BIOINFORMATICS

Vol. 27 ISMB 2011, pages i15–i23
doi:10.1093/bioinformatics/btr230

Environment specific substitution tables improve membrane protein alignment

Jamie R. Hill[1], Sebastian Kelm[1], Jiye Shi[2,3] and Charlotte M. Deane[1,*]

[1]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, [2]UCB Celltech, Branch of UCB Pharma S.A., 208 Bath Road, Slough SL1 3WE, UK and [3]Department of Biochemistry, School of Life Sciences, Shanghai University, Shanghai 200444, P.R. China

ABSTRACT

Motivation: Membrane proteins are both abundant and important in cells, but the small number of solved structures restricts our understanding of them. Here we consider whether membrane proteins undergo different substitutions from their soluble counterparts and whether these can be used to improve membrane protein alignments, and therefore improve prediction of their structure.

Results: We construct substitution tables for different environments within membrane proteins. As data is scarce, we develop a general metric to assess the quality of these asymmetric tables. Membrane proteins show markedly different substitution preferences from soluble proteins. For example, substitution preferences in lipid tail-contacting parts of membrane proteins are found to be distinct from all environments in soluble proteins, including buried residues. A principal component analysis of the tables identifies the greatest variation in substitution preferences to be due to changes in hydrophobicity; the second largest variation relates to secondary structure. We demonstrate the use of our tables in pairwise sequence-to-structure alignments (also known as 'threading') of membrane proteins using the FUGUE alignment program. On average, in the 10–25% sequence identity range, alignments are improved by 28 correctly aligned residues compared with alignments made using FUGUE's default substitution tables. Our alignments also lead to improved structural models.

Availability: Substitution tables are available at: http://www.stats.ox.ac.uk/proteins/resources.

Contact: deane@stats.ox.ac.uk

1 INTRODUCTION

Membrane proteins constitute ∼30% of human proteins (Almén *et al.*, 2009), and are important drug targets. Unfortunately, the structures of these proteins are hard to determine. As of January 2011, there are ∼70000 structures in the Protein Data Bank (PDB) (Berman *et al.*, 2000), of which fewer than 1500 are recognized as membrane proteins by the PDB_TM database (Tusnády *et al.*, 2005). Even these available structures are highly redundant. Thus, for most membrane proteins only a sequence is known: structural information must be inferred by modelling.

Structure modelling can broadly be divided into template-free and template based methods (Moult *et al.*, 2009). Template based modelling makes use of 'homologous' proteins that are identified as having similar structures to that of the target sequence. Proteins are normally considered homologous if they share significant sequence similarity.

Once a homologous protein with a known structure (the 'template') has been identified, it is aligned to the target sequence. This alignment between two sequences can be improved by the use of structural information from the template. This sequence-to-structure alignment forms the blueprint which coordinate generation programs such as MODELLER for soluble proteins (Sali, 1993), or MEDELLER for membrane proteins (Kelm *et al.*, 2010) use to build a model. The accuracy of a model is primarily determined by the quality of the initial alignment (Sánchez, 1997).

The membrane is a radically different environment from the aqueous environment of soluble proteins. Most membrane bilayers are composed of phospholipids with hydrophobic tail groups and charged head groups. This suggests for example that substitutions from charged residues to hydrophobic ones are unlikely in head-contacting regions, and conversely that substitutions from hydrophobic residues to charged ones are unlikely in tail-contacting regions.

Thus, membrane proteins will have unique patterns of substitutions (Mokrab *et al.*, 2010). However, due to lack of data, the alignment of membrane proteins is typically performed with substitution tables optimized for soluble proteins. Identifying appropriate tables for membrane environments is expected to improve methods that depend on them, particularly for sequence-to-structure alignment (Mokrab and Mizuguchi, 2005).

Environment-specific substitution tables (ESSTs) are one of the methods used to align target sequences with known structures. A substitution table tabulates the chances that an amino-acid in one protein is replaced by another in a second protein: for example that a glycine residue is replaced by an alanine. ESSTs are a set of substitution tables, each of which is to be used in a different environment. Environments are defined by features such as secondary structure and accessibility. ESSTs are used in sequence-to-structure alignment as the environments for each residue can be determined from the structure.

To create an ESST, substitutions in each environment must be counted between a number of related proteins. In early studies tables were constructed by counting substitutions between homologous proteins of known structure (Shi *et al.*, 2001). More recently, tables have been made by counting substitutions between one structure and many sequences (Mizuguchi *et al.*, 2007). The latter method is used here.

Environment independent substitution tables have previously been made for transmembrane regions. The JTT table (Jones *et al.*, 1994) appears to be the earliest example, with the PHAT

*To whom correspondence should be addressed.

(Ng *et al.*, 2000) and SLIM (Müller *et al.*, 2001) tables following. These tables were intended to be used in conjunction with a non-membrane table, such as BLOSUM62 (Henikoff and Henikoff, 1992). This 'bipartite' scheme requires a separate algorithm to decide where to use each table.

Sequence alignment has been attempted using PHAT, with both bad (Forrest *et al.*, 2006), and good (Pirovano *et al.*, 2008) results when compared with alignments using only BLOSUM62. The SLIM table is optimized for homology detection: its authors explicitly caution against using it for alignment.

Two problems complicate the construction of ESSTs for membrane proteins:

Firstly, it is difficult to determine the structural environment of residues even in known membrane structures—although secondary structure and accessible surface area can be determined as for soluble proteins, the location of a residue within the membrane bilayer cannot be inferred from the solved structure alone. Here we use the annotation program iMembrane (Kelm *et al.*, 2009) to determine these contacts.

Secondly, once an ESST is created, it is difficult to assess if it is representative of its environment across all membrane proteins. How can we tell if we have made a 'good' table? We describe a metric of ESST quality that is robust against perturbations in the observed frequencies of individual substitutions.

We create 'good' membrane ESSTs and analogous soluble ESSTs and make global comparisons between them. A dendrogram illustrates inter-table distances, and a principal component analysis is used to detect the dependence of substitution patterns on environment type. Membrane environments are found to be far more diverse than soluble protein environments. For example, any pair of lipid tail layer environments are more dissimilar than the corresponding pair of soluble environments.

FUGUE (Shi *et al.*, 2001) is a commonly used program to produce sequence-to-structure alignments with ESSTs. We compare several methods for sequence-to-structure alignment generation: default FUGUE, our membrane-based FUGUE, a bipartite PHAT/BLOSUM62 scheme, and the sequence-to-sequence alignment program MUSCLE (Edgar, 2004). Our membrane-specific ESSTs consistently improve alignment quality, especially at low sequence identity. In the 10–25% sequence identity range, compared with the next best method (the default FUGUE tables) 54 alignments are improved by at least 10 residues, whereas only 6 alignments are worsened by the same amount. In this range the average improvement per alignment is 28 more residues aligned correctly. These alignment improvements are found to lead to corresponding improvements in structure prediction.

2 METHODS

2.1 Environment descriptors

Secondary structure and accessible surface area are annotated using JOY (Mizuguchi *et al.*, 1998b). We use 'a' to label inaccessible residues and 'A' to label accessible ones. By convention, a residue is deemed accessible when more than 7% of its surface area is exposed (Hubbard and Blundell, 1987). The secondary structure types used are helix (H), β-strand (E), +ve ϕ angle (P), and coil (C).

In the case of membrane proteins, environments can also be defined depending on the part of the membrane a residue is contacting. iMembrane uses coarse-grained molecular dynamics simulation data from the CGDB database (Scott *et al.*, 2008) to annotate three contact environments. Residues that are in contact with the membrane for $< 10\%$ of the time are in the 'n' environment; of the remaining residues, those that spend more time in contact with the lipid heads are in the 'h' environment, and those that spend more time in contact with the lipid tails are in the 't' environment.

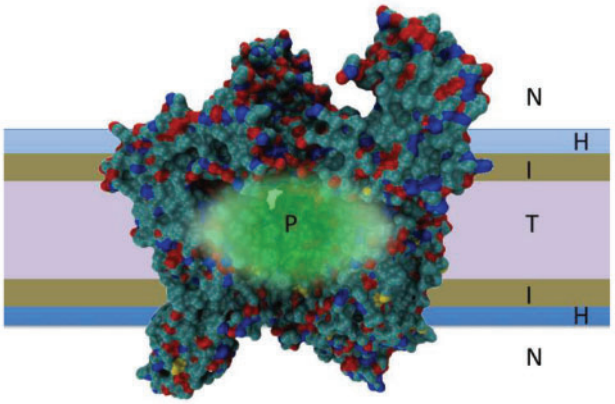

Fig. 1. A schematic slice through a membrane protein (1YEW) in the membrane indicating the layer types used. 'N' is the region outside the membrane, 'T' and 'H' span the tail and head groups of the membrane lipids respectively, 'P' is the area lining the pore, and 'I' is the interface region between the tail and head groups.

By taking a consensus of these contact annotations, a membrane protein can be divided into layers corresponding to the lipid heads (H), lipid tails (T), and not-in-membrane (N) regions. There are thus $72 = 2 \times 4 \times 3 \times 3$ distinct possible environments.

Larger numbers of environments lead to more specific substitution tables, at the cost of each table being constructed with less data. It is desirable to combine environments so as to find a minimal set that encompasses as much variation in substitution patterns as possible.

There are many possible valid minimal environment sets. Here we ignore the contact annotation, with two exceptions. Accessible residues that lie in the tail layer but rarely contact the membrane can be identified as residues that line a pore (P). Accessible residues that are annotated with head contacts but are in the tail layer, or with tail contacts but are in the head layer, define an interface region (I) spanning the hydrophilic and hydrophobic parts of the membrane. We add these to the existing layer types to give five labels: H(ead), N(ot in membrane), T(ail), P(ore) and I(nterface) regions (Figure 1).

It is convenient to refer to environments using a letter code e.g. 'IEA' = interface layer, β-strand, accessible residues. Letter codes will always be built in the order Layer:Secondary Structure:Accessibility. An asterisk '*' will be used when the exact letter does not matter. Under this system 'I**' refers to all interface layer environments, whereas 'I*A' refers to accessible interface layer environments.

2.2 Alignments for table generation

Transmembrane protein structures were identified from the PDB_TM database (Tusnády *et al.*, 2005) and downloaded from the PDB (Berman *et al.*, 2000). Each was then split into its component protein chains. Redundant chains—those with $> 80\%$ sequence similarity—were removed using Cd-hit (Li and Godzik, 2006). Chains without iMembrane search hits were also removed, leaving 328 chains.

For each chain, related sequences were obtained from 5 iterations of PSI-BLAST (Altschul *et al.*, 1997) using an *E*-value threshold of 1×10^{-3} for keeping a hit, and a threshold of 1×10^{-5} for including a hit in the sequence profile of the next iteration. PSI-BLAST searches were made against the NCBI nr database. These sequences were then aligned to their corresponding

Table 1. A glossary of membrane protein environments

First letter (layer)	Second letter (secondary structure)	Third letter (accessibility)
H – lipid head	H – helix	A – accessible
N – not in membrane	E – beta strand	a – inaccessible
T – lipid tail	C – coil	
P – pore-lining	P – +ve ϕ	
I – interface region		

All combinations of letters are possible. The exception being that Pore-lining and Interface-region layers cannot be inaccessible as the definition of these layers requires them to have contacts with either the membrane or solvent. Positive ϕ environments will later be merged into just two environments labelled 'NPA' and 'NPa' (see Section 3.2).

structures with MUSCLE (Edgar, 2004), and the alignments used to generate the membrane substitution tables.

Soluble tables were generated from four different alignment sets. The first of these was generated as above—that is, by aligning multiple homologous sequences with each structure. The structures were obtained by taking the first structure from each family in the HOMSTRAD database (Mizuguchi *et al.*, 1998a), and the sequences were found by searching the nr database. After filtering, this yielded 423 soluble chains which were used to produce our standard soluble tables. The other three alignment sets (SUB177, SUB371 and HOMSTRAD) are structure-to-structure alignments used only to validate our standard soluble tables.

SUB177 and SUB371 are described in the original FUGUE paper (Shi *et al.*, 2001). SUB177 is a set of 177 protein families comprising 706 structures used to build the default tables of FUGUE. SUB371 is a set of 371 protein families comprising 1357 structures used to test the stability of the SUB177-derived tables. The HOMSTRAD set comprises 1032 families and more than 3000 structures.

2.3 Table construction

Membrane ESSTs are constructed as follows. The number of times a particular substitution is observed in an environment is tabulated in an environment specific counts matrix A^E (where E labels the environment). Environments are determined by the annotations from iMembrane and JOY. For each structure in our set of 328 membrane protein alignments, every time a structure residue 'a' in environment E has a corresponding residue 'b' in one of the aligned sequences, the matrix element A^E_{ba} is increased by unity. The entries of the ESST S^E are obtained from the following formula:

$$S^E_{ba} = \frac{3}{log(2)} log\left(\frac{A^E_{ba} / \sum_b A^E_{ba}}{\sum_{a,E} A^E_{ba} / \sum_{a,b,E} A^E_{ba}} \right) \quad (1)$$

Given that the structure has a residue a, the numerator of the logarithm is the probability of a substitution $a \rightarrow b$ in the matched sequence. The denominator is the probability that any substitution in any environment will go to b rather than another residue. The prefactors (and the taking of the logarithm itself) are a standard rescaling. ESSTs are generally asymmetric ($S^E_{ba} \neq S^E_{ab}$), and are rounded to the nearest integer.

The program JSUBST (a java derivative of SUBST available at http://www.stats.ox.ac.uk/proteins/resources) was used to construct the counts matrices A^E. Counts were made between clusters of similar sequences (60% sequence identity) and the cluster containing the structure. Each cluster was weighted by the number of sequences it contained as described in Shi *et al.* (2001). Substitutions to and from gaps were not counted, but all columns in the alignments were included when constructing the matrices. A constant of 1/100 of a count was added to each entry A^E_{ba} to prevent S^E_{ba} evaluating to $-\infty$ in rare cases. All sequences in the same cluster as the structure were annotated with its structural annotation for the purposes of matrix construction. Soluble tables were built in an analogous manner for each of the four sets of soluble alignments.

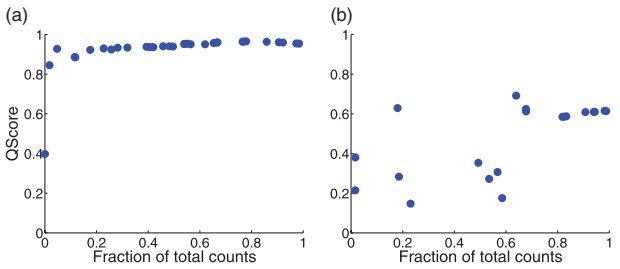

Fig. 2. A high-quality table (IHA, **a**) and low-quality table (TPa, **b**). Each point is the fraction of total counts and consistency of a table when constructed with 20 more alignments than the preceding point. Some points are superimposed.

2.4 Identifying consistent tables

How can we identify substitution tables that are unrepresentative of their environments? A crude method is to label as unrepresentative all those tables with fewer than a minimum number of counts. However, this method can run into problems—a rare environment might be extremely consistent in the substitutions it allows, such that the number of counts is small, but the data is representative.

Here we use a combination of a count threshold and a 'self-consistency' score. The latter is obtained as follows. By normalizing the columns of a counts matrix A^E_{ba}, we can interpret each entry as the probability that $a \rightarrow b$ in environment E.

When a vector of amino acid counts is multiplied by this matrix, it changes according to the mutation probabilities encoded in the matrix. After a large number of rounds of mutation (matrix multiplications), the resulting vector of amino acid counts is invariant under mutation. Mathematically, this vector is the eigenvector of the matrix with eigenvalue +1.

It is assumed that the distribution of amino acids in a given environment, averaged over all proteins, is stable over time. Thus, a representative table should have a limiting distribution of amino acids that is close to the distribution observed in the alignments used to construct it.

The self-consistency score, 'Q' is calculated according to Equation 2:

$$Q = 1 - \frac{1}{2} \sum_{i=1}^{20} |v_i - w_i| \qquad w_i = \sum_a A^E_{ia} \quad (2)$$

where v is the eigenvector of the probability matrix with eigenvalue +1, and w is a normalized vector of the observed amino acid frequencies, which can be estimated as shown. This has the desirable property of taking values between 0 (totally inconsistent) and 1 (identical).

A simple interpretation of this score exists. It is the maximum fraction of residues that could remain the same if substitutions occurred according to the probabilities encoded in the counts matrix over many iterations.

The self-consistency score is scale-invariant, so it provides a measure of table quality that is independent of the number of counts. Figure 2 shows a useful scheme for visually identifying poor tables. The fraction of the total number of counts and Q are plotted for each table with increasingly large subsets of the data. A stable counts matrix should tend to a stable level of Q as more data is included.

2.5 Table analysis and visualization

The relative similarity of tables was visualized in two ways. Firstly a dendrogram was constructed based on the Euclidean distance between ESSTs. The dendrogram was built using single linkage clustering—meaning that new branches join existing clades based on the smallest distance between

a member of the clade and the new branch. This linkage has the advantage that the dendrogram does not change under a rescaling of the data.

Secondly, following the example of Gong *et al.* (2009), a principal component analysis (Hotelling, 1933) in multi-dimensional 'substitution space' was performed. This selects a set of 2 or 3 orthogonal axes that explain the greatest amount of variation in the data, and thus projects substitution space down into 2D or 3D with minimal distortion.

2.6 Sequence-to-structure alignment

To test sequence-to-structure alignment, we take two homologous proteins of known structure and align the sequence of one (the target) to the structure of the other (the template). The alignments were made using FUGUE with the default tables, the PHAT/BLOSUM62 tables, and our membrane tables. The annotations from iMembrane (Kelm *et al.*, 2009) and JOY (Mizuguchi *et al.*, 1998b) determined where each table was to be applied. Pairwise alignments were also made using the sequence-to-sequence alignment program MUSCLE. The quality of these alignments was assessed against the implicit sequence alignments generated by the structure-to-structure alignment program TM-align (Zhang and Skolnick, 2005).

The MEDELLER test-set (Kelm *et al.*, 2010) consists of pairs of homologous membrane proteins of known structure. We use one element of each pair as the target sequence, and the other as the template structure. We filtered the set such that no two templates, and no two target sequences, had more than 80% sequence identity. This left 408 pairs of proteins ranging from 0 to 100% identity, with a median sequence identity of 14%.

The alignment of each template residue in the structure-to-structure alignment produced by TM-align was compared with the alignment of the same residue produced by one of the methods. A schematic of this procedure is given below. In this example, 9 residues are correctly aligned over a total alignment length of 10 residues.

```
TM-align
  Template Structure   --AGGA-CGPAA ...
    Target Structure   AAAGGAFCA-AL ...

Tested method
  Template Structure   --AGGA--CGPAA ...
     Target Sequence   AAAGGAFC-A-AL ...
            Correct?   --YYYYYNNYYYY
```

3 RESULTS

3.1 Validation of substitution tables

The tables used in FUGUE were obtained by counting substitutions between homologous structures. Due to the scarcity of membrane protein structures, we count substitutions between a structure and related sequences, following a similar method to that of Mizuguchi *et al.* (2007). To assess the validity of this procedure we compared eight soluble ESSTs generated by this method with those derived from the SUB177, SUB371 and HOMSTRAD structure sets (Table 2). From here onwards, soluble environments are labelled by a leading 's'.

Of the eight tables, larger differences are seen in the sEa (soluble, β-strand, inaccessible), sPa (soluble, +ve ϕ, inaccessible), and sPA (soluble, +ve ϕ, accessible) environments due to the greater number of rare substitutions in these environments. As the scores are logarithmic, small variations in the number of rare substitutions lead to disproportionately large effects on their log-odds scores.

The small number of differences between the structure/structure and structure/sequence derived tables, particularly when larger numbers of structures are used, suggests that structure/sequence derived tables are representative of substitution preferences. Below, only the structure/sequence derived tables are compared.

Table 2. The number of differences between soluble structure/sequence derived tables and their structure/structure derived counterparts

Dataset	sCa	sCA	sEa	sEA	sHa	sHA	sPa	sPA
SUB177	98	12	119	25	77	10	156	103
SUB371	57	7	78	19	41	1	138	99
HOMSTRAD	32	5	75	5	31	0	140	78

Number of table entries (out of a possible 400) that differ by more than 2 log-odds units from our soluble structure/sequence tables. This 2 unit threshold is chosen to be a reasonable measure of dissimilarity. Differences between structure/sequence and structure/structure derived tables decrease as the number of families included in the structure/structure set increases.

Table 3. Self-consistency scores and number of counts for each membrane environment specific substitution table

ESST	Q	Counts	ESST	Q	Counts
NPa	0.72	36 437	PCA	0.96	45 654
NPA	0.88	146 253	PHA	0.96	96 771
TCa	0.90	46 512	HHA	0.97	147 118
NEa	0.92	137 556	THA	0.97	265 566
NHa	0.92	118 306	HEA	0.97	109 238
PEA	0.92	60 677	THa	0.97	171 736
TCA	0.92	44 326	HCA	0.98	228 302
HCa	0.93	65 124	TEA	0.98	211 862
HHa	0.93	63 665	NCA	0.98	350 138
NCa	0.93	113 341	IHA	0.98	56 569
ICA	0.95	44 362	IEA	0.98	34 660
HEa	0.95	48 262	NEA	0.98	148 662
TEa	0.95	102 801	NHA	0.98	253 458

Environment labels are described in Section 2.1. Accessible environment tables (**A) tend to have higher self-consistencies than inaccessible environment tables (**a).

3.2 Membrane environment selection

Many +ve ϕ angle tables (*P*) suffer from low self-consistency scores, low count numbers, and poor stability. For example, the TPa environment of Figure 2 has a Q score of 0.64 from 10 060 counts. Low self-consistency scores are to be expected: the majority of substitutions in these environments involve glycine, and other substitutions may be too rare to be representative. To increase table quality, the +ve ϕ environments were merged into an accessible +ve ϕ environment (NPA) and an inaccessible +ve ϕ environment (NPa). These are labelled 'N' layer so as to maintain a consistent notation.

Self-consistency scores and total numbers of substitutions for each table in the resulting environment set are shown in Table 3. Each of our membrane tables has a three letter code of the form layer (H,N,T,P,I) : secondary structure (H,E,P,C) : accessibility (a,A).

3.3 Clustering of tables

The Euclidean distance between the log-odds tables is used to create a 'family-tree' of the different environments (Figure 3). Tables

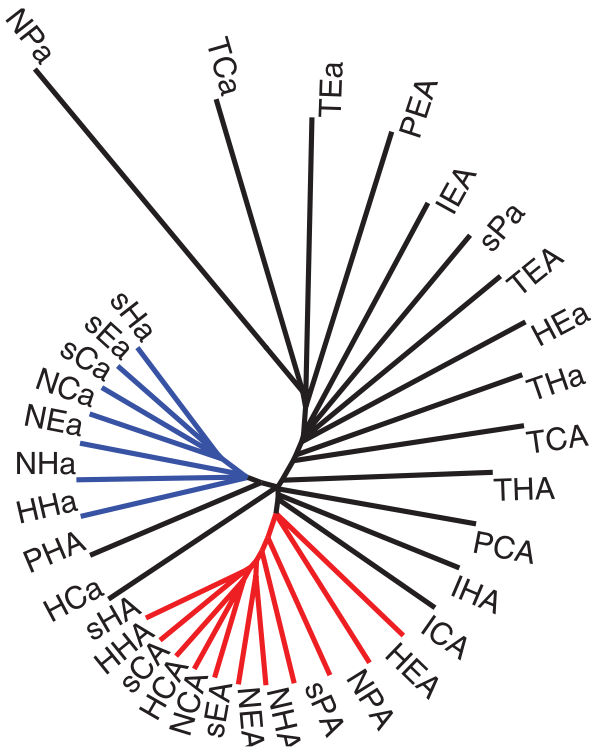

Fig. 3. Dendrogram of ESSTs. A split is seen between accessible (red) and inaccessible (blue) environments. Tail-layer environments (T**) appear not to cluster. Note that here, as elsewhere, 'NPa' and 'NPA' refer to combined +ve ϕ environments that include residues in the transmembrane regions (see Section 3.2).

for soluble proteins, labelled with a leading 's', are included for comparison. When calculating the distance, each substitution is normalized by its standard deviation across all the tables. This prevents the distance measure being dominated by a handful of extreme substitution changes.

It has been suggested that loops of membrane proteins that extend above and below the bilayer behave similarly to loops in soluble proteins (e.g. Tastan *et al.*, 2009). We also see this in our results, where each table of the form NC* clusters with its sC* counterpart. As might be expected, the not-in-membrane tables (N**) are most similar to their soluble equivalents (the notable exception being that sHA clusters with HHA rather than NHA).

The tail-contacting environments are clear outliers, and do not cluster. Environments of the form T*A are dissimilar to both s*a and s*A. This is consistent with a number of other studies (e.g. Stevens and Arkin, 1999) that have found little evidence for the early 'inside-out' hypothesis of membrane protein structure (Engelman and Zaccai, 1980).

Additional outliers are +ve ϕ environments (*P*), and some β-strand environments (*E*). This lasts may be because much of our β-strand data comes from outer-membrane porins of Gram-negative bacteria. In Gram-negative bacteria, the outer membrane is asymmetric: the inner leaflet is composed of phospholipids whereas the outer leaflet is composed of lipopolysaccharides. Additionally, inward-facing solvent-exposed residues are in contact with the periplasm rather than the cytosol. Evidence for the uniqueness of the β-strand environments can also be seen in their composition. Accessible β-strands within the membrane rarely contain cysteine, and the TEA environment is abundant in tyrosine.

The remaining environments separate by accessibility. Surprisingly, within the inaccessible clade (Figure 3, blue), the soluble secondary structure environments are more similar to each other than to their membrane equivalents. An accessible clade in Figure 3 is coloured red from the same level as the inaccessible clade. The pore-lining and interface environments lie just beyond these clades, suggesting that these environments have distinct properties, and therefore that their use is sensible.

A PCA plot allows patterns in substitutions to be discerned. Figure 4 accounts for 48% of the variation in the data with 3 principal components. Figure 4c shows that the differences between accessible and inaccessible environments cause most of the variation between tables—they are largely separated along the first principal component (the main exceptions being accessible tail-layer tables, T*A). This first component can broadly be identified as a measure of 'hydrophobicity'. Looking at the labelled points in Figure 4a, as the first principal component increases we move from tail layer to interface layer to head layer accessible environments, corresponding to decreasing hydrophobicity.

The second principal component appears to relate to secondary structure. Moving from left to right in Figure 4e, we encounter the labelled points in the order TCa, TEa, THa as the second component increases. The same ordering is found for other layer types within the membrane. However, for soluble and not-in-membrane environments the order instead runs coil tables, helix tables, β-strand tables (e.g. sCA, sHa, sEa).

The bottom row of plots shows that different secondary structure environments cluster in the second and third components. The third principal component appears to be dominated by the differences between β-strand environments.

3.4 Sequence-to-structure alignment

The previous section discussed the variations in substitution preferences in different environments. Now we demonstrate that a knowledge of these differences improves sequence-to-structure alignment.

Alignments were made with the sequence-to-sequence alignment program MUSCLE, and the sequence-to-structure alignment program FUGUE. Three different sets of substitution tables were used in the case of FUGUE (i) the default soluble tables, (ii) our membrane tables and (iii) the PHAT/BLOSUM62 tables in a bipartite scheme. In this last case, PHAT was applied to residues with a 'T' layer annotation (including pore-lining residues), and BLOSUM62 was used elsewhere.

As the same program, FUGUE, was used with each set of tables, fair comparisons can be made between them. Gap penalties were determined separately for each set of tables.

3.5 Gap penalty determination

In the case of FUGUE, the optimal alignment is that which maximizes the sum of the table entries S^E_{ba} for each pair of aligned residues. Not all residues will align, even between very similar proteins, and penalties to the alignment score must be determined for introducing gaps into the alignment. FUGUE distinguishes between

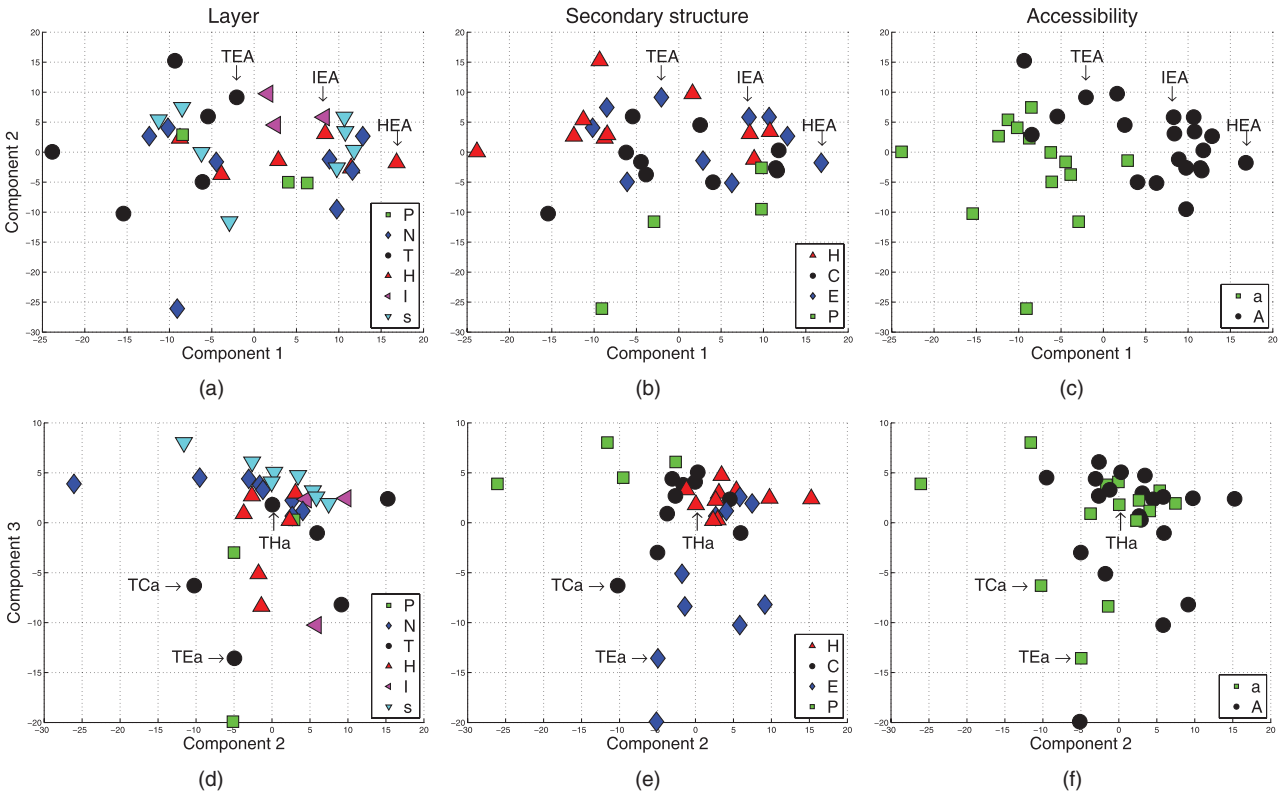

Fig. 4. Principal component analysis of ESSTs. The top row and the bottom row are views of the same data along different principal components. The columns colour-code the data-points by layer type, secondary structure and accessibility, respectively. This allows the three-letter table code of each point to be read off from left to right. The labelled tables are ordered by secondary structure in the second principal component—reading panel (**e**) from left to right we first encounter TCa, then TEa, then THa. A similar ordering holds for other layer and accessibility types.

several types of gaps (see Shi *et al.*, 2001 for details). Gaps are penalized in order of severity as follows:

(1) Gap within a secondary structure element (H)
(2) Gap at the end of a secondary structure element (L)
(3) Gap in a loop region (VL)
(4) Gap at a terminus (VVL)

There are actually 8 types of gap penalty: each of the above categories can initiate a gap or be an extension of an existing gap. Initiating a gap results in a larger penalty than continuing an existing gap: the alignment is thus biased to a small number of large insertion/deletion events rather than a larger number of smaller events.

A subset of 72 protein pairs was selected at random from the 408 pairs of proteins in the alignment dataset (see Section 2.6), and alignments made with perturbations of the default FUGUE gap penalties. Perturbations were made such that gap opening penalties were at least as large as gap extension penalties, and such that more 'severe' gaps had penalties at least as large as less 'severe' ones. The size of the perturbation steps ranged from 1 to 5 units and depended on the size of the default penalties. The alignment quality with the default FUGUE tables differed little as the penalties were changed. In view of this, and as most users are unlikely to change the gap penalties, the default penalties were kept.

Table 4. Gap penalties for each set of tables used with FUGUE

Tables	Initiation				Extension			
	H	L	VL	VVL	H	L	VL	VVL
Default	28	20	20	8	4	4	2	2
Membrane	35	25	20	8	4	4	2	2
PHAT/BLOSUM62	26	24	16	8	6	4	2	2

For the membrane tables, only two gap initiation penalties were found to substantially differ from the default values (Table 4). The increase of penalties in these cases is to be expected as the sizes of transmembrane secondary structure elements are constrained by the membrane thickness. The subsequent analysis uses these revised penalties, but membrane tables with the default gap penalties lead to similar results. The PHAT/BLOSUM62 tables are scaled differently from the others, so their penalties are not directly comparable.

3.6 Alignment accuracy

Alignments were made for the remaining 336 pairs of proteins in the alignment dataset. None of the methods performed well in

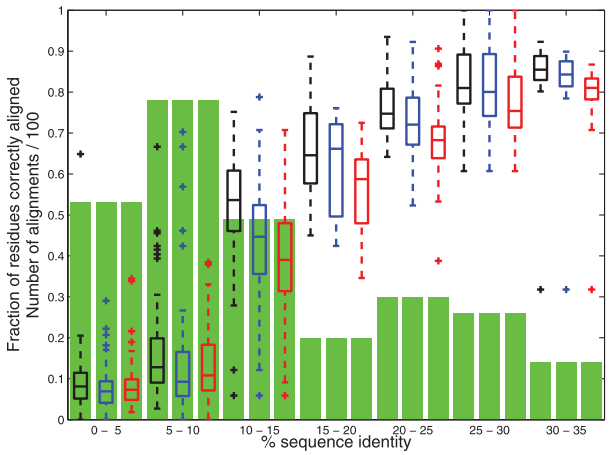

Fig. 5. Box plots of the fraction of residues aligned correctly as sequence identity increases. There are three boxes at each sequence identity, from left to right corresponding to membrane FUGUE (black), default FUGUE (blue) and MUSCLE (red). The green bars show the number of alignments divided by 100. For example, there are 78 alignments in the 5–10% sequence identity range.

the 0–10% sequence identity range, but beyond this the membrane tables gave a consistent alignment advantage. At >35% sequence identity, the alignments show few differences. Figure 5 compares the alignments of membrane specific tables to common alternatives. PHAT/BLOSUM62 is omitted for clarity—its performance is comparable to that of default FUGUE.

Consideration of the outliers in Figure 5 is informative. In the 30–35% sequence identity range, the three methods in the figure appear to have correctly aligned only ∼30% of the residues in one sequence-structure pair [PDB codes 1SU4A (sequence), 1XP5A (structure)]. In fact, these proteins are an identical rabbit Ca^{2+}ATPase in different conformations. In this case the structure-to-structure alignment (by rigid-body superposition) from TM-align does not capture local similarities, leading to the 30–35% sequence identity figure, and the low assessment of performance of the other alignment methods.

Outliers in the 0–10% sequence identity range are mostly due to short alignment lengths. Figure 5 gives a broad picture of performance differences, but does not distinguish between a small alignment improvement on a short protein and a much larger improvement on a bigger protein. Table 5 lists the number of times that membrane FUGUE correctly aligns at least 10 residues more (Win) or fewer (Loss) than another method.

Membrane FUGUE often improved alignment by more than 10 residues. Table 6 gives the number of correctly aligned residues across all the alignments in each sequence identity bracket. Membrane FUGUE outperforms all other methods in all brackets, except at >35% sequence identity where the differences between the methods are marginal.

If the alignment set is divided into α and β proteins the same trends in accuracy are seen for both, with membrane FUGUE outperforming the other methods. The principal difference is the scarcity of β-type alignment pairs at higher sequence identities.

Table 5. Alignment quality of membrane tables versus other methods

Identity (%)	Number of alignments	Membrane FUGUE versus					
		Default FUGUE		MUSCLE		PHAT/ BLOSUM62	
		Win	Loss	Win	Loss	Win	Loss
0–5	53	12	5	14	4	11	3
5–10	78	32	10	29	8	30	6
10–15	49	36	4	43	1	33	6
15–20	20	10	2	14	0	11	2
20–25	30	8	0	18	0	12	3
25–30	26	4	1	14	0	6	1
30–35	14	1	0	6	0	3	1
>35	66	1	2	3	1	1	0
Total	336	104	24	141	14	107	22

For each sequence identity range, the number of alignments where membrane FUGUE correctly aligns at least 10 more (Win) or 10 fewer (Loss) residues than the named alternative method. For example, in the 10–15% sequence identity range membrane FUGUE correctly aligns at least 10 more residues in 36 out of 49 alignments.

Table 6. Number of correctly aligned residues for each set of tables

Identity (%)	Number of residues	Membrane FUGUE	Default FUGUE	MUSCLE	PHAT/ BLOSUM62
0–5	12915	**1132**	872	910	1007
5–10	22734	**3571**	2555	3001	2721
10–15	24349	**12 915**	10819	9893	10427
15–20	7576	**5042**	4697	4145	4467
20–25	9156	**6900**	6607	6145	6565
25–30	5644	**4608**	4522	4300	4479
30–35	4792	**3448**	3403	3274	3402
>35	18881	17578	**17 586**	17547	17545
Total	106047	**55 194**	51061	49215	50613

The highest number of aligned residues for each sequence identity range is shown in bold.

3.7 Structure prediction

Models were built with MEDELLER for each of the 336 default FUGUE, and membrane FUGUE alignments. Models were also built for the implicit sequence alignments from TM-align.

MEDELLER provides different model-building options that prioritize accuracy or coverage. However, the relative quality of the models produced by different alignment methods showed little sensitivity to the model-building details. Results described below are for the default 'high-accuracy' models, but results for the 'naive' and complete models are similar.

Reasonable alignments are only achieved from 15% sequence identity upwards (Figure 5), and above 35% alignments differ little between methods. In the 15−35% sequence identity range the average RMSDs between the model and the native structure are: 3.4 Å (membrane FUGUE), 4.1 Å (default FUGUE), 2.0 Å (TM-align). The mean sequence identity is 24%.

4 DISCUSSION

We have constructed substitution tables for membrane proteins by aligning single structures to multiple homologous sequences. This

method, already used in the literature, allows a small number of structures to be leveraged to build tables at the cost of increased error in table construction. To address this problem, we suggest a method of assessing the quality of tables constructed in this way which allows us to build tables that are stable and consistent with the data used to construct them.

A principal component analysis of the individual tables revealed that residues in contact with lipid-tails have some substitution preferences typical of hydrophobic regions. However, the differences in other substitution preferences mean that membrane proteins are not simply 'inside out'.

Globally, it appears that accessibility is the primary determinant of membrane substitution preferences, followed by secondary structure. Position within the membrane has a less clearly-defined, but substantial effect. For example, membrane tables showed greater variability than their soluble equivalents. This suggests that an environment-specific approach to membrane protein modelling will yield greater improvements than did the environment-specific approach to soluble protein modelling.

Evidence for this supposition was found in a set of 336 alignments made by MUSCLE and FUGUE. MUSCLE is designed for sequence-to-sequence alignment, and so makes no use of structural information. It is unsurprising therefore that it performed worst at sequence-to-structure alignment. The default FUGUE tables and the bipartite PHAT/BLOSUM62 alignments performed better than MUSCLE and comparably to each other. Each makes use of different structural information—the default tables take into account the accessibility, secondary structure and hydrogen-bonding of a residue; whereas PHAT/BLOSUM62 distinguishes between residues inside and outside the membrane.

Conflicting accounts of the performance of PHAT have previously been reported. It has been suggested that this is due to bad alignments when PHAT is applied to non-transmembrane residues (Pirovano *et al.*, 2008). The good alignments here can most likely be attributed to the quality of the transmembrane annotation from iMembrane.

Our membrane tables distinguish between both membrane location, and secondary structure and accessibility. Compared with the best performing alternative tables, the use of the membrane tables led to 104 of the 336 alignments having at least 10 more correctly aligned residues, with only 24 alignments being worse by the same margin. These improved alignments translate into predicted structures with a lower average RMSD (3.4 Å membrane FUGUE, 4.1 Å default FUGUE) within the 15–35% sequence identity range.

These results represent only a proof-of-principle for this approach. Here, to demonstrate the method we have considered only pairwise alignment, but multiple sequence information should further improve results. Alignment quality might also be enhanced by changes to the definition of when a residue is in contact with the head or tail layers of a membrane, or from the introduction of a bipartite scheme of gap penalties in which insertions and deletions are punished more severely in the transmembrane region.

More radically, alignment might be improved by an iterative approach to table construction. The tables presented here were generated by counting substitutions between homologous sequences aligned to a single structure by MUSCLE. Instead, these alignments could be made by FUGUE using the membrane tables. The resulting improved substitution tables could then be used to realign the sequences. This procedure could be iterated until convergence.

Our substitution tables, which take into account the environments of residues in membrane proteins, substantially improve alignments between membrane protein sequences and structures. In turn, these improved alignments lead to better structural models of membrane proteins.

ACKNOWLEDGEMENT

Thanks go to the members of the Oxford Protein Informatics Group for useful discussion and feedback.

Funding: Engineering and Physical Sciences Research Council (to J.R.H. and C.M.D.); the Biotechnology and Biological Sciences Research Council (to S.K. and C.M.D.); and the University of Oxford Doctoral Training Centres (to C.M.D.).

Conflict of Interest: none declared.

REFERENCES

Almén,M. *et al*. (2009) Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. *BMC Biol.*, **7**, 50.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Engelman,D.M. and Zaccai,G. (1980) Bacteriorhodopsin is an inside-out protein. *Proc. Natl Acad. Sci. USA*, **77**, 5894–5898.

Forrest,L. *et al*. (2006) On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophy. J.*, **91**, 508–517.

Gong,S. *et al*. (2009) Structural and functional restraints in the evolution of protein families and superfamilies. *Biochemical Society Trans.*, **37**(Pt 4), 727–733.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Hotelling,H. (1933) Analysis of complex statistical variables into principal components. *J. Educ. Psychol.*, **24**, 417–441.

Hubbard,T. and Blundell,T. (1987) Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.*, **1**, 159–171.

Jones,D. *et al*. (1994) A mutation data matrix for transmembrane proteins. *FEBS Lett.*, **339**, 269–275.

Kelm,S. *et al*. (2009) iMembrane: homology-based membrane-insertion of proteins. *Bioinformatics*, **25**, 1086–1088.

Kelm,S. *et al*. (2010) MEDELLER: homology-based coordinate generation for membrane proteins. *Bioinformatics*, **26**, 2833–2840.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Mizuguchi,K. *et al*. (1998a) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.

Mizuguchi,K. *et al*. (1998b) JOY: protein sequence-structure representation and analysis. *Bioinformatics*, **14**, 617–623.

Mizuguchi,K. *et al*. (2007) Environment specific substitution tables for thermophilic proteins. *BMC bioinformatics*, **8** (Suppl. 1), S15.

Mokrab,Y. and Mizuguchi,K. (2005) Amino-Acid substitutions in membrane proteins: applications to homology recognition and comparative modelling. *BMC Bioinformatics*, **6**(Suppl. 3), S9.

Mokrab,Y. *et al*. (2010) A structural dissection of amino acid substitutions in helical transmembrane proteins. *Proteins*, **78**, 2895–2907.

Moult,J. *et al*. (2009) Critical assessment of methods of protein structure prediction - Round VIII. *Proteins*, **77** (Suppl. 9), 1–4.

Müller,T. *et al*. (2001) Non-symmetric score matrices and the detection of homologous transmembrane proteins. *Bioinformatics*, **17** (Suppl. 1), S182–S189.

Ng,P.C. *et al*. (2000) PHAT: a transmembrane-specific substitution matrix. *Bioinformatics*, **16**, 760–766.

Pirovano,W. *et al*. (2008) PRALINE: a strategy for improved multiple alignment of transmembrane proteins. *Bioinformatics*, **24**, 492–497.

Sali,A. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Sánchez,R. (1997) Advances in comparative protein-structure modelling. *Curr. Opin. Struct. Biol.*, **7**, 206–214.

Scott,K.A. *et al*. (2008) Coarse-grained MD simulations of membrane Protein-Bilayer self-assembly. *Structure*, **16**, 621–630.

Shi,J. *et al*. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.

Stevens,T.J. and Arkin,I.T. (1999) Are membrane proteins inside-out proteins? *Proteins: Struct. Funct. Genet.*, **36**, 135–143.

Tastan,O. *et al*. (2009) The effect of loops on the structural organization of alpha-helical membrane proteins. *Biophys. J.*, **96**, 2299–2312.

Tusnády,G. *et al*. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.*, **33**, D275–D278.

Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.