Replacement Matrices for Transmembrane Proteins

Master’s Thesis

Presented to
The Faculty of the Graduate School of Arts and Sciences
Brandeis University
Department of Biochemistry
Douglas Theobald, Advisor

In Partial Fulfillment
of the Requirements for the Degree
Master of Science

by
Marcus Kelly

May, 2011
Acknowledgments

I would like to thank the people of the Theobald Lab who worked there from November 2009 through May 2011. These are Jeff Boucher, Adam Drake, Mackenzie Gallegos, Nat Lazar, Kristine Mackin, Philip Steindel, and Marco Viera. Each of these people contributed to this project, whether by assisting me with programming and scripting or by helping revise this thesis and the defense after.

I would also like to thank Dr. Chris Miller for providing some of the annotations used to train the annotation method, as well as his support, experience, insight, and humor.

I would like to thank Dr. Douglas Theobald for providing more assistance with this project than I had any right to expect. I am honored that he trusted me with his time and resources. It has been a great privilege to work with him.

And finally, I would like to thank my family and friends for their unwa­vering support and encouragement, without which I could never have finished this work.
Abstract

Replacement Matrices for Transmembrane Proteins
A thesis presented to the Department of Biochemistry
Graduate School of Arts and Sciences
of Brandeis University
Waltham, Massachusetts
By Marcus Kelly

Recent advances in the estimation of amino acid exchangeability matrices have greatly improved the estimation of phylogenies [1, 2]. Here, we present an exchangeability matrix that specifically improves phylogenetic analysis of transmembrane proteins. We first use Bayesian inference to estimate replacement matrices for single families of transmembrane proteins. From these, we constructed a general replacement matrix, representative of all α-helical transmembrane proteins, in which the entries are the weighted harmonic means of the corresponding entries in the single-family replacement matrices. The best matrix improves the log-likelihood of transmembrane protein phylogenies by 0.002 points per site over a phylogeny made using the LG replacement matrix. Phylogenetic analysis of soluble proteins greatly benefits from the use of multiple exchangeability matrices that correspond to different structural features[3]. The software that performs this analysis depends upon annotated sequence alignment for input. Accordingly, we also present a means of automatically annotating individual residues of transmembrane protein structures. Using our structurally annotated alignments will allow construction of structure-specific exchangeability matrices for TM proteins, providing a means to incorporate structural information into phylogenetic inference of transmembrane proteins.
## Contents

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contents</td>
<td>v</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>Introduction</td>
<td>2</td>
</tr>
<tr>
<td>Materials and Methods</td>
<td>14</td>
</tr>
<tr>
<td>Estimation of Rate Matrices</td>
<td>14</td>
</tr>
<tr>
<td>Data Sets</td>
<td>14</td>
</tr>
<tr>
<td>Inference of Phylogenetic Trees</td>
<td>15</td>
</tr>
<tr>
<td>Estimation of Single-Family Rate Matrices</td>
<td>16</td>
</tr>
<tr>
<td>Averaging of Single-Family Rate Matrices</td>
<td>16</td>
</tr>
<tr>
<td>Annotation of Transmembrane Protein Structures</td>
<td>17</td>
</tr>
<tr>
<td>Estimation of Solvent Exposure</td>
<td>17</td>
</tr>
<tr>
<td>Determination of Secondary Structure</td>
<td>17</td>
</tr>
<tr>
<td>Prediction of Transmembrane Helices</td>
<td>18</td>
</tr>
<tr>
<td>Nearest Surface Neighbors</td>
<td>18</td>
</tr>
<tr>
<td>Annotation Decisions</td>
<td>19</td>
</tr>
<tr>
<td>Results</td>
<td>21</td>
</tr>
<tr>
<td>Data Set</td>
<td>21</td>
</tr>
<tr>
<td>Estimation of Rate Matrices</td>
<td>21</td>
</tr>
<tr>
<td>Prior Probabilities and Likelihoods of Annotations</td>
<td>26</td>
</tr>
<tr>
<td>Annotation of Transmembrane Protein Structures</td>
<td>26</td>
</tr>
<tr>
<td>Discussion</td>
<td>31</td>
</tr>
<tr>
<td>Appendix A: List of Variables</td>
<td>35</td>
</tr>
<tr>
<td>Appendix B: Structures in Data Set</td>
<td>36</td>
</tr>
</tbody>
</table>
List of Tables

1 Selected Substitution and Exchangeability Matrices . . . . . . . 8
List of Figures

1. Tree Topologies .................................................. 5
2. Exchangeability Matrix and Equilibrium Frequencies ........ 11
3. Decisions in Structural Annotations .......................... 20
4. Log-likelihood gains over $LG$ ................................. 23
5. Bubble plot of $KT$ .................................................. 24
6. Difference bubble plot ............................................. 25
7. Annotation probabilities by amino acid ....................... 27
8. Probabilities of neighbor counts ............................... 28
9. Annotations mapped on structures ............................. 30
**Introduction**

Since 1985, GenBank, the sequence database maintained by the National Center for Biotechnology Information, has doubled in size every 18 months [4]. At this pace, novel sequences cannot be characterized by experimental methods alone. The emergence of phylogenetics software in the past decades allows the rapid comparison of novel genes to ones already characterized.

Phylogenetics has become an essential and practical tool in biology. For example, novel species are often classified by phylogenetic analysis of the sequences of their large ribosomal subunits [5]. Analysis of the ancestry of single genes has also identified functionally critical features of proteins. For example, recent work by Kang et al. [6] demonstrated that putative orthologues of mammalian TRPA1 irritant-sensing channels in two model organisms, *Drosophila melanogaster* and *Caenorhabditis elegans*, were actually orthologous to Choanoflagellate and Anthozoan TRPA channels.

Phylogenetics software determines common ancestry by modeling evolutionary processes and accounting for the changes between observed sequences[7], typically using aligned sequences as input. In this study, we will make the following assumptions when discussing the generation and assessment of phylogenetic trees:

1. All branches of a tree evolve independently of one another.
2. All sites in a sequence alignment evolve independently of one another.

3. Residue substitution is a continuous-time Markov process. The probabilities for a transition from one state to another depend only on the current state and not on the state at any point in the past or future.

4. Evolution is time-reversible. The probability of one sequence evolving into another is exactly the same as that of the second sequence evolving into the first.

5. Sites do not evolve at identical rates. Instead, relative rates of sites follow a $\Gamma$ probability distribution function.

While rational objections might be raised to all five of these assumptions, they strike a balance between computation time and accuracy[7]. In general, these assumptions inform the calculations used to estimate the parameters that describe a phylogeny.

Arguably the most important feature of a phylogeny is its topology $\tau$. The topology of a tree determines which groups of taxa share more recent ancestors than do other groups. Trees may be rooted or unrooted depending on whether the last common ancestor of the included taxa is known. In some cases a distant relative may be included in the data as an “outgroup” to indicate the proper position of the root. Crucially, an unrooted phylogeny does not describe the direction of evolutionary changes. Trees may also vary in the number of branches radiating from each node. Those with three branches per node are “bifurcating”, and those with more are “multifurcating”. An unrooted
bifurcating tree of $N$ taxa has $2N - 3$ branches and

$$\frac{(2N - 4)!}{2^{N-2}(N-2)!}$$

possible topologies[7].

Each node is connected to another node or a branch tip by a branch of length $\nu$. The ensemble of branch lengths for a phylogeny is given as a vector $\vec{\nu}$. While any unrooted bifurcating tree of $N$ taxa will always have $2N - 3$ branches, the branches may connect different nodes or taxa under different topologies. Branch lengths are proportional to the time expected for the sequence at one end of a branch to evolve into the other [7, 8]. The possible topologies and their corresponding branch lengths for a tree of $N = 4$ taxa are shown in Figure 1.

In order to describe a phylogenetic tree, we need only $\tau$ and $\vec{\nu}$. To infer a phylogeny, however, we need estimates of many more parameters. For a phylogeny using data with $S$ different states ($S = 4$ for nucleotides, 20 for amino acids, etc.) that have indices $i$ and $j$, there are $S^2$ transition rates $q_{ij}$ between states and $S$ equilibrium frequencies $\pi_i$ of each state. We will refer to the rates collectively as a $S \times S$ matrix $Q$ and the frequencies as a diagonal $S \times S$ matrix $\Pi$. Finally, to account for different overall mutation rates across sites, we assume that rates follow a $\Gamma$-distribution, described by a single shape parameter $\alpha$[7, 9].

Together, these parameters are used to calculate the likelihood that we observe our data given our estimates of the parameters listed above. Consider some rooted tree of two taxa $x$ and $y$ that has at its root some unknown ancestor $z$. At some site $m$, we calculate the probability $p(x_m|\nu_{zx}, z_m)$ that
Figure 1: Possible unrooted bifurcating tree topologies where \(N=4\) taxa. Here, \(u, v, x\) and \(y\) are taxa with known sequences and \(z\) and \(w\) are their ancestors, whose sequences are unknown.

\(z_m\) evolved into \(x_m\) over some time \(\nu_{zx}\) and the probability \(p(y_m|\nu_{zy}, z_m)\) that \(z_m\) also evolved into \(y_m\) over some time \(\nu_{zy}\). However, \(z_m\) is unknown, so the likelihood of the tree \(p(x_m, y_m|\tau, \vec{\nu}, Q)\) using only information from this site is a sum of the probabilities over each possible state for \(z_m\):

\[
p(x_m, y_m|\tau, \vec{\nu}, Q) = \sum_{z_m} p(x_m|\nu_{zx}, z_m)p(y_m|\nu_{zy}, z_m)p(z_m) \tag{2}
\]

The prior probability \(p(z_m)\) of the root is often assumed to be the equilibrium frequency of \(z_m\) (i.e. \(\pi_{z_m}\))[7]. If \(z\) is not the root but instead represents the last common ancestor of \(x\) and \(y\) on some larger tree, \(p(z)\) is then conditional on the state of some more ancient ancestor and the branch length between that ancestor and \(z\). In general, then, we need to sum over all possible combinations of states of \(z_m\) and of the ancestor at site \(m\). Note that the probabilities of \(x_m\) and \(y_m\) are not conditional on each other. This reflects our assumption (assumption 1) that branches evolve independently of one another.

Because we have assumed that all sites evolve independently (assumption 2), the likelihood \(p(D|\tau, \vec{\nu}, Q)\) of the entire tree is the product of the likelihoods of the tree for each individual site:

\[
p(D|\tau, \vec{\nu}, Q) = \prod_m p(x_m, y_m|\tau, \vec{\nu}, Q) \tag{3}
\]

where \(x\) and \(y\) are the hypothetical taxa discussed above. Maximum-likelihood
(ML) algorithms search for the values of $\vec{\nu}$, $Q$ and the topology $\tau$ that maximize the left-hand term of equation (3). The tree likelihood is also a widely-accepted metric of the fit of a phylogeny to its data [3, 10, 11].

As will be shown below, matrix algebra is often necessary to calculate the likelihood, and giving each transition probability in terms of mutation rates, relative rates, branch lengths, etc. would be excessively tedious. Transition probabilities $p_{ii}$, which describe the probability of a state transitioning to itself, decrease with increasing $\nu$. The longer the time allowed, the more likely a state will evolve into another. Those that describe a transition from one state to another increase with time, and converge on some value between 0 and 1. For some branch along which some states are conserved and others not, there will be some $\nu$ that maximizes the product of the transition probabilities and the tree likelihood.

As an alternative to ML methods, Bayesian Inference (BI) methods calculate the Bayesian posterior probability from the likelihood in (3).

$$p(\tau, \vec{\nu}, Q | D) = \frac{p(D | \tau, \vec{\nu}, Q)p(\tau, \vec{\nu}, Q)}{p(D)}$$  \hspace{1cm} (4)

It may seem more intuitive to find the most likely parameters given the data (by Bayesian inference, as in equation (4)) than to find the parameters that make the data most probable (by maximum likelihood, as in equation (3)). However, estimating the prior probabilities $p(D)$ and $p(\tau, \vec{\nu}, Q)$ can pose some computational difficulty, and the choice of appropriate prior probabilities remains an open question[7, 12].

The maximum-likelihood ensemble of parameters cannot be found analytically except by finding the likelihood of each tree and choosing the tree with
the maximum likelihood. The problem of merely finding the optimal topology belongs to a class of computationally intensive programming problems, which includes the infamous “Travelling Salesman Problem[7].” As a result, we are forced to turn to heuristic algorithms that iterate over individual parameters to search for the maximum-likelihood ensemble rather than to find the best tree from all possible solutions. When using BI methods, we are equally challenged to find the maximum-posterior estimates.

Though the tree likelihood is always conditional on all of the tree parameters mentioned, we can fix certain parameters and search the others for their maximum-likelihood values[12]. When generating a phylogeny, we are chiefly interested in $\tau$ and $\vec{\nu}$. For nucleotide alignments, the independent $Q$ parameters do not contribute greatly to computation time and are often estimated concurrently with a phylogeny. For amino acids and codons, however, assuming a model with fixed rates greatly decreases computation time.

Fixing, rather than estimating, $Q$ also reduces the uncertainty of the log-likelihood. As more parameters are included in the model, the accuracy of the model improves, but so do the variances of the model’s estimates[12, 13]. The Akaike Information Criterion (AIC) has been proposed as a compromise between accuracy and precision. When choosing between models with different numbers of parameters, instead of choosing the model that maximizes the likelihood, we choose the model that maximizes the AIC. The AIC is defined as:

$$AIC(\hat{Q}, \hat{\tau}, \hat{\nu}) = \ln P(D|\hat{Q}, \hat{\tau}, \hat{\nu}) - K$$

(5)

where $\hat{\theta}$ indicates an estimate of a parameter $\theta$ and $K$ is the number of
<table>
<thead>
<tr>
<th>Name</th>
<th>Date</th>
<th>Data Set</th>
<th>Method</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAMn</td>
<td>1978</td>
<td>Broadly conserved soluble proteins</td>
<td>Counting</td>
<td>[14]</td>
</tr>
<tr>
<td>MtREV</td>
<td>1996</td>
<td>Mitochondrial Proteins</td>
<td>ML</td>
<td>[18]</td>
</tr>
<tr>
<td>MtMamm</td>
<td>1998</td>
<td>Mammalian mitochondrial proteins</td>
<td>ML</td>
<td>[19]</td>
</tr>
<tr>
<td>CpREV</td>
<td>2000</td>
<td>Chloroplast proteins</td>
<td>ML</td>
<td>[21]</td>
</tr>
<tr>
<td>LG</td>
<td>2008</td>
<td>General proteins</td>
<td>ML</td>
<td>[1]</td>
</tr>
<tr>
<td>MtZoa</td>
<td>2009</td>
<td>Animal mitochondrial proteins</td>
<td>BI</td>
<td>[2]</td>
</tr>
<tr>
<td>EX_EHO etc.</td>
<td>2010</td>
<td>Soluble proteins</td>
<td>ML</td>
<td>[3]</td>
</tr>
</tbody>
</table>

**Table 1**: Selected exchangeability and substitution matrices. Abbreviations for estimation methods: ML, maximum likelihood; BI, Bayesian inference. Here, $n$ designates a series of matrices (e.g. BLOSUM62, BLOSUM85, etc.).

parameters fit to the data. As $K$ increases, the AIC will decrease, representing the influence of parameter uncerainty (and less precision). If the inclusion of more parameters results in a significantly better likelihood, the first term will increase, corresponding to increased accuracy. The AIC can be viewed as a penalized likelihood.

With both computation time and penalties to the AIC in mind, rate matrices are often fixed rather than estimated. It is common practice to assume that mutation rates are not substantially different between protein families. Many substitution matrices and rate matrices have been estimated over the past forty years (see Table 1). As sequence data has accumulated and as computing power has improved, rate matrices have been estimated from larger sample sets and by more advanced algorithms[1, 11, 14–16].

The original amino acid substitution matrices predate ML inference of phylogenies[7, 14]. Dayhoff and Eck computed odds matrices, the PAM series, using sequence alignments of some protein families that were known at the
time to be broadly conserved. Later matrices derived from sequence alignments would use similar “counting methods” to estimate their entries. Though primarily intended for use with sequence alignments, homology searches, or maximum-parsimony (MP) phylogenies, the entries of BLOSUM, PAM, and JTT have been used to estimate ML phylogenies[11, 14–16]. Each can be treated as a transition matrix $P$ whose entries are $p(i \rightarrow j|t)$, where $i$ and $j$ are amino acids and $t$ is the time allowed for the transition to occur. For a vector of amino acid frequencies $\vec{f}$ that vary with time, we expect that

$$\vec{f}_t = P(t)\vec{f}_0$$

(6)

Because we assume a hidden Markov model, it follows that

$$P = e^{Qt}$$

(7)

The rate matrix $Q$ is:

$$Q = \begin{bmatrix}
-\sum_{i \neq 1} q_{i,1} & q_{1,2} & \cdots & q_{1,20} \\
q_{2,1} & -\sum_{i \neq 2} q_{i,1} & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
q_{20,1} & \cdots & \cdots & -\sum_{i \neq 20} q_{i,20}
\end{bmatrix}$$

(8)

All of the rates given in the rate matrix $Q$ are directional. However, in many cases we do not know in which “direction” evolution acts[7]. We therefore assume that evolution is time-reversible. For a character that is in state $i$ at one end of a branch and $j$ at another, the probability that $i$ evolved into $j$ must be equal to the probability that $j$ evolved into $i$. Where $\pi_i$ and $\pi_j$ are the equilibrium frequencies of $i$ and $j$ respectively,

$$q_{ij}\pi_j = q_{ji}\pi_i$$

(9)
The exchangeabilities on either side of equation (9) can be calculated by a simple decomposition of the rate matrix:

\[ Q = \Pi R \]  

(10)

where \( \Pi \) is a diagonal matrix of equilibrium frequencies \( \pi_i \) and \( R \) is an exchangeability matrix whose elements \( r_{ij} = q_{ij} \pi_j \). The likelihood calculation for an unrooted tree is identical to that for a rooted tree, and we must still assume the location of the root. However, using a time-reversible matrix, the likelihood for a tree stays constant if the location of the root is moved, and the particular position of the root is an arbitrary choice.

The assumption of time-reversibility also reduces the number of parameters to be estimated. Where \( S \) is again the number of possible states, instead of optimizing \( S^2 - S \) rates, we need now only optimize \( \frac{S^2 - S}{2} \) exchangeabilities. For these reasons, time-reversible rate matrices are conventionally represented by their exchangeabilities [2, 10, 11]. A bubble plot of an \( R \) matrix and its corresponding \( \Pi \) are shown in Figure 2.

Early models assumed that every site evolved at the same rate. However, we expect that functionally important regions of proteins change more slowly than unimportant ones. To accommodate rate variation across sites, we make the assumption (assumption 5) that relative rates of sites are distributed according to a \( \Gamma \)-distribution that can be described entirely by its shape parameter \( \alpha \)[9].

The various rate and substitution matrices published in the past 40 years differ not only in the methods of their estimation but also in the data sets from which they were estimated (see Table 1). Some have been estimated using data from only certain families of proteins in the expectation that these
Figure 2: R and Π, as shown by a bubble plot. Areas of circles are proportional to the exchangeabilities or equilibrium frequencies. The circles in the bottom left and top right of the figure provide scales for the equilibrium frequencies and exchangeabilities respectively. Values from the LG exchangeability matrix[1].
proteins are subject to different thermodynamic constraints on their structure and folding pathways. For example, phylogenies of mitochondrial proteins have greater likelihoods when estimated using an exchangeability matrix estimated from mitochondrial proteins [2, 18, 19].

We expect that a membrane-protein specific matrix would be useful for phylogenetic analysis of membrane proteins. Such a matrix, PHAT, has already been estimated, but is deficient in two regards[20]. First, a great deal more sequence data is available today than in 2000 when PHAT was estimated. Second, this paucity of data meant that the authors of PHAT had to use pseudocounts for some substitutions that were not observed at all (in all other respects, the estimation of PHAT was conducted analogously to that of BLOSUM[16, 20]). The result is that the rate matrix derived from PHAT has negative eigenvalues, which means that it cannot be exponentiated [D. Theobald, unpublished]. This, in turn, means that the PHAT matrix cannot be used to calculate phylogenies by ML or BI, although it remains useful in sequence alignments and homology searches[20].

In this study, we present a new substitution matrix for transmembrane proteins that generates ML and BI phylogenies with greater likelihood than does LG, the best soluble-protein matrix [1]. Our matrix, named KT after the authors, incorporates information from 45 families of transmembrane proteins, and improves likelihoods for trees not part of the training data set.

Matrices like KT and LG are intended to apply to every site of a sequence alignment. New methods improve upon this approach. Le and Gascuel have shown that accounting for structural features of proteins in phylogenetic in-
ference produces significant likelihood gains over general matrices [3]. Their best-performing model (EX_EHO_CONF_MIX) assigns each site in a sequence alignment one of six different structural categories according to a representative structure for each alignment. Likelihoods were calculated for each site using a rate matrix specific to that site’s category.

Unfortunately, the structural annotation scheme used by Le and Gascuel does not translate well to transmembrane proteins. The structural annotations used by Le and Gascuel treat all surface residues equally except with respect to their secondary structure. They do not account for exposure to a lipid bilayer, the defining feature of transmembrane proteins. Accounting for the bilayer was unnecessary for the soluble-protein sequences used to train EX_EHO_CONF_MIX. In addition, the software used by Le and Gascuel to make the annotations (DSSP) can not predict which residues might be exposed to the bilayer [3, 22].

We therefore also present a probabilistic method by which structures of transmembrane proteins can be annotated automatically such that their annotations may also be incorporated in future PhyML-Structure-like software. Our method incorporates both structural and evolutionary information, and produces reasonable annotations without any more input from the user than the structure itself.
Materials and Methods

Estimation of Rate Matrices

Data Sets

Representative X-ray crystallography structures, hereafter called “queries,” of $\alpha$-helical transmembrane proteins were selected from White’s membrane protein structure list [23]. Proteins were chosen for which we could confidently manually identify which residues were exposed to the lipid bilayer. For this reason, structures of SNAREs and ATP synthase were excluded, as well as any structures determined by NMR. While the transmembrane regions of photosystems and photosynthetic reaction centers were identifiable, many residues were exposed to chlorophyll or pheophytin rather than the lipid bilayer, so structures of these proteins were also excluded. The sequence identity of pairs of sequences was evaluated using MUSCLE[24].

In order to avoid redundant queries, primary structures were extracted from each potential query structure and aligned with each other. One in each pair of structures with $>40\%$ sequence identity was excluded.

In some cases where the structure of a protein could not be fully resolved and the authors were only able to model the amine of the terminal residue, that residue was removed from the primary sequence to resolve conflicts between other programs.
The remaining sequences were used to query the refseq[25] protein sequence database using BLAST[26]. A maximum of 2,000 sequences were collected for each query, with a threshold expectation \( E = \exp(-10) \). This expectation value was found to return homologues with relatively moderate sequence identity to the query but few proteins completely unrelated to the query.

Highly similar sequences contribute little information to the estimation of phylogenetic trees, but their inclusion increases the computation time as much as does the inclusion of any other sequence. Therefore, one in each pair of taxa with \( >95\% \) sequence identity to each other was removed, in order to ensure high-confidence alignments.

By contrast, sequences dissimilar to the query might be only very distant homologs, if related at all. Any sequence in a collection under 67\% of the length of the query or over 133\% of the length of the query was removed. Finally, any sequence with under 50\% identity to all other sequences was removed.

After having removed individual sequences in this way, the remaining sequences in each collection were aligned by MUSCLE[24].

**Inference of Phylogenetic Trees**

All phylogenetic trees of these data were generated by PhyML3[10]. Equilibrium frequencies were estimated from the frequencies of amino acids in the alignment. This approach produces higher-likelihood trees than would be made using the equilibrium frequencies given in the substitution model. Relative rates were assigned to one of six possible categories to approximate a \( \Gamma \)-distribution.
whose shape parameter $\alpha$ was estimated concurrently with the tree. The tree topology was searched using nearest-neighbor interchanges [7].

**Estimation of Single-Family Rate Matrices**

Rate matrices were estimated for each sequence alignment using *phylobayes* [27, 28]. The tree topology was held fixed to that of a tree estimated by *PhyML3* with the LG rate matrix for its substitution model. Relative rates were assigned using a continuous $\Gamma$ distribution. Two sets of parameters were estimated in parallel chains, and estimation was stopped when all relative differences of each parameter between the two chains (in this case, the branch lengths, $\alpha$, and each rate) were below 0.1. The relative differences were evaluated every 20 cycles after the first 100 cycles. The “burnin” factor was 1/5, meaning that only data from the last 4/5 of cycles was used to calculate the relative differences.

Once the chains were stopped, the expectation of the exchangeability matrix and the equilibrium frequencies were estimated using a conservative burnin factor of 1/2.

**Averaging of Single-Family Rate Matrices**

Each exchangeability matrix estimated by *phylobayes* incorporated information from only one sequence alignment. To produce a single exchangeability matrix generally applicable to $\alpha$-helical transmembrane proteins, we calculated an average matrix.

Each single-family exchangeability matrix $R$ was multiplied by its corresponding matrix of equilibrium frequencies $\Pi$ to produce a rate matrix $Q$ as in equation (9). Each rate matrix was then normalized such that the average
number of mutations per unit time was 1 and the matrix satisfied the condition that

\[- \text{trace}(Q\Pi) = 1 \quad (11)\]

Exchangeabilities were then re-calculated from each such rate matrix. Because each single-family matrix was not estimated with an equal amount of data, each matrix was given a weight \( w \) according to one of four possible schemes. Where \( N_k \) is the number of taxa in a training alignment \( k \) and let \( M_k \) is the number of sites in that alignment, weights \( w_k \) were assigned by the number of characters in that alignment (CHAR; \( w_k = N_k M_k \)), the square root of the number of characters (SQRT; \( w_k = \sqrt{N_k M_k} \)), the number of sites (COL; \( w_k = M_k \)), or equally, regardless of the size of the alignment (FLAT; \( w_k = 1 \)).

The harmonic mean of each entry in all matrices \( k \) with weights \( w_k \) was calculated as:

\[ r_{ij}^* = \frac{\sum_k w_k}{\sum_k w_k r_{ij,k}} \quad (12) \]

Background frequencies \( \pi_{i,k} \) were also averaged analogously.

**Annotation of Transmembrane Protein Structures**

**Estimation of Solvent Exposure**

The relative solvent accessibility (RSA) of each residue in each query structure was calculated with NACCESS[29] using a probe radius of 1.4 Å.

**Determination of Secondary Structure**

The secondary structure of all proteins was determined using DSSP[22] to analyze the bond angles of the polypeptide chain in each query structure. Because DSSP recognizes subtleties of helices that, in this case, only complicate
annotations, any residue whose neighbors in the protein primary sequence were both designated \(\alpha\)-helical was also considered \(\alpha\)-helical.

**Prediction of Transmembrane Helices**

The location of transmembrane helices in the sequence of each query was predicted by TMHMM2.0\(^{[30]}\). TMHMM2.0 uses a hidden Markov model to predict the location of transmembrane helices solely from sequence alignment data. Unlike DSSP, it attempts to distinguish between parts of helices that are buried in the lipid bilayer and parts that are not.

**Prediction of Transmembrane Regions by Nearest Surface Neighbors**

Six transmembrane protein structures chosen to represent diverse \(\alpha\)-helical transmembrane proteins were annotated by hand. These structures were bovine rhodopsin (RCSB id.: 1F88), *H. salinarum* bacteriorhodopsin (RCSB id.: 1QHJ), *A. aeolicus* LeuTaa (RCSB id.: 2A65), Human \(\beta\)-2 adrenergic receptor (RCSB id.: 2RH1), *C. glutamicum* BetP (RCSB id.: 2WIT) and *S. enterica* AdiC (RCSB id.: 3NCY).

Each residue was annotated “L” for “lipid-exposed,” “S” for “solvent-exposed” or “I” for “interior.” Using these annotations, the following probabilities were calculated for residues annotated “S” or “L” only:

- \(p(\text{Ala}, \text{Arg}, \ldots)\), the frequency of each amino acid
- \(p(L, S)\), the frequency of each non-“I” annotation
- \(p(L, S|\text{Ala}, \text{Arg}, \ldots)\), the probability of a non-“I” annotation given a type of amino acid
- \(p(\phi, \psi)|L, S)\), the probability of \(\phi\) hydrophobic neighbors or \(\psi\) hydrophilic neighbors given a non-“I” annotation.

Residues were considered neighbors if any atom in one residue was within 7 Å of another. MOLEMAN2\(^{[31]}\) was used to calculate inter-residue distances.
Residues were considered hydrophobic if \( \log \left( \frac{p(L|\text{a.a.})}{p(S|\text{a.a.})} \right) > 0.3 \) and hydrophilic if \( \log \left( \frac{p(L|\text{a.a.})}{p(S|\text{a.a.})} \right) < -0.3 \).

Keeping these probabilities fixed, a posterior odds ratio was calculated for the annotation of each residue, derived using Bayes’ equation:

\[
\frac{p(L|\phi, \psi, \text{a.a.})}{p(S|\phi, \psi, \text{a.a.})} = \frac{p(\phi|L)p(\psi|L)p(\text{a.a.}|L)p(L)}{p(\phi|S)p(\psi|S)p(\text{a.a.}|S)p(S)}
\]  

A shell script, `scoreres`, has been written to calculate these ratios for each residue in a structure.

**Annotation Decisions**

Any residue with relative sidechain accessibility (RSA) < 0.1 was annotated “I” (interior). Any non-I residue not considered \( \alpha \)-helical by DSSP was assumed to be outside the lipid bilayer, and annotated “S” (solvent-exposed). To determine which residues in the remaining \( \alpha \)-helices should be annotated “L” (lipid-exposed), the log-probabilities of belonging in a transmembrane helix provided by TMHMM and `scoreres` were summed. If the sum was > 0, the residue was annotated “L”. Otherwise, the residue was annotated “S.” These decisions are illustrated in Figure 3.
Figure 3: A flowchart showing the decisions made in structural annotations. Here, as elsewhere, RSA is the relative solvent exposure returned by NACCESS.
Results

Data Set

One hundred and four structures listed on White’s membrane protein structure list were eligible for inclusion in the training data after removing SNAREs, ATP synthase subunits, photosystems, NMR structures and one in each pair of highly similar structures. Of these, 45 were successfully aligned and annotated. These 45 alignments and structures form the data set for all remaining results. Their corresponding query proteins are listed in Appendix B.

The remaining 59 proteins were excluded from the database for a range of reasons. For some, the sequence alignment produced after removing highly similar or dissimilar sequences had fewer than 4 taxa, so no phylogeny could be produced. In other cases, some feature peculiar to the protein structure caused an error in DSSP, NACCESS, TMHMM, or scoreres, and structural annotations could not be produced. For example, the number of chain fragments in 1q90 prevented DSSP from producing any output.

Estimation of Rate Matrices.

An initial tree for each alignment was inferred using PhyML3 with the LG matrix. Holding the topology from this tree fixed, rate matrices were estimated for each individual protein alignment with phylobayes.

The harmonic mean of the single-alignment matrices was then found ac-
cording to one of the four weighting schemes (CHAR, SQRT, COL and FLAT).

Finally, the WAG matrix was also used to infer phylogenies using PhyML3.

Let $L$ be the log-likelihood of a tree generated from an alignment $k$ and $M_k$ and $N_k$ the number of sites and taxa in that alignment respectively. The mean log-likelihood gain per character of a tree over one estimated using LG

$$\frac{L_k - L_{LG,k}}{N_k M_k}$$

is shown in Figure 4(a). Each weighting scheme produced log-likelihood gains for some alignments and losses for others. A histogram for trees made using the overall best weighting scheme, SQRT, is shown in Figure 4(b). We will refer to this matrix as KT after the authors.

A bubbleplot of KT analogous to the one of LG in Figure 2 is shown in Figure 5. A bubble plot of the log-ratios of exchangeabilities between the two matrices is shown in Figure 6. As test of a sequence alignment not in our training data, trees were made of the TRPA1 sequence alignment used by Kang et al[6]. Trees were also generated of sites 432-986, which did not contain a large soluble domain found in the full-length alignment. Trees of these alignments made with LG and KT had identical topologies but different branch lengths. The tree of the full-length alignment generated with KT was 6.22 log-likelihood points worse than that generated with LG. By contrast, the tree made with KT and the truncated sequence alignment improved upon that made with LG and the truncated sequence alignment by 35.45 log-likelihood points. Conventionally, an improvement of 3-5 log-likelihood points is considered a significant improvement.
(a) Mean log-likelihood gain per site of trees over trees made with LG. Shown are the four different weighting schemes and the WAG matrix.

(b) Histogram of Log-likelihood gains per site using the KT matrix.

Figure 4: Log-likelihood differences between trees made with KT and those made with LG. Here, $M$ is the number of sequences in an alignment, $N$ is the number of taxa in that alignment, and $L$ is the log-likelihood of the tree generated from that alignment.
Figure 5: A bubble plot of $KT$. Areas of circles are proportional to the corresponding exchangeability or equilibrium frequency.
**Figure 6:** A bubble plot where areas of circles are proportional to $\log \frac{KT}{LG}$. Circles representing positive values, where the value in $KT$ is greater, are filled in black. Those representing negative values, for which the value in $LG$ is greater, are represented in red.
Prior Probabilities and Likelihoods of Annotations

The probabilities used to calculate the posterior odds ratio in equation (13) are shown in Figures 7 and 8. For the purpose of calculating the number of hydrophobic ($\phi$) and hydrophilic ($\psi$) neighbors of a residue, amino acids were considered hydrophobic, hydrophilic, or neutral. Amino acids for which $\log \left( \frac{p(L|a.a.)}{p(S|a.a.)} \right) > 0.3$, (Phe, Ile, Val, Trp, and Tyr) were considered hydrophobic. Those for which $\log \left( \frac{p(L|a.a.)}{p(S|a.a.)} \right) < -0.3$ (Asp, Glu, His, Lys, Asn, Pro, Gln, and Arg) were considered hydrophilic. The remaining six amino acids were considered neutral and did not count toward either $\phi$ or $\psi$ for a particular neighbor. This approach made the distributions of $p(\phi|S)$ and $p(\phi|L)$ distinct from each other and the distributions of $p(\psi|S)$ and $p(\psi|L)$ distinct from each other. Distributions of $p(\phi|S)$, $p(\phi|L)$, $p(\psi|S)$ and $p(\psi|L)$ are shown in Figure 8(a) and Figure 8(b).

Annotation of Transmembrane Protein Structures

The 45 query structures were annotated according to the scheme described in Figure 3. For each structure, any residues determined by NACCESS to have relative solvent accessibility (RSA) < 10% were annotated “I” for “interior.” Of the remaining residues, any not in $\alpha$-helices were annotated “S” for “solvent-exposed.” Each column in each sequence alignment derived from a query structure was given a probability of belonging to a transmembrane helix by TMHMM2.0. Each residue in a structure was also given a probability of being exposed to the lipid bilayer (and annotated “L”). The log-probabilities of belonging to a transmembrane helix and of exposure to the lipid bilayer were
Figure 7: The probability of a “solvent-exposed” (S, blue) or “lipid-exposed” (L, red) annotation for each amino acid. These probabilities were taken only from residues determined by NACCESS to be exposed, so all pairs of probabilities sum to \( \approx 1.0 \).
(a) The probability of \( \phi \) hydrophobic neighboring residues given annotations S (blue) or L (red).

(b) The probability of \( \psi \) hydrophilic neighboring residues given annotations S (blue) or L (red).

Figure 8: Probabilities of neighbor counts given annotations.
summed for each non-I residue. Those residues for which the sum was positive were annotated “L”, and the rest were annotated “S”.

A random sample of 16 structures is shown in Figure 9. Residues annotated “L” are colored red, those annotated “S” are colored blue, and those annotated “I” are colored gray. Some structures consist of multiple chains, and only unique transmembrane chains were annotated.
Figure 9: A sample of sequence annotations mapped on their query structures. Residues annotated “L” are colored in red, those annotated “S” are colored in blue, and those annotated “I” are colored gray. Many structures consisted of multiple chains, and only unique transmembrane chains were annotated.
Discussion

The standard measure of the fit of a phylogeny to its data is its log-likelihood \( L \), which is calculated as shown in equations (2) and (3). In creating a rate matrix for \( \alpha \)-helical transmembrane proteins, we expect that phylogenies of these proteins will have greater log-likelihoods when generated using our rate matrix than when generated using a more general one, such as \( LG \).

Trees of the 45 sequence alignments used to train the transmembrane protein matrices gain, on average, 0.0023 log-likelihood points per site when generated with \( KT \) rather than \( LG \). For a typical alignment with 267,800 characters, this corresponds to a log-likelihood gain of 53 points (see Figure 4). Again, a log-likelihood gain of 3-5 points is conventionally considered significant.

The \( KT \) matrix performs well using TRPA proteins as a test. For the truncated alignment, \( KT \) improves upon the log-likelihood of the TRPA tree by 0.002 points per site, which is close to the average improvement for a tree that was part of the training data. Moreover, we are given some confidence that the exchangeabilities in \( KT \) are specifically relevant to transmembrane domains. If the soluble domain of TRPA is included in the alignment, \( LG \), which was trained with data from both soluble and transmembrane proteins, improves upon \( KT \) by 6.8 log-likelihood points, or 0.0003 points per site.

Some alignments of transmembrane proteins in the training data still pro-
duce better trees with LG, which is expected (see Figure 4(b)). The training data of the LG matrix also occasionally produced better trees using WAG[1].

Our estimates of the exchangeabilities can be improved in several ways. Many alignments had to be excluded from our data set because they included large soluble domains. We expect that LG will reflect how these regions evolve better than any of the matrices that we estimated. We are currently developing a method to remove the soluble domains from the sequence alignments in an automated way to improve the “purity” of the data and to increase the amount of training data.

In addition, many sequence alignments were excluded because of errors in their annotations. The process outlined in Figure 3 requires that the output from several different programs in very different formats be combined to reach a decision. Non-biological features of crystal structures (e.g. selenomethionine, lysosome fusion structures) are handled variously by different programs. For example, DSSP counts the number of residues in a protein by the number of atoms named “CA” (for \(\alpha\)-carbon) in a PDB file, while NACCESS counts how many different residue numbers (RESNs) are given for each chain[22, 29]. When only part of a residue (such as the amine) is modelled, DSSP and NACCESS will disagree on the number of residues in a structure. These particular features of the programs have been identified and we now remove incomplete residues from our query structures. Other such discrepancies between programs remain to be identified and reconciled.

Bubbleplots of the LG matrix (shown in Figure 2) and of KT (in Figure 5) show that the matrices are qualitatively similar. In both LG and KT, amino
acids with similar chemical properties exchange with each other frequently. Figure 6 emphasizes the difference between the exchangeabilities of the two.

Two trends are clearly visible. The first is that serine exchanges much less frequently with other polar amino acids, and slightly more with valine and leucine. Serine is frequently found in the lipid-exposed regions of transmembrane proteins but less in the hydrophobic cores of soluble proteins, so this change is not wholly unexpected. The second is that the three uncharged aromatic residues exchange with each other much less in KT than in LG.

While some improvements to our method might improve a general matrix, Le and Gascuel [3] have shown that trees can be improved a great deal by using different matrices for different structural features. In order to incorporate structural information into sequence alignments of transmembrane proteins, we require some method of confidently annotating those sequence alignments. Here we have presented a means of annotating sequence alignments using structural and phylogenetic information.

Without experimentally determining whether every annotated residue is in fact buried in the lipid bilayer, our annotations are difficult to validate. However, the annotations are intuitively reasonable and are consistent with the biophysical properties of membrane proteins. As seen in Figure 9, the membrane exposed regions are all of the same depth – roughly 30 Å, the thickness of the lipid bilayer. No geometric information, such as the size of the lipid bilayer, was included in our method. It is an encouraging result of our annotation method that we recapitulate the bilayer dimensions. Admittedly, the width of the lipid-exposed regions are irregular, but transmembrane proteins are known
to remodel the lipid bilayer. Some of the structures shown in Figure 9 have odd shapes, such as lone helices that protrude into the bilayer at an angle. This is because only unique transmembrane chains of each of these proteins are shown, and in many cases members of the biologically relevant complex are not shown.

While both our general transmembrane protein matrix and our annotation schemes will be useful in their own right, they also provide groundwork for software that can implement structure-specific information in much the same way as PhyML-structure[3]. Such software would face two major problems in its realization. The first is the estimation of a single set of parameters from multiple unrelated gene families. We have shown how such a matrix may be estimated in the context of transmembrane proteins. The second is the determination of annotations that correspond to distinct patterns of structural evolution (and therefore distinct rate matrices). We have presented a method by which such annotations can be generated. Work is in progress to incorporate this structural information into phylogenetic inference of transmembrane proteins.
### Appendix A: List of Variables

<table>
<thead>
<tr>
<th>Description</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indices within a matrix (correspond to states)</td>
<td>$i, j$</td>
</tr>
<tr>
<td>Index of a matrix in an averaging scheme</td>
<td>$k$</td>
</tr>
<tr>
<td># model parameters (used in AIC)</td>
<td>$K$</td>
</tr>
<tr>
<td>Log-likelihood of a tree</td>
<td>$L$</td>
</tr>
<tr>
<td>Index of a site in an alignment</td>
<td>$m$</td>
</tr>
<tr>
<td># sites (=columns) in an alignment</td>
<td>$M$</td>
</tr>
<tr>
<td>Index of a sequence in an alignment</td>
<td>$n$</td>
</tr>
<tr>
<td># sequences in an alignment, taxa in a tree</td>
<td>$N$</td>
</tr>
<tr>
<td>Transition probability matrix</td>
<td>$P$</td>
</tr>
<tr>
<td>Rate matrix</td>
<td>$Q$</td>
</tr>
<tr>
<td>Exchangeability matrix</td>
<td>$R$</td>
</tr>
<tr>
<td># possible characters in a model</td>
<td>$S$</td>
</tr>
<tr>
<td>Topology</td>
<td>$\tau$</td>
</tr>
<tr>
<td>Branch length</td>
<td>$\nu$</td>
</tr>
<tr>
<td>Weight of a matrix</td>
<td>$w$</td>
</tr>
<tr>
<td>Sequence names</td>
<td>$x, y$</td>
</tr>
</tbody>
</table>
## Structures in Replacement Matrix Data Set

<table>
<thead>
<tr>
<th>RCSB id.</th>
<th>Chain(s)</th>
<th>Organism</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1ehk</td>
<td>A</td>
<td><em>Thermus thermophilus</em></td>
<td>Cytochrome-C oxidase</td>
</tr>
<tr>
<td>1f88</td>
<td>A</td>
<td><em>Bos taurus</em></td>
<td>Rhodopsin</td>
</tr>
<tr>
<td>1fft</td>
<td>A</td>
<td><em>Escherichia coli</em></td>
<td>Ubiquinol oxidase</td>
</tr>
<tr>
<td>1fx8</td>
<td>A</td>
<td><em>Escherichia coli</em></td>
<td>Glycerol facilitator</td>
</tr>
<tr>
<td>1h68</td>
<td>A</td>
<td><em>Natromononas Pharaonis</em></td>
<td>Sensory rhodopsin II</td>
</tr>
<tr>
<td>1l0v</td>
<td>C,D</td>
<td><em>Escherichia coli</em></td>
<td>Quinol-fumarate reductase</td>
</tr>
<tr>
<td>1l7v</td>
<td>A</td>
<td><em>Escherichia coli</em></td>
<td>ABC transporter</td>
</tr>
<tr>
<td>1nek</td>
<td>C</td>
<td><em>Escherichia coli</em></td>
<td>Succinate dehydrogenase</td>
</tr>
<tr>
<td>1pw4</td>
<td>A</td>
<td><em>Escherichia coli</em></td>
<td>Glycerol-3-phosphate transporter</td>
</tr>
<tr>
<td>1rc2</td>
<td>A</td>
<td><em>Escherichia coli</em></td>
<td>Aquaporin Z</td>
</tr>
<tr>
<td>1rhz</td>
<td>A</td>
<td><em>Methanococccus janaschii</em></td>
<td>Sec61/SecY</td>
</tr>
<tr>
<td>1u7g</td>
<td>A</td>
<td><em>Escherichia coli</em></td>
<td>Ammonia channel AmtB</td>
</tr>
<tr>
<td>1vgo</td>
<td>A</td>
<td><em>Halobacterium</em></td>
<td>Archaeorhodopsin-2</td>
</tr>
<tr>
<td>1xlo</td>
<td>A</td>
<td><em>Anabaena</em></td>
<td>Sensory rhodopsin</td>
</tr>
<tr>
<td>1z98</td>
<td>A</td>
<td><em>Spinacia oleracea</em></td>
<td>Aquaporin SoPIP2</td>
</tr>
<tr>
<td>1zcd</td>
<td>A</td>
<td><em>Escherichia coli</em></td>
<td>Na⁺/H⁺ antiporter</td>
</tr>
<tr>
<td>2a65</td>
<td>A</td>
<td><em>Aquifex aeolicus</em></td>
<td>Neurotransmitter homologue LeuT</td>
</tr>
<tr>
<td>2b2f</td>
<td>A</td>
<td><em>Archaeoglobus fulgatus</em></td>
<td>Ammonium transporter</td>
</tr>
<tr>
<td>2f2b</td>
<td>A</td>
<td><em>Methanothermobacter marburgensis</em></td>
<td>Aquaporin M</td>
</tr>
<tr>
<td>2gfp</td>
<td>A</td>
<td><em>Escherichia coli</em></td>
<td>Multidrug transporter EmrD</td>
</tr>
<tr>
<td>2m2q</td>
<td>A</td>
<td><em>Haemophilus influenzae</em></td>
<td>ABC transporter</td>
</tr>
<tr>
<td>2q7m</td>
<td>A</td>
<td><em>Homo sapiens</em></td>
<td>Flap</td>
</tr>
<tr>
<td>2r6g</td>
<td>F,G</td>
<td><em>Escherichia coli</em></td>
<td>Maltose transporter</td>
</tr>
<tr>
<td>2rh1</td>
<td>A</td>
<td><em>Homo sapiens</em></td>
<td>β-2 adrenergic GPCR</td>
</tr>
<tr>
<td>2w2e</td>
<td>A</td>
<td><em>Pichia pastoris</em></td>
<td>Aquaporin</td>
</tr>
<tr>
<td>2wit</td>
<td>A</td>
<td><em>Corynebacterium glutamicum</em></td>
<td>Na-coupled glycine/betaine symporter</td>
</tr>
<tr>
<td>2z73</td>
<td>A</td>
<td><em>Todarodes pacificus</em></td>
<td>Rhodopsin</td>
</tr>
<tr>
<td>2zjs</td>
<td>Y</td>
<td><em>Thermus thermophilus</em></td>
<td>SecY translocon</td>
</tr>
<tr>
<td>2zt9</td>
<td>A</td>
<td><em>Nostoc sp. pcc 7120</em></td>
<td>Cytochrome b6b complex</td>
</tr>
<tr>
<td>2zw3</td>
<td>A</td>
<td><em>Homo sapiens</em></td>
<td>Connexin-26 gap junction channel</td>
</tr>
<tr>
<td>3a7k</td>
<td>A</td>
<td><em>Natromonas pharaonis</em></td>
<td>Halorhodopsin</td>
</tr>
<tr>
<td>3b9y</td>
<td>A</td>
<td><em>Nistroosomas europa</em></td>
<td>Rh protein</td>
</tr>
<tr>
<td>3e02</td>
<td>A</td>
<td><em>Plasmodium falciparum</em></td>
<td>Aquaglyceroporin</td>
</tr>
<tr>
<td>3d1l</td>
<td>C</td>
<td><em>Methanosarcina acetivorans</em></td>
<td>ABC transporter</td>
</tr>
<tr>
<td>3d4a</td>
<td>A</td>
<td><em>Vibrio parahaemolyticus</em></td>
<td>Na⁺/Sugar symporter</td>
</tr>
<tr>
<td>3dh4</td>
<td>A</td>
<td><em>Escherichia coli</em></td>
<td>Methionine importer</td>
</tr>
<tr>
<td>3eml</td>
<td>A</td>
<td><em>Homo sapiens</em></td>
<td>Adenosine receptor</td>
</tr>
<tr>
<td>3gd8</td>
<td>A</td>
<td><em>Homo sapiens</em></td>
<td>Aquaporin 4</td>
</tr>
<tr>
<td>3gla</td>
<td>A</td>
<td><em>Methanococccus janaschii</em></td>
<td>Na⁺-independent amino acid transporter</td>
</tr>
<tr>
<td>3hd6</td>
<td>A</td>
<td><em>Homo sapiens</em></td>
<td>Rh Glycoprotein RhCG</td>
</tr>
<tr>
<td>3k3f</td>
<td>A</td>
<td><em>Desulfovibrio vulgaris</em></td>
<td>Urea transporter</td>
</tr>
<tr>
<td>3ncy</td>
<td>A</td>
<td><em>Salmonella enterica</em></td>
<td>Arg/agmatine antiporter</td>
</tr>
</tbody>
</table>

## Structures Annotated by Hand to Estimate Annotation Probabilities

<table>
<thead>
<tr>
<th>RCSB id.</th>
<th>Chain(s)</th>
<th>Organism</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1f88</td>
<td>A</td>
<td><em>Bos taurus</em></td>
<td>Rhodopsin</td>
</tr>
<tr>
<td>1qhv</td>
<td>A</td>
<td><em>Halobacterium salinarum</em></td>
<td>Bacteriorhodopsin</td>
</tr>
<tr>
<td>2a65</td>
<td>A</td>
<td><em>Aquifex aeolicus</em></td>
<td>Neurotransmitter homologue LeuT</td>
</tr>
<tr>
<td>2rh1</td>
<td>A</td>
<td><em>Homo sapiens</em></td>
<td>β-2 adrenergic GPCR</td>
</tr>
<tr>
<td>2wit</td>
<td>A</td>
<td><em>Corynebacterium glutamicum</em></td>
<td>Na-coupled glycine/betaine symporter</td>
</tr>
<tr>
<td>3ncy</td>
<td>A</td>
<td><em>Salmonella enterica</em></td>
<td>Arg/agmatine antiporter</td>
</tr>
</tbody>
</table>
References


