Using Protein Structural Information in Evolutionary Inference: Transmembrane Proteins

Pietro Liò and Nick Goldman*†*

*Department of Genetics and †Isaac Newton Institute for the Mathematical Sciences, University of Cambridge, Cambridge, U.K.

Corresponding author:

Nick Goldman

Department of Genetics

University of Cambridge

Downing Street

Cambridge CB2 3EH

UK

tel: + 44 - (0)1223 - 333981

fax: + 44 - (0)1223 - 333992

e-mail: N.Goldman@gen.cam.ac.uk

Key words: hidden Markov model, maximum likelihood, molecular evolution, phylogeny, protein structure, transmembrane proteins

Running head: Structure and Evolution of Transmembrane Proteins

Abstract

We present a model of amino acid sequence evolution, based on a hidden Markov model, that extends to transmembrane proteins previous methods that incorporate protein structural information into phylogenetics. Our model aims to give a better understanding of processes of molecular evolution, and to extract structural information from multiple alignments of transmembrane sequences and use such information to improve phylogenetic analyses. This should be of value in phylogenetic studies of transmembrane proteins; for example, mitochondrial proteins have acquired a special importance in phylogenetics and are mostly transmembrane proteins. The improvement in fit to example data sets of our new model relative to less complex models of amino acid sequence evolution is statistically tested. To further illustrate the potential utility of our method, phylogeny estimation is performed on primate CCR5 receptor sequences, sequences of 'l' and 'm' subunits of the light reaction centre in purple bacteria, guinea pig sequence with respect to lagomorph and rodent sequences of calcitonin receptor and K-substance receptor, and cetacean sequences of cytochrome-b.

Introduction

Recent phylogenetic analyses of DNA and protein sequences have been improved by incorporating structural and functional properties into inferential models. A first approach has considered information only indirectly related to structure, such as physicochemical properties (e.g. hydrophobicity, charge, size) of amino acids in mitochondrial proteins (Naylor and Brown 1997). Rzhetsky (1995) introduced a model that takes into account rRNA secondary structure elements, namely stem and loop regions, to estimate base substitution in ribosomal RNA genes and to infer phylogenetic relationships. Phylogenetic studies could gain great advantage from the comparison of tertiary structures of homologous proteins belonging to different species but it seems very unlikely that structural biologists will soon fill the gap between the explosion in protein sequences and the relatively slow speed at which experiments can reveal protein structures.

There are several reasons to use protein secondary structure information in evolutionary models. Selection pressure acts on protein function that, in turn, is closely related to structure. Hence, incorporating structure information into evolutionary analyses can assist in incorporating selective constraints. With respect to the primary structure, secondary structure of homologous proteins persists long after the statistically significant sequence similarity has vanished; sequences with 25% sequence amino acid identity are very likely to have the same secondary structure organization (Chothia and Lesk 1986; Russell et al. 1997). Secondary structure is expected to be more useful than tertiary structure in evolutionary studies since secondary structure elements are more conserved than the precise atomic structure (Mizuguchi and Go 1995): changes mainly occur at boundaries of α-helices and β-sheets, and replacements of hydrophobic residues in the core of the protein are usually accommodated by small shifts in secondary structure positions rather than compensatory

amino acid substitutions (Lesk and Clothia 1982; Heinz et al. 1993).

Probabilistic evolutionary models that incorporate structural information for globular proteins have been developed by Thorne, Goldman, and co-workers (Thorne, Goldman, and Jones 1996; Goldman, Thorne, and Jones 1996, 1998; Liò et al. 1998). Those models extract evolutionary and structural information contained in a multiply aligned set of homologous amino acid sequences, and use this information for both reconstructing phylogenies and predicting secondary structure of globular proteins. The models have two main features: Markov processes that describe amino acid replacements and a Markov chain that describes features of the secondary structure of globular proteins. The replacement models are related to those of Dayhoff and coworkers (Dayhoff, Eck, and Park 1972; Dayhoff, Schwartz, and Orcutt 1978) but whereas the Dayhoff approach simply considers the 'average' structural environment for each amino acid residue, Thorne, Goldman, and co-workers use a different Markov process model of amino acid replacement for each of the different categories of structural environment they want to describe. Underlying (but typically unobserved—hence 'hidden') transitions between the different categories along a protein-coding sequence are described with the Markov chain of structure. The resulting hidden Markov models (HMMs) permit the simultaneous inference of phylogeny and protein structure, allowing information about each to contribute to and improve inference of the other. For globular proteins, these models have been shown to fit real data considerably better than models which ignore structural constraints on evolution and treat all protein residues as equal and independent (Goldman, Thorne, and Jones 1998).

The model presented here, referred to as the TM126 model, represents a substantial extension of the models proposed by Thorne, Goldman, and Jones (1996) and Goldman, Thorne, and Jones (1996, 1998). We have focused our attention on transmembrane proteins. These constitute a very

large and important class of proteins, including for example mitochondrial proteins and membrane receptors. In particular, we note that mitochondrial proteins are widely used in phylogenetic analysis both for phylogeny assessment and evolutionary model testing. Thus, we think it is important to understand if the structure of these proteins should also be considered in phylogenetic analyses.

In the following sections, we describe in detail the components of the new TM126 model and their implementation to permit phylogenetic inference. We then apply the model to a number of example transmembrane protein alignments, both to estimate phylogenetic relationships and, more importantly, to evaluate statistically the improvement in fit to the data of the new model relative to other models.

Methods

Structural Categories

We have considered 10 different categories of structural environment: residues buried or exposed to solvent in globular α-helices (categories denoted H*b*, H*e* respectively), β-sheets (E*b*, E*e*), turns (T*b*, T*e*), and coil (C*b*, C*e*), and residues in transmembrane helices (TM) and short (typically less than 10 residues) cytoplasmic loops connecting two transmembrane domains (CL). The 10 categories seem to have different evolutionary dynamics, as described by their amino acid replacement models (see below), and so are expected to be useful in evaluating evolutionary and structural information.

The first 8 of these structural categories (all but TM and CL) characterize globular cytoplasmic or extracytoplasmic domains. These categories are identical to their counterparts in the '$4/2/38^{+}$' model of globular protein secondary structure described by Goldman, Thorne, and Jones (1998), to which the reader is referred for details. The remaining two structural categories, TM and CL, characterize the structure and topology

of transmembrane domains.

Hidden Markov Model

We adopt a HMM to describe the secondary structure state along a transmembrane protein amino acid sequence. The states of the model correspond to the underlying but unobserved ('hidden') topology, each state representing residues belonging to one of the 10 structural categories. Transitions among the states are modelled with a Markov chain. We have designed our model to be a probabilistic description of the secondary structure organization of a wide variety of transmembrane proteins, such as receptors and integral membrane proteins. The way in which states identified with the 10 structural categories are connected into a HMM represents our understanding of the typical or likely secondary structures of transmembrane proteins. We start the description of the HMM by indicating how multiple hidden states identified with the same structural categories are connected to model local structures (α-helices, β-sheets, turns, coils, transmembrane helices, short cytoplasmic loops). As shown by Goldman, Thorne, and Jones (1998), this can be useful in order to describe better the typical lengths of secondary structure elements.

As in that study, α-helices were modelled by concatenating 10 linked pairs of Hb and He states in a manner which ensured that α-helices are of length at least four residues and have a distribution of lengths that closely matches the empirically observed distribution. Also as in Goldman, Thorne, and Jones (1998), β-sheets were modelled with 6 pairs of Eb and Ee states (minumum sheet length of two residues), turns with two pairs of Tb and Te states (minimum turn length two residues) and coils with one pair of Cb and Ce states (no minimum length).

A choice of 24 TM states to represent transmembrane helices was made with a likelihood ratio testing procedure that took into account the number of parameters being estimated and the improvement in goodness of fit as

the number of position-specific states increased, as in Goldman, Thorne, and Jones (1998). In the following we illustrate the organization of these 24 TM states. We denote the ith of these states by TM_i, for $i \in \{1, 2, \ldots, 24\}$. We constrain the HMM to enter a transmembrane helix only through the TM_1 state. Once in state TM_i, for $i \in \{1, 2, \ldots, 23\}$, the HMM must continue by progressing to either TM_{i+1} (with probability $1 - p_i^{\mathrm{TM}}$) or by leaving the TM states and entering a different category (with probability p_i^{TM}). Once in the state TM_{24}, the HMM can remain there (with probability $1 - p_{24}^{\mathrm{TM}}$) or can leave the transmembrane helix (with probability p_{24}^{TM}). Taking into account the observed distribution of transmembrane helix lengths, the model enforces a minimum length of 14 states for a transmembrane helix (by setting $p_1^{\mathrm{TM}} = p_2^{\mathrm{TM}} = \ldots = p_{13}^{\mathrm{TM}} = 0$), and exactly matches the observed and expected distributions for lengths from 14 to 24 (by appropriate choice of p_{14}^{TM}, p_{15}^{TM}, ..., p_{24}^{TM}). This arrangement of 24 TM states is illustrated in figure 1.

The CL category describes residues in short cytoplasmic loops (their length is generally less than 10 residues) that often join successive transmembrane helices and have high propensity for positive amino acids. This category is important mainly for topology assignment, because coil and turn states in globular domains can also describe loops. Because of steric constraints on the shortest loop which could join two transmembrane helices, we used two CL states to represent cytoplasmic loops, which were assigned a minimum length of two residues.

The majority of transmembrane proteins have more than one transmembrane domain; a large proportion have 6 or 7 transmembrane domains, generally connected by short loops. At the proteins' extremities, but also intercalating the transmembrane domains in large proteins (e.g. the majority of receptor proteins), there can be globular domains: in large proteins containing both globular and transmembrane domains, the numbers of transmembrane helices and α-helices in globular domains can

be of the same order. The local structures described above are now linked in order to make a general model describing the structure of a wide variety of transmembrane proteins. Figure 2 also indicates how this is done.

We model a transmembrane protein as having one or more transmembrane helices (24 TM states) spanning the membrane, interspersed with extra- or intra-cytoplasmic regions. Extra-cytoplasmic regions are assumed to conform to the '$4/2/38^{+}$' HMM for globular proteins described by Goldman, Thorne, and Jones (1998) (a total of $10{\times}2 + 6{\times}2 + 2{\times}2 + 2 = 38$ hidden states). Intra-cytoplasmic regions can conform to this model (a further 38 states), or can adopt a short cytoplasmic loop conformation (two states). The prevalence of positively charged residues (Arg and Lys) in short cytoplasmic loops is known as the 'positive inside' rule (Gavel and von Heijne 1992; von Heijne 1992). This pattern may assist in the assessment of protein topology, for example the direction of insertion of transmembrane helices.

Two copies of each of the $4/2/38^{+}$ globular protein and transmembrane helix structures are needed to maintain appropriate directionality (see fig. 2), e.g. to forbid inadmissable sequences such as cytoplasmic loop–transmembrane helix–cytoplasmic loop. The complete model consequently has a total of 126 states ($24{\times}2 + 38{\times}2 + 2$), and we refer to it as the TM126 model. We emphasise that our model requires that all sites with a particular secondary structure (H, E, T, C, TM, CL) and accessibility status (*b* or *e* suffix for states H, E, T, C) experience the same amino acid replacement process, regardless of relative position within their secondary structure element. Thus, each of the 126 HMM states corresponds to a particular one of the 10 amino acid replacement categories.

To the best of our knowledge, our TM126 model is the first that attempts to describe the full variety of transmembrane protein structures, including transmembrane helices, short cytoplasmic loops, and the possible existence of extra- or intra-cytoplasmic globular regions.

HMM Parameter Estimation

The most natural way to estimate transition probabilities (ρ_{ij}, the probability that a residue is in hidden state j given that the preceding residue in in state i) among the 126 HMM states is to examine sequences of known structure, count how many times a site in state i is followed by a site in state j, and divide this count by the number of times sites in state i are followed by any site. Previous experience with globular proteins has suggested that we have insufficient data to make reliable estimates of all (up to $126^2 = 15876$) transition probabilities between hidden states. To reduce the number of parameters, we have made simplifying assumptions in the manner of Goldman, Thorne, and Jones (1998).

To estimate transition probabilities among states belonging to globular domains we use exactly the $4/2/38^+$ model of Goldman, Thorne, and Jones (1998), except for minor modifications where the globular domains meet a transmembrane domain (see below). Transition probabilities within transmembrane helices and short cytoplasmic loops have been described above. Finally, transition probabilities between these regions are estimated subject to the following simplifications:

- The transition from a globular domain to a transmembrane domain (and *vice versa*) occurs only through a coil state (C*b* or C*e*).
- After exploratory tests comparing predicted and actual structures, we assign to the transition from coil in a globular extra-cytoplasmic domain to a transmembrane domain the same probability as that of moving from coil to globular α-helices. This probability is divided equally between C*b* and C*e* states. Because transmembrane domains have strong signals (e.g. a stretch of hydrophobic residues), we note that even small, underestimated probabilities of transition to these domains appear sufficient for their detection.

- Analysis of the flanking regions of transmembrane domains revealed that the 'positive inside' rule is not ubiquitous. We assign values $\frac{1}{4}$, $\frac{3}{8}$, $\frac{3}{8}$ respectively to the probabilities, conditional on leaving a transmembrane helix, of moving to a short positive loop (CL) or the C*b* or C*e* states of a cytoplasmic globular domain. Although these values are probably not correct (in 80% of cases contiguous transmembrane domains are connected by short loops), we found that by giving such a high probability to entering a cytoplasmic globular domain we could increase the sensitivity of the model for detecting structured cytoplasmic domains. Relative to the $4/2/38^{+}$ model, we thus adjusted the coil transition probabilities in order to have the same probabilities of moving from coil to α-helix, β-sheet and turn in both the extra-cytoplasmic and cytoplasmic globular domains.

- The sum of the probabilities of moving from the cytoplasmic globular coil and CL states to a transmembrane helix is set equal to the probability of moving from the extra-cytoplasmic coil states to a transmembrane helix.

- We set the probability of moving from the membrane to the extra-cytoplasmic globular region equal to the sum of the probabilities of moving from a transmembrane helix to the CL or cytoplasmic globular coil states. The C*b* and C*e* coil states are reached with equal probability.

The HMM transition probabilities are conveniently considered as a 126×126 matrix (ρ), although the structure of the TM126 model and the restrictions detailed above mean that $\rho_{ij} = 0$ for many pairs (i,j). The matrix ρ is available from the authors on request.

Amino Acid Replacement Models

Our models of amino acid replacement are Markovian with respect to

time; each of the 10 secondary structure categories ($k \in \{$TM, CL, Hb, He, Eb, Ee, Tb, Te, Cb, C$e\}$) is associated with a different amino acid replacement process. The process for category k is specified by parameters α_{ij}^k, the rates of instantaneous change from amino acid i to j within category k. Writing $\alpha_{ii}^k = -\sum_{j \neq i} \alpha_{ij}^k$, the parameters α_{ij}^k are conveniently written as matrices α^k which define amino acid replacement probabilities in the standard manner, according to $P^k(t) = \exp(t\alpha^k)$ (Liò and Goldman 1998).

For the globular protein categories ($k \in \{$Hb, He, Eb, Ee, Tb, Te, Cb, C$e\}$), the α^k used are those described by Goldman, Thorne, and Jones (1998), estimated from the BRKALN database (see below) using a method that is a slight modification (Jones, Taylor, and Thornton 1992; Goldman, Thorne, and Jones 1996) of the approach of Dayhoff and collaborators (Dayhoff, Eck, and Park 1972; Dayhoff, Schwartz, and Orcutt 1978). For the TM and CL categories, amino acid replacement rates (α^{TM} and α^{CL}) were estimated by using exactly the procedure of Goldman, Thorne, and Jones (1996), applied to the TMALN database described below.

Two issues arise regarding scaling of the matrices α^k. The first relates to the derivation of the α^k from two different databases (BRKALN and TMALN, described below) which differ in their values of an otherwise arbitrary constant ($\sum_{m=1}^{M} N_m t_m / \sum_{m=1}^{M} N_m$, as described by Goldman, Thorne, and Jones 1996: eq. 3). Fortunately, it is possible to relate the two databases by comparing the combined Cb and Ce categories (BRKALN) with short extra-cytoplasmic regions neighbouring transmembrane helices in TMALN. We find good correspondence between amino acid replacement rates for these two categories, subject to an overall scaling factor representing the ratio of the two databases' values of $\sum_{m=1}^{M} N_m t_m / \sum_{m=1}^{M} N_m$ (results not shown). This scaling factor can subsequently be applied to normalize the matrices α^{TM} and α^{CL} to make them comparable with the matrices corresponding to the 8 globular protein categories.

Secondly, as in Goldman, Thorne, and Jones (1996, 1998), the stationary distribution (Ψ) of the HMM transition matrix ρ is used to scale all of the matrices α^k so that the mean rate of amino acid replacement, with each structure category weighted according to its expected frequencies of appearance, equals 1. This effectively means that branch lengths are measured in expected numbers of replacements per residue, averaged over all structure categories.

Data Sets Used for Model Parameter Estimation

For the components of the HMM derived from the globular protein model of Goldman, Thorne, and Jones (1998), amino acid replacement rates and HMM transition probabilities were estimated from the BRKALN database described in that paper. The BRKALN database consists of 207 families of easily aligned globular protein sequences. The tertiary structure of at least one member of each protein family has been experimentally determined and secondary structure assignments are then made using the DSSP program (Kabsch and Sander 1983).

Parameters of the new model specific to transmembrane proteins were estimated as described above from the TMALN database (P. Liò, unpubl. results). This database consists of 181 families of easily aligned transmembrane domains sequences. Only a few complete transmembrane tertiary structures are known: problems with their determination include these proteins' large size and the hydrophobicity of their membrane spanning regions. Nevertheless there are experimental methods that allow investigation of which residues are buried in the membrane (see for example Ben-Efraim, Bach, and Shai 1993; Jones et al. 1996; Spruijt et al. 1996). In the TMALN database, each transmembrane domain sequence is recorded together with the two flanking regions consituted by a number of residues (10 or more) belonging to intra- or extra-cytoplasmic loop regions.

The BRKALN and TMALN databases are disjoint, but not completely

unrelated: as noted above, the extra-cytoplasmic flanking regions in TMALN are comparable with the globular protein coil category (considering buried and exposed coils together) of the BRKALN database.

Likelihood Calculations

The amino acid replacement rates (α^k) and the HMM transition probabilities (ρ) estimated from the TMALN and BRKALN databases as described above are fixed for all subsequent analyses, and are assumed representative of all transmembrane proteins. The secondary structure of the particular transmembrane protein under study is assumed to be unknown, and the phylogenetic tree relating a set of sequences under study is assumed unknown and is estimated by maximum likelihood (ML) methods. We calculate the likelihood of a candidate phylogenetic tree T (representing both topology and branch lengths) as follows. We denote the aligned data set by S, its length (number of amino acids) by N, the first i columns of the data set by S_i, and the ith column itself by s_i. Gaps in the alignments are considered as missing information, as in the ML programs of the PHYLIP package (Felsenstein 1995). In the equations below many of the probabilities are actually conditional upon the α^k and ρ, but for the sake of clarity we omit α^k and ρ when this is feasible. The likelihood of the tree T is given by $\Pr(S \mid T)$, and this is calculated via the terms $\Pr(S_i, c_i \mid T)$ for each possible HMM state c_i at site i using the iteration:

$$\Pr(S_i, c_i \mid T) = \sum_{c_{i-1}} \Pr(S_{i-1}, c_{i-1} \mid T)\, \rho_{c_{i-1}c_i} \Pr(s_i \mid c_i, T) \qquad (1)$$

for $i > 1$.

The terms $\Pr(s_i|c_i, T)$ are evaluated using the Markov process replacement models (defined by the matrix α^{c_i}) appropriate for each secondary structure c_i and the 'pruning' algorithm of Felsenstein (1981). Because the site at the N-terminus tends to be exposed coil (Ce), we

assume this is the case and start the iteration according to:

$$\Pr(S_1, c_1 \mid T) = \Pr(s_1 \mid c_1, T) \cdot \frac{1}{2}\delta_{c_1,\mathrm{Ce}}, \tag{2}$$

where $\delta_{c_1,\mathrm{Ce}} = 1$ if c_1 is either of the two Ce states (exposed coil in cytoplasmic or extra-cytoplasmic globular domains) in the 126-state HMM, and 0 otherwise. This form assumes that the cytoplasmic and extra-cytoplasmic exposed coil states are equally likely for the N-terminal residue; other possibilities (Rost, Fariselli, and Casadio 1996) can easily be accomodated as required. When completed, the iteration gives the required $\Pr(S \mid T)$ because

$$\Pr(S \mid T) = \sum_{c_N} \Pr(S_N, c_N \mid T). \tag{3}$$

Numerical optimization routines are used to find the ML tree topology and branch lengths ($\widehat{T}$). Calculation of *a posteriori* probabilities of HMM states for each site of the protein, $\Pr(c_i|S, \widehat{T})$, allows prediction of the secondary structure for each site, as described by Goldman, Thorne, and Jones (1996).

Statistical Tests

ML methods evaluate competing hypotheses (trees, models and parameter values) by selecting those with the highest likelihood, as it is these which render the observed data most plausible. Likelihood also provides a natural means of hypothesis testing. Suitable choice of null and alternative hypotheses, for example differing in their model of sequence evolution, and comparisons of ML scores under these hypotheses in relation to the distribution expected under the null hypothesis, permit a statistical assessment of the hypotheses tested. Likelihood ratio tests (LRTs) are a class of powerful statistical tests that compare the ML values of competing hypotheses and have proven useful in phylogenetics (Goldman 1993; Yang, Goldman, and Friday 1994; Huelsenbeck and Rannala 1997).

To better understand the extent to which incorporation of structural information is responsible for improved fits of model to data, we use a LRT with test statistic Δl calculated as

$$\Delta l = \hat{l}_{\text{TM126}} - \hat{l}_0 \tag{4}$$

where $\hat{l}_{\text{TM126}}$ is the maximum log-likelihood for the alternative hypothesis, i.e. our new model of transmembrane protein evolution, and $\hat{l}_0$ is the maximum log-likelihood for the null hypothesis, e.g. some other model that does not use structural information.

In the absence of reliable asymptotic results, we can use simulations to estimate the distribution of Δl under the null hypothesis (Goldman 1993). The null model of sequence evolution is used, along with the maximum likelihood topology and branch lengths ($\widehat{T}_0$) estimated under the null hypothesis for a data set, to generate each simulated data set. The simulated data sets have the same number of taxa and are the same length as the original data set. For each simulated data set, e.g. $i \in \{1, 2, \ldots, 100\}$, Δl_i can be calculated via likelihood maximization under the null and alternative hypotheses. Further details are given by Goldman, Thorne, and Jones (1998).

If the value of Δl observed for the original data set is sufficiently extreme relative to the distribution of simulated values $\{\Delta l_i\}$, then the null hypothesis can be rejected in favour of the TM126 model. One measure of extremity is the proportion of simulated test statistic values that exceed the actual value, giving an estimated P-value for the observed Δl. Sufficiently low values (e.g. ≤ 0.05) imply rejection of the null hypothesis in favour of TM126. Since often we find very low P-values (e.g. ≤ 0.01, meaning none of 100 values of Δl_i exceed Δl), we have also found a z-score to be a useful measure of extremity. The z-score is calculated by subtracting the mean of the Δl_i from the observed value Δl and then dividing by the sample standard deviation of the Δl_i. This gives a rough estimate of how many standard deviations the observed test statistic is from the mean of the

expected distribution. Assuming an approximately Normal distribution of Δl under H_0, then for example a z-score of 2.33 corresponds to a P-value of 0.01, $z \geq 3.09$ corresponds to $P \leq 0.001$, etc. (one-sided test). For further details of these tests, see Goldman, Thorne, and Jones (1998).

Results and Discussion

Primate HIV Co-receptor CCR5 Phylogeny

The chemokine receptor type 5 (CCR5) is a G-coupled receptor that transduces a signal by increasing the level of intracellular calcium ions. In HIV-infected cells, CCR5 acts as a co-receptor with CD4 for the envelope glycoproteins of HIV. A number of intra- and inter-species studies have established its importance in HIV and SIV transmission (e.g. Deng et al. 1996; Dragic et al. 1996; Kuhmann et al. 1997; Zhang et al. 1997). The topology of CCR5 is known and contains seven transmembrane domains. We have evaluated the phylogenetic information provided by CCR5 amino acid sequences in the following species: mouse, chimpanzee, gorilla, human, rhesus monkey, pig-tailed macaque, baboon, African green monkey, sooty mangabey. The multiple alignment of these 9 sequences comprised 354 amino acid residues. Figure 3a shows the ML estimate of phylogeny derived from our HMM method using the TM126 model for this set of CCR5 homologs; table 1 gives the maximum log-likelihood for this tree. The resulting phylogeny reflects how close the green monkey co-receptors are to those of the old world monkeys and thus it concurs with the similarity of evolution of the simian and human immunodeficiency viruses (see also Kuhmann et al. 1997).

In order for our model to be able to improve over models which do not incorporate knowledge of transmembrane protein structure, it is evident that it must be able at least to recognize the major secondary structure elements of transmembrane proteins. To demonstrate that this is the case, figure 3b shows a graphical representation of the observed locations of the

seven transmembrane domains of the CCR5 (human) protein (see SWISS-PROT P51681) and a comparison between our model and several others currently used for transmembrane protein prediction. This shows that our model correctly predicts the locations of all the transmembrane helices. Of the four methods that correctly identify all seven transmembrane helices, three (TM126, PHDhtm and HMMTOP; see fig. 3b for details) use multiple aligned sequences. We believe that the results of the comparison suggest that programs using multiple aligned sequences seem to perform generally better than programs using single sequences.

Horizontal Gene Transfer: Additional Insights from Secondary Structure

Purple bacteria (proteobacteria) can be classified into three major groups, the α, β and γ classes; the α class is further classified into four subclasses, α-1, α-2, α-3 and α-4 (Woese 1987). Both 16S rRNA- and cytochrome-c-based phylogenetic analyses show a clear division between the α subclasses and the β and γ classes. Using nucleotide sequences coding for the 'l' and the 'm' subunits of the light reaction centre, Nagashima et al. (1997a) found discordances with the rRNA and cytochrome-c phylogenies, showing that the γ class clusters with the α-1 subclass and the β class clusters with the α-2 subclass. They suggested that these discordances could be explained by assuming a horizontal transfer of genes coding for the photosyntethic reaction centre among purple bacteria.

We evaluated the phylogenetic and structural information provided by the amino acid sequence of subunits 'l' and 'm' of the light reaction centre in the following 7 purple photosyntethic bacteria: *Allochromatium vinosum* (γ class), *Rhodobacter sphaeroides* (α-3 subclass), *Rubrivivax gelatinosus* (β), *Rhodospirillum molischianum* (α-1), *Erythrobacter longus* (α-4), *Rhodomicrobium vannielii* (α-2) and *Acidophilium rubrum* (α-1 subclass, based on 16S rRNA phylogeny, but not forming a tight cluster (Woese et al. 1984; Woese 1987; Lane et al. 1992; Nagashima et al. 1997b). While all

known photosynthetic organisms have a chlorophyll complexed with a magnesium (Mg) atom, *A. rubrum* has a zinc (Zn)-containing bacteriochlorophyll as its major photosynthetic pigment (Wakao et al. 1996; Nagashima et al. 1997b).

Our estimated phylogenies are shown in figures 4a and 4b (maximum log-likelihood values in table 1). We found that the 'm' subunit tree has the same topology obtained by Nagashima et al. (1997a), while the 'l' subunit tree shows the γ class closer to the α-3 subclass than to α-1. *A. rubrum* is clustered closer to the α-4 subclass in the 'l' subunit phylogeny and closer to the α-2 subclass in the 'm' subunit phylogeny.

The fact that the trees for the 'l' and 'm' subunits are different suggests that the related genes have been subjected to different evolutionary events. It is known that the Zn-containing bacteriochlorophyll binds the 'l' and 'h' subunits and not the 'm' subunit (Nagashima et al. 1997b). *A. rubrum* contains Glu instead of His at position 168 of the 'l' subunit; this position is known to interact with the chromophore. Probably, because the Zn-chlorophyll binds the 'l' subunit and not the 'm' subunit, *A. rubrum* can exchange the 'm' subunit gene with Mg-chlorophyll purple bacteria more easily than the 'l' subunit gene. Thus differences in phylogenies constructed using 'l' and 'm' subunit amino acid sequences may suggest that these genes not only underwent horizontal gene transfer but also that the transfer may have occurred through independent recombination and transfer events. Branch lengths in the 'l' subunit tree are about four times greater than branch lengths in the 'm' subunit tree. This is in good agreement with the hypothesis of a larger degree of exchange of the 'm' subunit gene than of the 'l' subunit gene among purple bacteria. The high sequence identity among the 'm' and 'l' subunit amino acid sequences in related species could be both the result of common stringent selection control and of recombination events. Thus the fact that these genes are very conserved among the purple bacteria also suggests that exchange of genetic material

through recombination could have happened several times in evolution. Recent work shows that horizontal transfer of rRNA genes is not precluded (Asai et al. 1999).

The Guinea Pig Challenge

The phylogenetic position of the guinea pig and the monophyly of the rodents is a matter of debate. Different markers and different methods have led to different conclusions (see, for instance, Hervé 1997 and references therein). We estimated the phylogenetic position of guinea pig with respect Lagomorpha and Rodentia from the amino acid sequences of the calcitonin receptor and K-substance receptor (neurokinin-2 receptor).

The calcitonin and K-substance receptor genes are both G proteins with seven transmembrane domains. The proteins have low sequence identity, are functionally unrelated and do not interact each other, and are preferentially expressed in different cells and tissues. The K-substance receptor gene is intronless and is located on chromosome 10 in humans; the calcitonin receptor gene contains introns and in humans is located on chromosome 7 (although several paralogous genes are known to exist). Therefore, these genes have evidently been subject to different mutation events and different structural and functional selection pressures, and give completely independent information for phylogenetic analyses. Completely unrelated sequences, or at least sequences that have very little evolutionary history in common, may represent the best choice for phylogenetic analysis based on a multi-sequence approach.

Figure 4c shows the ML estimate of phylogeny among human, rabbit, guinea pig, pig and rat for the calcitonin receptor sequences, and figure 4d the ML phylogeny for human, rabbit, guinea pig, hamster and rat from K-substance receptor sequences, estimated using our new TM126 model. Tree topologies and branch lengths are similar for these proteins; the guinea pig is placed among the rodents, in general agreement with other ML

results (e.g. Cao, Okada, and Hasegawa 1997). Maximum log-likelihood values are shown in table 1. Simultaneously with phylogeny estimation, our method predicts seven transmembrane domains, in agreement with experimental evidence (results not shown).

Cetaceans in a Sea of Trees

The relationships among the major group of cetaceans is still matter of controversy. The analysis of the cytochrome-b gene sequence using parsimony methods allowed Arnason and Gullberg (1994) to distinguish five primary evolutionary lineages of extant cetaceans, one representing the baleen whales (Mysticeti, denoted M) and four (Platanistoidea, Physeteroidea, Ziphioidea and Delphinida, respectively denoted Pl, Ph, Z and D) representing the toothed whales (Odontoceti). They proposed the evolutionary relationship (M,(Pl,(Z,(Ph,D)))), and suggested that the cetacean lineages had diverged almost simultaneously.

Hasegawa, Adachi, and Milinkovitch (1997) illustrated that often the analysis of a single gene does not resolve ambiguity of phylogenetic relationships. For instance, for cetaceans their analysis of myoglobin amino acid sequence supported the branching order (D,((Z,M),Ph)), whereas analysis of cytochrome-b amino acid sequence supported (D,((Z,Ph),M)). Using ML methods, Hasegawa, Adachi, and Milinkovitch (1997) established an alternative phylogeny of cetaceans, (D,(Z,(M,Ph))), on the basis of the total maximum likelihood given by the combined analyses of myoglobin amino acid sequences, 12S and 16S rRNA sequences, and either amino acid or DNA sequences of cytochrome-b.

We analyzed a data set of amino acid sequences of cytochrome-b, a mitochondrial membrane protein. Analysis under the TM126 model of 8 sequences from spinner dolphin (D), bottlenose dolphin (D), gray whale (M), minke whale (M), Peruvian beaked whale (Z), Gervais beaked whale (Z), pigmy sperm whale (Ph) and sperm whale (Ph), suggested that

Ziphioidea and Physeteroidea are not monophyletic (fig. 4e). Our phylogeny is in agreement with Hasegawa, Adachi, and Milinkovitch's (1997) analysis of cytochrome-b amino acid sequences, other than regarding the monophyly of the Ziphioidea and Physeteroidea.

Note that we have used sequences different from the ones used by Hasegawa, Adachi, and Milinkovitch (1997). Also, since we only have available incomplete cytochrome-b sequences for Peruvian beaked whale and Gervais beaked whale (134 amino acids each, cf. 379 amino acids for spinner dolphin) this is possibly a source of error in the assignment of the Ziphioidea and Physeteroidea divergence. Analyses with more sequences from the Ziphioidea and Physeteroidea are needed to resolve this issue. The cetacean controversy shows the unsatisfactory results that phylogenetic methods sometimes present, even when, as in the relationships among cetaceans, there are large morphological differences that indicate that time intervals separating the species cannot have been short.

Statistical Tests

We performed statistical tests comparing the goodness of fit of the new transmembrane protein HMM (TM126) with two other models of evolutionary change in amino acid sequences. These models are the following:

- JTT model: this is the model of Jones, Taylor, and Thornton (1992), which is based on the analysis of a large database of globular protein amino acid sequences. It is essentially an update of the models of Dayhoff, Eck, and Park (1972) and Dayhoff, Schwartz, and Orcutt (1978), and incorporates no structural information: each residue of sequences analyzed is assumed to conform to some 'average' evolutionary dynamics. Such a model is evidently inappropriate for the analysis of transmembrane proteins, but is included here as a

convenient baseline and because it is a widely used model which, in the absence of structurally better-informed models, has in the past been applied to transmembrane proteins.

- TmHEM model: this is an alternative model we have devised for transmembrane protein evolution. It incorporates no structural information, but represents the average dynamics of amino acid replacements in transmembrane proteins by a single instantaneous rate matrix (α^{TmHEM}) formed by averaging the 10 structure-specific matrices (α^k) of the TM126 model. To be precise,

$$\alpha^{\mathrm{TmHEM}} = \sum_k w_k \alpha^k \tag{5}$$

where the weights w_k are given by

$$w_k = \sum_{i=1}^{126} \Psi_i \cdot \mathrm{I}(\text{HMM state } i \text{ corresponds to structure category } k), \tag{6}$$

Ψ is the stationary distribution of ρ, and the indicator function $\mathrm{I}(\cdot)$ equals 1 if the statement in parentheses is true, and equals 0 otherwise. The TmHEM model is a transmembrane analogue of the JTT model for globular proteins. Since the TM126 and TmHEM models were derived from the same sequence databases, but with only TM126 making use of structural information, comparisons of these models test the significance of the representation of transmembrane protein structure in the TM126 model. By analogy with the 'THEM' model of Thorne, Goldman, and Jones (1996), devised for a similar comparison of globular protein models, TmHEM stands for 'TransMembrane Homogeneous Evolutionary Model'.

The statistical comparisons between TM126 and these two models were designed to test the importance of the use of both transmembrane protein structural information, as embodied in the 126-state HMM described above, and of different evolutionary models of amino acid replacement dynamics

depending on structural context. These tests mirror those of Thorne, Goldman and Jones (1996) for testing HMMs of globular protein evolution. Tests of the TM126 model versus the JTT model and of TM126 vs. TmHEM were performed for all six data sets described above, and the results are presented in table 1.

The comparison between the TM126 and JTT models clearly indicates the superiority of the former. In all cases, P-values are less than 0.01, and the smallest z-score is 3.43 (approximately corresponding to $P = 0.0003$). These results are not surprising; the JTT model was devised for application in circumstances very different from the analysis of transmembrane proteins.

Similarly, the comparisons of the TM126 and TmHEM models indicate the superiority of TM126. The largest P-value is 0.03; all others are less than 0.01 and have z-scores of at least 2.83 (approximately, $P \leq 0.002$). It appears that the new HMM is incorporating appropriate and useful information about the constraints on sequence evolution imposed by the maintainance of transmembrane protein structure, relative to a model based on average transmembrane protein amino acid replacement dynamics but incorporating no site-specific structural information.

In addition note that of the analogous 'structure-less' models TmHEM and JTT, the one which was designed for transmembrane proteins generally performs better (TmHEM model having smaller Δl than JTT, meaning $\hat{l}_{\text{TmHEM}} > \hat{l}_{\text{JTT}}$, and smaller z-score, meaning it is rejected in favour of TM126 less strongly than is JTT). It is encouraging that basing a model on average transmembrane protein replacement dynamics is helpful, even without consideration of site-specific structural information. The only possible exception is in the analysis of the calcitonin receptor sequences, for which the z-scores are approximately equal.

Taking all results together, it is evident that for transmembrane protein sequences the TM126 model may be significantly better than many other

available models. This may lead to it being preferred for phylogenetic analyses of transmembrane proteins, and also leads us to hope that it may be used to aid in the interpretation of the function of sequences with unknown structure.

Conclusions

Earlier work (Thorne, Goldman, and Jones 1996; Goldman, Thorne, and Jones 1998) has indicated that the effects of secondary structure and solvent accessibility on the dynamics of (evolutionary) amino acid replacement are generally very important. In this paper we have shown that the incorporation of structural information relevant to transmembrane proteins can lead to a very much improved fit of the model to transmembrane protein amino acid sequence data sets. Our results (table 1) show significantly higher likelihoods for the new TM126 model, with P-values and z-scores suggesting rejection of the JTT or TmHEM models in favour of the TM126 model. In addition, good agreement between the predicted and the observed secondary structures (e.g. CCR5, fig. 3b) suggests that the new model is able to interpret structural information available in multiple sequence alignments.

A significant part of any success of a structure-based model depends on its ability to recognize the structural elements of proteins. Some amino acids have very similar propensities for different structural categories, as for instance Ile, Val and Thr in β-sheets in globular domains and in transmembrane helices. Some transmembrane domains may have a weak transmembrane 'signature', being somewhat atypical themselves but stabilized by neighbouring transmembrane domains. Moreover, the majority of methods for secondary structure prediction show a general lack of power in β-sheet prediction; this may be related to underestimation of β-sheet sequence heterogeneity.

The differences in transmembrane domain prediction performance

reported above and in figure 3b stimulate further consideration of several limitations in the prediction of transmembrane helices. The transmembrane pattern may be weaker for some sequences in a multiple alignment and stronger for others and in these cases, the averaged signal obtained from the multiple alignment is generally stronger, over all the transmembrane domain, than that for a single sequences. In multiple spanning segment proteins, contacts between helices affect stabilization and homologous sequences may show differences in inter-helix stabilization. A potential drawback in algorithms that use a set of multiple aligned sequences may concern the possible unreliability in the structure assignment at the boundaries of secondary structure elements due to small variations in the location of transmembrane domains among proteins belonging to distantly related species. Moreover, it is known that transmembrane helix cap regions have slightly different amino acid replacement dynamics (Jones, Taylor, and Thornton 1994).

Light reaction centre subunits 'l' and 'm' contain transmembrane helices and large globular α-helix and β-sheet domains. It is in cases like these that the TM126 model, which incorporates both transmembrane and globular domain components, may be especially successful. We find TM126 to be strongly supported (table 1) in preference to both JTT and TmHEM (e.g. all z-scores at least 6.37; approx. $P \leq 0.0001$).

Some of the α-helices of 'm' subunit of the light reaction centre lie along the membrane surface and thus they are amphipathic helices with an average hydrophobicity that resembles transmembrane helices. Since no amphipathic property is included in any of our 10 structural categories this may induce inaccuracies that are reflected by the model comparisons in table 1, where TM126 is less strongly preferred (lower z-scores) over the other models for the 'm' subunit than for the 'l' subunit.

Cytochrome-b is a mitochondrial inner membrane protein (Xia et al. 1997). This characteristic makes the structural categories related to

globular domains inefficient in describing any part of the protein. Therefore, it is not surprising that the JTT model is very strongly rejected in favour of TM126 for these sequences (table 1; $P < 0.01$; $z = 5.98$). TmHEM also appears preferable to JTT ($z = 2.83$ for TmHEM, cf. 5.98 for JTT).

When a statistical comparison indicates that one evolutionary model is superior to a simpler model, this implies that features possessed by the more complex model and absent in the simpler model may be evolutionarily important. The examples we report show that the new model of evolution of transmembrane protein amino acid sequences, based on secondary structure and topology information and implemented in a ML framework, performs better than the JTT model (which ignores protein structure and assumes all residues are subject to identical evolutionary dynamics inferred from globular proteins) and the TmHEM model (which is similar to JTT but assumes amino acid replacement dynamics inferred from transmembrane protein data). It is in itself interesting to have improved understanding of evolutionary pressures and constraints, in this case due to structural (and hence functional) constraints in transmembrane proteins. Improved models are expected to give improved results in phylogenetic analyses (Yang, Goldman, and Friday 1994, 1995).

Additionally, in the case that little or nothing is known about a gene under study, statistical tests such as those above may be used to decide the most appropriate model to use.

At the same time, it has been shown that phylogeny can be an important tool in investigating protein structure (Goldman, Thorne, and Jones 1996). Algorithms that use multiple alignments of biological sequences as data should take into account the evolutionary relationships among the sequences in the alignment in order to allow correctly for the correlations in the data due to shared ancestry. Current methods for secondary structure prediction of proteins seem not to be able to pass an upper threshold of 70–80% in prediction accuracy of globular proteins and

90% for transmembrane proteins; probably this limit can also be attributable to natural variation of secondary structure among homologous proteins. Thus a deeper understanding of phylogenetic relationships and historical changes of secondary structure may suggest ways to overcome this limit.

There are, of course, numerous assumptions and approximations made in our model, in order to make it practical. We describe some of these here, to illustrate directions in which this research could be continued and improved. While in globular proteins long stretches of hydrophobic amino acids are quite rare, transmembrane domains usually have a long stretch of hydrophobic amino acids in order to span the lipid layer. Our model can allow for this, in its amino acid replacement model specific to transmembrane helices (α^{TM}). However, this pattern is not always well conserved; for example, charged amino acids are usually absent in the middle of the transmembrane helices but two residues of opposite charges may be found close together inside the membrane, neutralizing each other's charge. Moreover, in many cases it is not easy to distinguish between transmembrane helix and transmembrane β-sheet structures (as found in porins). We have not yet been able to consider a replacement matrix for transmembrane β structures because of lack of data. Generally transmembrane β structures form membrane proteins by the cooperative effect of distant β-sheets along the sequence and the assembling of several subunits (e.g. Kreusch and Schulz 1994); this kind of process may make prediction very difficult.

Our model currently assumes that there has been no change in protein structure or accessibility status since sequence divergence. Advanced models that explicitly address the evolution of structure would be of great interest for the study of evolutionary processes, for phylogenetic estimation, and for structure prediction. The step from secondary to tertiary structure (in some ways analogous to understanding correlations between molecular

and morphological evolution, e.g. Pagel 1994; Omland 1997) will be yet more difficult to implement in evolutionary models (see for instance Qu et al. 1993; Grishin 1997). Eventual understanding of links between evolution and 3-D structure may allow us to extend our understanding of homologous proteins to those that have diverged to the so-called 'twilight zone', i.e. that have almost completely lost their primary structure homology.

Acknowledgments

P.L. is supported by an EPSRC/BBSRC Bioinformatics Initiative grant. N.G. was supported by a Wellcome Trust Fellowship in Biodiversity Research and by a Visiting Fellowship at the Isaac Newton Institute for the Mathematical Sciences, Cambridge, funded by EPSRC Grant GR K99015, during the time when this work was performed. Software, parameter estimates, and data sets used for the analyses described above are available from the authors and via
`http://ng-dec1.gen.cam.ac.uk/hmm/contents.html`.

LITERATURE CITED

ARNASON, U., and A. GULLBERG. 1996. Cytochrome-b nucleotide sequences and the identification of five primary lineages of extant cetaceans. Mol. Biol. Evol. **13**:407–417.

ASAI, T., D. ZAPOROJETS, C. SQUIRES, and C. L. SQUIRES. 1999. An Escherichia coli strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. Proc. Natl Acad. Sci. USA **96**:1971–1976.

BEN-EFRAIM, I., D. BACH, and Y. SHAI. 1993. Spectroscopic and functional characterization of the putative transmembrane segment of the minK potassium channel. Biochemistry **32**:2371–2377.

CAO, Y., N. OKADA, and M. HASEGAWA. 1997. Phylogenetic position of guinea pigs revisited. Mol. Biol. Evol. **14**:461–464.

CHOTHIA, C., and A. M. LESK. 1986. The relation between the divergence of sequence and structure in proteins. EMBO J. **5**:79–94.

CSERZO, M., E. WALLIN, I. SIMON, G. VON HEIJNE, and A. ELOFSSON. 1997. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Prot. Eng. **10**:673–676.

DAYHOFF, M. O., R. V. ECK, and C. M. PARK. 1972. A model of evolutionary change in proteins. Pp. 89–99 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure, Vol. 5. National Biomedical Research Foundation, Washington DC.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure, Vol. 5, suppl. 2. National Biomedical Research Foundation, Washington DC.

DENG, H., R. LIU, W. ELLMEIER, S. CHOE, D. UNUTMAZ, M. BURKHART, P. DI MARZIO, S. MARMON, R. E. SUTTON,

C. M. Hill, C. B. Davis, S. C. Peiper, T. J. Schall, D. R. Littman, and N. R. Landau. 1996. Identification of a major co-receptor for primary isolates of HIV-1. Nature **381**:661–666.

Dragic, T., V. Litwin, G. P. Allaway, S. R. Martin, Y. Huang, K. A. Nagashima, C. Cayanan, P. J. Maddon, R. A. Koup, J. P. Moore, and W. A. Paxton. 1996. HIV-1 entry into CD4(+) cells is mediated by the chemokine receptor CC-CKR-5. Nature **381**:667–673.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

Felsenstein, J. 1995. PHYLIP (Phylogenetic Inference Package), Ver. 3.57. Department of Genetics, University of Washington, Seattle WA.

Gavel, Y., and G. von Heijne. 1992. The distribution of charged amino acids in mitochondrial inner membrane proteins suggests different modes of membrane integration for nuclearly and mitochondrially encoded proteins. Eur. J. Biochem. **205**:1207–1215.

Goldman, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. **36**:182–198.

Goldman, N., J. L. Thorne, and D. T. Jones. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. J. Mol. Biol. **263**:196–208.

Goldman, N., J. L. Thorne, and D. T. Jones. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics **149**:445–458.

Grishin, N. V. 1997. Estimation of evolutionary distances from protein spatial structures. J. Mol. Evol. **45**:359–369.

Hasegawa, M., J. Adachi, and M. C. Milinkovitch. 1997. Novel phylogeny of whales supported by total molecular evidence. J. Mol. Evol. **44**:s117–s120.

VON HEIJNE, G. 1992. Membrane protein structure prediction. J. Mol. Biol. **225**:487–494.

HEINZ, D. W., W. BAASE, F. D. DAHLQUIST, and B. W. MATTHEWS. 1993. How amino acid insertions are allowed in an α-helix of T4 lysozyme. Nature **361**:561–564.

HERVÉ, P. 1997. Rodent monophyly: pitfalls of molecular phylogenies. J. Mol. Evol. **45**:712–715.

HIROKAWA, T., S. BOON-CHIENG, and S. MITAKU. 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics **14**:378–379.

HOFMANN, K., and W. STOFFEL. 1993. TMbase: a database of membrane spanning proteins segments. Biol. Chem. Hoppe-Seyler **347**:166–171.

HUELSENBECK, J. P., and B. RANNALA. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276**:227–232.

JONES, D. T. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry **33**:3038–3049.

JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. CABIOS **8**:275–282.

JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1994. A mutation data matrix for transmembrane proteins. FEBS Letts **339**:269–275.

JONES, P. C., A. SIVAPRASADARAO, D. WRAY, and J. B. FINDLAY. 1996. A method for determining transmembrane protein structure. Mol. Membr. Biol. **13**:53–60.

JURETIC, D., and A. LUCIN. 1998. The preference functions method for predicting protein helical turns with helical propensity. J. Chem. Inf. Comput. Sci. **38**:575–585.

KABSCH, W., and C. SANDER. 1983. Dictionary of protein secondary

structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers **22**:2577–2637.

KREUSCH, A., and G. E. SCHULZ. 1994. Refined structure of the porin from *Rhodopseudomonas blastica*—comparison with the porin from *Rhodobacter capsulatus*. J. Mol. Biol. **243**:891–905.

KUHMANN, S. E., E. J. PLATT, S. L. KOZAK, and D. KABAT. 1997. Polymorphisms in the CCR5 genes of African green monkeys and mice implicate specific amino acids in infections by simian and human immunodeficiency viruses. J. Virol. **71**:8642–8656.

LANE, D. J., A. P. HARRISON, J. R. STAHL, B. PACE, S. J. GIOVANNONI, G. J. OLSEN, and N. R. PACE. 1992. Evolutionary relationships among sulphur- and iron-oxidizing eubacteria. J. Bacteriology. **174**:269–278.

LESK, A. M., and C. CHOTHIA. 1982. Evolution of proteins formed by beta-sheets. II. The core of the immunoglobulin domains. J. Mol. Biol. **160**:325–342.

LIÒ, P., and N. GOLDMAN. 1998. Models of molecular evolution and phylogeny. Genome Res. **8**:1233–1244.

LIÒ, P., N. GOLDMAN, J. L. THORNE, and D. T. JONES. 1998. Combining protein secondary structure prediction and evolutionary inference. Bioinformatics **14**:726–733.

MIZUGUCHI, K., and N. GO. 1995. Comparison of spatial arrangements of the secondary structure elements in proteins. Prot. Eng. **8**:353–362.

NAGASHIMA, K. Y. P., A. HIRAISHI, K. SHIMADA, and K. MATSUURA. 1997a. Horizontal transfer of genes coding for the photosynthetic reaction centers of purple bacteria. J. Mol. Evol. **45**:131–136.

NAGASHIMA, K. Y. P., K. MATSUURA, N. WAKAO, A. HIRAISHI, and K. SHIMADA. 1997b. Nucleotide sequences of genes coding for photosynthetic reaction centers of *Acidiphilium rubrum* and related aerobic acidophilic bacteria. Plant Cell Physiol. **38**:1249–1258.

NAYLOR, G., and W. M. BROWN. 1997. Structural biology and phylogenetic estimation. Nature **388**:527–528.

OMLAND, K. E. 1997. Correlated rates of molecular and morphological evolution. Evolution **51**:1381–1393.

PAGEL, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc. R. Soc. London B **255**:37–45.

QU, C., L. LAI, X. XU, and Y. TANG. 1993. Phyletic relationships of protein structures based on spatial preference of residues. J. Mol. Evol. **36**:67–78.

ROST, B., R. CASADIO, P. FARISELLI, and C. SANDER. 1995. Transmembrane helices predicted at 95% accuracy. Prot. Sci. **4**:521–533.

ROST, B., P. FARISELLI, and R. CASADIO. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. Prot. Sci. **5**:1704–1718.

RUSSELL, R. B., M. A. S. SAQI, R. A. SAYLE, P. A. BATES, and M. J. E. STERNBERG. 1997. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J. Mol. Biol. **269**:423–439.

RZHETSKY, A. 1995. Estimating substitution rates in ribosomal RNA genes. Genetics **141**:771–783.

SONNHAMMER, E. L. L., G. VON HEIJNE, and A. KROGH. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. ISMB **6**:175–182.

SPRUIJT, R. B., C. J. WOLFS, J. W. VERVER, and M. A. HEMMINGA. 1996. Accessibility and environment probing using cystein residues introduced along the putative transmembrane domain of the major coat protein of bacteriophage M13. Biochem. **35**:10383–10391.

THORNE, J. L., N. GOLDMAN, and D. T. JONES. 1996. Combining protein evolution and secondary structure. Mol. Biol. Evol. **13**:666–673.

TUSNADY, G.E. and I. SIMON. 1998. Principles governing amino acid

composition of integral membrane proteins: application to topology prediction. J. Mol. Biol. **283**:489–506.

WAKAO, N., N. YOKOI, N. ISOYAMA, A. HIRAISCI, K. SHIMADA, M. KOBAYASHI, H. KISE, M. IWAKI, S. ITOH, S. TAKAICHI, and Y. SAKURAI. 1996. Discovery of natural photosynthesis using Zn-containing bacteriochlorophyll in an aerobic bacterium *Acidiphilium rubrum*. Plant Cell Physiol. **37**:889–893.

WOESE, C. R. 1987. Bacterial evolution. Microbiol. Rev. **51**:221–271.

WOESE, C. R., E. STACKEBRANDT, W. G. WEISBURG, B. J. PASTER, M. R. MADIGAN, V. J. FOWLER, C. M. HAHN, P. BLANZ, R. GUPTA, K. H. NEALSON, and G. E. FOX. 1984. The phylogeny of purple bacteria: the alpha subdivision. Sys. Appl. Microbiol. **5**:315–326.

XIA, D., C. YU, H. KIM, J. XIA, A. M. KACHURIN, L. ZHANG, L. YU, and J. DEISENHOFER. 1997. Crystal structure of the cytochrome bc1 complex from Bovine heart mitochondria. Science 277:60–66.

YANG, Z., N. GOLDMAN, and A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. **11**:316–324.

YANG, Z., N. GOLDMAN, and A. FRIDAY. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. **44**:384–399.

ZHANG, L., C. D. CARRUTHERS, T. HE, Y. HUANG, Y. CAO, G. WANG, B. HAHN, and D. D. HO. 1997. HIV type 1 subtypes, coreceptor usage, and CCR5 polymorphism. AIDS Res. Hum. Retroviruses **13**:1357–1366.

Table 1

Statistical Tests Comparing the TM126 Model with Two Other Evolutionary Models

	TM126	JTT			TmHEM		
dataset	$\hat{l}_{\text{TM126}}$	Δl	P	z	Δl	P	z
CCR5	-1449.49	15.21	0.00	4.75	10.30	0.00	3.21
LRC 'l'	-2534.97	52.37	0.00	7.83	33.44	0.00	6.72
LRC 'm'	-3127.18	41.02	0.00	7.21	35.46	0.00	6.37
Calc R	-2685.12	9.55	0.00	5.02	37.28	0.00	5.14
K-sub R	-2002.48	8.70	0.00	3.43	2.04	0.03	2.42
Cyt-b	-1623.55	40.96	0.00	5.98	10.93	0.00	2.83

NOTE.—As described in the text, we have performed statistical tests comparing the TM126 model with the JTT and TmHEM models for each of six data sets (CCR5: primate chemokine receptor; LRC 'l' and LRC 'm': purple bacteria 'l' and 'm' subunits of the light reaction centre; Calc R and K-sub R: lagomorph and rodent calcitonin and K-substance receptors; Cyt-b: cetacean cytochrome-b). Table entries are $\hat{l}_{\text{TM126}}$, the observed maximum log-likelihood under the TM126 model and, for the tests of TM126 vs. each null hypothesis model JTT and TmHEM, Δl, the maximum log-likelihood under TM126 minus the maximum log-likelihood under the null model; P, the proportion of times (from 100 simulations) that a value of Δl simulated under the null hypothesis exceeded the observed value; and a z-score, described in the text, which is the estimated number of standard deviations by which the observed value of Δl exceeds the mean of the simulated values.

FIGURE LEGENDS

Figure 1.—Examples of permitted transitions among the 24 hidden TM states comprising a transmembrane helix. Arrows indicate the permitted transitions among the states illustrated, and are labelled with the parameter combinations that define the transition probabilities ρ_{ij}. After entry via state TM_1 (upper part of fig: from extra-cytoplasmic region'ext.'; lower part: from cytoplasmic region 'cyt.'), the HMM must progress through the transmembrane domain until TM_{14}; it may then remain in the transmembrane helix, progressing through the subsequent TM states (TM_{15} to TM_{24}), or may at any stage leave the transmembrane helix and enter (upper part of fig.) a cytoplasmic region, via the states CL, C*b* or C*e*, or (lower) an extra-cytoplasmic region, via the states C*b* or C*e*.

Figure 2.—General features of the 126-state transmembrane HMM model. Arrows indicate the permitted transitions among the regions of the model. Structures/domains are labelled (in parentheses) with the number of HMM states comprsing them. The cartoon to the left represents a hypothetical transmembrane protein with seven transmembrane helices and three globular domains.

Figure 3.—(a) The ML estimate of phylogeny derived from the TM126 model for the set of CCR5 homologs from mouse, chimpanzee, gorilla, human, rhesus monkey, pig-tailed macaque, baboon, African green monkey, sooty mangabey. The scale bar refers to the branch lengths, measured in units of expected numbers of replacements per site. (b) Comparison of CCR5 observed transmembrane domain locations and the predictions of our TM126 model and other methods currently used in transmembrane protein prediction, namely PHDhtm (Rost et al. 1995), HMMTOP (Tusnady and Simon 1998), TMHMM (Sonnhammer, von Heijne, and Krogh 1998),

MEMSAT (Jones 1994), SOSUI (Hirokawa, Boon-Chieng, and Mikatu 1998), Split35 (Juretic and Lucin 1998), TopPred2 (von Heijne 1992), TMpred (Hofmann and Stoffel 1993) and DAS (Cserzo et al. 1997). Methods marked with an asterisk use multiple aligned sequences; other methods use only single sequence input.

Figure 4.—Other ML phylogenies derived from the TM126 model. Branch lengths are measured in units of expected numbers of replacements per site. Trees are for data sets of (a) purple bacteria 'l' and (b) 'm' subunits of the light reaction centre; (c) calcitonin receptor homologs from human, rabbit, guinea pig, pig and rat; (d) K-substance receptor homologs from human, rabbit, guinea pig, hamster and rat; and (e) cytochrome-b homologs from spinner dolphin, bottlenose dolphin, gray whale, minke whale, Peruvian beaked whale, Gervais beaked whale, pigmy sperm whale and sperm whale.

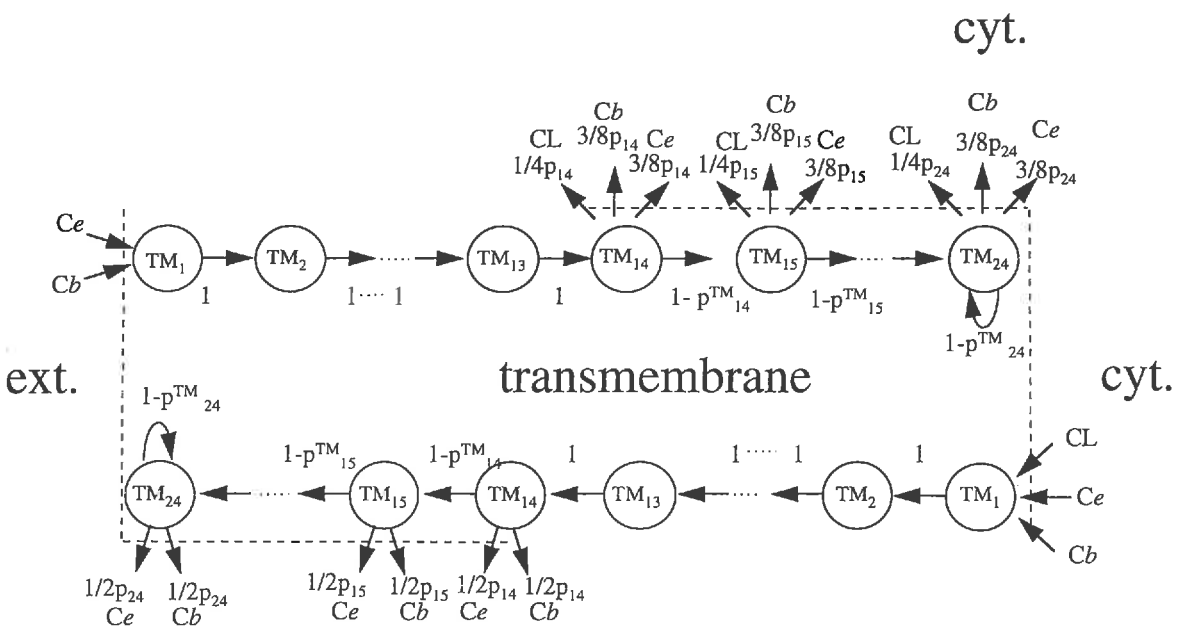

cyt.
Cb
CL 3/8p14 Ce
3/8p14
CL 1/4p14
Cb
CL 3/8p15 Ce
3/8p15
CL 1/4p15
Cb
CL 3/8p24
Ce
3/8p24
CL 1/4p24
Ce
Cb
TM1
TM2
TM13
TM14
TM15
TM24
1
1
1
1- pTM14
1-pTM15
1-pTM24
ext.
transmembrane
cyt.
1-pTM24
1-pTM15
1-pTM14
1
1
1
CL
Ce
Cb
TM24
TM15
TM14
TM13
TM2
TM1
1/2p24 Ce
1/2p24 Cb
1/2p15 Ce
1/2p15 Cb
1/2p14 Ce
1/2p14 Cb
ext.

Fig 1

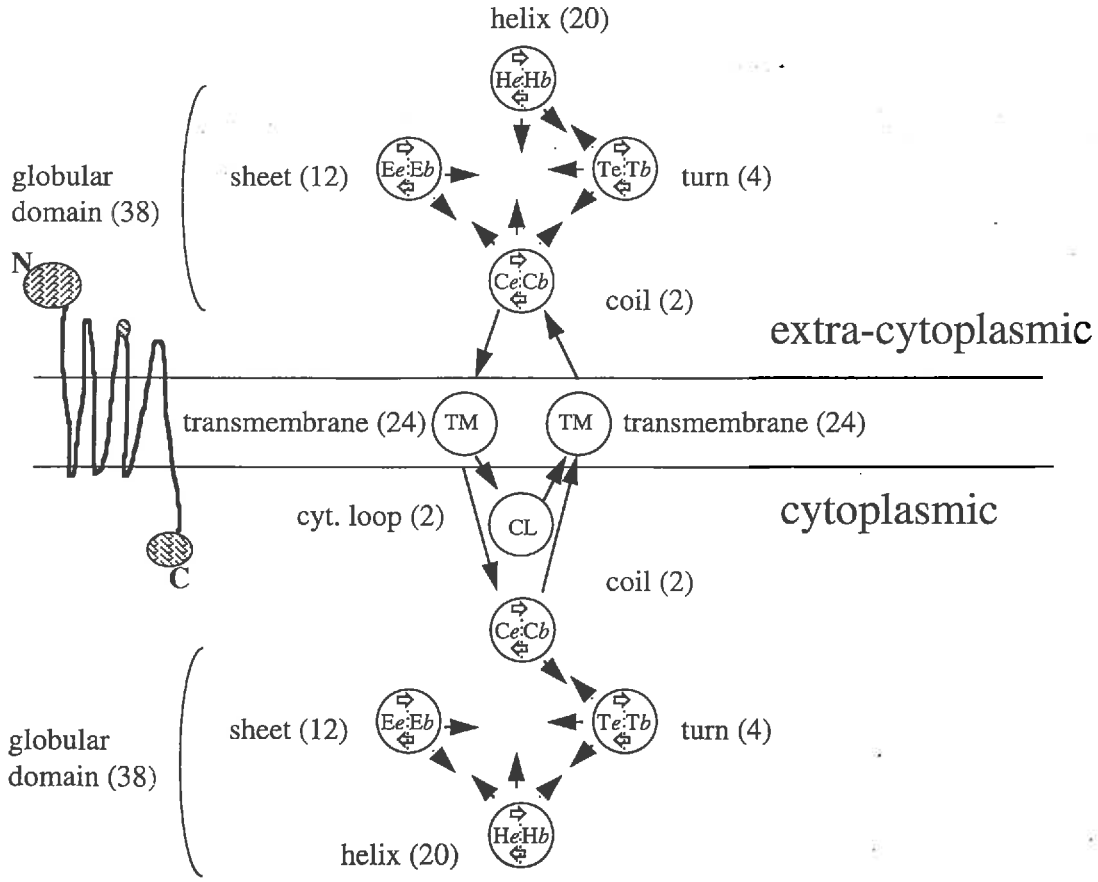

helix (20)
He:Hb
globular domain (38)
sheet (12)
Ee:Eb
Te:Tb
turn (4)
N
Ce:Cb
coil (2)
extra-cytoplasmic
transmembrane (24)
TM
TM
transmembrane (24)
cyt. loop (2)
CL
cytoplasmic
C
coil (2)
Ce:Cb
globular domain (38)
sheet (12)
Ee:Eb
Te:Tb
turn (4)
He:Hb
helix (20)

Fig 2

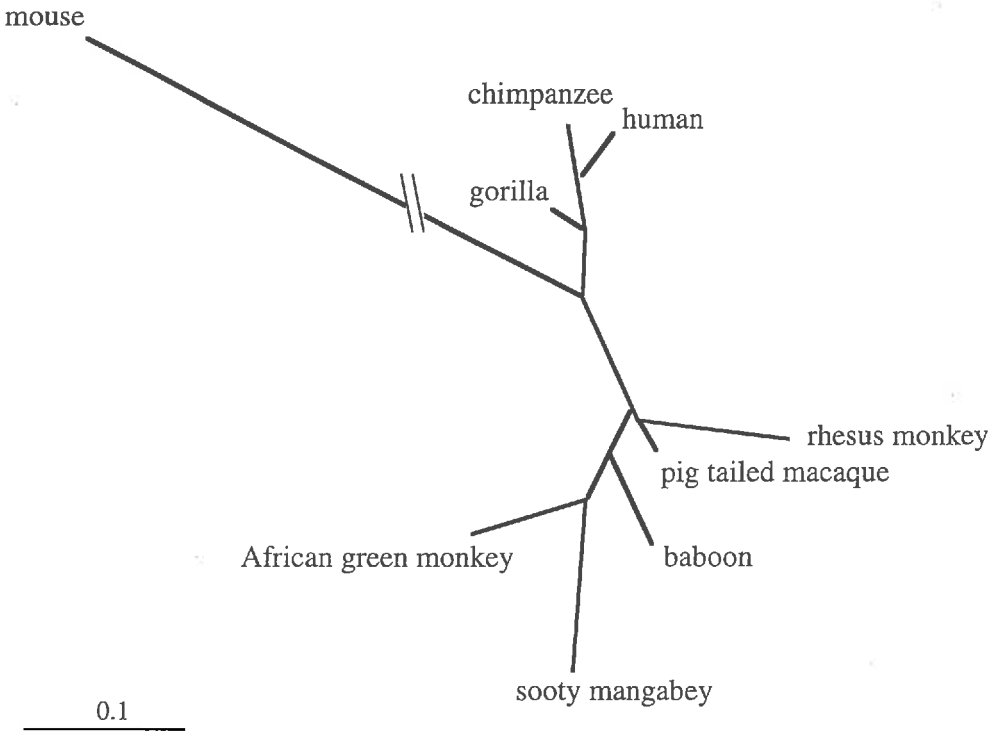

mouse
chimpanzee
human
gorilla
rhesus monkey
pig tailed macaque
African green monkey
baboon
sooty mangabey
0.1

Fig 3a

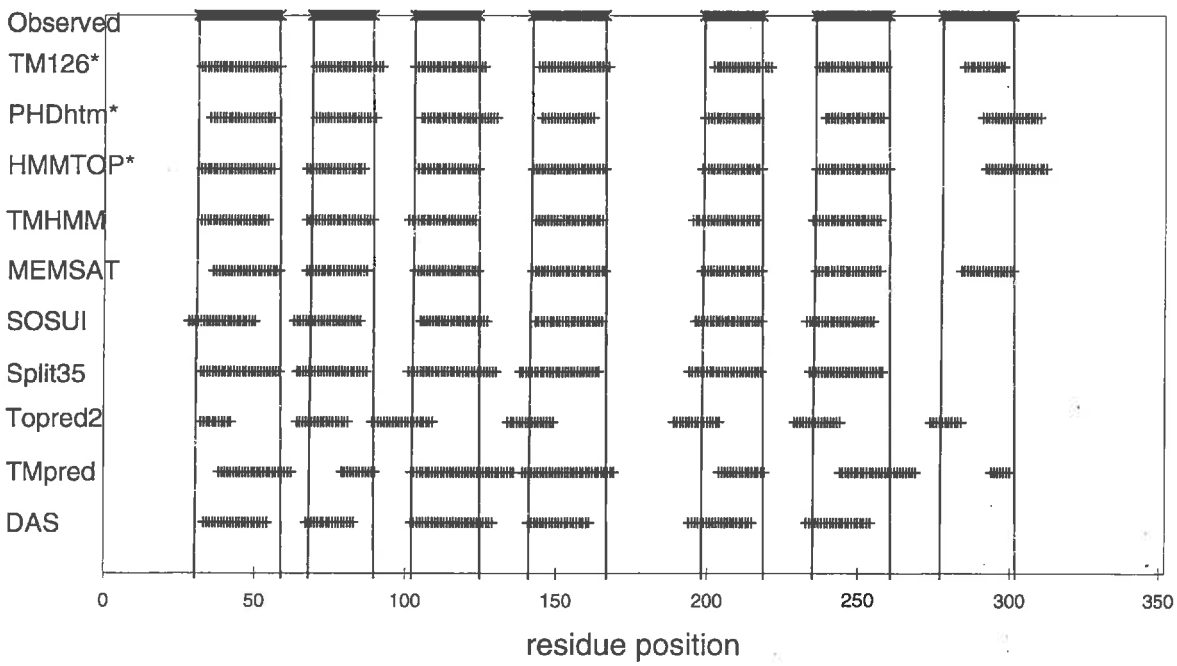

Observed
TM126*
PHDhtm*
HMMTOP*
TMHMM
MEMSAT
SOSUI
Split35
Topred2
TMpred
DAS
0
50
100
150
200
250
300
350
residue position

Fig 3b

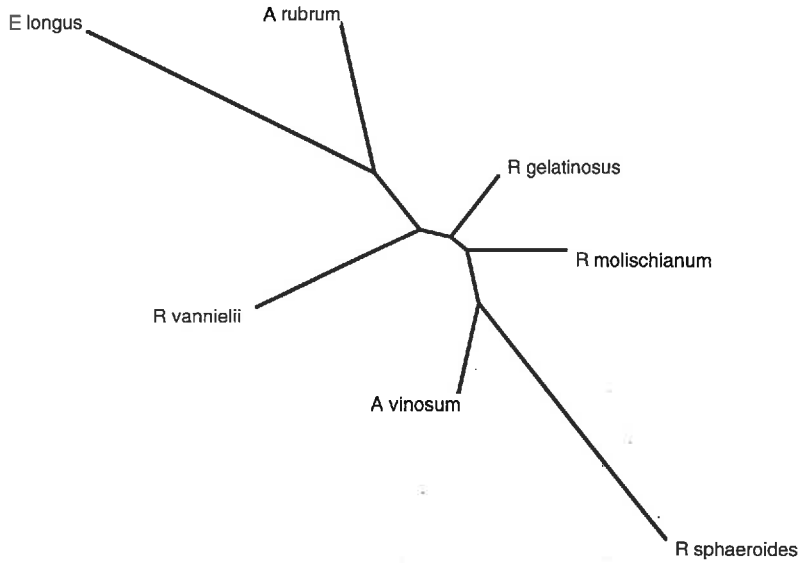
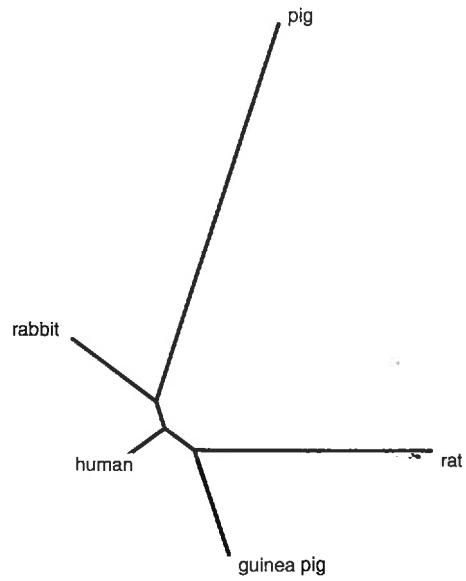

E longus
A rubrum
R gelatinosus
R molischianum
R vannielii
A vinosum
R sphaeroides
0.1

Fig. 4a

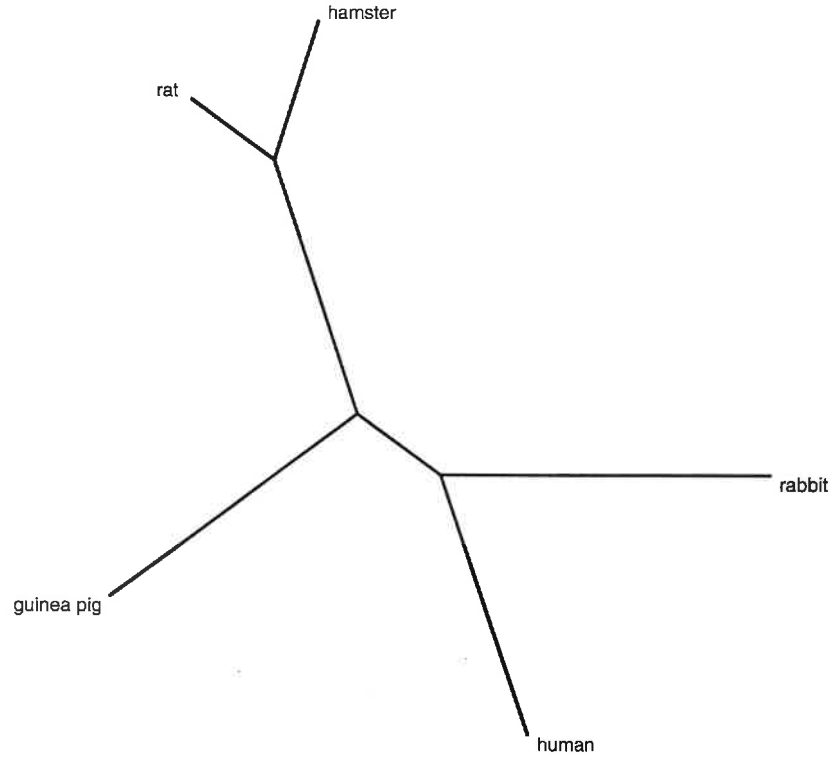

R molischianum
A vinosum
R gelatinosus
R vannielii
A rubrum
E longus
R sphaeroides
0.1

Fig. 4b



0.1

Fig 4c



0.1

Fig 4d

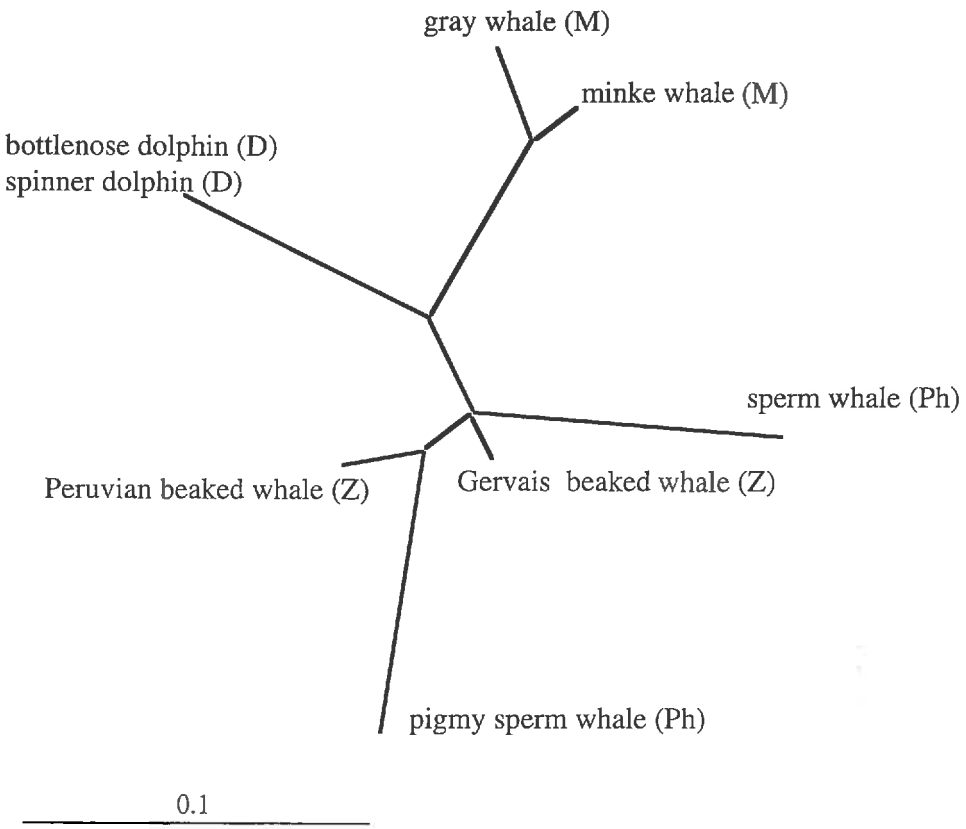

gray whale (M)
minke whale (M)
bottlenose dolphin (D)
spinner dolphin (D)
sperm whale (Ph)
Peruvian beaked whale (Z)
Gervais beaked whale (Z)
pigmy sperm whale (Ph)
0.1

Fig 4e