
Details of our procedure for exploring the conformational space of membrane protein folds matching distance constraints are provided in [43] and are summarized in the methods section. Briefly, the procedure generates an exhaustive set of helix bundles within a specified RMSD by positioning the helices 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 penalty function developed in this paper and a Monte Carlo simulated annealing (MCSA) protocol to search the local conformational space of the bundle.

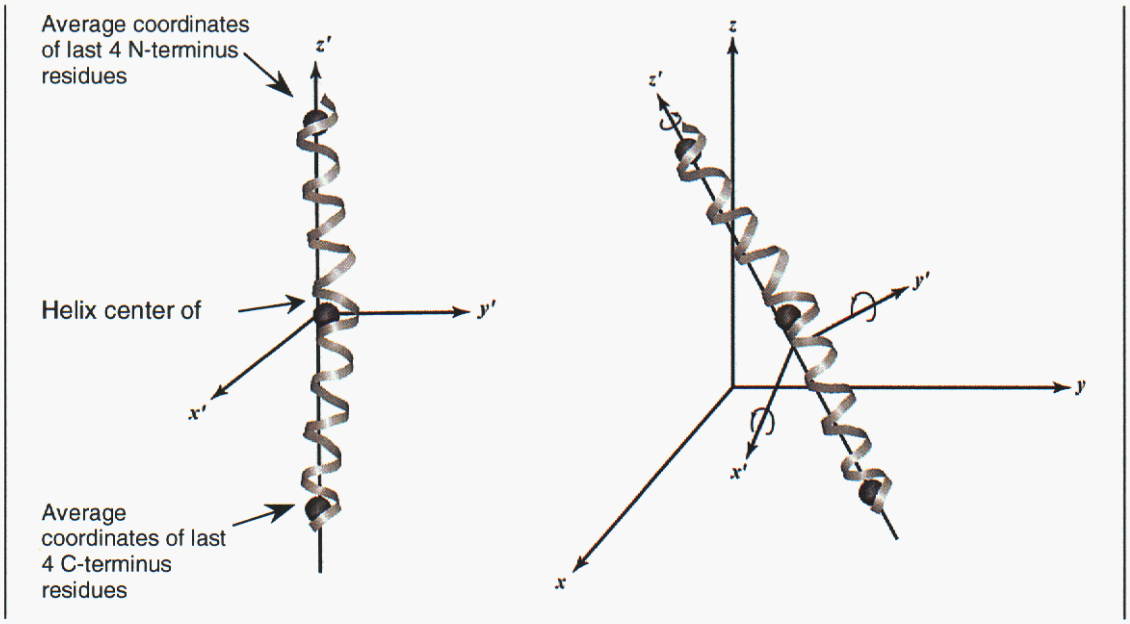
Helix bundle definition

A helix bundle is defined as any arrangement of the helices in Cartesian coordinate space. The helix z-axis (z′ in



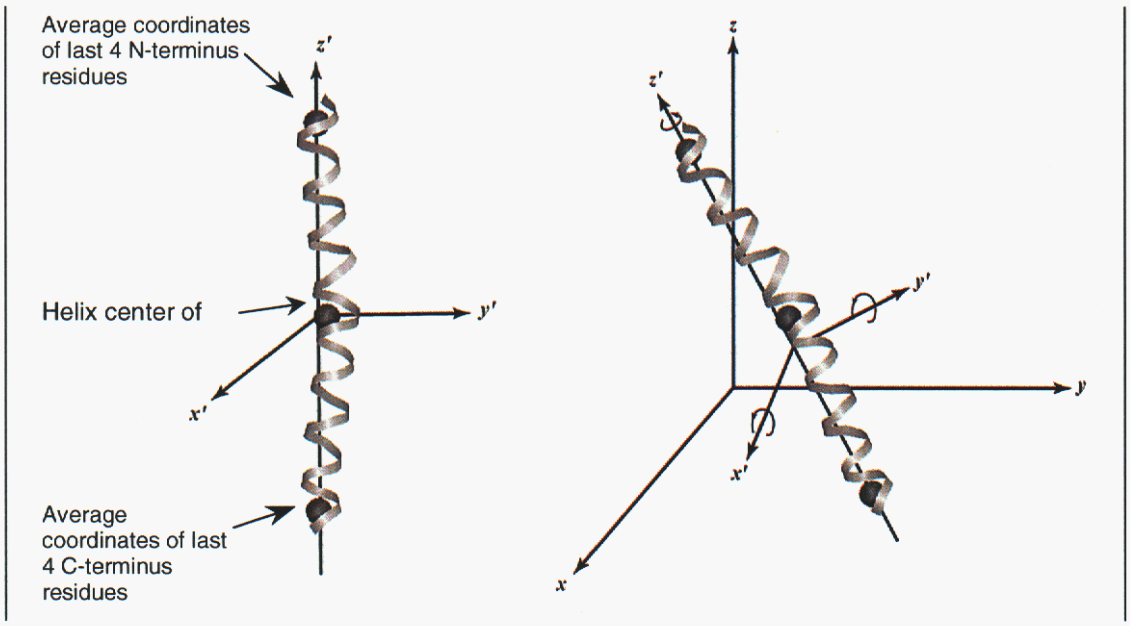
) is defined as the line segment connecting the average coordinates of the N and C termini for each helix

(



). 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

(

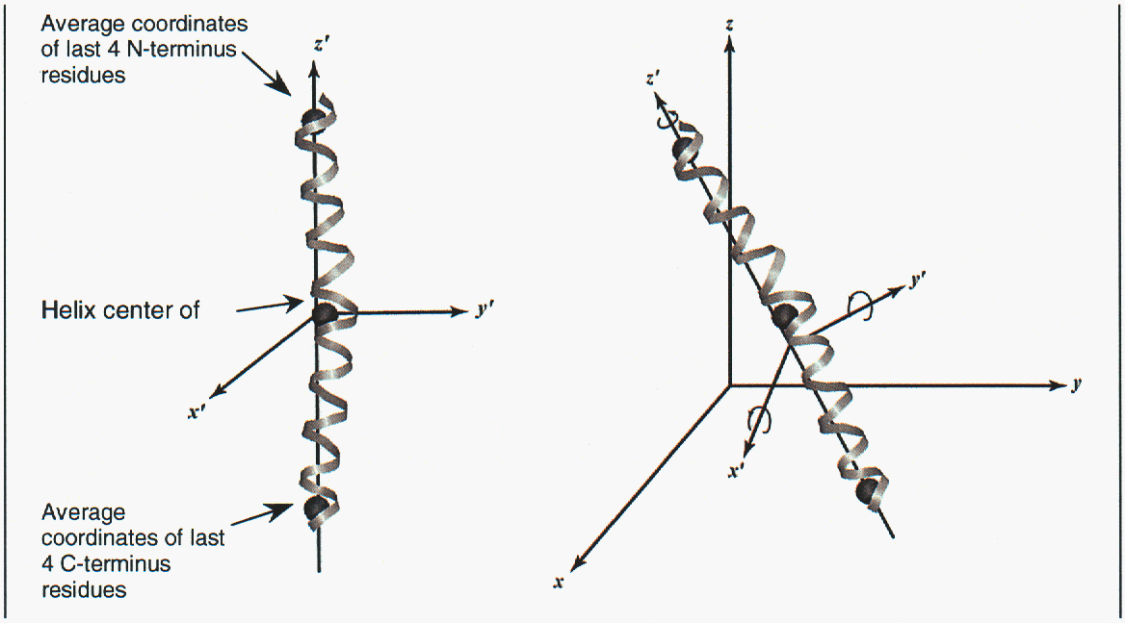


), giving a system wide total of 6*n* degrees of freedom, where *n* is the number of helices.

Monte Carlo sampling

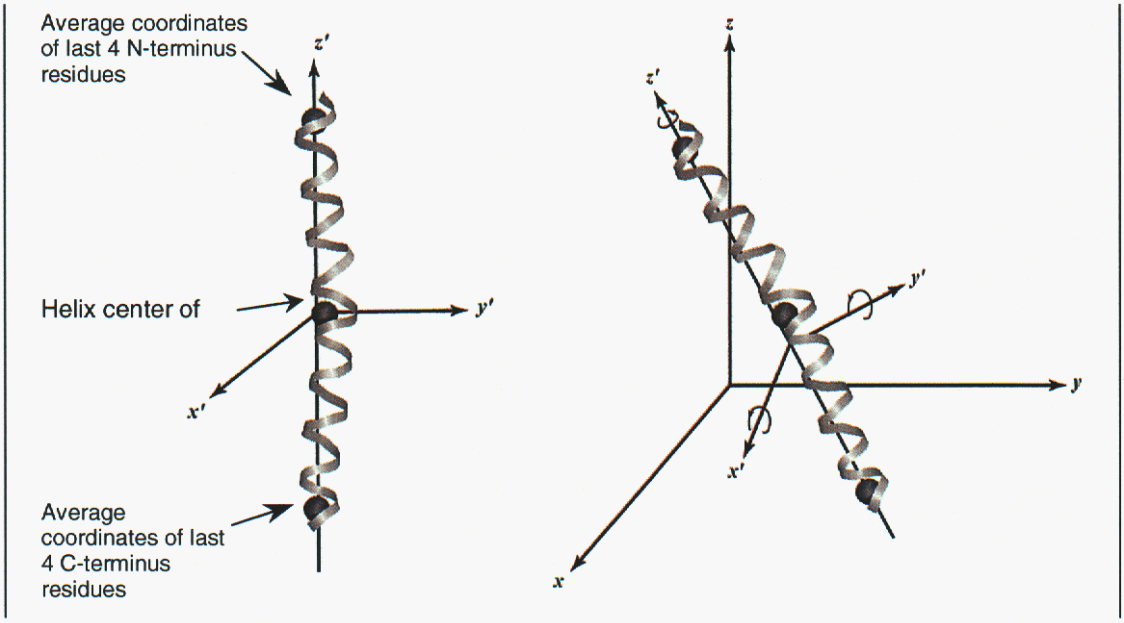
Starting from the last accepted arrangement, a new helical bundle is generated by randomly selecting one of the secondary structural elements (SSEs) and randomizing its position by either translation in the global axis system (x, y, z) or rotation in the local axis system (x′, y′, z′)

(



). Similar to those used by Hertzyk and Hubbard [42], four moves are possible

(



): (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 amount of either translation or rotation are chosen randomly within a user defined limits. If the 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 [44]. The probability factor, p, is compared to a random number, r, from a uniform [0,1] distribution. If $q < r$, the new configuration is accepted as the new best structure; otherwise, the new bundle is rejected [45].

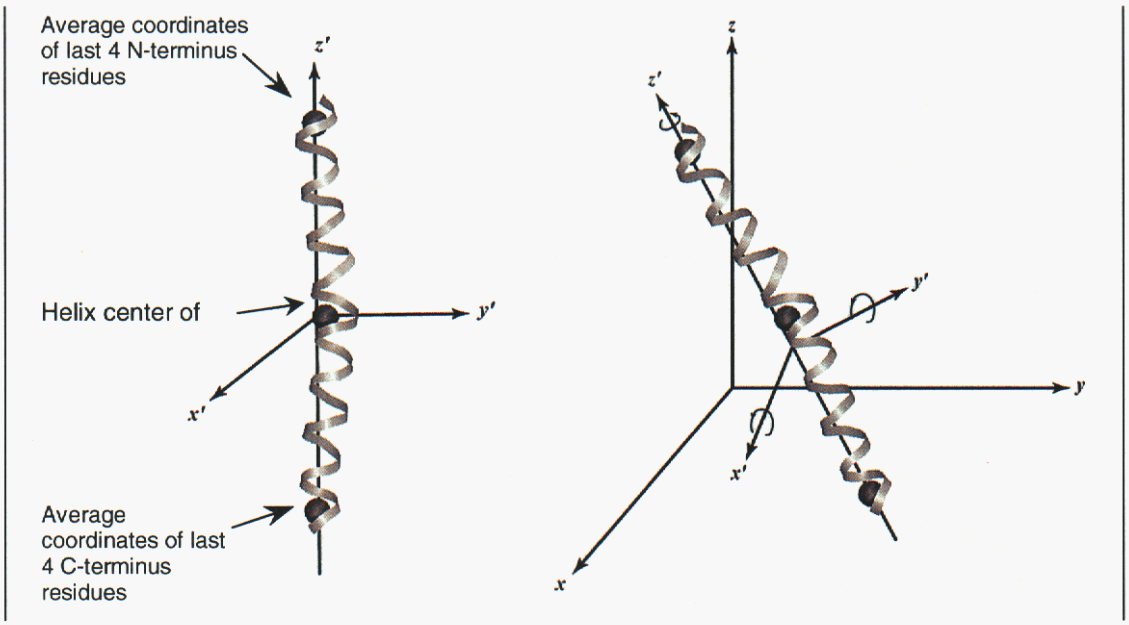


Figure 15. Definition of helix axis system (left) and helix degrees of freedom (right). The helix z-axis is defined as the vector connecting the average coordinates of the last four residues of the helix N and C termini. Helix degrees of freedom include translations in the global (x, y, z) axis system and x′, y′ and z′ rotations around the helix axes.

Cooling schedule

The cooling schedule used for refinements started at $T = 30$ and reduced T 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 100 candidate structures were accepted.

Structural analysis and data processing

Root mean square deviation calculations and various manipulations of pdb files were performed using the the Multiscale Modeling Tools in Structural Biology, MMTSB, tool set [46]. Molecular visualization and renderings were obtained using VMD [47]. All analysis of the penalty data was done using programs written in MATLAB 6.5 (The Math Works Inc., Natick, MA).

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 our scoring or penalty function, which incorporates data from both experimental and statistical analyses of known structures. The function was validated on a set of six membrane proteins for which crystal structures have been deposited in the PDB. Validation is done using distances corresponding to pairs of amino acids (K-K, K-D, K-E, K-C and C-C) that could potentially be cross-linked using chemical cross-linkers. Clearly, this method is not limited to the consideration of distances derived from cross-linking experiments, and so, in the last section, we also demonstrate our method on the structure of rhodopsin using a set of distance constraints taken from the literature.

Statistical analysis of membrane protein structures

A set of 14 membrane proteins with all-alpha helical transmembrane domains was examined to extract statistical information about their helix packing distances, angles and number of nearest neighbors. Table 4 lists the representative set used in this study. BUNDLER constructs and optimizes idealized helical bundles so we concluded that collecting statistics on an idealized set of the 14 proteins would result in more 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 on the helices in the transmembrane domains. The resultant helical Cα RMSDs of the idealized vs. real structures ranged from 0.56 Å (1PRC, 17 aa) to 4.07 Å (1QLAC, 35 aa) and the transmembrane domain-level Cα RMSDs ranged from 1.15 Å (1FQY, 136 aa) to 2.37 Å (1QLAC, 160 aa).

Table. 4. Structures used to derive statistical characterization of membrane protein bundles

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

Table. 5. Statistics describing membrane protein bundles.

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
θ_{pack}	30.9°	16.4 °	336
n_{neigh}	3.4	1.4	102

Statistics that were collected on the 14 idealized representative structures are listed in Table 5. 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 (θ_{pack}) and the number of helical neighbors with a minimum pairwise approach distance (n_{neigh}) $\leq$ 15 Å. N indicates the sample size.

Penalty function

The goal of this work was to create a scoring (or penalty) function that incorporates distance constraints determined from experimental methods such as chemical cross-linking, dipolar EPR, FRET and NMR. This function would assess a possible helical bundle and assign it a score as a measure of how similar it is to the actual structure. Given enough experimental distance constraints, such a function would require no additional considerations. However, measuring distances in membrane proteins can be difficult and thus, we expect only a sparse number of distance constraints to be available. Moreover, we expect that the available distances will not be error free. Therefore, our scoring function includes penalties for violating a set of experimental distance constraints as well as penalties for structures that do not satisfy a number of helix packing parameters determined by analysis of a set of 14 membrane protein structures from the PDB. Thus, the total penalty, *P*, is the sum of a distance constraint penalty and the structure-based penalties:

$$P = P_{\text{distance constraints}} + P_{\text{structure}} \quad (1)$$

Distance Constraints Penalty (P_{dist})

As previously noted, distance constraints are a crucial component in modeling helical membrane proteins [43], and thus we incorporate a penalty for violating distance constraints in our scoring function. Structures are penalized for violating distance constraints according to a soft square well potential defined as

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

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.

In addition to some experimentally determined distances, we include distances generated by correlating loop lengths to helix-end to helix-end distances. 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 [41]. Using the set of 16 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 16). Across the span of loop lengths, this correlation is quite low (R^2 = 0.4). However, we divided this sample into two groups: loops with seven or fewer amino acids (R^2 = 0.8) and loops with eight or more amino acids (R^2 = 0.2). This 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:

$$\begin{aligned} \#AA \le 7 \quad & \begin{cases} LB = 0.7x + 2.9 \\ UB = 1.6x + 6.9 \end{cases} \\ \#AA \ge 8 \quad & \begin{cases} LB = 5 \\ UB = 25 \end{cases} \end{aligned} \quad (3)$$

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 [42].

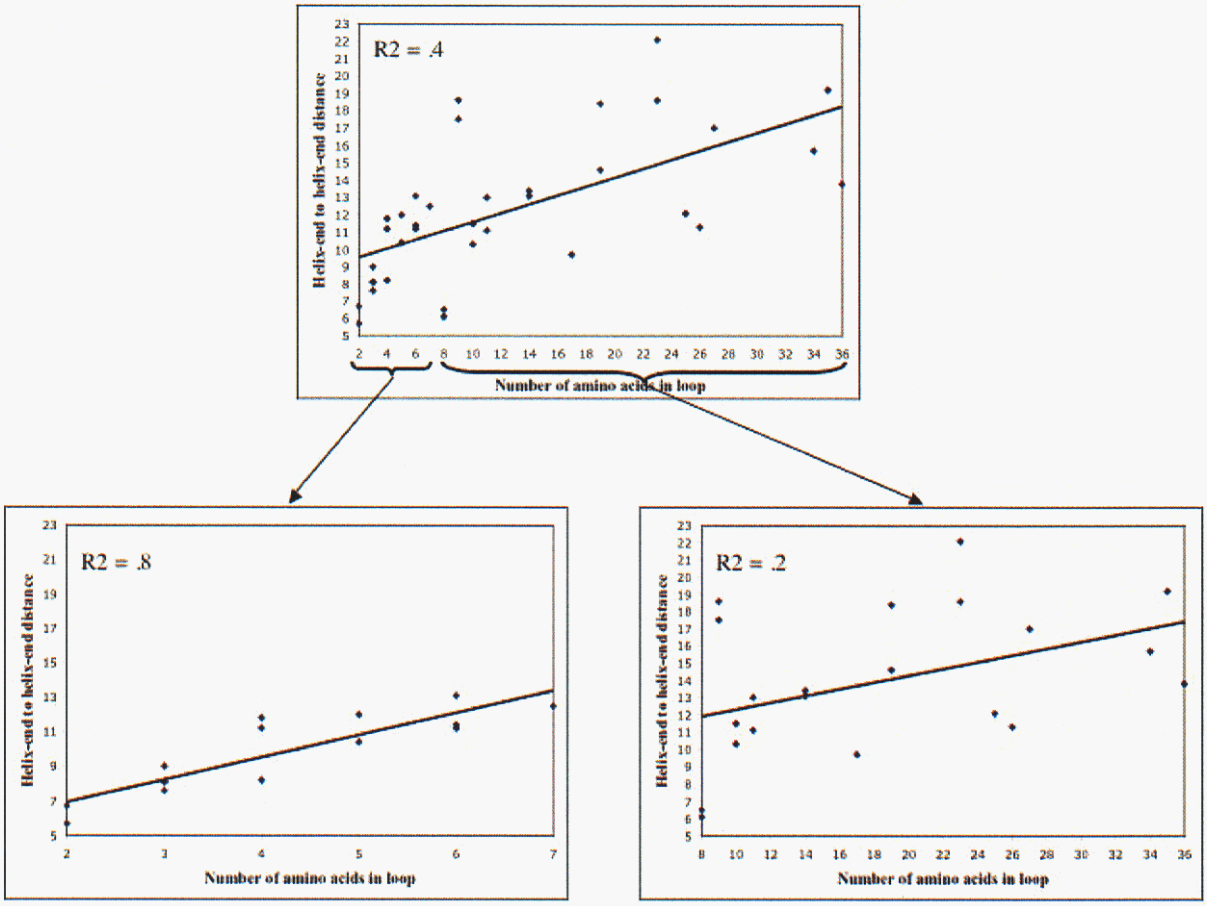


Figure 16. Correlation of helix-end to helix-end distance and number of amino acids in the loop.

Structure based penalties

The structure based scoring function consists of penalties for helical bundles having packing angles, packing distances and/or packing densities outside the ranges determined from our evaluation of 14 non-redundant helical transmembrane proteins. 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

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

We now describe each of these terms in detail.

Packing Distance Penalty (P_{pdist})

The mean distance between the centers of mass of consecutive helices, as derived from a set of 14 non-redundant helical transmembrane protein structures, 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 functional form of packing distance penalty is the "soft" square well potential,

$$P_\delta = k_\delta \begin{cases} (\delta_{ij} - \delta_l)^2, & \delta_{ij} < \delta_l \\ 0, & \delta_l \le \delta_{ij} \le \delta_u, \\ (\delta_u - \delta_{ij})^2, & \delta_{ij} > \delta_u \end{cases} \quad \delta_l = \bar{\delta} - 1.5 s_\delta \text{ and } \delta_u = \bar{\delta} + 1.5 s_\delta, \tag{5}$$

where $\bar{\delta}$ and s_δ are the mean and standard deviation of the interhelical distance; δ_{ij} is the distance between the centers of mass of helix *i* and helix *j* in the current structure; and k_δ is a force constant, which is 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 [48]. We analyzed a non-redundant set of 28 membrane proteins and found the mean backbone packing density to be 37.1 ± 2.5 Å. Hence, we penalize those structures with a packing density more than 1.5 standard deviations away from the mean value. Again, we use a soft square potential and define the packing density penalty as

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

where $\bar{\rho}$ and s_ρ are the mean and standard deviation of the packing density; and k_ρ is a force constant, which is 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 a non-redundant set of 16 "idealized" helical transmembrane proteins, is 30.9 ± 16.3 Å. Packing angle violations are penalized according to a soft square well potential,

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

where $\bar{\theta}$ and s_θ are the mean and standard deviation of the packing angles; and θ_{ij} is the angle between helix *i* and helix *j*. The force constant is k_θ = 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 [49],

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

to prevent interhelical clashes. Here, s is a predetermined van der Waals scaling factor; r_{ij} is the distance between Cβ 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. This piece of the penalty function is summed over the set of all pairs of Cβ atoms. For computing efficiency, we look for only Cβ – Cβ clashes.

Contact Penalty ($P_{contact}$)

In helical membrane protein bundles, it is known that consecutive helices are in contact. Thus, each helix must have at least two neighboring helices. We apply a simple linear penalty to any structure containing a helix that is not in contact with at least two of its neighbors and define the contact penalty as

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

Here, $c < 2$ is the number of helices whose center of mass is helices with a center of mass ≤ □(COM) – 1.5□(COM) of the center of mass of a given helix; and $k_{contact}$= 500. A contact penalty is calculated for each helix in the bundle.

Side-Chain Interaction Preference Penalty ($P_{contact}$)

The amino acids in membrane proteins show a preference for the amino acids with which they interact on neighboring helices [36, 50, 51]. To incorporate this into our scoring function, we use the membrane helical interfacial pairwise (MHIP) amino acid interaction propensity matrix of Adamian and Liang [50]. We adjusted the entries to represent 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 scoring pair.

$$P_{sc} = k_{sc}P_{ij} \quad (10)$$

Total Score

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

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

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 expect to have a penalty function for which the least penalized bundle was necessarily the native structure. 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 our penalty function. The principal requirement of our penalty function 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 our penalty function 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 protein data bank, 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) that only partially insert into the membrane were excluded. For each structure, we derived a set of distance constraints corresponding to pairs of amino acids (K-K, K-D, K-E, K-C and C-C) that could potentially be cross-linked using commercially available chemical cross-linkers and 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 17 displays the results for all six test cases as plots of the penalty function value versus distance from the known structure measured using the RMSD across the Cα atoms (Cα-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 penalty. Moreover, the general trend is for bundles closer in Cα-RMSD to the known structure to have lower penalties than those farther from the known structure. In the case of Aquaporin, this trend is not as strong. Although the known structure does have the lowest penalty, the correlation between distance from the known structure and penalty was not strong. This lack of correlation for Aquaporin is not unexpected considering that we are including only the transmembrane helices that span the membrane and omitting the two short helices.

To further test the robustness of the penalty function 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. These results are shown in the bar charts in Figure 17. Overall, the lower scoring structures correspond to structures closer to the target or native structure. Thus, it is reasonable to expect that the lowest scoring models 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 higher scoring bundle is not close to the native-like bundle and the bundles with the lowest penalty represent the most native-like bundles amongst the set of possible models. Exluding Aquaporin, these results also provide sufficient evidence that a maximum penalty can be used to pick a subset of models for further refinement. For example, model bundles with a penalty of less than 2000, or more conservatively 3000, appear to be 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 main goal of this work is to develop a technique for building the transmembrane helix 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 [43] with Monte Carlo simulated annealing (MCSA) of the empirical scoring function previously described in this paper. 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 helical bundles to find those satisfying a given set of distance constraints and then by searching the local conformational space of each of these candidate models. Each step is detailed in the Methods section.

We demonstrate the method 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., [52]. These included dipolar EPR distances [53-56], disulfide mapping distances [57-61] and distances from electron cryo-microscopy [62]. These distance constraints have an average error of ± 3.75 Å.

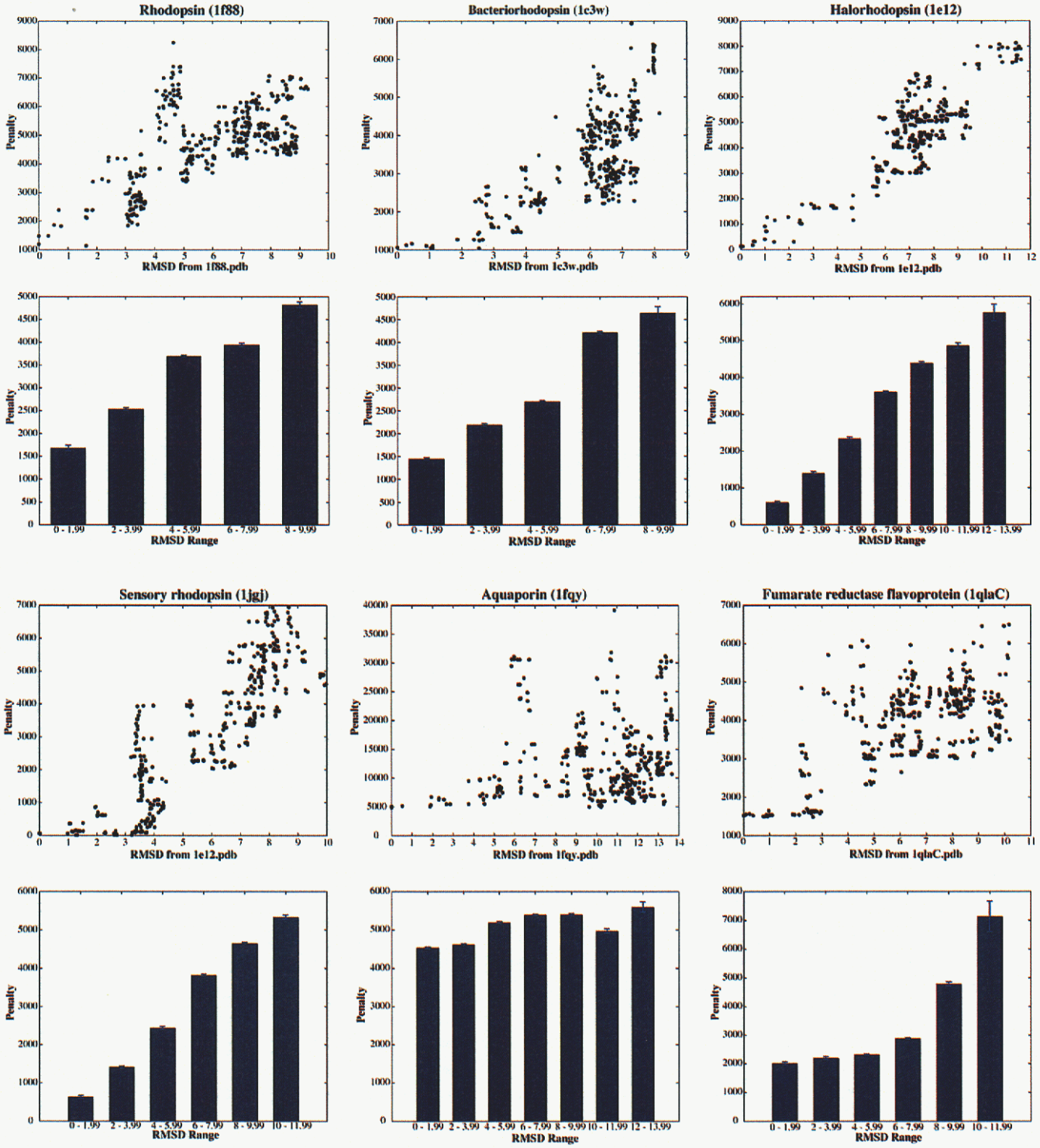


Figure 17. Penalty as a function of root mean square deviation from the x-ray structure for six integral membrane proteins. Sets of 500 structures were generated using a Monte Carlo simulated annealing algorithm at a single high temperature as described in the text. Scatter plots show the results for 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.

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 Cα-Cα 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) [53, 54, 63], four from Human Carbonic Anhydrase II [64, 65], four from T4-lysozyme (3LZM) [66, 67], and one each from Maltose-binding protein Liganded form (1MDP) [68, 69] and Maltose-binding protein unliganded form (1DMB) [70]. 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 Cβ 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 [43], we developed 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 x 10^{11} possible seven-helix configurations to only 87 helical bundles with Ca – RMSD ranging from 4.3 to 9.5 Å [43]. 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 penalty function value. Since our goal is only to search the local conformational space of each bundle, 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 1000 total structures were generated or 100 structures were accepted, whichever occurred first. The MCSA simulations were run for 34 temperature steps.

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 18, the penalty on this structure is much lower than those of the lowest RMSD helix assemblies, which indicate that models with penalties in the range of 1000 to 2000 should be native-like bundles. Among the 87 refined bundles several have minimized penalties around 1000.

Figure 18. Comparison of predicted helical bundle (black) to the native 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 bundle among these has a penalty of 1003.3, a Cα – RMSD of 3.2 Å (**Error! Reference source not found.**). This result again provides evidence that simply minimizing an empirical structure based penalty function will 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 penalties between approximately 1000 and 2000, which provides a better stopping criteria for our MCSA refinement protocol. For example, we can anneal the structure, possibly with faster cooling, until reaching a penalty of 2000 and then slow the cooling to more thoroughly sample those conformations with scores between 1000 and 2000. The search will ultimately be stopped when the penalty drops below 1000.

Table. 6. Experimental distances used for the Rhodopsin structure[1]

Helix1	Helix2	Residue1	Residue2	Minimum Distance	Maximum Distance	Experimental Method	Reference
C	F	139	248	7	19	Dipolar SDSL-EPR	[54]
C	F	139	249	8	27	〃	〃
C	F	139	250	8	27	〃	〃
C	F	139	251	4	22	〃	〃
C	F	139	252	8	27	〃	〃
A	G	65	316			〃	[53]
E	F	204	276	4	8	Disulfide Mapping	[60]
C	E	140	222	4	8	〃	[61]
C	E	140	225	4	8	〃	〃
C	F	135	250	4	8	〃	〃
C	E	136	222	4	8	〃	[57]
C	E	136	225	4	8	〃	〃
B	C	71	134	7	15	Electron Diffraction	[71]
B	C	90	116	5	10	〃	〃
B	D	71	153	4	11	〃	〃
B	D	86	172	15	20	〃	〃
C	E	136	226	5	10	〃	〃
C	E	125	215	5	10	〃	〃
D	E	152	225	18	22	〃	〃
E	F	216	258	9	13	〃	〃
F	G	253	305	5	9	〃	〃
F	G	264	298	5	9	〃	〃
A	G	39	286	9	14	〃	〃
C	F	114	268	14	18	〃	〃
D	F	171	268	16	21	〃	〃
B	F	73	250	10	15	〃	〃
A	F	62	250	16	20	〃	〃
A	F	47	264	15	20	〃	〃

[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.

Discussion of BUNDLER Results

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

In this work we have presented a penalty function designed to assist in modeling membrane proteins. The function penalizes helical arrangements that violate distance constraints and that violate constraints derived from a statistical analysis of 15 non-redundant membrane protein structures from the PDB. The function is validated across a set of helical bundles from a variety of membrane proteins. Because the majority of known structures are seven helix bundles, it is not surprising that the penalty function works very well for this class of membrane proteins. However, we have also illustrated that our function can be useful for modeling other classes of helical bundles (e..g. Aquaporin, Fumarate reductase flavoprotein).

In Figure 17, we show that for the six helical bundle of Aquaporin (with the two short non-membrane-spanning helices removed), our function identifies structures having penalties around 4500 as being closer to the native structure than those having a higher score.

This chapter presents our attempt at building a penalty function for refining helical bundles, and it stands as a proof of concept of the functional form that incorporates structural based penalties. 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. Whether or not a truly general function useful for refining helix bundles with a range of secondary structural elements can be developed remains to be seen. While it is likely that the form of the penalty function presented in this chapter 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 the penalty functional easily adapted to new statistics based parameterization.

In summary, we have presented a two-step approach to modeling transmembrane helical bundles and demonstrated this approach using the structure of rhodopsin as a test case.

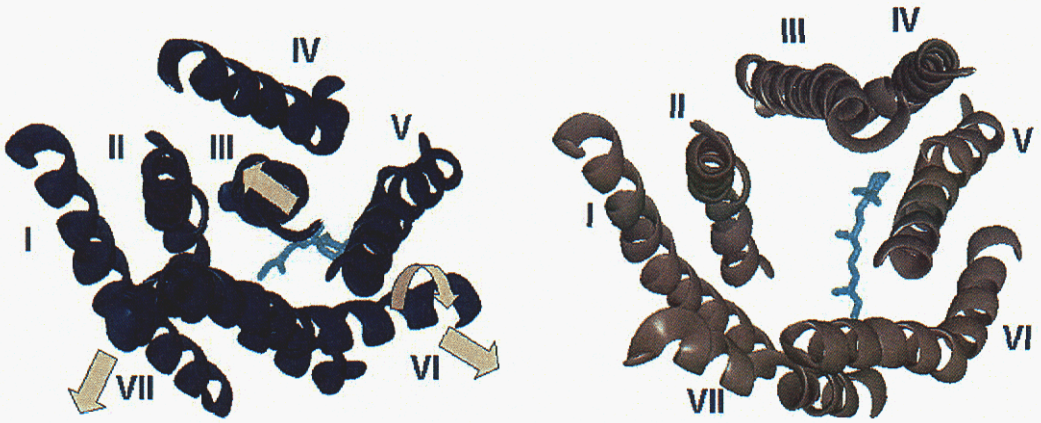


Figure 19. Proposed model of rhodopsin light activation. Left: Ribbon diagram of the rhodopsin crystal structure helical bundle (1f88). The arrows indicate the predicted helical movements. Right: Optimized light-adapted rhodopsin model generated by a distance geometry calculation using 24 literature-derived experimental constraints.

Given a set of 27 distance constraints extracted from the literature, we modeled the helical bundle of dark-adapted rhodopsin to within 3.2 Å of the X-ray structure deposited in the PDB.

Light-adapted rhodopsin model

We constructed a backbone-level model of the light-adapted rhodopsin structure using the known dark-adapted structure as a starting point. The 7 transmembrane helices from the dark-adapted structure (1f88) and the all-trans form of the rhodopsin chromophore, retinal, were represented as rigid objects in a distance geometry calculation. Twenty-four distance constraints compiled from published spin labeling, ligand cross-linking, metal binding and disulfide mapping experiments were included in a distance geometry calculation, along with additional constraints imposed by the topology of the rhodopsin structure. The resultant light-adapted structure (Figure 18) is consistent with the helical movements proposed in the literature [72].

Mass Spectrometry Data Reduction and Analysis

Data reduction

We developed a macro program to be used within the XMASS MS spectrum analysis program. The macro automatically processes spectra acquired from LC/MS or direct-infusion MS experiments, and picks monoisotopic peaks for further analysis. The macro program compiles information about the m/z values, charge state(s), MH+, intensity, and (if relevant) the scan number for each species observed in a MS spectrum. This output file is used as input for the next step in the analysis procedure.

MS spectrum assignment

The Automated Spectrum Assignment Program (ASAP), originally developed at the University of California, San Francisco [5] and improved under the current work was used to suggest possible structures for both cross-linked and non-cross-linked peptides resulting from the proteolytic digestion of cross-linked proteins. Datasets of mass spectra obtained from FT-MS experiments were searched with ASAP using a mass error of ±5-10 ppm.

MS2Assign was developed in this current work to assign tandem mass spectra of unmodified, labeled and/or cross-linked peptides. Given information about the identity of the molecular ion and the cross-linking reagent, MS2Assign generates a theoretical library containing all of the possible fragmentation products. The theoretical library is constructed based on common peptide fragmentation pathways that result in a,b,c-type, x,y,z-type, internal and immonium ions with associated common losses of H_2O, NH_3, CO, and CO_2. In addition, MS2Assign calculates all of the fragments generated from a list of user-defined peptide mass modifications (for example, carbamidomethyalted cysteines) and/or a defined intra- or inter-peptide crosslink. The number and type of user-defined modifications used in the library calculation is completely up to the user's discretion. The current version of MS2Assign only supports 1 crosslink per peptide or pair of peptides, and does not calculate the fragmentation products generated from cleavage(s) within the crosslinker itself. The additional fragments due to user-defined modifications or crosslinks are stored in the theoretical library.

MS2Assign then attempts to assign each product ion peak obtained in a MS/MS experiment from a given protonated molecular ion (MH_+) to a species in the fragmentation library to within a user-defined error threshold (usually ±50-100 ppm). The MS2Assign output consists of a list of assigned peaks, with information about the observed and theoretical masses, the experimental error, the ion-type name, and sequence information for each assigned species. MS2Assign summarizes the number of successfully assigned peaks at the end of the assignment calculation. For peaks with multiple possible assignments within the given error range, all assignments are listed in the output.

MS2Assign is a C program that is currently compiled under IRIX. Assignment calculations for a typical set of cross-linked peptides take on the order of seconds to perform, but the runtime of the program scales linearly with the length of the input mass list. Web-based versions of MS2Assign and ASAP are available for beta testing at http://roswell.ca.sandia.gov/~mmyoung.

MS/MS spectrum assignment

Assignment of the fragmentation spectra was performed by our in-house software package MS2Links. MS2Links accepts user input on the protein or peptide sequence(s), the mass modification and amino acid specificity of the cross-linking reagent and calculates a complete theoretical fragmentation library. The fragmentation library contains all possible cross-linking possibilities for the b-type, y-type and internal fragment ions. Given an input m/z fragment list, MS2PRO assigns the m/z peaks that are within a defined ppm error to species in the theoretical library. When the input list indicates that the monoisotopic peak may have been or definitely was not observed, MS2Links also checks for matches to the first ^{13}C peak to the species in the theoretical library.

References

1. Palczewski, K., et al., *Crystal structure of rhodopsin: A G protein-coupled receptor.* Science, 2000. **289**(5480): p. 739-45.
2. Noel, J.P., H.E. Hamm, and P.B. Sigler, *The 2.2 A crystal structure of transducin-alpha complexed with GTP gamma S.* Nature, 1993. **366**(6456): p. 654-63.
3. Gaudet, R., A. Bohm, and P.B. Sigler, *Crystal structure at 2.4 angstroms resolution of the complex of transducin betagamma and its regulator, phosducin.* Cell, 1996. **87**(3): p. 577-88.
4. Sondek, J., et al., *GTPase mechanism of Gproteins from the 1.7-A crystal structure of transducin alpha-GDP-AIF-4.* Nature, 1994. **372**(6503): p. 276-9.
5. Young, M.M., et al., *High-Throughput Structure Determination: Rapid Identification of Protein Folds Using Mass Spectrometry and Intramolecular Cross-linking.* Proc Natl Acad Sci U S A, 2000. **97**(11): p. 5802-6.
6. Peters, K. and F.M. Richards, *Chemical Cross-linking: Reagents and Problems in Studies of Membrane Structure.* Ann. Rev. Biochem., 1977. **46**: p. 523-551.
7. Brunner, J., *New photolabeling and cross-linking methods.* Ann. Rev. of Biochem, 1993. **62**: p. 483-514.
8. Fancy, D.A., *Elucidation of protein-protein Interactions using chemical cross-linking or label transfer techniques.* Curr. Opin. in Chem. Biol., 2000. **4**: p. 28-33.
9. Havel, T.F., G.M. Crippen, and I.D. Kuntz, *Effects of Distance Constraints on Macromolecular Conformation. II. Simulation of Experimental Results and Theoretical Predictions.* Biopolymers, 1979. **18**: p. 73-81.
10. Crippen, G.M. and T.F. Havel, *Distance Geometry and Molecular Conformation.* Taunton, England, Research Studies Press John Wiley, 1988.
11. HŠnggi, G. and W. Braun, *Pattern recognition and self-correcting distance geometry calculations applied to myohemerythrin.* FEBS Letters, 1994. **344**: p. 147-153.
12. Asz—di, A.M. and W.R. Taylor, *Hierarchic inertial projection: a fast distance matrix embedding algorithm.* Computers Chem, 1996. **21**: p. 13-23.
13. Asz—di, A.M. and W.R. Taylor, *Homology Modeling by Distance Geometry.* Folding & Design, 1996. **1**: p. 325-334.
14. Asz—di, A.M., R.E.J. Munro, and W.R. Taylor, *Protein Modeling by Multiple Sequence Threading and Distance Geometry.* Proteins: Struct. Func. & Genetics S1, 1997: p. 38-42.
15. Huang, E.S., R. Samudrala, and J.W. Ponder, *Ab Initio Fold Prediction of Small Helical Proteins using Distance Geometry and Knowledge-based Scoring Functions.* J. Mol. Biol., 1999. **290**: p. 267-281.
16. Pearlman, D.A., et al., *AMBER, a Package of Computer Programs for Applying Molecular Mechanics, Normal Mode Analysis, Molecular*

Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. Comp. Phys. Lett., 1995. **91**: p. 1-41.
17. Zheng, D., et al., *Automated protein fold determination using a minimal NMR constraint strategy.* Protein Sci, 2003. **12**(6): p. 1232-46.
18. Service, R.F., *Structural genomics. Tapping DNA for structures produces a trickle.* Science, 2002. **298**(5595): p. 948-50.
19. Gerstein, M., et al., *Structural genomics: current progress.* Science, 2003. **299**(5613): p. 1663.
20. Terwilliger, T.C., *Structural genomics in North America.* Nat Struct Biol, 2000. **7 Suppl**: p. 935-9.
21. Sali, A., *100,000 protein structures for the biologist.* Nature Struct. Biol., 1998. **5**: p. 1029-1032.
22. Schilling, B., et al., *MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically cross-linked peptides.* J Am Soc Mass Spectrom, 2003. **14**(8): p. 834-50.
23. Faulon, J., M. Rintoul, and M. Young, *Constrained walks and self-avoiding walks: Implications for protein structure determination.* J Phys A, 2002. **35**: p. 1-19.
24. Bowie, J.U., *Helix-bundle membrane protein fold templates.* Protein Sci, 1999. **8**(12): p. 2711-9.
25. Meng, E.C. and H.R. Bourne, *Receptor activation: What does the rhodopsin structure tell us?* Trends Pharmacol. Sci., 2001. **22**(11): p. 587-93.
26. Kim, S. and T.A. Cross, *Uniformity, ideality, and hydrogen bonds in transmembrane alpha-helices.* Biophys J, 2002. **83**(4): p. 2084-95.
27. White, S.H. and W.C. Wimley, *Membrane protein folding and stability: physical principles.* Annu Rev Biophys Biomol Struct, 1999. **28**: p. 319-65.
28. Engelman, D.M., T.A. Steitz, and A. Goldman, *Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins.* Annu Rev Biophys Biophys Chem, 1986. **15**: p. 321-53.
29. Jacobs, R.E. and S.H. White, *The nature of the hydrophobic binding of small peptides at the bilayer interface: implications for the insertion of transbilayer helices.* Biochemistry, 1989. **28**(8): p. 3421-37.
30. Popot, J.L. and D.M. Engelman, *Membrane protein folding and oligomerization: the two-stage model.* Biochemistry, 1990. **29**(17): p. 4031-7.
31. Rose, G.D., *Prediction of chain turns in globular proteins on a hydrophobic basis.* Nature, 1978. **272**(5654): p. 586-90.
32. Jayasinghe, S., K. Hristova, and S.H. White, *Energetics, stability, and prediction of transmembrane helices.* J Mol Biol, 2001. **312**(5): p. 927-34.
33. Jayasinghe, S., K. Hristova, and S.H. White, *MPtopo: A database of membrane protein topology.* Protein Sci, 2001. **10**(2): p. 455-8.
34. White, S.H. and W.C. Wimley, *Hydrophobic interactions of peptides with membrane interfaces.* Biochim Biophys Acta, 1998. **1376**(3): p. 339-52.
35. Bowie, J.U., *Helix-bundle membrane protein fold templates.* Protein Science, 1999. **8**: p. 2711-2719.

36. Nikiforovich, G.V., et al., *Novel approach to computer modeling of seven-helical transmembrane proteins: Current progress in the test case of bacteriorhodopsin.* Acta Biochimica Polinica, 2001. **48**: p. 53-64.
37. Vaidehi, N., et al., *Prediction of structure and function of G protrein-couple receptors.* PNAS, 2002. **99**: p. 12622-12627.
38. Kim, S., A.K. Chamberlain, and J.U. Bowie, *A simple method for modeling transmembrane helix oligomers.* Journal of Molecular Biology, 2003. **329**: p. 831-840.
39. Bowie, J.U., *Helix packing in membrane proteins.* J. Mol. Biol, 1997. **272**: p. 780-789.
40. Dobbs, H., et al., *Optimal potentials for predicting inter-helical packing in transmembrane proteins.* Proteins, 2002. **49**(3): p. 342-9.
41. Fleishman, S.J. and N. Ben-Tal, *A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices.* J Mol Biol, 2002. **321**(2): p. 363-78.
42. Herzyk, P. and R.E. Hubbard, *Automated method for modeling seven-helix transmembrane receptors from experimental data.* Biophys J, 1995. **69**(6): p. 2419-42.
43. Faulon, J.-L., K. Sale, and M. Young, *Exploring the conformational space of membrane protein folds matching distance constraints.* Protein Science, 2003. **12**: p. 1750-1761.
44. Kirkpatrick, S., C.J. Gerlatt, and M. Vecchi, *Optimization by simulated annealing.* Science, 1983. **220**: p. 671-680.
45. Metropolis, N., et al., *Equations of state calculations by fast computing machines.* Journal of Chemical Physics, 1958. **21**: p. 1087 - 1092.
46. Feig, M., J. Karanicolas, and C.L.I. Brooks, *MMTSB Tool Set.* 2001, MMTSB NIH Research Resource, The Scripps Research Institute.
47. Humphrey, W., A. Dalke, and K. Schulten, *VMD - Visual Molecular Dynamics.* Journal of Molecular Graphics, 1996. **14**: p. 33-38.
48. Richards, F., *The interpretation of protein structures: total volume, group volume distributions and packing density.* Journal of Molecular Biology, 1974. **82**: p. 1-14.
49. Brunger, A.T., A. Krukowski, and J.W. Erickson, *Slow-cooling protocols for crystallographic refinement by simulated annealing.* Acta Crystallogr A, 1990. **46 (Pt 7)**: p. 585-93.
50. Adamian, L. and J. Liang, *Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins.* J Mol Biol, 2001. **311**(4): p. 891-907.
51. Adamian, L., et al., *Higher-order interhelical spatial interactions in membrane proteins.* J Mol Biol, 2003. **327**(1): p. 251-72.
52. Yeagle, P.L., G. Choi, and A.D. Albert, *Studies on the structure of the G-protein-coupled receptor rhodopsin including the putative G-protein binding site in unactivated and activated forms.* Biochemistry, 2001. **40**: p. 11932-11937.
53. Yang, K., et al., *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, 1996. **35**: p. 14040 - 14046.

54. Farrens, D.L., et al., *Requirement of rigid-body motion of transmembrane helices for light activation of rhodopsin.* Science, 1996. **274**: p. 768 - 770.
55. Albert, A.D., et al., *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, 1997. **1328**: p. 74 - 82.
56. Galasco, A., R.K. Crouch, and D.R. Knapp, *Intrahelic arrangement in the integral membrane protein rhodopsin investigated by site-specific chemical cleavage and mass spectrometry.* Biochemistry, 2000. **39**: p. 4907 - 4914.
57. Cai, K., et al., *Structure and function in rhodopsin: topology of the C-terminal polypeptide chain in relation to the cytoplasmic loops.* Proc. Natl. Acad. Sci. USA, 1997. **94**: p. 14267 - 14272.
58. Cai, K., et al., *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, 1999. **38**: p. 12893 - 1898.
59. Sheikh, S.P., et al., *Rhodopsin activation blocked by metal-ion-binding sites linking transmembrane helices C and F.* Nature, 1996. **383**(6598): p. 347 - 350.
60. Yu, H., et al., *A general method for mapping tertiary contacts between amino acid residues in membrane embedded proteins.* Biochemistry, 1995. **34**: p. 14963 - 14969.
61. Yu, H., M. Kono, and D.D. Oprian, *State-dependent Disulfide Cross-linking in Rhodopsin.* Biochemistry, 1999. **38**: p. 12028 - 12032.
62. Unger, V.M. and G.F. Schertler, *Low resolution structure of bovine rhodopsin determined by electron cryo-microscopy.* Biophysical Journal, 1995. **68**: p. 1776-1786.
63. Palcewski, K., et al., *Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor.* Science, 2000. **289**: p. 739.
64. Hakansson, K., et al., *Structure of native and apo carbonic anhydrase II and structure of some of its anion-ligand complexes.* Journal of Molecular Biology, 1992. **227**: p. 1192.
65. Persson, M., et al., *Comparison of electron paramagnetic resonance methods to determine distances between spin labels on human carbonic anhydrase II.* Biophys J, 2001. **80**(6): p. 2886-97.
66. Matsumura, M., et al., *Structural studies of mutants of T4 lysozyme that alter hydrophobic stabilization.* J Biol Chem, 1989. **264**(27): p. 16059-66.
67. McHaourab, H.S., et al., *Conformation of T4 lysozyme in solution. Hinge-bending motion and the substrate-induced conformational transition studied by site-directed spin labeling.* Biochemistry, 1997. **36**(2): p. 307-16.
68. Sharff, A.J., et al., *Refined structures of two insertion/deletion mutants probe function of the maltodextrin binding protein.* Journal of Molecular Biology, 1995. **246**: p. 8.
69. Hall, J.A., et al., *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, 1997. **272**(28): p. 17610-4.

70. Sharff, A.J., L.E. Rodseth, and F.A. Quiocho, *Refined 1.8-A structure reveals the mode of binding of beta- cyclodextrin to the maltodextrin binding protein.* Biochemistry, 1993. **32**: p. 10553.
71. Baldwin, J.M., G.F. Schertler, and V.M. Unger, *An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors.* J Mol Biol, 1997. **272**(1): p. 144-64.
72. Bourne, H.R. and E.C. Meng, *Structure. Rhodopsin sees the light.* Science, 2000. **289**(5480): p. 733-4.

Distribution

1	MS 9004	Rick Stulen, 8100
1	MS 9951	Len Napolitano, 8100
1	MS 9951	Malin Young, 8130
1	MS 9951	Joe Schoeniger, 8130
1	MS 0323	D. Chavez, LDRD Office, 1011
3	MS 9018	Central Technical Files, 8945-1
1	MS 0899	Technical Library, 9616
1	MS 9021	Classification Office, 8511, for Technical Library, MS 0899, 9616 DOE/OSTI via URL

This page intentionally left blank