W370–W375 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 7 May 2012*
doi:10.1093/nar/gks379

Transmembrane helix: simple or complex

Wing-Cheong Wong[1,*], Sebastian Maurer-Stroh[1,2], Georg Schneider[1] and Frank Eisenhaber[1,3,4,*]

[1]Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01, Matrix, Singapore 138671, [2]School of Biological Sciences (SBS), Nanyang Technological University (NTU), 60 Nanyang Drive, Singapore 637551, [3]Department of Biological Sciences (DBS), National University of Singapore (NUS), 8 Medical Drive, Singapore 117597 and [4]School of Computer Engineering (SCE), Nanyang Technological University (NTU), 50 Nanyang Drive, Singapore 637553

Received January 26, 2012; Revised April 4, 2012; Accepted April 12, 2012

ABSTRACT

Transmembrane helical segments (TMs) can be classified into two groups of so-called 'simple' and 'complex' TMs. Whereas the first group represents mere hydrophobic anchors with an overrepresentation of aliphatic hydrophobic residues that are likely attributed to convergent evolution in many cases, the complex ones embody ancestral information and tend to have structural and functional roles beyond just membrane immersion. Hence, the sequence homology concept is not applicable on simple TMs. In practice, these simple TMs can attract statistically significant but evolutionarily unrelated hits during similarity searches (whether through BLAST- or HMM-based approaches). This is especially problematic for membrane proteins that contain both globular segments and TMs. As such, we have developed the transmembrane helix: simple or complex (TMSOC) webserver for the identification of simple and complex TMs. By masking simple TM segments in seed sequences prior to sequence similarity searches, the false-discovery rate decreases without sacrificing sensitivity. Therefore, TMSOC is a novel and necessary sequence analytic tool for both the experimentalists and the computational biology community working on membrane proteins. It is freely accessible at http://tmsoc.bii.a-star.edu.sg or available for download.

INTRODUCTION

The 'modus operandi' of the sequence homology concept is governed by two principles. First is the inference of evolutionary history from sets of homologous protein sequences for building believable phylogenetic trees (1,2) [e.g. 1964, fibronopeptides (3); 1967, cytochrome c (4)]. Second is the inference of sequence–structure–function relationships from well-studied proteins to uncharacterized sequences [e.g. 1967, lactalbumin (5); 1986, angiogenin (6,7)]. The overall concept can be formally rationalized as similarity in amino acid sequence implies, to a certain degree, similarity in 3D structure and, hence, biological function where the conservation of the hydrophobic pattern in amino acid sequence of globular proteins is required to form the tightly packed hydrophobic core of the tertiary structure (8–11). High level of sequence similarity is thought to have originated from common ancestry under the pressure of selection at each step of mutational divergence with rare, alternative instances of convergent evolution (12,13).

When applying the sequence homology concept, there are two important caveats. First, homology (as a hypothesis about common ancestry) can only be inferred via similarity measures. While similarity by chance can be eliminated through strict statistical criteria (e.g. *E*-value cutoff), ambiguity remains between convergent evolution and common ancestry for the high similarity scores (14–16). In practice, alignment tools [e.g. BLAST (17), HMMER (18–20)] do not differentiate between common ancestry and convergent evolution for high similarity scores. Therefore, one must be mindful in distinguishing between long stretches of similarity versus local resemblances that are physiologically constrained (e.g. membrane-spanning stretches from non-polar residues; linkers between globular domains from polar ones) (12). Second, proof of the sequence homology concept stems from cases of globular sequence segments and it is not directly applicable to non-globular ones. In particular, signal-peptides (SP) and transmembrane helices (TM) belong to a special class of non-globular sequences. Their mimicry of hydrophobic core patterns in similarity searches can attract unrelated spurious hits with impressive similarity scores. Essentially, these hits are unrelated to the seed sequence other than some hydrophobic pattern

*To whom correspondence should be addressed. Tel: +65 64788305; Fax: +65 64789047; Email: wongwc@bii.a-star.edu.sg
Correspondence may also be addressed to Frank Eisenhaber. Tel: +65 64788338; Fax: +65 64789047; Email: franke@bii.a-star.edu.sg

matches via their SP/TM segments (21). As collateral damage, such unjustified application of the sequence homology concept to infer homology will result in wrongful annotation, especially in automated annotation pipelines.

With regard to the SPs, their necessary exclusion from seed sequences prior to similarity searches is uncontested since these segments are cleaved away from the mature proteins. However, the exclusion of all TMs is unsatisfactory due to the diverse architecture of membrane proteins (from the single-spanning TM proteins with some globular segments to the multi-spanning TM ones that are connected via loops with essentially no globular segments). In fact, not all TM helices need to be excluded. This is because a TM helix can either be simple or complex (22). Specifically, simple TMs have low sequence complexity but high hydrophobicity and are enriched in aliphatic hydrophobic residues. They merely serve as membrane anchors and can be a result of convergent evolution. In contrast, the complex TMs have higher sequence complexity, lower hydrophobicity and are enhanced with structural, charged and aromatic residues. They have additional functional roles (e.g. ligand binding, active sites, signal transduction) aside from membrane insertion and are likely derived from common ancestry (22). Most importantly, the simple TMs which can be present in membrane proteins regardless of any topology cause spurious hits in similarity searches. This necessitates for their identification and exclusion from the seed sequences prior to similarity searches.

To provide a simple way of identifying and masking simple TMs within a membrane protein sequence, we provide a user-friendly web-interface transmembrane helix: simple or complex (TMSOC). In a nutshell, TMSOC first predicts any TM segments within the sequence if they are not defined by the users. Next, based on the sequence complexity and hydrophobicity of each TM segment, TMSOC will identify the simple TM segments [in accordance with criteria in (22)] and mask them in the fasta-formatted protein sequence that can serve as an input to the BLAST (17) suite or other sequence similarity search routines.

THE WEBSERVER

Input description

TMSOC requires: (i) a fasta-formatted sequence as a mandatory input and (ii) the associated TM segments as an optional input.

Output description

TMSOC produces four sections in the output. First, TMSOC displays the sequence with complex, twilight and simple TMs colored in red, orange and blue, respectively (see Figure 1A). Next, a summary table that contains: (i) the indices and (ii) sequences of the TM segments, (iii) the positions of the predicted/user-defined TM segments, (iv) the sequence complexity, (v) hydrophobicity, (vi) *z*-score and (vii) classification [simple/twilight/complex based on (22)] for each TM segment, is given (see Figure 1B). The third section outputs a sequence complexity/hydrophobicity plot of the predicted/user-defined TM segments (in black) against the background of membrane anchors (in blue), functional TMs (in red) and α-helices (in green) from the SCOP (23,24) database (see Figure 1C). Finally, the last section displays the fasta-formatted input sequence with the masked simple TMs (replaced by a continuum of 'X'). This output sequence serves as an input into any appropriate similarity search routines (see Figure 1D).

Workflow description

Behind the web-interface, TMSOC is comprised of two main computational steps (see 'Materials and Methods' section for detail). In the first step, if the user does not input any TM segments, the presence and length of any TM helices within the input protein sequence will be derived from a set of five TM predictors [DASTM (25,26), TMHMM (27), HMMTOP (28), SAPS (29), PhobiusTM (30,31)] where the TM prediction results are statistically combined as described in (21). In most situations, a predicted TM segment will correspond to a TM helix. However, it is possible that the predicted TM segment may contain more than one TM helices in situations where the various TM predictors output varying TM helix borders. It is strongly recommended to the users to enter the TM segments and to use the TM prediction option only as the next best alternative since the predicted TM boundaries might be inaccurate. In the second step, each user-defined or predicted TM segment will be assigned a *z*-score that is calculated from the sequence complexity and hydrophobicity of each segment in accordance with Equations (1–3) in (22). A *z*-score criterion, that is associated to some preset false-negative rates (FNRs), will then be applied to determine if each TM segment is simple, twilight or complex (22). Subsequently, only the simple TMs will be masked (replaced by a continuum of 'X') in the input protein sequence.

If the application of Phobius (30,31) generates a predicted signal peptide that is non-overlapping with a TM region, this signal peptide will be removed from the masked sequence. In the case of an overlap, a warning will be issued.

The backend code for TMSOC was developed as PERL modules while the web-interface was written as a CGI script. The WWW server is available via http://tmsoc.bii.a-star.edu.sg or http://mendel.bii.a-star.edu.sg/METHODS/TMSOC/cgi-bin/. Alternatively, the TMSOC program is freely available for download as a command-line version at the WWW server site. Note that the command-line TMSOC will contain only the TM classification module.

MATERIALS AND METHODS

The algorithms used in TMSOC are described in our previous publications (21,22) in detail. For brevity, only a simplistic outline is given here.

1. Sequence overview

A

MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIP
EGMQCSCGIDYYTPHEETNNESFVIYMFVVHFIIPLIVIFFCYGQLVFTVKEAAAQQQES
ATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSAV
YNPVIYIMMNKQFRNCMVTTLCCGKNPLGDDEASTTVSKTETSQVAPA

(TM classification: complex-TM, twilight-TM, simple-TM)

2. TM segment(s) summary

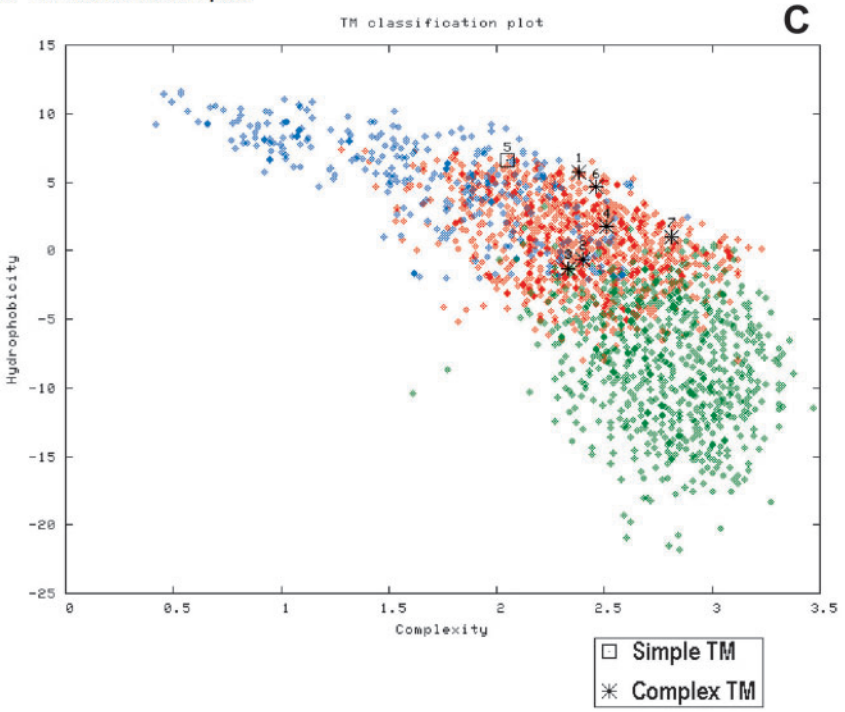
B

Index	Sequence	Predicted TM segment	Complexity	Hydrophobicity	Zscore	Class
1	FSMLAAYMFLLIMLGFPINFLTLYVTV	37,63	2.38	5.75	-3.23	complex
2	YILLNLAVADLFMVFGGFTTTLYTSLHG	74,101	2.41	-0.89	0.29	complex
3	NLEGFFATLGGEIALWSLVVLAIERYV	111,137	2.33	-1.29	0.50	complex
4	AIMGVAFTWVMALACAAPPLVGW	153,175	2.51	1.72	0.27	complex
5	SFVIYMFVVHFIIPLIVIFFCYGQLVFTV	202,230	2.05	6.57	-5.70	simple
6	RMVIIMVIAFLICWLPYAGVAFYIFT	252,277	2.46	4.67	-2.04	complex
7	FGPIFMTIPAFFAKTSAVYNPVIYIMM	283,309	2.81	1.03	1.91	complex

FW: aromatic residues RDEH: charged residues GP: structurally important residues (based on ClustalX color code)

3. TM classification plot

C



(TM types: Membrane anchors, Functional TM helices, SCOP Alpha helices , your predicted TMs)

4. Masked FASTA sequence

D

>P02699_Bovine_Rhodopsin
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEPWQFSMLAAYMFLLIMLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVFGPTGCNLEGFFATLG
GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLVGWSRYIP
EGMQCSCGIDYYTPHEETNNEXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXKEAAAQQQES
ATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSAV
YNPVIYIMMNKQFRNCMVTTLCCGKNPLGDDEASTTVSKTETSQVAPA

Figure 1. Example output of TMSOC analysis for the bovine rhodopsin (P02699) sequence. Generally, TMSOC produces four sections (see **A–D** of Figure 1) for each analysis. In Figure 1A, the sequence of the bovine rhodopsin reveals six complex TMs (in red) and one simple TM (in blue). There are no twilight TMs in this case, otherwise they will be colored in orange. In Figure 1B, a summary table that contains: (i) the indices and (ii) sequences of the TM segments, (iii) the positions of the predicted or user-defined TM segments, (iv) the sequence complexity, (v) hydrophobicity,

(continued)

Statistical quantification of TM segments

The input sequence is first analyzed by five TM predictors [DASTM (25,26), TMHMM (27), HMMTOP (28), SAPS (29), PhobiusTM (30,31)]. For every j-th position in the sequence, the total logarithmic probability for M predictors is given as:

$$\log \hat{p}_{j,\text{total}} = \sum_{m=1}^{M} \log \hat{p}_{j,m}$$

where $\hat{p}_{j,m}$ is a Bernoulli random variable and it takes either 1 for positive TM detection or 0 (in the implementation, it is set as 0.01 so that logarithm can be evaluated) for negative TM prediction. Then, for each TM segment, the average logarithmic probability is given as:

$$\log\langle\hat{p}\rangle = \frac{1}{R}\sum_{r}^{r+R} \log \hat{p}_{r,\text{total}}$$

where R is the total number of predicted residues for the TM segment and r is the starting position of the TM segment. The cutoff criterion for a valid TM segment is set at $\log\langle\hat{p}\rangle \geq -12$ which corresponds to an approximate false-positive rate of 5% and FNR of 8% (21).

Quantitative criteria for identifying simple and complex TM segments

The z-score of each TM segment is calculated from its associated sequence complexity and hydrophobicity (22). It is given as:

$$z(x_\Phi, x_c) = (-1)^s \left[\frac{(x_\Phi - \mu_\Phi)^2}{\sigma_\Phi^2} + \frac{(x_c - \mu_c)^2}{\sigma_c^2}\right]$$

where x_c and x_Φ are the moving window averages for sequence complexity c (32) and hydrophobicity Φ (33,34) for a given segment, μ and σ are the mean and SD of sequence complexity and hydrophobicity for the functional TM set [defined as TMs containing active residues. See 'Methods and Materials' section in (22) for details]. The exponent s is set to one if:

$$\frac{x_\Phi - \mu_\Phi}{\sigma_\Phi} \geq -\frac{1}{\rho_{c,\Phi}}\frac{x_c - \mu_c}{\sigma_c}$$

and zero otherwise. $\rho_{c,\Phi}$ is the correlation between sequence complexity and hydrophobicity for the set of functional TMs. For determining simple, twilight or complex TMs, the cutoff criterion is given as:

$$z_{\text{threshold}}(\mu_\Phi + f\sigma_\Phi, \mu_c + f\sigma_c)$$

where $f = 0.840, 1.000, 1.282, 1.645, 1.980$ [corresponding to FNRs of 20, 16, 10, 5 and 2.5%]. Note that simple TMs are declared at FNR of 5% and below, twilight TMs at FNR of between 5% and 10% and complex TMs at FNR of 10% and above.

PROOF OF CONCEPT

The workflow in TMSOC was previously applied to domain and sequence databases for generating the results in Refs (21,22), collectively showing: (i) the importance of identifying 'simple' and 'complex' TMs, (ii) the necessity of 'simple' TM removal prior to similarity searches without sacrificing sensitivity and (iii) the expected number of simple/complex TMs per protein. Specifically, simple and complex TMs were successfully identified by TMSOC in the 7-TM rhodopsin (P02699), 6-TM bacterial rhomboid protease (P09391), *E. coli* aspartate receptor (P07010) and colicin (PDB:1COL) where only the complex ones were experimentally shown to be functionally important (22).

Here, we further illustrate useful insights that TMSOC can provide for the bovine rhodopsin sequence (P02699). To recapitulate, the functional role of TM-5 in the latter has not been established whereas the Gly51 in TM-1 and Gly89 in TM-2 have been linked to the retinal degenerative disease autosomal dominant retinitis pigmentosa (35) while Glu113 in TM-3, Ala169 in TM-4, Trp265 in TM-6 and Lys296 in TM7 are functionally important (36,37). Indeed, only TM-5 [positions 200–225 (38); z-score of −6.12] in bovine rhodopsin is considered simple by TMSOC and was masked in the sequence. The PSI-BLAST search results of the original and masked bovine rhodopsin (P02699) were then generated (see Supplementary Data S1 and S2) for further investigations.

TMSOC detects heterogeneity in transmembrane helix 5 among the rhodopsins

Orthologous rhodopsin hits (107) were detected in our PSI-BLAST runs (five iterations with standard parameters against Swiss-Prot). Consequently, they were analyzed by TMSOC for simple and complex TMs as summarized in Table 1. Based on the results, TM-5 shows the highest percentage of simple TM at ~10%, followed by TM-1 at >1% while the rest contain no simple TMs. In a nutshell, besides bovine, many species of fishes (e.g. OPSD_DANRE, OPSD_CYPCA, OPSD_CARAU, OPSD_LEOKE) and amphibians (e.g. OPSD_RANCA, OPSD_RANPI, OPSD_BUFMA) also possess simple TM-5. Essentially, these collective findings from TMSOC suggests heterogeneity in TM-5 among rhodopsins and this is in agreement with previous report that

Figure 1. Continued
(vi) z-score and (vii) classification [simple/twilight/complex based on (22)] for each TM segment in the bovine rhodopsin sequence is shown. In addition, enriched functional residues (aromatic/charged/structurally related) in the complex TMs are coded with the ClusterX color scheme. Figure 1C depicts the sequence complexity/hydrophobicity plot of the predicted TM segments of the bovine rhodopsin (in black) against the background of membrane anchors (in blue), functional TMs (in red) and α-helices (in green) from the SCOP (23,24) database. Figure 1D shows the fasta-formatted bovine rhodopsin sequence with its simple TM masked by a continuum of 'X' which can serve as an input into any appropriate similarity search routines.

Table 1. Percentage of simple TMs found within each TM helix (1–7) among the 108 (including bovine seed sequence) PSI-BLAST rhodopsin hits

	TM-1	TM-2	TM-3	TM-4	TM-5	TM-6	TM-7
Complex	76	108	108	105	80	99	104
Twilight	31	0	0	3	18	9	0
Simple	1	0	0	0	10	0	0
Percentage of simple TMs	0.93	0.00	0.00	0.00	9.25	0.00	0.00

The first column describes the type of TM helices and the percentage of simple TMs for each of the seven helices across all 108 rhodopsins. Specifically, the second to last columns detail the specific numbers and percentages of each TM helix. Note that TM-7 of some hits was undetected.

TM-5 exhibits the highest level of sequence divergence among the seven TM helices in GPCR (39). Indeed, a comparison of crystal structures between bovine and squid rhodopsin reveals that TM-5 can be different. Notably, TM-5 and TM-6 of squid rhodopsin (P31356) extends into the cytoplasmic medium. This unusual structure, that is not observed in bovine, is regarded as an important structural motif for coupling with G_q-type G protein. In particular, TM-5 is divided into a membrane embedded region and a medium-exposed region that has motional freedom due to a flexible joint at Ser266 (40). This TM helix [positions 195–239 (38); *z*-score is 0.41] is considered complex by TMSOC.

Exclusion of simple TM in bovine rhodospin clarifies the sequence similarity distance between rhodopsin and cholecystokinin-1 receptors

A comparison between the PSI-blast output hit lists of the original and of the masked bovine rhodopsin sequence revealed a cluster of cholecystokinin-1 receptors (CCKAR_CAVPO, CCKAR_RAT, CCKAR_MOUSE, CCKAR_HUMAN, CCKAR_CANFA, CCKAR_RABIT) that was present in the original list of top 500 hits, was excluded from that of the masked one. This cluster of CCKAR was analyzed by TMSOC and all the respective sequences have in common simple TM-1, TM-5, TM-6 but complex TM-2, TM-3, TM-4 and TM-7. These findings suggest that the sequence similarity between rhodopsins and cholecystokinin-1 receptors has been overestimated if their simple TM-5 s are included into the alignment. As it turns out, classical homology modeling of CCKAR using rhodopsin as the reference structure leads to a model that cannot correctly accommodate the cholecystokinin (CCK) ligand in its binding site due to obvious structural divergence. Notably, the binding site of CCK requires a number of residues located in the extracellular surface as well as the upper third of the TM helices whereas most binding residues in rhodopsin are buried in the TM helices (41,42).

CONCLUSION

TMSOC enables researchers to identify simple and complex TMs in membrane proteins for differentially treating them in sequence similarity searches and for planning further functional characterization of membrane proteins.

SUPPLEMENTARY DATA

Supplementary Data is available at NAR Online: Supplementary Data 1 and 2.

ACKNOWLEDGEMENTS

The authors thank Hong-Sain Ooi for his technical advice.

FUNDING

Agency of Science, Technology and Research (A*STAR). Funding for open access charge: Biomedical Research Council (A*STAR).

Conflict of interest statement. None declared.

REFERENCES

1. Dayhoff,M.O. (1969) Computer analysis of protein evolution. *Sci. Am.*, **221**, 86–95.
2. Jardine,N., van Rijsbergen,C.J. and Jardine,C.J. (1969) Evolutionary rates and the inference of evolutionary tree forms. *Nature*, **224**, 185.
3. Doolittle,R.F. and Blombach,E. (1964) Amino-acid sequence investigations of fibrinopeptides from various mammals: evolutionary implications. *Nature*, **202**, 147–152.
4. Fitch,W.M. and Margoliash,E. (1967) Construction of phylogenetic trees: a method based on mutational distances as estimated from cytochrome c sequences is of general applicability. *Science*, **155**, 279–284.
5. Brew,K., Vanaman,T.C. and Hill,R.L. (1967) Comparison of the amino acid sequence of bovine alpha-lactalbumin and hens egg white lysozyme. *J. Biol. Chem.*, **242**, 3747–3749.
6. Allen,S.C., Acharya,K.R., Palmer,K.A., Shapiro,R., Vallee,B.L. and Scheraga,H.A. (1994) A comparison of the predicted and X-ray structures of angiogenin. Implications for further studies of model building of homologous proteins. *J. Protein Chem.*, **13**, 649–658.
7. Palmer,K.A., Scheraga,H.A., Riordan,J.F. and Vallee,B.L. (1986) A preliminary three-dimensional structure of angiogenin. *Proc. Natl Acad. Sci. USA*, **83**, 1965–1969.
8. Bork,P. and Gibson,T.J. (1996) Applying motif and profile searches. *Methods Enzymol.*, **266**, 162–184.
9. Devos,D. and Valencia,A. (2000) Practical limits of function prediction. *Proteins*, **41**, 98–107.
10. Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
11. Todd,A.E., Orengo,C.A. and Thornton,J.M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, **307**, 1113–1143.
12. Doolittle,R.F. (1994) Convergent evolution: the need to be explicit. *Trends Biochem. Sci.*, **19**, 15–18.
13. Gough,J. (2005) Convergent evolution of domain architectures (is rare). *Bioinformatics.*, **21**, 1464–1471.
14. Doolittle,R.F. (1981) Similar amino acid sequences: chance or common ancestry? *Science*, **214**, 149–159.
15. Doolittle,R.F. (1989) Similar amino acid sequences revisited. *Trends Biochem. Sci.*, **14**, 244–245.
16. Reeck,G.R., de,H.C., Teller,D.C., Doolittle,R.F., Fitch,W.M., Dickerson,R.E., Chambon,P., McLachlan,A.D., Margoliash,E., Jukes,T.H. *et al.* (1987) "Homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell*, **50**, 667.

17. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
18. Eddy,S.R. (2004) What is a hidden Markov model? *Nat. Biotechnol.*, **22**, 1315–1316.
19. Eddy,S.R. (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput. Biol.*, **4**, e1000069.
20. Wong,W.C., Maurer-Stroh,S. and Eisenhaber,F. (2011) The Janus-faced E-values of HMMER2: extreme value distribution or logistic function? *J. Bioinform. Comput. Biol.*, **9**, 179–206.
21. Wong,W.C., Maurer-Stroh,S. and Eisenhaber,F. (2010) More than 1,001 problems with protein domain databases: transmembrane regions, signal peptides and the issue of sequence homology. *PLoS Comput. Biol.*, **6**, e1000867.
22. Wong,W.C., Maurer-Stroh,S. and Eisenhaber,F. (2011) Not all transmembrane helices are born equal: Towards the extension of the sequence homology concept to membrane proteins. *Biol. Direct.*, **6**, 57.
23. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
24. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
25. Cserzo,M., Eisenhaber,F., Eisenhaber,B. and Simon,I. (2002) On filtering false positive transmembrane protein predictions. *Protein Eng*, **15**, 745–752.
26. Cserzo,M., Eisenhaber,F., Eisenhaber,B. and Simon,I. (2004) TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. *Bioinformatics.*, **20**, 136–137.
27. Sonnhammer,E.L., von,H.G. and Krogh,A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.
28. Tusnady,G.E. and Simon,I. (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.*, **283**, 489–506.
29. Brendel,V., Bucher,P., Nourbakhsh,I.R., Blaisdell,B.E. and Karlin,S. (1992) Methods and algorithms for statistical analysis of protein sequences. *Proc. Natl Acad. Sci. USA*, **89**, 2002–2006.
30. Kall,L., Krogh,A. and Sonnhammer,E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **338**, 1027–1036.
31. Kall,L., Krogh,A. and Sonnhammer,E.L. (2007) Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. *Nucleic Acids Res.*, **35**, W429–W432.
32. Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.
33. White,S.H. and Wimley,W.C. (1998) Hydrophobic interactions of peptides with membrane interfaces. *Biochim. Biophys. Acta*, **1376**, 339–352.
34. White,S.H. and Wimley,W.C. (1999) Membrane protein folding and stability : physical principles. *Annu. Rev. Biophys. Biomol. Struc.*, **28**, 365.
35. Bosch,L., Ramon,E., Del Valle,L.J. and Garriga,P. (2003) Structural and functional role of helices I and II in rhodopsin. A novel interplay evidenced by mutations at Gly-51 and Gly-89 in the transmembrane domain. *J. Biol. Chem.*, **278**, 20203–20209.
36. Borhan,B., Souto,M.L., Imai,H., Shichida,Y. and Nakanishi,K. (2000) Movement of retinal along the visual transduction path. *Science*, **288**, 2209–2212.
37. Li,J., Edwards,P.C., Burghammer,M., Villa,C. and Schertler,G.F. (2004) Structure of bovine rhodopsin in a trigonal crystal form. *J. Mol. Biol.*, **343**, 1409–1438.
38. Shimamura,T., Hiraki,K., Takahashi,N., Hori,T., Ago,H., Masuda,K., Takio,K., Ishiguro,M. and Miyano,M. (2008) Crystal structure of squid rhodopsin with intracellularly extended cytoplasmic region. *J. Biol. Chem.*, **283**, 17753–17756.
39. Bywater,R.P. (2005) Location and nature of the residues important for ligand recognition in G-protein coupled receptors. *J. Mol. Recognit.*, **18**, 60–72.
40. Murakami,M. and Kouyama,T. (2008) Crystal structure of squid rhodopsin. *Nature*, **453**, 363–367.
41. Archer,E., Maigret,B., Escrieut,C., Pradayrol,L. and Fourmy,D. (2003) Rhodopsin crystal: new template yielding realistic models of G-protein-coupled receptors? *Trends Pharmacol Sci.*, **24**, 36–40.
42. Archer-Lahlou,E., Tikhonova,I., Escrieut,C., Dufresne,M., Seva,C., Pradayrol,L., Moroder,L., Maigret,B. and Fourmy,D. (2005) Modeled structure of a G-protein-coupled receptor: the cholecystokinin-1 receptor. *J. Med. Chem.*, **48**, 180–191.