Missense Mutations in Transmembrane Domains of Proteins: Phenotypic Propensity of Polar Residues for Human Disease

Anthony W. Partridge, Alex G. Therien, and Charles M. Deber*

Division of Structural Biology and Biochemistry, Research Institute, Hospital for Sick Children, Toronto, and Department of Biochemistry, University of Toronto, Toronto, Ontario, Canada

ABSTRACT **Previous experiments on the cystic fibrosis transmembrane conductance regulator suggested that non-native polar residues within membrane domains can compromise protein structure/function. However, depending on context, replacement of a native residue by a non-native residue can result either in genetic disease or in benign effects (e.g., polymorphisms). Knowledge of missense mutations that frequently cause protein malfunction and subsequent disease can accordingly reveal information as to the impact of these residues in local protein environments. We exploited this concept by performing a statistical comparison of disease-causing mutations in protein membrane-spanning domains versus soluble domains. Using the Human Gene Mutation Database of 240 proteins (including 80 membrane proteins) associated with human disease, we compared the relative phenotypic propensity to cause disease of the 20 naturally occurring amino acids when removed from—or inserted into—native protein sequences. We found that in transmembrane domains (TMDs), mutations involving polar residues, and ionizable residues in particular (notably arginine), are more often associated with protein malfunction than soluble proteins. To further test the hypothesis that interhelical cross-links formed by membrane-embedded polar residues stabilize TMDs, we compared the occurrence of such residues in the TMDs of mesophilic and thermophilic prokaryotes. Results showed a significantly higher proportion of ionizable residues in thermophilic organisms, reinforcing the notion that membrane-embedded electrostatic interactions play critical roles in TMD stability. Proteins 2004;54:648–656.** © **2004 Wiley-Liss, Inc.**

Key words: helix; hydrophobic; bilayer; statistics; bioinformatics; database; folding

INTRODUCTION

Genetic diseases are caused primarily by mutations of a single nucleotide within protein-encoding genes, resulting in the specific substitution of one amino acid for another. Such missense mutations manifest their deleterious effects by compromising protein structure and/or function. At the molecular level, this results from an inability of the mutant residue to fulfill the roles of the wild-type amino acid. Consequently, trends obtained from examining databases of human disease-causing mutations can potentially yield information regarding the roles of specific amino acid residues and the general exchangeability between the 20 commonly occurring residues. Insights derived from this type of study would be particularly useful in the case of transmembrane (TM) proteins, because very few high-resolution structures are available for this class of proteins.[1] Furthermore, membrane proteins carry out a variety of critical cellular functions,[2] are often associated with human pathologies,[3] and represent most current therapeutic targets.[4]

The Human Gene Mutation Database[5] (HGMD; www.hgmd.org) contains thousands of phenotypic (disease-causing) mutations associated with hundreds of proteins, about 80 of which are TM proteins (as of July 2002). Therefore, this database lends itself particularly well to the analysis of some of the general and specific trends of disease-causing mutations. By comparing these trends in transmembrane domains (TMDs) versus their corresponding occurrences in soluble domains (SDs) of proteins, we were able to circumvent some mutational variability issues at the DNA level. Our analysis gives the "relative phenotypic propensity" of missense mutations—defined as the relative tendency for which gain or loss of an amino acid residue leads to a disease state.

We discuss the implications of our findings in terms of the structural and functional roles of amino acids in these domains, with emphasis on TMDs. In general, we find a high relative phenotypic propensity for adding or removing polar, in particular, ionizable residues from TM domains. On the other hand, mutations to (or from) nonpolar residues are found to have a low relative phenotypic

Grant sponsor: Canadian Institutes for Health Research; Grant sponsor: National Institutes of Health-National Institute of Diabetes and Digestive and Kidney Diseases.

Anthony W. Partridge and Alex G. Therien contributed equally to the work.

A.G. Therien's present address is Merck Frosst Canada & Co., P.O. Box 1005, Pointe-Claire, Quebec, H9R 4P8, Canada.

*Correspondence to: Charles M. Deber, Division of Structural Biology and Biochemistry, Research Institute, Hospital for Sick Children, Toronto M5G 1X8, Ontario, Canada. E-mail: deber@sickkids.ca

Received 24 April 2003; Accepted 8 August 2003

propensity within TMDs, suggesting that collectively these residues have a more dispensable role within the membrane.

MATERIALS AND METHODS

Data Compilation

A list of genes and proteins related to disease was compiled from the locus-specific databases section of the HMGD database. The complete DNA and protein sequences for these genes were obtained from the Expasy Web site (www.expasy.org). For proteins in which the presence and identity of signal cleavage sequences are known, these sequences were excluded from this analysis. Putative TM segments in these genes were identified by using TM-Finder (www.bioinformatics-canada.org/TM/login.html; parameters used were as follows: N-terminal window = 6; C-terminal window = 5; minimum core length = 6; close gaps of length = 4; minimum segment length = 15). The DNA sequences corresponding to TMDs (equivalent to the compiled identified TM segments) and those corresponding to SDs (non-TM segments) were compiled separately. Compilation of individual codons was conducted by using a program coded by the Centre for Computational Biology at the Hospital for Sick Children. Codons were compiled: 1) as tetranucleotides that include the 5′-flanking nucleotide, to account for dinucleotide affects on DNA mutability at the first codon position; 2) as tetranucleotides that include the 3′-flanking nucleotide, to account for dinucleotide affects on DNA mutability at the third codon position; and 3) as the usual trinucleotides for calculations involving mutations to the second codon position. The missense DNA mutations and corresponding codons for each gene were obtained from the main HGMD Web site (not from the locus-specific section of the database); for mutations occurring at the first and third codon positions these included, respectively, the 5′- or 3′-flanking nucleotide. The mutations were then divided into two groups, those within TMDs, and those within SDs (see above). Additional TMD mutations were compiled from the locus-specific databases linked to the HGMD site.

Calculations of Relative Phenotypic Propensities

For calculation of relative phenotypic propensity for native residue loss, mutations were first grouped according to the identity of the altered native amino acid residue. For each residue, the mutations were further subdivided according to the specific mutated codon; for mutations occurring at the first or third codon position, this included the appropriate flanking nucleotides. For mutations in SDs, the frequency of mutation for a specific amino acid residue (F_{SD}) was generated by dividing the sum of all the individual mutations for that residue (SD_{mut}) by the total number of codons encoding that residue (SD_{cod}) in these domains:

$$F_{SD} = SD_{mut}/SD_{cod}. \qquad (1)$$

For TM mutations, a similar approach was taken except that the values were normalized in two ways. First, the calculated number of mutations for each codon (including flanking nucleotides, when appropriate) was normalized according to the frequency of that codon in TMDs compared with SDs to account for differences in mutability due to codon usage. Specifically, we multiplied the actual number of TM mutations for a specific codon (TM_{mut}) by the ratio of the codon usage in SDs (CU_{SD}) versus codon usage in TM domains (CU_{TM}):

$$TM'_{mut} = TM_{mut} \times CU_{SD}/CU_{TM} \qquad (2)$$

The normalized numbers of mutations (TM'_{mut}) were then used to calculate frequencies of mutations in TMDs (F'_{TMD}):

$$F'_{TMD} = TM'_{mut}/TM_{cod} \qquad (3)$$

The second normalization was conducted to account for the more frequent overall rate of mutation within TMDs compared to SDs (about twofold in the HGMD database). We first calculated a normalization factor (NF) as follows:

$$NF = (N_{TMmut}/N_{SDmut})/(N_{TMcod}/N_{SDcod}), \qquad (4)$$

where N_{TMmut} is the total number of mutations in TMDs, N_{SDmut} is the total number of mutations in SDs, N_{TMcod} is the total number of codons in TMDs, and N_{CDcod} is the total number of codons in SDs. All calculated mutation frequencies for TMDs (F'_{TMD} from Eq. 3) were divided by NF to yield a mutation frequency in TMDs (F_{TMD}) that is directly comparable to the mutation frequency in SDs:

$$F_{TMD} = F'_{TMD}/NF \qquad (5)$$

The relative phenotypic propensity of native residue loss is reported as the ratio of the mutation frequency of a particular amino acid residue in TMDs (F_{TMD} from Eq. 5) over the mutation frequency of the same residue in SDs (F_{SD} from Eq. 1). Thus, values that deviate from unity indicate lower or higher relative phenotypic propensities in TMDs. Statistical significance was determined by using a standard large-sample *z* test.

For calculation of relative phenotypic propensity for non-native residue gain, mutations were grouped according to the nature of the non-native residue that is gained after mutation. Subsequent calculations were identical to those for relative phenotypic propensity of native residue loss except that the total number of codons for each individual residue (SD_{cod} and TM_{cod} in Eqs. 1 and 3, respectively) is the sum of all native codons that can potentially yield the non-native residue. For calculation of relative phenotypic propensity for specific mutations, mutations that involved identical changes at the amino acid level were grouped together and analyzed as described above for relative phenotypic propensity of native residue loss. For calculation of relative phenotypic propensity for general mutations, mutations involving nonpolar (A, F, G, I, L, M, V), polar (C, D, E, H, K, N, Q, R, S, T, W, Y), strongly polar (D, E, N, Q), or charged (D, E, H, K, R) amino acid residues were grouped together and analyzed as described above for relative phenotypic propensity of native residue loss.

Analysis of Bacterial Proteomes

The complete protein sequences of the following organisms were obtained from www.ebi.ac.uk/proteome/: mesophiles: *Methanosarcina mazei, Methanosarcina acetovorans, Agrobacterium tumefaciens, Bacillus subtilis, Lactococcus lactis, Staphylococcous aureus, Clostridium acetobutylicum, Eschericia coli, Fusobacterium nucleatum, Neisseria meningitides, Chlamydophila pneumoniae, Salmonella typhi, Vibrio Cholrae, Streptomyces coelicolor,* and *Rhizobium meliloti;* Thermophiles: *Methanobacterium thermoautotrophicum, Methanococcus jannaschii, Archaeoglobus fulgidus, Sulfolobus tokodaii, Thermotoga maritime, Thermoanaerobacter tengcongensis,* and *Methanococcus jannaschii;* Hyperthermophiles: *Aquifex aeolicus, Aeropyrum pernix, Methanopyrus kandleri, Pyrobaculum aerophilum, Pyrococcus abyssi, Pyrococcus furiosus,* and *Pyrococcus horikoshii.* Organisms were selected from both archea and eubacteria domains at random. TMDs were identified by using TM-finder (same parameters as above) and analyzed for amino acid composition by using a program written by the Hospital for Sick Children Centre for Computational Biology. Proteomes were classified as mesophiles, thermophiles, or hyperthermophiles, defined as organisms that grow at temperatures between 20 and 45°C, between 65 and 90°C, and above 90°C, respectively.

RESULTS

Relative Phenotypic Propensity of Residues in Disease-Related Mutations

An initial approach toward understanding the distinct roles of amino acid residues within TM domains (TMD) might involve comparing a calculated rate of mutation of individual amino acid residues with an "expected" theoretical distribution. However, calculation of the latter distribution is prevented by the nonrandom nature of nucleotide mutation rates. Specifically, it has been shown that the rate of mutation of a specific nucleotide is strongly affected by the nature of the nucleotides that flank it; thus, different di- and trinucleotide sequences have varying mutability profiles.[6] For example, the dinucleotide CG has a particularly high mutation rate because the high methylation rate of cytosine in this sequence often causes it to undergo spontaneous deamination to thymine.[7] Thus, CG codons often mutate to TG (or to CA on the opposite strand) even in the absence of mutagen, hampering a quantitative understanding of amino acid mutabilities per se. However, by comparing the mutation rates of amino acids in transmembrane domains with those in soluble domains (the latter including both soluble proteins and the soluble domains of membrane proteins), one can obtain the "relative phenotypic propensities" of mutations in membrane-spanning regions of proteins. Here, we used this approach because it has the advantage of normalizing nonrandom mutabilities at the di- and trinucleotide level, making it possible to extract useful information.

Data Compilation

To compare the rate of mutation within TMDs and SDs, we compiled the gene sequences, as well as the disease-causing missense mutations, of all the proteins found in the HMGD. Note that >99.9 % of these mutations are a product of a single base pair change; any mutations that resulted from multiple base pair changes were discarded. To identify TMDs, and consequently TM proteins, in the database, we used TM-Finder,[8] a program developed in our laboratory that recognizes TM sequences based on both hydrophobicity and helical propensity of residues. TM-Finder is well suited to this type of analysis because of its low rate of "false-positives" and its demonstrated ability to accurately predict the TM regions of membrane proteins.[8] By this method, we found that—as a percent—the rate of mutation in TMDs was slightly higher (~2-fold) than in SDs. However, because the total number of TM-based mutations was limited (due to the intrinsically lower number of TMD residues vs SD residues), we supplemented our data with TMD mutations from the literature and other databases that were not included in the HMGD (see Materials and Methods). Increasing the number of mutations in the TMD set in this manner was not a concern because our subsequent comparison of the two datasets involves normalization of the overall rate of mutations between domain classes.

We then compared equivalent mutations at both the amino acid and nucleotide level. Therefore, the data presented here represent ratios of the mutation rates in TMDs to those in SDs. Mutations involving a base change at the second nucleotide position in a codon were compared directly because the identities of both the 5′- and 3′-flanking codons are contained within the codon itself. However, for mutations involving base changes within the first or third nucleotide position, it was necessary to take into account nucleotides in neighboring codons. For example, for Ala-to-Thr mutations involving a change at the first position, we compared cGCX-to-cACX mutations (where X is any nucleotide; bases written in lower case denote nucleotides from flanking codons) in TMDs to cGCX-to-cACX mutations in SDs.

Relative Phenotypic Propensity for Native Residue Loss

Those residues that have critical roles (structural and/or functional) within TMDs are also those that, if mutated, are ostensibly more likely to lead to disease states. Accordingly, we calculated the relative mutability of specific wild-type residues (see Materials and Methods), which we term "relative phenotypic propensity for native residue loss," in TMDs compared to SDs (Fig. 1). For this and subsequent analyses, a value statistically greater than unity corresponds to mutations that more often result in a disease state when occurring in TMDs rather than SDs, whereas one that is less than unity corresponds to the opposite situation.

As shown in Figure 1, some residues indeed appear to be statistically more, or less, frequently involved in disease-causing mutations in TMDs versus SDs and accordingly serve as a measure of their relative importance in these domains. The residues that have a high relative phenotypic propensity for native residue loss are H, P, R, Q, and

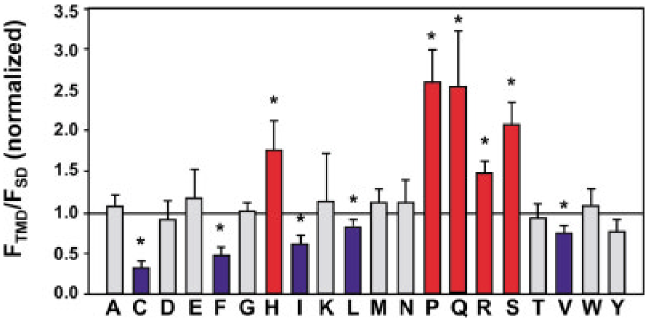

Fig. 1. Relative phenotypic propensities for native residue loss. Frequencies of disease-causing mutations that involve changes from each of the 20 naturally occurring amino acids in TMDs (F_{TMD}) and in SDs (F_{SD}) were calculated as indicated in Materials and Methods. Data are expressed as the ratio F_{TMD}/F_{SD}. Values that statistically deviate from unity (*$p < 0.05$) indicate disease-causing mutations that occur more (red bars) or less (blue bars) frequently in TMDs than in SDs.

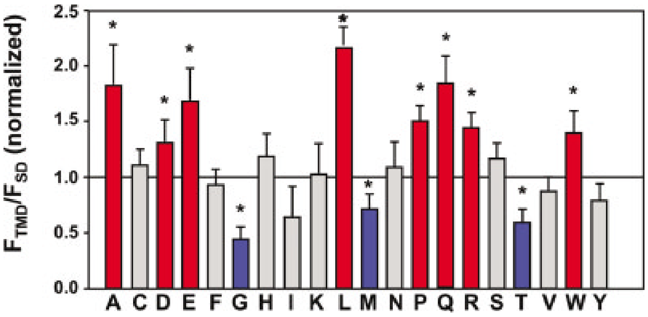

Fig. 2. Relative phenotypic propensities for non-native residue gain. Frequencies of disease-causing mutations that involve changes to each of the 20 naturally occurring amino acids in TMDs (F_{TMD}) and in SDs (F_{SD}) were calculated as indicated in Materials and Methods. Data are expressed as the ratio F_{TMD}/F_{SD}. Values that statistically deviate from unity (*$p < 0.05$) indicate disease-causing mutations that occur more (red bars) or less (blue bars) frequently in TMDs than in SDs.

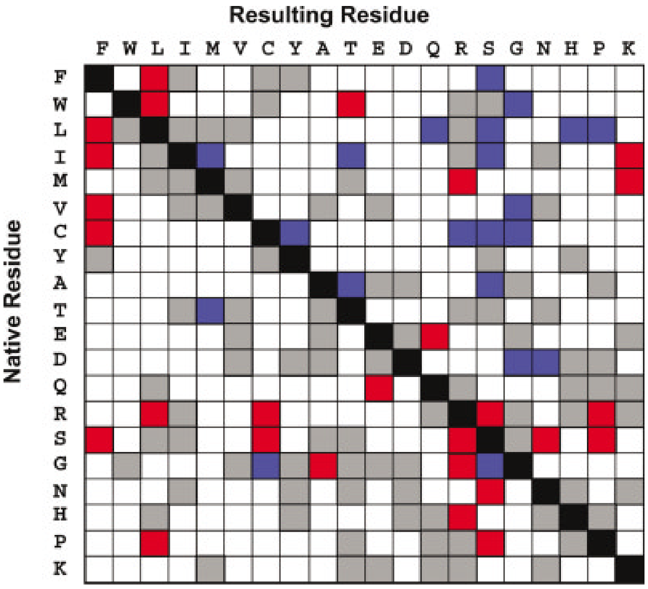

Fig. 3. Relative phenotypic propensities of specific mutations. Frequencies of disease-causing mutations that involve changes from each amino acid (rows) to every other amino acid (columns) in TMDs (F_{TMD}) and in SDs (F_{SD}) were calculated as indicated in Materials and Methods. Data are expressed as the ratio F_{TMD}/F_{SD}. Values that statistically deviate from unity ($p < 0.05$) indicate disease-causing mutations that occur more (red rectangles) or less (blue rectangles) frequently in TMDs than in SDs. Gray rectangles indicate mutations that are not possible after a single-nucleotide change. White rectangles indicate cases where no difference or statistically nonsignificant differences ($p > 0.05$) arose.

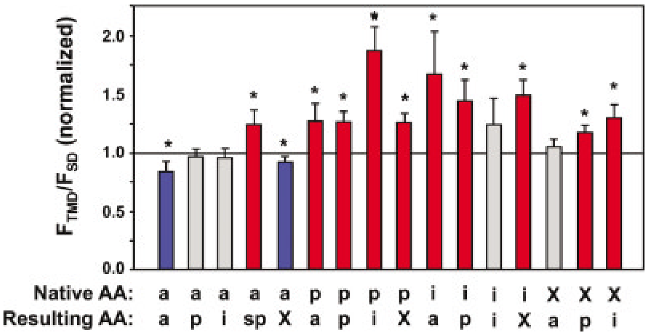

Fig. 4. General trends in relative phenotypic mutation propensities. Frequencies of disease-causing mutations that involve changes from apolar ("a"), polar ("p"), ionizable ("i"), or all ("X") amino acid residues to apolar, polar, strongly polar ("sp"), ionizable, or all amino acid residues, in TMDs (F_{TMD}) and in SDs (F_{SD}). Values were calculated as described in Materials and Methods. Data are expressed as the ratio F_{TMD}/F_{SD}. Values that statistically deviate from unity (*$p < 0.05$) indicate disease-causing mutations that occur more (red bars) or less (blue bars) frequently in TMDs than in SDs. For the purposes of this figure, apolar residues correspond to A, F, G, I, L, M, and V; polar residues correspond to C, D, E, H, K, N, Q, R, S, T, W, and Y; strongly polar residues correspond to D, E, N, and Q; and ionizable residues correspond to D, E, H, K, and R. Pro is excluded from this analysis because of its large overrepresentation in disease-causing mutations in TMDs.

S. All of these residues are polar and, therefore, capable of participating in interhelical H-bonds; although P is not classically regarded as a polar residue, it does offer the possibility of polar interactions with the backbone carbonyl moiety of the residue at both the i-3 and i-4 positions.[9–12] Notwithstanding possible polar involvement of Pro, the high phenotypic propensity of Pro in TMDs may be more related to its acknowledged capacity to induce functionally required kinks in the normally canonical TM α-helix.[12–14] In contrast, mutations of residues such as C, F, I, L, and V are underrepresented, suggesting they play less important roles in TM domains. It is of interest that all of these residues are strongly apolar, with the exception of Cys. The reason for the low disease-causing potential of mutations of Cys residues in TM domains is discussed briefly below; for a discussion of the structural importance of Cys residues in soluble proteins, see Ref. 46.

Relative Phenotypic Propensity for Non-native Residue Gain

The compiled data set can also be used to determine which amino acids more frequently result in a disease state when introduced into TMDs versus soluble domains, a trend that we term "relative phenotypic propensity for non-native residue gain." Results from this analysis (Fig. 2) show that several residues with H-bonding potential (D, E, P, Q, R, W) are overrepresented as molecular causes of disease when they replace a native TMD residue. In addition to these polar residues, two apolar residues (Ala and Leu) apparently have a higher relative phenotypic propensity for non-native residue gain. However, this is likely due to their frequent conversion from residues that were shown to have a high relative phenotypic propensity of native residue *loss* (Pro and Ser for Ala; Pro, Gln, and Arg for Leu) as described in the previous section. Indeed, if these specific mutations are not taken into account, the relative phenotypic propensities for non-native residue gain of Ala and Leu are no longer higher in TMDs. In other words, Ala and Leu are likely more often disease-causing when introduced into TMDs because they are replacing important native residues, not because they are themselves deleterious.

Trends Observed for Disease-Causing Mutations

Although the relatively low number of known disease-causing mutations in TMDs precludes a statistically meaningful analysis of many specific mutations, some important information can be gleaned from those mutations that are significantly different in TMDs compared with SDs (Fig. 3). This type of analysis reveals that some specific residues do not adequately fulfill the roles of other specific residues in TMDs versus SDs. For example, the amino acids Glu and Gln can apparently have different functions [i.e., one cannot perform the role(s) of the other] as evidenced by the high relative phenotypic propensity of both the E to Q and Q to E mutations. In contrast to this, certain residues cannot be replaced by any other residue. Thus, mutation from a native Ser residue to any other residue generally appears to be deleterious in TMDs. However, the most striking observations involve arginine: mutations involving the loss or gain of Arg residues, almost without exception, more often produce disease in TMDs versus SDs. The importance of Arg residues in TMDs overall, discussed below, is underscored by the fact that Arg is involved in 9 of the 25 (36%) specific disease-causing mutations that are significantly more frequent in TMDs versus SDs.

Both the horizontal and vertical axes in Figure 3 are arranged in the order of increasing polarity (according to the Liu–Deber scale[15]). This presentation affords the observation of trends related to hydrophobic nature of the native and mutant residues. Specifically, disease-causing mutations occur more frequently within TMDs when they involve the mutation of a native polar residue (there are more red blocks on the lower half of the grid) or when they result in a polar amino acid (there are more red blocks on the right-hand half of the grid). SDs tend to be affected by mutations that involve the mutation of a native nonpolar residue or (similar to TMDs) by mutations that result in a polar residue. This latter result is due largely to the high phenotypic propensity for non-native gain of Ser residues in SDs. This observation suggests that gain of a mutant Ser residue is relatively more damaging to SDs than TMDs—a result consistent with experimental results showing that Ser residues can *only* drive the association of TM helices when multiple Ser side-chains are present on a given helical face.[16]

The overall trends related to amino acid polarity in the propensity for disease-causing mutations in TMDs compared with SDs were examined with the analysis presented in Figure 4. In general, apolar residues are seen to be largely exchangeable in TMDs, because apolar-to-apolar mutations are significantly underrepresented in these domains compared with SDs. Furthermore, the replacement of an apolar residue with any other residue, on the whole, is seen to have a more deleterious effect in SDs, indicating that these residues have correspondingly less important roles in TM helices. On the other hand, mutations involving polar residues are more deleterious in TMDs than in SDs. One exception to this trend is the class of mutations involving a change from an apolar to a polar residue, which occur with equal frequency in TMDs and SDs. However, if the analysis is restricted to mutations *from* apolar residues *to* "strongly" polar residues (residues whose side-chain chemistry allows two or more H-bonding sites, viz., D, E, N, Q, and R), a greater phenotypic propensity is found for TMDs than for SDs. These findings are consistent with observations in model TM peptides that only strongly polar residues (specifically D, E, N, and Q) can drive H-bond-mediated association of polyleucine or randomly hydrophobic helices, whereas weaker polar residues cannot.[17–20]

A clear trend in Figure 4 is the high relative phenotypic propensity of mutations involving ionizable residues. With the exception of the apolar-to-ionizable class of mutation, all general mutation classes involving ionizable residues are highly overrepresented as causing disease in TMDs versus SDs. Even though ionizable residues are often the least common species in these domains, these residues ostensibly play vital roles in TM helices.

Amino Acid Composition of TMDs in Thermophiles

The results obtained with use of the HMGD are indicative of an important role of polar and ionizable residues in stabilizing TMDs. Is it the role of such TMD residues to stabilize membrane domain structure through native interhelical electrostatic cross-linking? To address this possibility, we compared the amino acid compositions of TMDs of a variety of mesophilic prokaryotes with those that have adapted to survive in high-temperature environments. Organisms were divided into three categories (mesophiles, thermophiles and hyperthermophiles) based on their optimal growth temperatures (see Materials and Methods). We hypothesized that thermophilic and hyperthermophilic species would require increased stability of their TM domains, possibly imparted by a greater number of electrostatic links within these domains, which act to prevent protein unfolding in a high-heat environment.

Although a significant increase in polar residues per se was not detected in the TMDs of hyperthermophiles and thermophiles compared with mesophiles, there was an increase when the comparison was limited to ionizable residues [D, E, K, H, and R; see Fig. 5(A)]. These results agree well with a similar study published by Schneider et al.,[21] which found both E and D are each more abundant in thermophiles than in mesophiles. Highlighting this obser-

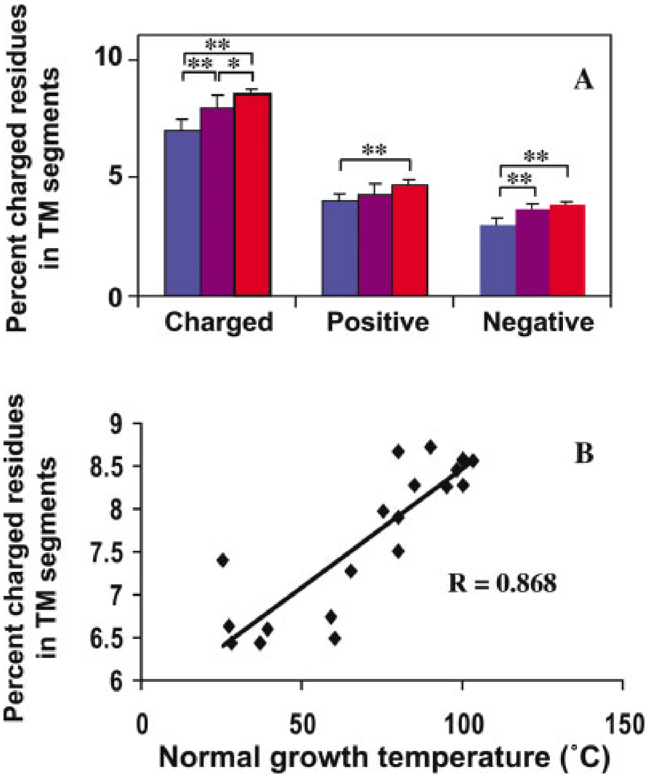

Fig. 5. Amino acid composition of TMDs of mesophilic, thermophilic and hyperthermophilic prokaryotes. **A:** Comparison of the percent charged, positively charged, and negatively charged residues in TMDs of mesophiles (blue bars; *M. mazei, M.a acetovorans, A. tumefaciens, B. subtilis, L. lactis, S. aureus, C. acetobutylicum, E. coli, F. nucleatum, N. meningitides, C. pneumoniae, S. typhi, V. Cholerae, S. coelicolor,* and *R. meliloti*), thermophiles (purple bars; *M. thermoautotrophicum, M. jannaschii, A. fulgidus, S. tokodaii, T. maritime, T. tengcongensis,* and *M. jannaschii*), and hyperthermophiles (red bars; *A. aeolicus, A. pernix, M. kandleri, P. aerophilum, P. abyssi, P. furiosus,* and *P. horikoshii*). Indicated differences are significantly different (*$p < 0.05$, **$p < 0.01$). **B:** Correlation between percent charged residue in the TMDs of prokaryotes (as above) and the temperature at which these prokaryotes optimally grow.

vation, a moderate correlation was found between the average percent ionizable residues in TMDs and the optimal growth temperature of a number of individual organisms [Fig. 5(B)]. These results further suggest that polar residues serve the function of stabilizing TMDs through electrostatic interactions.

DISCUSSION

The substitution of a native residue by a non-native residue can result in either benign effects (e.g., polymorphisms) or in genetic disease. Scanning mutagenesis studies on both TM[22–24] and soluble[25,26] proteins has revealed that most positions in proteins can be altered without serious effects on protein structure or function. Therefore, mutations that are phenotypic for disease are, in effect, "broadcasting" information regarding the importance of certain residues at specific locations in the protein sequence. In this report, we sought to exploit this concept through examination of 240 proteins associated with human disease, as collected in the HMGD, to gather information about the roles of specific amino acid residues in TMDs based on the disease-causing propensities of mutations involving those residues.

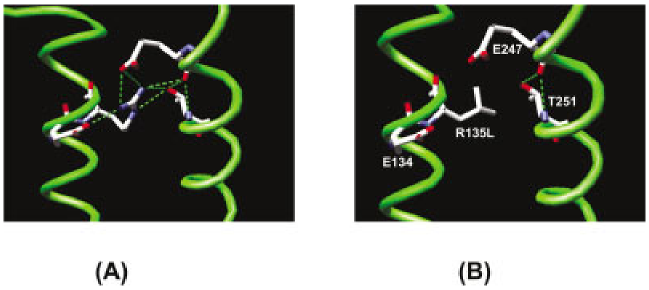

Fig. 6. Embedded arginine residues can act as electrostatic loci that stabilize TM helical packing. This specific example shows TM helices 3 and 6 from the bovine rhodopsin protein with R135 acting as the central player in the electrostatic network connecting the two helices in the wild-type structure. **A:** R135 (TM3) participates in electrostatic interactions with E247 and T251 (TM6) and is tethered back onto TM3 via an interaction with E134. **B:** Mutations of R135 would result in a loss of this extensive electrostatic network; consequences of the R135L phenotypic mutant are depicted here in this model structure. Images were generated by using the rhodopsin crystal structure[47] and SwissPDB viewer.[48]

Although a direct comparison of observed rates of mutation to "expected" rates of mutation was obviated by the aforementioned variability in DNA mutability, we were able to calculate relative phenotypic propensities of mutations in TMDs by comparing mutability patterns to the exact same mutations in SDs. One important validation of our results is that an analysis of mutations involving only the first nucleotide in a codon results in very similar general trends compared with a similar analysis of mutations involving the middle nucleotide, even though the populations of mutations in each of these two sets are completely independent of each other (not shown). Note that although TMDs consist almost exclusively of lipid embedded α-helices, SDs are made up of helical, strand, turn, and unordered structures in both surface and core regions. Thus, it is likely more meaningful to discuss the data from a TMD perspective, as done herein.

Loss of Polar Residues in TMDs

The present results point to particularly important roles for polar and ionizable residues in the membrane. This concept is supported by recent data that suggest membrane-inserted H-bonds form unusually strong links, which play dominant roles in helix–helix interactions[17–20,27,28] and, by inference, in protein folding within membrane domains. Indeed, a recent survey of known TM protein crystal structures revealed that side-chain–side-chain H-bonds exist widely in TMDs.[29] In the present study, we show that mutating ionizable residues causes disease more frequently in TMDs than in SDs (Fig. 4). TMD ionizable side-chains have the potential to form electrostatic interactions that are possibly even stronger than H-bonds (i.e., membrane-embedded salt-bridges). It is unlikely that unpaired ionizable side-chains would exist as membrane-embedded charged species due to the enthalpic penalties involved with burying these highly polar species in the hydrophobic environment. Lipid-embedded opposite charges placed at appropriate distances should form ion pairs, a situation that has in fact been observed in model systems[30] and has been proposed to underlie the stability of the T cell receptor–CD3 complex.[31] It is important that this energetically favorable interaction is up to 20 times stronger in a nonpolar environment (dielectric constant = 4) versus aqueous solution (dielectric constant = 80).[32] In addition to offsetting the desolvation penalty of inserting a polar species into the bilayer, this linkage would help direct helix–helix packing of TMDs, particularly if the residues involved in the salt bridge are in nonsequential helices. Our work supports the idea of a stabilizing role for membrane-embedded salt bridges in two ways. First, mutations involving ionizable residues are significantly overrepresented in TMDs than in SDs (Fig. 4). Second, thermophilic, and especially hyperthermophilic bacteria, have a significantly higher proportion of ionizable residues in their TMDs, suggesting that these organisms have adapted to their extreme environments by increasing the number of stabilizing electrostatic cross-links within their TMDs. It is noteworthy that studies on soluble proteins have revealed that salt bridges can serve to enhance their thermal stability as well.[32–36]

Further support for an important stabilizing role of membrane-embedded salt bridges comes from work on the cystic fibrosis transmembrane conductance regulator (CFTR). Welsh and coworkers[37] found that mutation of residue R347 in TM6 to Glu results in spurious fluctuations in channel conduction states but normal anion conductance, suggesting that R347 plays a role in TMD structure but not in ion conductance.[37] In addition, these workers noted that the deleterious behavior of the R347E mutation can be reversed if complemented by the mutation D924R (TM-8), suggesting an important interhelical salt bridge between R347 and D924 in the native protein. This notion is supported by the fact that mutations of R347 to L, C, P, or H are phenotypic for CF disease (http://www.genet.sickkids.on.ca/cftr/).

Gain of Polar Residues in TMDs

Our analysis indicates that diseases can occur not only after the disruption of native membrane-embedded electrostatic links, as discussed above, but also from the formation of non-native electrostatic links. This concept is supported experimentally by 1) work showing that a CF phenotypic mutation V232D in TM4 can give rise to non-native H-bonds that likely compromise the structure/function of the CFTR protein[28]; and 2) the cancer-inducing mutation V to E in the rat neu receptor that induces non-native interhelical H-bonds.[38,39]

The introduction of an additional polar residue into TM helix increases the free energy of transferring the segment from the aqueous phase into the transmembrane position. Therefore, apolar to polar mutations might cause disease by altering the insertion of a TM segment. However, because most native TM segments have a total hydrophobicity well above the membrane insertion threshold,[40] this phenotypic mechanism is likely restricted to those few helices whose hydrophobicity is close to this threshold.

Functional Roles of Polar Residues in TMDs

In addition to the demonstrable role of polar residues in membrane protein structure, these residues are often involved in important functional mechanisms as well. As such, a higher prevalence in TMDs of disease-causing mutations in which polar residues are lost can be partially explained by the loss of function attributable to such a mutation, independent of structural effects. For example, in rhodopsin the K296 residue plays a functional role by acting as the retinal binding site. The phenotypic mutation K296M prevents retinal binding, thus producing a severe form of the disease retinitis pigmentosa.[41] Apolar to polar mutations can also result in disruption of protein function through nonstructural means. One example is the congenital night blindness phenotype that results from mutation of Gly90 in rhodopsin to Asp.[42] In this case, the mechanism of disease is believed to be related to non-native electrostatic interactions that affect *cis-trans* isomerization of the retinal moiety of the protein, without gross alterations in protein structure.[3,42,43]

Role of Arginine in the Membrane

The present analysis also pinpoints certain specific residues as having particularly important roles in TMDs and/or to be deleterious when inserted into these domains. Arg stands out as one residue to which both situations apply. One may readily appreciate why addition of this residue into a TMD can be problematic: its large size and high polar character (it uniquely contains three polar atoms) will frequently disrupt native structure through steric and/or electrostatic means, regardless of the wild-type residue it is replacing.

The high phenotypic propensity of the reverse situation (mutation of Arg to another residue) indicates that native Arg residues, in turn, play a special role in TMDs. The residue R135 in rhodopsin illustrates how a native Arg residue can be rigorously required for membrane protein structure/function. An analysis of the three-dimensional structure of rhodopsin shows that R135 is involved in an extensive membrane-embedded electrostatic, structure-stabilizing network [Fig. 6(a)]. Specifically, this side-chain interacts with E134, E247, and T251 to hold TM helices 3 and 6 in close proximity. The specific requirement of an Arg residue at this site is demonstrated by the fact that the mutations R135L/W/P/G are all phenotypic for retinitis pigmentosa (www.hgmd.org). Replacing Arg135 with Leu (or any of the indicated residues) would result in the loss of the interactions that hold helices 3 and 6 together, leading to structural destabilization of the protein [Fig. 6(b)].

Another interesting observation is that most of the membrane-buried Arg residues are located at the extremities of the membrane helices (not shown). This observation likely relates, in part, to a role of arginine in anchoring TMDs to the membrane. For example, the simultaneous mutation of the Arg residues in the conserved motif R-X-G-R-R located near the ends of helices 2 and 8 in the Glut1 glucose transporter impairs proper TM topology.[44] However, given the redundancy in basic residue content near TM helical ends, one might predict that mutation of a single TM anchoring residue may not be sufficient to result in human disease.

Mutation of Hydrophobic Residues in TMDs

Apolar residues are generally underrepresented as disease-causing mutations in TMDs compared to SDs (Fig. 4). Because both the hydrophobic insertion and packing of TM helices (the generally accepted roles of these residues) are dependent on the cumulative properties of the full TM segment, altering the hydrophobicity or van der Waals specificity at one site is unlikely to result in major structural/ functional disruptions (although specific exceptions do occur, e.g., the G83I mutation in GpA disrupts TM dimerization[45]). Furthermore, from the reverse perspective, these residues play some critical roles in SDs as the mediators of hydrophobic collapse, which is the main driving force behind soluble protein folding. As well, mutations to the resulting tightly packed protein core may create potentially destabilizing cavities.

Low Phenotypic Propensity of Cys Residues in TMDs

Data in Figure 1 show that loss of native Cys residues is significantly underrepresented in terms of their relative phenotypic propensity. This finding is almost certainly a manifestation of the important role this residue plays in disulfide bonding for the stability of soluble domains.[46] We note that membrane-embedded disulfide bonds have not been observed in native systems, further supporting the stabilizing role of interhelical cross-links between pairs of polar residues in TM domains.

TM-Embedded Ionizable Residues in Thermophiles

The latter portion of the work described here demonstrates that ionizable residues are increased in the TM helices of bacteria living at extreme temperatures (Fig. 5). A similar analysis has been published recently by Schneider et al.[21] It is encouraging, that although the two studies used different methods for determining the putative TM segments and analyzed different sets of organisms, similar results were obtained in that both studies found that acidic residues are increased in thermophiles. Our study also found that if basic residues (Lys, Arg, and including His) are considered as a group for statistical purposes, these species are increased as well in hyperthermophiles versus mesophiles, although the difference is not significant when comparing thermophiles with mesophiles (see also Ref. 21).

CONCLUSIONS

Although polar residues are of modest occurrence in TMDs, they have the potential to form strong electrostatic cross-links, a situation that is commonly observed in known crystal structures. However, these observations alone did not reveal if the individual removal of these cross-linking residues would substantially alter protein activity and/or be implicated in human disease states. The statistical analysis presented here confirms the suspected phenotypic role of polar residues in membrane-spanning regions of proteins: the removal of single polar side-chains often compromises TMD structure/function so extensively that a disease state ensues. As well, mutations involving the gain of polar residues into TMDs are relatively likely to result in disease due to their ability to drive the formation of non-native electrostatic crosslinks (e.g. Ref. 28) The findings also highlight the unique capacity of Arg side-chains to participate in H-bonded "networks" whose loss or gain can influence membrane protein structure profoundly. The overall results will not only further our understanding of the role(s) of individual amino acids in membrane proteins but may also serve as useful predictors of disease associated with this class of proteins.

ACKNOWLEDGMENTS

This work was supported, in part, by grants to C.M.D. from the Canadian Institutes for Health Research (CIHR) and NIH-NIDDK under a SCOR project. A.G.T. held a CIHR Postdoctoral Fellowship and A.W.P. held a CIHR Doctoral Award. We thank Theodore Chiang and Edgar Crowdy of the Hospital for Sick Children Centre for Computational Biology for writing programs that were used for data compilation. We also thank Roman Melnyk of this laboratory for useful discussions.

REFERENCES

1. Popot JL, Engelman DM. Helical membrane protein folding, stability, and evolution. Annu Rev Biochem 2000;69:881–922.
2. Gennis RB. Biomembranes: molecular structure and function. New York: Springer-Verlag; 1989.
3. Partridge AW, Therien AG, Deber CM. Polar mutations in membrane proteins as a biophysical basis for disease. Biopolymers 2002;66:350–358.
4. Gurrath M. Peptide-binding G protein-coupled receptors: new opportunities for drug design. Curr Med Chem 2001;8:1605–1648.
5. Krawczak M, Cooper DN. The human gene mutation database. Trends Genet 1997;13:121–122.
6. Krawczak M, Ball EV, Cooper DN. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 1998;63:474–488.
7. Antonarakis SE, Krawczak M, Cooper DN. Disease-causing mutations in the human genome. Eur J Pediatr 2000;159 Suppl 3:S173–178.
8. Deber CM, Wang C, Liu LP, Prior AS, Agrawal S, Muskat BL, Cuticchia AJ. TM Finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. Protein Sci 2001;10:212–219.
9. Woolfson DN, Williams DH. The influence of proline residues on alpha-helical structure. FEBS Lett 1990;277:185–188.
10. Visiers I, Braunheim BB, Weinstein H. Prokink: a protocol for numerical evaluation of helix distortions by proline. Protein Eng 2000;13:603–606.
11. MacArthur MW, Thornton JM. Influence of proline residues on protein conformation. J Mol Biol 1991;218:397–412.
12. Cordes FS, Bright JN, Sansom MS. Proline-induced distortions of transmembrane helices. J Mol Biol 2002;323:951–960.
13. Tieleman DP, Shrivastava IH, Ulmschneider MR, Sansom MS. Proline-induced hinges in transmembrane helices: possible roles in ion channel gating. Proteins 2001;44:63–72.
14. Sansom MS, Weinstein H. Hinges, swivels and switches: the role of prolines in signalling via transmembrane alpha-helices. Trends Pharmacol Sci 2000;21:445–451.
15. Liu LP, Li SC, Goto NK, Deber CM. Threshold hydrophobicity dictates helical conformations of peptides in membrane environments. Biopolymers 1996;39:465–470.
16. Dawson JP, Weinger JS, Engelman DM. Motifs of serine and threonine can drive association of transmembrane helices. J Mol Biol 2002;316:799–805.
17. Choma C, Gratkowski H, Lear JD, DeGrado WF. Asparagine-mediated self-association of a model transmembrane helix. Nat Struct Biol 2000;7:161–166.

18. Gratkowski H, Lear JD, DeGrado WF. Polar side chains drive the association of model transmembrane peptides. Proc Natl Acad Sci USA 2001;98:880–885.
19. Zhou FX, Merianos HJ, Brunger AT, Engelman DM. Polar residues drive association of polyleucine transmembrane helices. Proc Natl Acad Sci USA 2001;98:2250–2255.
20. Zhou FX, Cocco MJ, Russ WP, Brunger AT, Engelman DM. Interhelical hydrogen bonding drives strong interactions in membrane proteins. Nat Struct Biol 2000;7:154–160.
21. Schneider D, Liu Y, Gerstein M, Engelman DM. Thermostability of membrane protein helix-helix interaction elucidated by statistical analysis. FEBS Lett 2002;532:231–236.
22. Zhou Y, Wen J, Bowie JU. A passive transmembrane helix. Nat Struct Biol 1997;4:986–990.
23. Wen J, Chen X, Bowie JU. Exploring the allowed sequence space of a membrane protein. Nat Struct Biol 1996;3:141–148.
24. Frillingos S, Sahin-Toth M, Wu J, Kaback HR. Cys-scanning mutagenesis: a novel approach to structure function relationships in polytopic membrane proteins. FASEB J 1998;12:1281–1299.
25. Heinz DW, Baase WA, Matthews BW. Folding and function of a T4 lysozyme containing 10 consecutive alanines illustrate the redundancy of information in an amino acid sequence. Proc Natl Acad Sci USA 1992;89:3751–3755.
26. Matthews BW. Structural and genetic analysis of the folding and function of T4 lysozyme. FASEB J 1996;10:35–41.
27. Partridge AW, Melnyk RA, Deber CM. Polar residues in membrane domains of proteins: molecular basis for helix-helix association in a mutant CFTR transmembrane segment. Biochemistry 2002;41:3647–3653.
28. Therien AG, Grant FE, Deber CM. Interhelical hydrogen bonds in the CFTR membrane domain. Nat Struct Biol 2001;8:597–601.
29. Adamian L, Liang J. Interhelical hydrogen bonds and spatial motifs in membrane proteins: polar clamps and serine zippers. Proteins 2002;47:209–218.
30. Wimley WC, Gawrisch K, Creamer TP, White SH. Direct measurement of salt-bridge solvation energies using a peptide model system: implications for protein stability. Proc Natl Acad Sci USA 1996;93:2985–2990.
31. Call ME, Pyrdol J, Wiedmann M, Wucherpfennig KW. The organizing principle in the formation of the T cell receptor-CD3 complex. Cell 2002;111:967–979.
32. Kumar S, Nussinov R. Close-range electrostatic interactions in proteins. Chembiochem 2002;3:604–617.
33. Vieille C, Zeikus GJ. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev 2001;65:1–43.
34. Ladenstein R, Antranikian G. Proteins from hyperthermophiles: stability and enzymatic catalysis close to the boiling point of water. Adv Biochem Eng Biotechnol 1998;61:37–85.
35. Chakravarty S, Varadarajan R. Elucidation of determinants of protein stability through genome sequence analysis. FEBS Lett 2000;470:65–69.
36. Bogin O, Levin I, Hacham Y, Tel-Or S, Peretz M, Frolow F, Burstein Y. Structural basis for the enhanced thermal stability of alcohol dehydrogenase mutants from the mesophilic bacterium Clostridium beijerinckii: contribution of salt bridging. Protein Sci 2002;11:2561–2574.
37. Cotten JF, Welsh MJ. Cystic fibrosis-associated mutations at arginine 347 alter the pore architecture of CFTR. Evidence for disruption of a salt bridge. J Biol Chem 1999;274:5429–5435.
38. Smith SO, Smith C, Shekar S, Peersen O, Ziliox M, Aimoto S. Transmembrane interactions in the activation of the Neu receptor tyrosine kinase. Biochemistry 2002;41:9321–9332.
39. Smith SO, Smith CS, Bormann BJ. Strong hydrogen bonding interactions involving a buried glutamic acid in the transmembrane sequence of the neu/erbB-2 receptor. Nat Struct Biol 1996;3:252–258.
40. Deber CM, Liu LP, Wang C. Perspective: peptides as mimics of transmembrane segments in proteins. J Pept Res 1999;54:200–205.
41. Sullivan JM, Scott KM, Falls HF, Richards JE, Sieving PA. A novel rhodopsin mutation at the retinal binding site (Lys-296-Met) in ADRP. Invest Ophthalmol Vis Sci 1993;34(Suppl):1149.
42. Rao VR, Cohen GB, Oprian DD. Rhodopsin mutation G90D and a molecular mechanism for congenital night blindness. Nature 1994;367:639–642.
43. Sieving PA, Richards JE, Naarendorp F, Bingham EL, Scott K, Alpern M. Dark-light: model for nightblindness from the human rhodopsin Gly-90→Asp mutation. Proc Natl Acad Sci USA 1995; 92:880–884.
44. Sato M, Mueckler M. A conserved amino acid motif (R-X-G-R-R) in the Glut1 glucose transporter is an important determinant of membrane topology. J Biol Chem 1999;274:24721–24725.
45. Lemmon MA, Flanagan JM, Treutlein HR, Zhang J, Engelman DM. Sequence specificity in the dimerization of transmembrane alpha-helices. Biochemistry 1992;31:12719–12725.
46. Matsumura M, Signor G, Matthews BW. Substantial increase of protein stability by multiple disulphide bonds. Nature 1989;342: 291–293.
47. Palczewski K, Kumasaka T, Hori T, Behnke CA, Motoshima H, Fox BA, Le Trong I, Teller DC, Okada T, Stenkamp RE, Yamamoto M, Miyano M. Crystal structure of rhodopsin: A/G protein-coupled receptor. Science 2000;289:739–745.
48. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723.