The use of soluble protein structures in modeling helical proteins in a layered membrane

Hong Wing Lee\textsuperscript{a,1}, Hong Ching Lee\textsuperscript{a,1}, Lawrence K. Lee\textsuperscript{a,b}, Erdahl T. Teber\textsuperscript{a,c} and W. Bret Church\textsuperscript{a,*}

\textsuperscript{a}Faculty of Pharmacy, Group in Biomolecular Structure and Informatics, University of Sydney, Sydney, NSW 2006, Australia; \textsuperscript{b}Computational and Structural Biology Division, Victor Chang Cardiac Research Institute, Sydney, NSW 2010, Australia; \textsuperscript{c}Children’s Medical Research Institute, 214 Hawkesbury Road, Westmead, NSW 2145, Australia

Communicated by Ramaswamy H. Sarma

(Received 4 January 2011; final version received 9 January 2013)

Major advances have been made in the prediction of soluble protein structures, led by the knowledge-based modeling methods that extract useful structural trends from known protein structures and incorporate them into scoring functions. The same cannot be reported for the class of transmembrane proteins, primarily due to the lack of high-resolution structural data for transmembrane proteins, which render many of the knowledge-based method unreliable or invalid. We have developed a method that harnesses the vast structural knowledge available in soluble protein data for use in the modeling of transmembrane proteins. At the core of the method, a set of transmembrane protein decoy sets that allow us to filter and train features recognized from soluble proteins for transmembrane protein modeling into a set of scoring functions. We have demonstrated that structures of soluble proteins can provide significant insight into transmembrane protein structures. A complementary novel two-stage modeling/selection process that mimics the two-stage helical membrane protein folding was developed. Combined with the scoring function, the method was successfully applied to model 5 transmembrane proteins. The root mean square deviations of the predicted models ranged from 5.0 to 8.8 Å to the native structures.

Keywords: protein structure; polytopic membrane proteins; knowledge-based modeling; structural bioinformatics

Introduction

Helical integral membrane proteins (IMPs) consist of one or more helices that traverse the membrane. They facilitate vital processes in cells, acting as signal receptors, channels, and pumps. They are also major pharmaceutical targets, with 60–70% of current drugs targeting membrane proteins (Lundstrom, 2004).

Despite their importance, attempts to experimentally determine the three-dimensional structures of helical membrane proteins are often unsuccessful, largely due to the difficulties in obtaining sufficient quantities of pure protein for crystallization trials. This is reflected in the Protein Data Bank (PDB), in which (as of 3 April 2012) there are 80,550 protein structures (Berman et al., 2000), but no more than 1000 of these are IMPs (White & Wimley, 1999), only approximately 350 of which are helical IMPs (Lomize, Lomize, Pogozheva, & Mosberg, 2006). This number of IMPs equates to about 1% of all known protein structures, in contrast with the 20–30% estimated in the proteome (Almén, Nordström, Fredriksson, & Schiöth, 2009; Gough, Karplus, Hughey, & Chothia, 2001; von Heijne, 2007).

An alternate avenue to obtaining the 3D structure of proteins experimentally is via computational protein modeling, which is routinely utilized to generate structures of soluble proteins. The most accurate models rely on the existence of homologous protein structures, and hence, the success of this approach greatly depends on the large number of soluble protein structures elucidated. In contrast, the lack of experimentally determined helical IMP structures restricts the modeling of most IMPs to \textit{ab initio} approaches, which are generally less accurate, but allow for the modeling of novel folds.

The most successful \textit{ab initio} methods for predicting helical IMP structures are the folding in lipid membranes (FILM) and Rosetta-membrane methods, which have produced moderately accurate models. FILM is modified from a modeling method designed for soluble proteins (Pellegrini-Calace, Carotti, & Jones, 2003), but with an additional scoring function for membrane-associated
helices. For fragments containing up to two transmembrane helices, FILM-generated models with root mean square deviations (RMSDs) of 3.5–6.5 Å from crystal structures (Pellegrini-Calace et al., 2003). The Rosetta-membrane algorithm generated models of proteins containing three to seven transmembrane helices with RMSDs of 3.1–16 Å from their crystal structures, with three structures under 5 Å RMSD (Yarov-Yarovoy, Schonbrun, & Baker, 2006). This utilized a small set of 28 existing IMP structures to determine energy functions that were used to model 12 membrane protein structures or subdomains. Rosetta-membrane implementations have already been strategically applied to illuminate the molecular mechanisms of the potassium channels, such as for the voltage-dependent potassium channel gating (Pathak et al., 2007; Yarov-Yarovoy, Baker, & Catterall, 2006) and stabilization of pore domains of the human Ether-α-go-go-Related Gene (Lees-Miller et al., 2009).

In this study, we explore the possibility of utilizing the large number of soluble protein structures for the prediction of IMP structures. We characterize common stereochemical trends found in soluble structures, define these as scoring functions and demonstrate that these can be re-parameterized for the folding and validation modeling of helical IMP structures. To validate these properties, we created decoy sets for transmembrane proteins and tested the scoring functions against them. Finally, structures containing three to five transmembrane helices were modeled with comparable efficacy as the existing methods.

Materials and methods

Nonredundant dataset of high-resolution soluble protein structures

This nonredundant set of 1810 experimental X-ray structures was compiled using the PISCES server (Wang & Dunbrack, 2003) and contained only structures with a sequence identity less than or equal to 25%, resolution better than 1.8 Å and R-value less than .25.

Energy components

The energies, in this study, were derived from the analyses of high-resolution protein structures. Each energy component describes features observed in native structures of proteins and is expressed as a probability density function based on the statistical potential of mean force method (Sippl, 1990).

Atom pairing ($E_{pair}$)

The atom pairing score describes the energy from atom–atom interactions such as salt bridges, charge–charge, and polar interactions (Hendlich et al., 1990). For a pair of atoms of types $a_i$ and $a_j$, at a distance $d$ Å, the score is:

$$E_{pair} = -\log \frac{P(d|a_i, a_j)}{P(d)}$$

where $P(d|a_i, a_j)$ is the probability of $a_i$ and $a_j$ being $d$ Å apart and $P(d)$ is the probability of any atom types being $d$ Å apart.

Agreement with secondary structure prediction ($E_{SS}$)

The predicted probability of a residue $r$ adopting a secondary structure of type $ss$ is converted into a score:

$$E_{ss} = -\log P(ss|r)$$

The probability $P(ss|r)$ is provided by the sequence-based prediction methods PSIPRED (Jones, 1999) and ConPred II (Arai et al., 2004), and for each residue $r$, the probability that $r$ belongs to each secondary structure type ($P_{\text{helix}}, P_{\text{sheet}}$ and $P_{\text{ coil}}$) sums to unity.

Compactness ($E_{\text{compact}}$)

To capture the compactness feature of folded IMPs, we sum the interactions made by each atom ($cn_{sh}$), which is defined by the number of atoms within a radius shell of atom $i$ ($sh$).

$$E_{\text{compact}} = \sum_i n \cdot cn_{sh}$$

Steric repulsion ($E_{\text{steric}}$)

Steric repulsion models the repulsion of electron clouds between atoms:

$$E_{\text{steric}} = \sum_i \sum_j \left( \frac{r_{ij}^2 - d_{ij}^2}{r_{ij}} \right) ; \ d_{ij} < r_{ij}$$

where $i$ and $j$ are the indices of the atoms, $r_{ij}$ is the sum of the van der Waals radii of the $i$th and $j$th atoms, $d_{ij}$ is the distance between the two atoms.

Atom packing ($E_{\text{atompacking}}$)

In order to improve efficiency, three-dimensional structures are represented using a unified atom model allowing groups of side chain atoms to be approximated by a sphere. Despite the significant efficiency gains, the accuracy of some atom groups can be poorly represented by a sphere, such as a phenyl ring of phenylalanine. For such atom groups, a unified atom model can result in overcrowding and severe atom clashes when side chains are expanded back to full atomic representation. In order
to overcome this problem, an atom packing score is created based on a measure called the contact number (CN) (Karchin, Cline, & Karplus, 2004), which is the count of neighboring atoms within the radius shell of an atom. For a unified atom of type \( t \), contact number \( cn_{sh} \) in radius shell \( sh \) and fraction of area buried \( ba \), the atom density score is as follows:

\[
E_{\text{atomPacking}} = -\log \frac{P(cn_{sh}/ba|t, sh, ba)}{P(cn_{sh}/ba = 1|t, sh, ba)}
\]

where \( P(cn_{sh}/ba|t, sh, ba) \) is the probability of the ratio between the CN and the fraction of area buried. There are five radius shells of 0–6 Å, 6–7 Å, 7–8 Å, 8–10 Å, and 10–12 Å.

**Membrane environment (\( E_{\text{environment}} \))**

The interaction with lipid and water is modeled by an environment score. We extended the two-environment Reverse-Environment Prediction of Integral Membrane Protein Structure (REPIMPS) method (Dastmalchi, Morris, & Church, 2001), in which the contact environment is divided into polar and nonpolar environments, corresponding to the membrane model described in the following section. The contact score for an atom with type \( a \) with contact area \( A \) in contact with environment type \( env \) is given by:

\[
E^a_{env}(A) = -c^a_{env} A
\]

where \( c^a_{env} \) is the contact preference to environment type \( env \) given by:

\[
c^a_{env} = -\log \frac{P_{obs}(env|a)}{P_{exp}(env|a)}
\]

**Membrane burial propensity (\( E_{\text{membrane}} \))**

This scoring component describes the tendency for a particular residue type to be buried inside the membrane and is a variation on the study by Ulmschneider, Sansom, and Di Nola (2005). The propensities are determined from 32 unique helical IMP structures obtained from the OPM database (Lomize et al., 2006) (see Table S2), using only structures solved by X-ray crystallography and filtered with BLAST_CLUST (Altschul, Gish, Miller, Myers, & Lipman, 1990) with a sequence identity cutoff of 30% to reduce redundancy. For a residue of type \( r \) at a membrane depth of \( z \), the membrane burial propensity is

\[
E_{\text{membrane}} = -\log \frac{P(r|z)}{P(r|z = 0)}
\]

\( P(r|z) \) is the probability of residue type \( r \) being found at membrane depth \( z \) and \( P(r|z = 0) \) is the probability of \( r \) existing at \( z = 0 \). The \( z \) value is defined as the depth along the membrane normal and starts with 0 at the water and membrane interface boundary defined below.

**Gaussian smoothing**

Continuous distributions are discretized in the process of sampling, such as with the radius shells of \( E_{\text{atomPacking}} \). This leads to discretization errors that are minimized by Gaussian smoothing. For a sample with \( n \) observations, the probability of observing a certain value \( s \) is given by the following equation:

\[
P(s) = \frac{\sum_{i=1}^{n} G^a_{si}(s)}{n}
\]

where \( s_i \) is the value of the \( i \)th observation in the sample and \( G^a_{si}(s) \) is the Gaussian function centred at \( s_i \) with standard deviation \( \sigma \):

\[
G^a_{si}(s) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(s-s_i)^2/2\sigma^2}
\]

**Decoy sets**

A decoy structure was generated from the native structure of an IMP, with each phi and psi dihedral angle in the helical and loop regions randomly altered by ±10° and ±30°, respectively. Following this, 50,000 random backbone dihedral substitutions were performed and assessed with \( E_{\text{steric}} \) and \( E_{\text{compact}} \) to remove steric clashes, while retaining the compactness of the structure. In a subset of runs, all changes that decrease the RMSD from the native structure are retained, while changes that increase the RMSD are randomly rejected.

**Membrane layers**

The membrane was represented by two parallel planes and was further divided into five horizontal layers based on the study by White and Wimley (White, 1999). These were two aqueous layers; two membrane interface layers and the membrane core or hydrophobic layer. The environment of the aqueous and membrane interface layers was treated as polar, while the environment of the membrane core layer was treated as nonpolar. The membrane core layer was assumed to be 30 Å thick, and the interface layers were 15 Å each.

**Side chain modeling**

Side chains were modeled using Dead End Elimination (Canutescu, Shelenkov, & Dunbrack, 2003) and Monte Carlo simulation to search for the best rotamer
conformations, with side chain conformations published in the Dunbrack backbone-dependent rotamer library (Dunbrack & Cohen, 1997).

**Statistical evaluation**

An evaluation of the predictive abilities of the overall implementation is required, but this can only be objectively performed with a test set protein structures that have not been utilized in the training of the method. The overall approach is shown in Figure 1. Thus, both training set and testing sets of structures are assigned randomly and are independent. For the structures in this work, we have employed decoy sets to provide structures that give distributions of each of the energies, and by which the method’s performance can be reported by z-scores, allowing comparisons. A negative z-score reflects discriminating power in recognizing native-like structures and the lower the z-score, the better the discriminating power. The limited number of protein structures available requires us to choose native-like substructures in which the α-helices are substantially packed against each other, providing only 11 suitable independent structures composed of 3–5 helices and their adjoining loops. In a similar manner to the requisite independence and comparable size of the training and testing sets, the decoys sets were also derived by methods independent of the final prediction calculations.

**Results**

**Scoring functions**

To generate structural models, multicomponent scoring functions guide a Monte Carlo search of torsion angles to fold proteins toward the native state. To achieve this, the scoring functions assess structural features and assign a lower energy to those that are more similar to those found in native structures. These structural features are encapsulated in the individual probabilistic components, which are defined in the Materials and methods. These are the following:

1. *Atom pairing* ($E_{pair}$) to assess the likelihood that two atoms are found within a certain distance. This encompasses direct interactions such as salt bridges or polar interactions.
2. *Secondary structure (SS) prediction* ($E_{SS}$), integrates SS predictions from PSIPRED (Jones, 1999) for soluble regions and ConPred II (Arai et al., 2004) for the prediction of transmembrane helices.
3. *Compactness* ($E_{compact}$) to reflect the compact nature of proteins.
4. *Steric repulsion* ($E_{steric}$) to avoid clashes.
5. *Membrane burial propensity* ($E_{membrane}$) describing the tendency for a particular residue type to be closer to the periphery or the center of the membrane bilayer.
6. *Atom Packing* ($E_{atomPacking}$) reflects the propensity for different residue types to be in contact with a certain number of atoms.
7. *Membrane Environment* ($E_{environment}$) to determine the likelihood of a particular residue type to be exposed to an aqueous or a lipid environment.

All scoring components, except $E_{membrane}$, were derived from a nonredundant set of 1810 high-resolution crystal structures of soluble proteins. $E_{membrane}$, on the other hand, was derived from 32 unique helical IMP structures.

**Overall scoring functions**

Two scoring functions were constructed from different combinations of these scoring components. These were used sequentially to generate protein structures in two stages. The first was a coarse grain scoring function ($E_{coarseFold}$) that incorporates steric repulsion ($E_{steric}$), secondary structure and membrane topology predictions ($E_{SS}$) and residue packing ($E_{atomPacking}$) to guide assembly of secondary structure and establish the approximate relative position and orientation of the membrane helices, which could be performed quickly due to the simplicity of the function. This was defined as follows with weighting parameters $a$–$c$:

$$E_{coarseFold} = aE_{steric} + bE_{SS} + cE_{atomPacking}$$

(1)
The second scoring function \( E_{\text{refine}} \) was used to refine the coarse model generated using \( E_{\text{coarseFold}} \) and required the most accurate, and therefore, the most computationally expensive score to provide the final models. In addition to \( E_{\text{steric}} \) and \( E_{\text{atomPacking}} \), structure refinement incorporated statistics on interatomic distances \( (E_{\text{pair}}) \), residue affinity for hydrophobic or hydrophilic environments \( (E_{\text{environment}}) \) and the tendency for residues to be in the middle or periphery of the membrane bilayer \( (E_{\text{membrane}}) \). During refinement, secondary structure that was determined from \( E_{\text{coarseFold}} \) was maintained, and hence, \( E_{\text{SS}} \) was not used in \( E_{\text{refine}} \). \( E_{\text{refine}} \) is defined as follows with the weighting terms \( d–h \) that are optimized to maximize the predictive power:

\[
E_{\text{refine}} = dE_{\text{steric}} + eE_{\text{atomPacking}} + fE_{\text{pair}} + gE_{\text{environment}} + hE_{\text{membrane}}
\]  

(2)

Optimizing the scoring functions

The components and the overall scoring functions were optimized in two steps to maximize their applicability to IMP structures: First, each energy component was individually optimized. This involved optimizing the \( \sigma \) value of the Gaussian (see Materials and methods) to remove artificial local energy minima resulting from the discrete statistics, which are susceptible to outliers, and to reduce the energy landscape when it improves the efficiency of the conformational search. Inadequate smoothing leads to the introduction of sampling noise, whereas over-smoothing can remove genuine structural features. The statistics were transformed into a continuous function using Gaussian smoothing for the \( E_{\text{pair}}, E_{\text{membrane}} \) and \( E_{\text{atomPacking}} \) components (see Materials and Methods). Moreover, as statistics for the energy components were collected from soluble proteins, some components required re-parameterization to be applied to IMP structures. Second, ideal relative weights of the independently optimized components of \( E_{\text{coarseFold}} \) and \( E_{\text{refine}} \) were calculated to maximize their predictive power. These were achieved using large sets of \textit{in silico} generated helical IMP structure decoy sets (Gilis, 2004).

Decoy sets

For this study, we were interested in modeling IMP domains of 3–5 helices and compiled all such domains that did not contain bound cofactors and appeared to be structurally independent. As there were only 11 such domains available (Table 1), it remains unknown if these are sufficiently representative of all helical IMPs. Therefore, optimization on these structures alone is likely to bias the predictive power of the method to similar protein folds. To overcome this restriction, scoring functions were optimized and tested on sets of decoy protein structures produced \textit{in silico} that have the same primary sequence but different conformations to the native structure. For each of the 11 structures, 1100–1700 decoys were generated as described in the Materials and Methods. All decoys were within 10 Å RMSD of the corresponding crystal structure and were characterized as ‘native-like’ or ‘non-native’ if they were less or greater than 4 Å RMSD, respectively.

Optimal \( \sigma \) values, parameters, and weights were those that gave scoring components the best discriminative power between native-like and non-native decoys. These were determined using a hill-climbing search. For each energy component, the \( \sigma \) value and the optimizable parameters were each multiplied by a random number in the range (.9 and 1.1). The energy score was then re-derived with the new parameter values, and these were accepted if and only if the energy score had improved discriminative power. The relative weights of \( E_{\text{coarseFold}} \) and \( E_{\text{refine}} \) (‘d–c’ and ‘d–h’ in [1] and [2], respectively) were similarly optimized without re-deriving the energy score.

### Table 1. Helical IMP decoy subdomains.

| PDB ID* | Protein                        | Resolution (Å) | Number of helices | Length | Subdomain residues | Number of decoys |
|---------|--------------------------------|----------------|-------------------|--------|--------------------|------------------|
| 117v    | Bacterial ABC transporter      | 3.2            | 3                 | 97     | A84–A180           | 1624             |
| 1pw4    | Glycerol-3-phosphate transporter | 3.3           | 4                 | 119    | A27–A145           | 1702             |
| 1v55    | Cytochrome C Oxidase           | 1.9            | 4                 | 143    | C119–C261          | 1379             |
| 2bl2    | V-type ATPase                  | 2.1            | 4                 | 145    | A12–A156           | 979              |
| 2gif    | Multidrug efflux pump          | 2.9            | 4                 | 139    | A359–A497          | 1337             |
| 2ic8    | Protease gpLG                  | 2.1            | 5                 | 127    | A146–A272          | 1272             |
| 1j4n    | Aquaporin 1 water channel      | 2.4            | 4                 | 116    | A4–A119            | 1628             |
| 1py6    | Bacteriorhodopsin              | 1.8            | 4                 | 123    | A75–A197           | 1138             |
| 1zcd    | Sodium/proton antipporter NhaA | 3.4            | 3                 | 66     | A175–A240          | 1594             |
| 2b2f    | Ammonium transporter NhaA      | 1.7            | 4                 | 113    | A211–A323          | 1153             |
| 2gfp    | Multidrug transporter EmrD     | 3.5            | 4                 | 116    | A9–A124            | 1310             |

*Protein decoys are subdomains of the structure in the PDB, yet the PDB ID has been adopted as the identifying name.
components. In this study, discrimination power was quantified using z-scores:

\[ z\text{-score} = \frac{\text{mean}(\text{Score}_{\text{native-like}}) - \text{mean}(\text{Score}_{\text{all}})}{\text{stdDev}(\text{Score}_{\text{all}})} \]

where \( \text{mean}(\text{Score}_{\text{native-like}}) \) refers to the mean energy score of native-like decoy structures, \( \text{mean}(\text{Score}_{\text{all}}) \) and \( \text{stdDev}(\text{Score}_{\text{all}}) \) refer to the mean and standard deviation of the energy scores of all decoy structures, respectively. Thus, a negative z-score reflects discriminating power in recognizing native-like structures and the lower the z-score, the better the discriminating power.

Six decoy sets (training sets) were used for the optimizations, and the ideal \( \sigma \) values for the Gaussian smoothing and relative weights for the terms in \( E_{\text{coarseFold}} \) and \( E_{\text{refine}} \) were determined. The components \( E_{\text{compact}} \) and \( E_{\text{steric}} \) were left unchanged as the rules governing steric repulsion from Van der Waals interactions are unlikely to differ in the membrane environment. Similarly, \( E_{\text{aa}} \) was also unchanged as IMP information is incorporated in its use of the ConPred II algorithm. By reversing the exposed environment in the membrane bilayer from aqueous hydrophilic to hydrophobic in the case of soluble proteins, the statistics collected for \( E_{\text{environment}} \) from soluble protein structures was then directly applicable to IMP protein structures, consistently producing negative z-scores across all training sets with an average z-score of \(-.5\) and ranking the native structure on average within the top 16% of decoys (Table 2).

In contrast, \( E_{\text{pair}} \) and \( E_{\text{atomPacking}} \) required re-parameterization to be applicable to membrane protein prediction. We found that the maximum distance for which atom-atom pairing was useful in the modeling of IMP structures was 8.5 Å, whereas in soluble proteins, this parameter remains useful at distances greater than 20 Å (Melo, Sánchez, & Sali, 2002). This may suggest that atoms exhibit different long-range interactions in helical IMPs and soluble proteins. Furthermore, it was found that the best z-score was achieved when the score for each pairwise interaction was capped between \(-1.9\) and 2.0. This presumably prevents \( E_{\text{pair}} \) from being dominated by over-represented or rare interactions seen in soluble protein structures. Similarly, the maximum score allocated for \( E_{\text{atomPacking}} \) was set to 20.

After optimization, each individual scoring component favorably discriminated between the native-like and non-native decoys alone, each producing negative z-scores in all training sets with the exception of \( E_{\text{atomPacking}} \) and \( E_{\text{pair}} \) in one decoy set (Table 2). Furthermore, the individual components all ranked the native crystal structure among the most native-like of all the decoys, with \( E_{\text{atomPacking}} \) and \( E_{\text{pair}} \) identifying the crystal structure as the most accurate. Furthermore, \( E_{\text{refiner}} \), which combines several of these components and with optimal weighting (Equation (2)), outperformed any individual component and achieved an average z-score of \(-1.01\) for the six training sets (Table 2). This illustrates the predictive power of the individual optimized scoring components and that these are complementary with improved discriminatory power when utilized in the combination specified by the overall scoring functions (Equations (1) and (2)). Finally, to ensure that the scoring components were not biased to the training sets, the remaining five decoy sets were used to assess their discriminative power blindly. In all test sets, each scoring component alone and in combination unanimously produced negative z-scores (Table 2) and showed no significant differences between the variances and mean z-scores from the training and testing decoy sets.

| PDB ID | Type | Number | \( E_{\text{environment}} \) | Native rank | \( E_{\text{membrane}} \) | Native rank | \( E_{\text{atomPacking}} \) | Native rank | \( E_{\text{pair}} \) | Native rank | \( E_{\text{refiner}} \) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 117v | Train | 1642 | .63 | 51 | .65 | 40 | .05 | 16 | .25 | 16 | -2.8 | 6 |
| 1pw4 | Train | 1702 | -.59 | 334 | -.77 | 41 | -.06 | 1 | -1.22 | 1 | -1.39 | 1 |
| 1v55 | Train | 1379 | -.1 | 287 | -.52 | 135 | -.98 | 1 | -.49 | 1 | -.9 | 1 |
| 2bl2 | Train | 979 | -.79 | 75 | -.8 | 24 | -1.01 | 1 | -.58 | 1 | -1.1 | 1 |
| 2gif | Train | 1337 | -.46 | 328 | -.81 | 27 | -.76 | 1 | -.83 | 1 | -1.21 | 1 |
| 2ic8 | Train | 1272 | -.43 | 223 | -.8 | 116 | -.62 | 1 | -1.19 | 1 | -1.17 | 1 |
| 1j4n | Test | 1628 | -.25 | 53 | -.56 | 52 | -.44 | 3 | -.73 | 3 | -.82 | 1 |
| 1py6 | Test | 1138 | -.24 | 646 | -.53 | 195 | -.84 | 1 | -.36 | 1 | -.75 | 1 |
| 1zcd | Test | 1594 | .43 | 1 | -.73 | 1 | -.92 | 29 | -.97 | 29 | -1.26 | 1 |
| 2h2f | Test | 1153 | -.33 | 429 | -.65 | 5 | -.66 | 1 | -.78 | 1 | -.78 | 6 |
| 2gfp | Test | 1310 | -.02 | 430 | -.11 | 481 | -.28 | 38 | -.4 | 38 | -.47 | 20 |

Table 2. The z-scores and rankings of the scoring function components for the decoy sets.
**Model building**

Proteins were folded using a Monte Carlo simulation that randomly altered backbone phi or psi dihedral angles according to the operator scheme (see Supplementary material). Each search is guided by either $E_{\text{coarseFold}}$ or $E_{\text{refine}}$ to assess whether a round of simulation lowered the energy of the structure, in which case the change was accepted. If a round of simulation did not improve the structure, the change was randomly either accepted or rejected. This provided the means to escape from local energy minima, while still biasing the search towards the native-like structures.

Side chains are included from the initial stages but are represented using a simplified model based on work by Herzyk and Hubbard (1993) in which side chains are represented by up to three spheres known as unified atoms. This greatly improves the efficiency of protein folding while still ensuring that sufficient room is left for assignment of side chain rotamers, which occurs in the final stage of the model generation.

The models were produced in two stages as shown in Figure 2. The first stage entailed 50,000 rounds of Monte Carlo simulations that were guided by $E_{\text{coarseFold}}$ to assign secondary structure and the relative position of transmembrane helices. A restraint was randomly imposed on a subset of rounds such that changes were localized and did not alter the overall structure of the protein. During the first 5000 rounds, only the $E_{\text{steric}}$ and $E_{\text{SS}}$ components of $E_{\text{coarseFold}}$ are used to generate the secondary structure with the ‘c’ weight set to zero (Equation (1)) so that there is no influence from $E_{\text{atomPacking}}$. In the next 15,000 rounds, the weight of $E_{\text{atomPacking}}$ is linearly increased from zero to full weight. During the remainder of the simulation, the weights of the scoring function are unchanged. This search is repeated until 1000 initial models are collected and the 60 models with the lowest scores according to $E_{\text{refine}}$ (Equation (2)) are selected as the seed population for the second stage of model building.

In the second stage of model building, a model is randomly selected from the seed population for refinement. The model is first altered by small random movement of the helices and is then subject to 30,000 rounds of Monte Carlo simulation that are guided by $E_{\text{refine}}$. If the score of the refined model is lower than the highest scoring model in the seed population, the new model replaces the worst scoring model in the population. In total, 1500 model refinement runs are performed, producing 1500 models. The seed population size is kept constant at 60, but the scores of the models inside the population are gradually lowered. The stages of the method provide for a method generating a set of structures of a suitable population from which to select and refine good models with acceptable energies and structures.

To correctly classify the local environment of exposed residues as aqueous, interfacial or hydrophobic, according to the membrane layers defined in the Materials and Methods, the location of the membrane is estimated after every round of refinement as follows. First, the center of the membrane was initially placed at the center of the protein, which was defined as the average coordinates of the $C^\beta$ of all residues. The membrane normal was the average of the axes of all helices, and the two parallel planes representing the membrane were fitted 15 Å either side of the protein center along the membrane normal. The initial estimate of the location was refined using 20,000 rounds of Monte Carlo optimization in which the angle of the membrane normal was varied.
as well as the position of the membrane center to minimize $E_{\text{environment}}$.

The final set of models was selected via clustering as the structures that were most highly represented (Shortle, Simons, & Baker, 1998) from the ensemble of 1500 generated models. Any cluster containing more than 50 models was considered as an over-represented fold, and the model with the lowest average RMSD to any other structure within the cluster, the center model, was reported as a final model.

**Prediction of helical IMPs from sequence**

The helical IMP prediction approach successfully modeled the structures of five helical IMP subdomains (Table 1), resulting in models with RMSDs of 5.0–8.8 Å from the native structures (Figure 3).

**4-helix subdomain of aquaporin-1 water channel AQP1 (1j4n)**

This subdomain contains three long transmembrane helices and another short helix (helix-3) that only spans half of the membrane (Figure 3(a), column ii). The lowest RMSD model created has an RMSD of 5.8 Å, the top cluster contains 245 models, and the center model has an RMSD of 6.9 Å (Figure 3(a), columns i and iii). Despite the relatively high RMSD, the top cluster center exhibits many of the native features. All helices are packed in the correct order, including helix-3, which is also modeled as a shorter helix that spans only half the membrane. However, the tilt angles of the helices, the deviation of axis relative to the membrane normal, deviate from the native structure. This is most noticeable on the C-terminal section of helix-4, which is curved in the model. The short helix-3 also deviates from the native structure by more than 8 Å for most residues.

**4-helix subdomain of bacteriorhodopsin (1py6)**

Simulations of the 4-helix subdomain of bacteriorhodopsin produced models with an RMSD of 4.3 Å from the native structure (Figure 3(b), ii). Only one cluster was formed containing 560 models with a center model that had a 5.0 Å RMSD from the native structure (Figure 3(b), iii). In this model, the relative position of the helices was correctly predicted. In particular, helix-3 and helix-4 protrude further along the z-axis in the native structure, which is correctly modeled. However, the tilt angles of the helices in the model differ to the native structure and are smaller in the model.

**3-helix subdomain of Na+/H+ antiporter NhaA (1zcd)**

Three clusters were found for the three-helix subdomain of the Na+/H+ antiporter NhaA, each with the helices in different relative positions (Figure 3(c)). The cluster with the correct arrangement of helices had a center model with the lowest RMSD of 4.7 Å from the native structure but contained the fewest models. The most accurate model had an RMSD of 3.8 Å from the native structure.

**4-helix subdomain of ammonium transporter Amt-1 (2b2f)**

The 4-helix subdomain of the ammonium transporter Amt-1 contains four helices that are tilted at approximately 30° to the membrane normal. Two clusters were found: the largest cluster contains 239 models, with RMSD of 9.9 Å; the second cluster contains 74 models, with RMSD of 7.8 Å (Figure 3(d)). The tilt angles of the helices were again smaller than the native. The center models from both clusters had their helices roughly parallel to the membrane normal. In the first cluster model, the positions of helix-3 and helix-4 are swapped. The second cluster model has the four helices arranged in the correct order. The lowest RMSD model was 5.2 Å from the native structure.

**4-helix subdomain of multidrug transporter EmrD (2gfp)**

The 4-helix subdomain of Multidrug Transporter EmrD (2gfp) has a relatively complex topology. Helix-4 is 'wedged' between the other three helices, unlike the simple clockwise or anticlockwise arrangements of the previous structures. The lowest RMSD model generated had an RMSD value of 6.5 Å, and the first cluster contained 329 models with RMSD value of 12 Å (Figure 3(e)). On closer inspection, it is observed that helix-1 contains several hydrophilic residues (Gln 21, Gln 24, and Thr 25), which would restrain the simulation from modeling of these residues into the membrane. Also the N-terminal side of helix-1 contains small hydrophobic residues with no charged residues or tyrosines or tryptophans to assist in anchoring of the helix into membrane. It appears that the hydrophilicity of helix-1, coupled with the lack of membrane anchoring residues, is the reason half of helix-1 is modeled protruding away from the membrane. Further, the native sequence of EmrD contains 8 extra residues (MKRQRNVN) in the N-terminal region that are missing in the X-ray crystal structure, which contains two arginines and a lysine. We were able to produce a more accurate model when these eight residues were included in another simulation run, producing a cluster with 266 models with the center model being an RMSD of 8.8 Å. Even though the model RMSD did not improve significantly over the model without the 8 N-terminal residues, most of helix-1 was correctly positioned inside the membrane.
Figure 3. Structure prediction of test set, subdomains of: (a) 1j4n; (b) 1py6; (c) 1zcd; (d) 2b2f and (e) 2gfp. Except for (e), column (i) displays the predicted structure with the lowest RMSD; (ii) the native crystal structure of the subdomain; (iii) the center model from the largest populated cluster, except in the case of 2gfp; (iv) the center model from the second largest populated cluster; (v) the center model from the third largest populated cluster. For (c), (iii) displays the predicted subdomain structure upon the inclusion of the MKRQRNVN sequence (orange colored region) at the N terminus. The displayed Ångström values refer to the RMSD values relative to the respective native crystal structures. Starting from the N-terminus, the helices are colored blue, green, yellow, and red.
Discussion

We have shown that significant structural information can be obtained from the wealth of soluble protein data and adapted to IMP protein modeling by implementing a comprehensive energy function using decoy training. Some physical principles, such as steric repulsion, were presumed to be unchanged between the soluble membrane environment. Furthermore, we have demonstrated that by reversing the exposed environment within the bilayer to be hydrophobic, statistics describing the local environment of residues in soluble structure can be directly applied to IMP structures. This is consistent with the REPIMPS method of validating the structures of IMPs (Dastmalchi et al., 2001). In contrast, the requirement to re-parameterize other scoring components suggests that there may be some intrinsic differences to long-range electrostatics and the nature of atomic interactions within the membrane compared with the aqueous environment.

We have demonstrated that this method produces accurate models, but several refinements can be made. In most cases, the method produced and selected models with helices correctly inserted into the membrane and arranged in the correct sequence and with clearly defined membrane boundaries. In future work, we will address the issue that the helices of the predicted models are not sufficiently tilted compared to the native structures. Transmembrane helices cross the membrane and hence the thickness of the membrane could restrict their tilt angles, despite there being no specific scoring function for this property. It is therefore likely that by allowing localized variation in membrane thickness the helices could more easily adopt the correct tilt angle while remaining in their correct environment. Further, the largest clusters were not necessarily the cluster with the correct helix arrangement. This may be due to the relatively uniform properties of transmembrane helices, which are primarily hydrophobic and packed together by van der Waals attraction. The identification of the correct helix arrangement may be aided by a specific score that recognizes surface complementarity. Finally, our prediction results can be improved by incorporating algorithms that correctly model loop regions, as the current restraints imposed by our scoring functions are likely to be satisfiable by multiple loop conformations.

We have developed a complete knowledge-based modeling strategy to relieve the scarcity of structural information on helical IMPs (Lomize et al., 2006). Data from the extensive library of soluble proteins were utilized to generate unbiased models of five helical subdomains. Further, the elucidation of more experimentally determined transmembrane protein structures and the availability of more powerful computing resources will continue to increase the potential of this method to generate more accurate models. As more structures are experimentally determined and if the current general principles of membrane protein structures hold, we anticipate our method will also improve, in part due to the improved likelihood of training and testing on entirely native protein structures. Although there were limitations imposed by the dataset, under the circumstances the overall approach was well founded, and the evaluation was completely systematic. It is envisaged that a future application of our modeling method would be to aid in the independent interpretation and planning of techniques such as fluorescence spectroscopy, electron microscopy or small angle X-ray scattering on challenging membrane proteins.

Acknowledgments

This study was supported by Australian Postgraduate Awards from the Australian Government to HWL, LKL, and ETT and a University of Sydney Postgraduate Award to HCL. We thank Vimal Kapoor, Phillip Crawford, and Noeris Salam for many helpful discussions. The award of scholarships to HWL, LKL, ETT, and HCL are gratefully acknowledged.

Supplementary material

The supplementary material for this paper is available online at http://dx.doi.10.1080/07391102.2013.765808.

References

Almén, M. S., Nordström, K. J., Fredriksson, R., & Schiöth, H. B. (2009). Mapping the human membrane proteome: A majority of the human membrane proteins can be classified according to function and evolutionary origin. BMC Biology, 7, 50.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215, 403–410.

Arai, M., Mitsuke, H., Ikeda, M., Xia, J. X., Kikuchi, T., Satake, M., & Shimizu, T. (2004). ConPred II: A consensus prediction method for obtaining transmembrane topology models with high reliability. Nucleic Acids Research, 32, W390–W393.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Research, 28, 235–242.

Canutescu, A. A., Shelenkov, A. A., & Dunbrack, R. L. Jr (2003). A graph-theory algorithm for rapid protein side-chain prediction. Protein Science, 12, 2001–2014.

Dastmalchi, S., Morris, M. B., & Church, W. B. (2001). Modeling of the structural features of integral-membrane proteins reverse-environment prediction of integral membrane protein structure (REPIMPS). Protein Science, 10, 1529–1538.
Dunbrack, R. L. Jr, & Cohen, F. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science, 6*, 1661–1681.

Gilis, D. (2004). Protein decoy sets for evaluating energy functions. *Journal of Biomolecular Structure and Dynamics, 21*, 725–735.

Gough, J., Karplus, K., Hughey, R., & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology, 313*, 903–919.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., … Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models. *Journal of Molecular Biology, 216*, 167–180.

Herzyk, P., & Hubbard, R. E. (1993). A reduced representation of proteins for use in restraint satisfaction calculations. *Proteins: Structure, Function, and Genetics, 17*, 310–324.

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

Karchin, R., Cline, M., & Karplus, K. (2004). Evaluation of local structure alphabets based on residue burial. *Proteins: Structure, Function, and Bioinformatics, 55*, 508–518.

Lees-Miller, J. P., Subbotina, J. O., Guo, J., Yarov-Yarovoy, V., Noskov, S. Y., & Duff, H. J. (2009). Interactions of H562 in the S5 helix with T618 and S621 in the pore helix are important determinants of hERG1 potassium channel structure and function. *Biophysical Journal, 96*, 3600–3610.

Lomize, M. A., Lomize, A. L., Pogozheva, I. D., & Mosberg, H. I. (2006). OPM: Orientations of proteins in membranes database. *Bioinformatics, 22*, 623–625.

Lundstrom, K. (2004). Structural genomics on membrane proteins: The MePNet approach. *Current Opinion in Drug Discovery & Development, 7*, 342–346.

Melo, F., Sánchez, R., & Sali, A. (2002). Statistical potentials for fold assignment. *Protein Science, 11*, 430–448.

Pathak, M. M., Yarov-Yarovoy, V., Agarwal, G., Roux, B., Barth, P., Kohout, S., … Isacoff, E. Y. (2007). Closing in on the resting state of the shaker K+ channel. *Neuron, 56*, 124–140.

Pellegrini-Calace, M., Carotti, A., & Jones, D. T. (2003). Folding in lipid membranes (FILM): A novel method for the prediction of small membrane protein 3D structures. *Proteins: Structure, Function, and Bioinformatics, 50*, 537–545.

Shortle, D., Simons, K. T., & Baker, D. (1998). Clustering of low-energy conformations near the native structures of small proteins. *Proceedings of the National Academy of Sciences, 95*, 11158–11162.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mena force. *Journal of Molecular Biology, 213*, 859–883.

Ulmschneider, M. B., Sansom, M. S., & Di Nola, A. (2005). Properties of integral membrane protein structures: Derivation of an implicit membrane potential. *Proteins: Structure, Function, and Bioinformatics, 59*, 252–265.

von Heijne, G. (2007). The membrane protein universe: What’s out there and why bother? *Journal of Internal Medicine, 261*, 543–557.

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

White, S. H. & Wimley, W. C. (1999). Membrane protein folding and stability: Physical principles. *Annual Review of Biophysics and Biomolecular Structure, 28*, 319–365.

http://blanco.biomol.uci.edu/mpstruc/listAll/list

Yarov-Yarovoy, V., Baker, D., & Catterall, W. A. (2006). Voltage sensor conformations in the open and closed states in ROSETTA structural models of K+ channels. *Proceedings of the National Academy of Sciences, 103*, 7292–7297.

Yarov-Yarovoy, V., Schonbrun, J., & Baker, D. (2006). Multi-pass membrane protein structure prediction using Rosetta. *Proteins, 62*, 1010–1025.