RHYTHM—a server to predict the orientation of transmembrane helices in channels and membrane-coils

Alexander Rose[1], Stephan Lorenzen[2], Andrean Goede[3], Björn Gruening[4] and Peter W. Hildebrand[1,*]

[1]Institute for Medical Physics and Biophysics, Charité, University Medicine Berlin, Ziegelstrasse 5-9, 10098 Berlin, [2]Institute for Theoretical Biology, Humboldt University, Invalidenstraße 43, 10115 Berlin, [3]Institute for Biochemistry, Charité, University Medicine Berlin, Monbijoustrasse 2, 10117 Berlin and [4]Institute of Physiology, Charité, University Medicine Berlin, Arnimallee 22, 14195 Berlin, Germany

ABSTRACT

RHYTHM is a web server that predicts buried versus exposed residues of helical membrane proteins. Starting from a given protein sequence, secondary and tertiary structure information is calculated by *RHYTHM* within only a few seconds. The prediction applies structural information from a growing data base of precalculated packing files and evolutionary information from sequence patterns conserved in a representative dataset of membrane proteins ('Pfam-domains'). The program uses two types of position specific matrices to account for the different geometries of packing in channels and transporters ('channels') or other membrane proteins ('membrane-coils'). The output provides information on the secondary structure and topology of the protein and specifically on the contact type of each residue and its conservation. This information can be downloaded as a graphical file for illustration, a text file for analysis and statistics and a PyMOL file for modeling purposes. The server can be freely accessed at: URL: http://proteinformatics. de/rhythm

INTRODUCTION

About one third of the presently mapped gene sequences encode for membrane proteins, which are also major targets for pharmaceutical products (1,2). In contrast, only a minor fraction (February 2009, 1.8%) of the protein structures deposited in the protein data bank (PDB) belongs to this structural class (3,4). Due to difficulties in over expression and crystallization, their tertiary structure is often evaluated using computational methods (5–8).

Homology modeling may be applied when an appropriate template structure is available (9). In other cases, *ab initio* or knowledge-based tertiary structure modeling comes into play. There is a high level of predictability regarding secondary structure elements (10–12). New approaches deal with the prediction of the exact lengths of the transmembrane helices (13). Finally, transmembrane topology prediction was optimized applying consensus predictions also identifying signal peptides (14). However, tools that perform or assist in low resolution tertiary structure modeling of helical membrane proteins are still rare (15–19).

The growing data on high-resolution structures of helical membrane proteins provide an appropriate base for structural analysis, statistics and the development of knowledge-based prediction methods (12,15–32). The type of packing of α-helices is fundamental for the stabilization and function of all helical membrane proteins (33–36). Residues involved in helix–helix interactions are therefore regularly more conserved than others and are often arranged in specific sequence motifs that reflect the type of packing (23,25,35,37). Right-handed parallel and anti-parallel interactions are typically found in channels (membrane proteins with a functional pore). These interactions are mainly accomplished by weakly polar amino acids $(G > S > T > F)$ that preferably create contacts every fourth residue (23,37,38). Left-handed anti-parallel interactions are predominantly found in membrane-coils. There, large and polar residues $(D > S > M > Q)$ create characteristic contacts every 3.5th residues (23,37,39).

The higher conservation of residues involved in helix–helix contacts was applied in methods predicting tertiary structure contacts (40,41). Such applications can be further improved combining conservation criteria with amino acid propensity scales (18,24,32,42). The combination of statistical potentials with fragment-based modeling and energy minimizations were applied for *de novo* modeling

*To whom correspondence should be addressed. Tel: +49 30 450 524 155; Fax: +49 30 450 524 952; Email: peter.hildebrand@charite.de
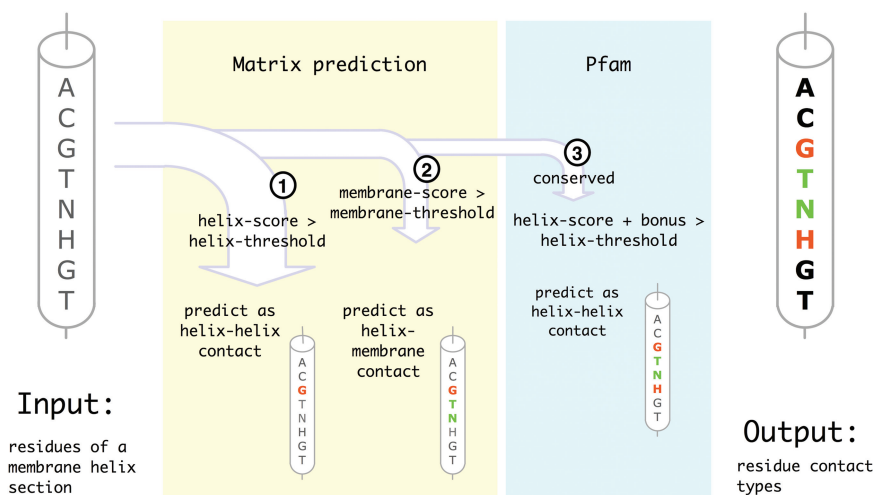
Figure 1. Workflow of *RHYTHM*: the prediction is performed in three steps including (1) matrix prediction of helix–helix contacts; (2) matrix prediction of helix–membrane contacts and (3) prediction of helix–helix contacts by conservation criteria (Pfam domains) (52).

approaches (28,43–45). There are some tools available to predict buried versus exposed regions of transmembrane helices. ProperTM (18), LIPS (16), RANTS (15) and TMX (19) depend on multiple sequence alignments to produce predictions about transmembrane helix orientations or solvent accessibility. However, the quality of prediction by these tools largely depends on the quality of the multiple sequence alignment provided by the user. Due to the small size of several transmembrane protein families, such alignments are not always at hand. Moreover, the output is not always presented in a user-friendly format and thus cannot be directly used for modeling purposes.

RHYTHM is the first server that predicts the exposure or burial of transmembrane residues incorporating the structural specificities of channels. The quality of prediction (expressed by AUC-values) of helix–helix contacts rises by 16% to an average value of 76% when the sequence motifs typical for channels are applied, compared to the same approach when a non-specific matrix is taken (23). For our web service, the position-specific matrices were updated using an enlarged data set of input structures. To optimize the sensitivity of helix–helix contact predictions at high specificity thresholds, the matrix prediction method is now combined with a prediction directly applying evolutionary information from 'Pfam-domains' (46,47). *RHYTHM* also integrates the secondary structure prediction tool HMMTOP (48). Thus, after the upload of a single sequence file and the specification of the position specific matrix type ('channel' or 'membrane-coil') the prediction for tertiary structure contacts is started.

METHODS

Matrix prediction method

The prediction of buried versus exposed residues is based on two different sets of propensity matrices derived from representative and non-redundant datasets of 21 channels and 14 membrane-coils containing 310 and 179

transmembrane helices, respectively (see website for details). The data were analyzed as described in detail in earlier analyses (23,49). Shortly, helical sections were defined by the Kabsch and Sander algorithm (50). Only those residues were defined as transmembrane helixes with their Cα-atoms lying between the two membrane planes. The membrane planes were calculated applying the output of the TMDET algorithm (51). The type of contact of a specified residue was determined counting the atomic contacts to residues of another helix, to the virtual membrane or to virtual water (23). Structures with helix pairs too far apart were removed after visual inspection.

The matrices (which will be regularly updated due to the growing data set of high resolution membrane protein structures) store the propensities of residues to contact another helix or the membrane. To account for sequence motifs, propensities of all neighboring amino acids are stored in the same matrix [see website for details or ref. (23)]. Scores are calculated by summation of the residual propensities at positions 0 to ± 4 (channels) or 0 to ± 7 (membrane-coils). These windows account for the different *RHYTHM* of contacts in channels and membrane-coils (23,25,38). An amino acid is predicted to be buried (step 1, see Figure 1) or exposed (step 2, see Figure 1), when this score is above a certain threshold specified by the user. The advantage of this approach is that the prediction is thus much less affected by variations of single amino acid propensities. However, amino acids at the helix termini are not recorded by our method.

Conservation criteria

The Pfam database is an extensive set of protein domains and families currently covering 72% of known protein sequences (46). The families consist of multiple alignments of functionally or evolutionary-related protein sequences (47). These alignments also reproduce evolutionary relationships that would otherwise not be detected (9). To search the Pfam database, HMMER (version 2.3.2) is applied (52). HMMER allows for sensitive

searching in a database of the consensus sequences of various protein families using Hidden Markov Models. To speed up the search the Pfam database was restricted to the 691 membrane protein families provided in February 2009. A bonus is added to the helix–helix score of fully conserved residues, according to the finding that conserved residues are often involved in helix–helix contacts (37,53). The value of the bonus depends on the selected specificity and is optimized for highest accuracy.

Three step prediction

A three-step approach was applied to predict buried versus exposed residues (Figure 1):

(1) Matrix prediction of helix–helix contacts: Amino acids predicted by HMMTOP (48) or specified by the user to be part of a transmembrane helix are scored. The prediction matrix has to be chosen by the user. In order to do this, the user must know whether the protein of the uploaded sequence has a functional pore (channels) or not (membrane-coil). The residues above the selected specificity threshold for helix–helix contacts (medium, high, very high and highest) are predicted as helix–helix contacts. The specificities for a single contact type range from about 75% for medium to 90% for very high thresholds.
(2) Matrix prediction of helix–membrane contacts: The remaining residues are predicted analogously as helix–membrane contacts using the threshold specified by the user at the beginning of the procedure. The specificities for that prediction also range from about 75–90%, respectively. In conjunction, a maximum of 70% (medium specificity) of the residues are recorded at the moment by the matrix prediction method.
(3) Pfam prediction: To optimize the sensitivity of helix–helix contact predictions at high specificity thresholds, residues not recorded by the matrix prediction method may be verified using conservation criteria. This means that a bonus optimized for the positive predictive value at a selected specificity threshold is added to the helix–helix score from matrix prediction. A residue is predicted as buried when the combined score is above the defined threshold. As a result a plus of 10–20% residues are additionally assigned to be part of a helix–helix contact. The specificity of prediction is not significantly affected by the Pfam prediction.

RESULTS AND DISCUSSION

Performance of the combined prediction

The prediction quality of *RHYTHM* improved compared to our previous analysis (23). This is due to the enlarged data set of helical membrane proteins and the combination of the matrix prediction method with the prediction from evolutionary conservation. The average AUC-values (from a leave-one-out cross validation) for the prediction

of helix–helix contacts are 0.72 for channels [as in our previous analysis (23)] and 0.68 for membrane–coils, respectively. The corresponding values for the prediction of helix–membrane contacts are 0.75 and 0.73. Best predictions were obtained for helix–helix contacts of the translocon channel (PDB-entry: 1rh5, AUC-value: 0.78) and for helix–membrane contacts of the ABC-transporter protein (PDB-entry: 2qi9, AUC-value: 0.86). To receive high quality predictions with *RHYTHM*, we suggest selecting the default specificity threshold 'very high'. This threshold may then be reduced if too few contacts are predicted. Besides tertiary structure contact types, the output assigns the secondary structure and topology of the protein (51). This information is provided as a graphical file for illustration (Figure 2), a text file for analysis and statistics and a PyMOL file for modeling purposes (Figure 3).

### Complexity of tertiary contact predictions

The quality of prediction will further improve as the data set of non-homologous high resolution membrane protein structures grows. At the moment the prediction is limited for several reasons: A significant number of buried residues is close to internal cavities (37,54). Such residues are not judged in our analysis to be part of a helix–helix contact due to insufficient contacts to other residues and are thus often evaluated as false positives in our prediction. Large packing defects regularly account for structural flexibilities (36,55–57). The separate prediction of residues involved in packing defects could therefore enhance the prediction of tertiary structure contacts. Moreover, about one quarter of the residues is in contact with both another helix and the membrane. These residues are frequently not recorded at high specificity thresholds but will be predicted as buried or exposed at lower thresholds. This ambiguity clearly complicates the prediction, as well as the fact that many channels are highly flexible. Residues that are buried in one functional state may become exposed in another (45,58). Finally, residues that appear to be exposed to lipid may become (and may also be predicted to be) buried in quaternary complexes (59). With more structural data of channels a prediction of residues that are exposed or buried depending on their functional state will be possible.

Technical details

All computations are done on our server including optional prediction of membrane helix sections and searches for Pfam domains. Modern web technologies (AJAX, JavaScript, PHP, CSS) were used to create a fast and intuitively usable web application.
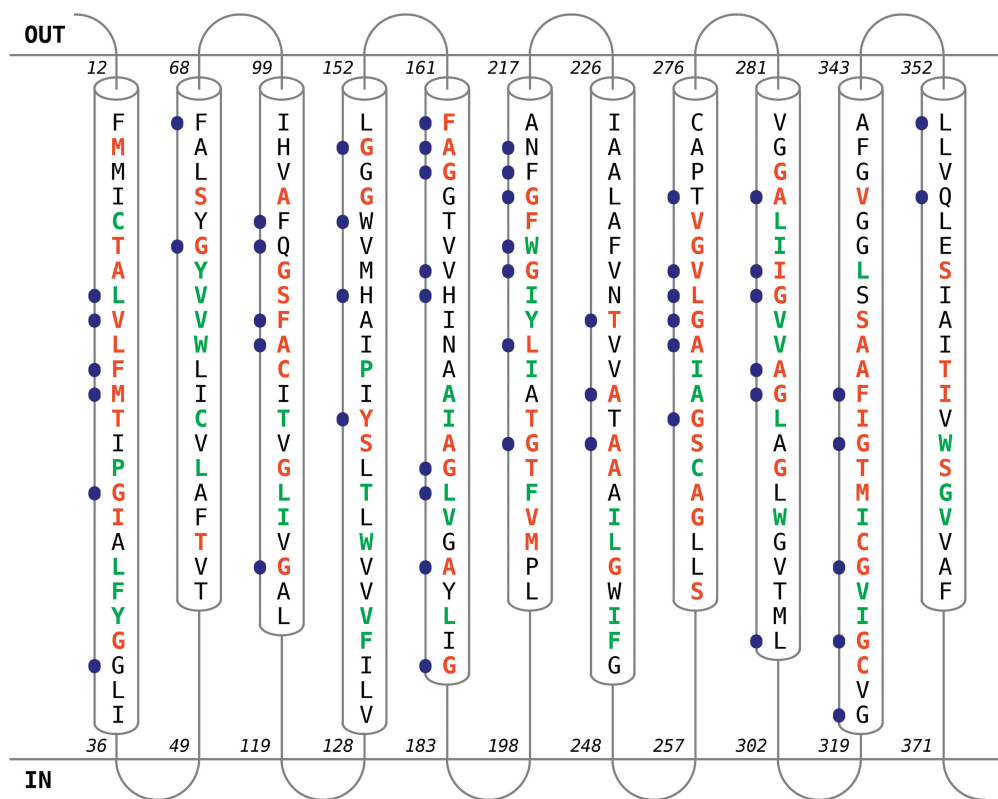
Figure 2. Example graphical output of *RHYTHM*: topology of the ammonium transporter predicted with HMMTOP (51). Tertiary structure contacts predicted as helix–helix contacts (red) or helix–membrane contacts (green). Highly conserved residues are denoted with blue dots.
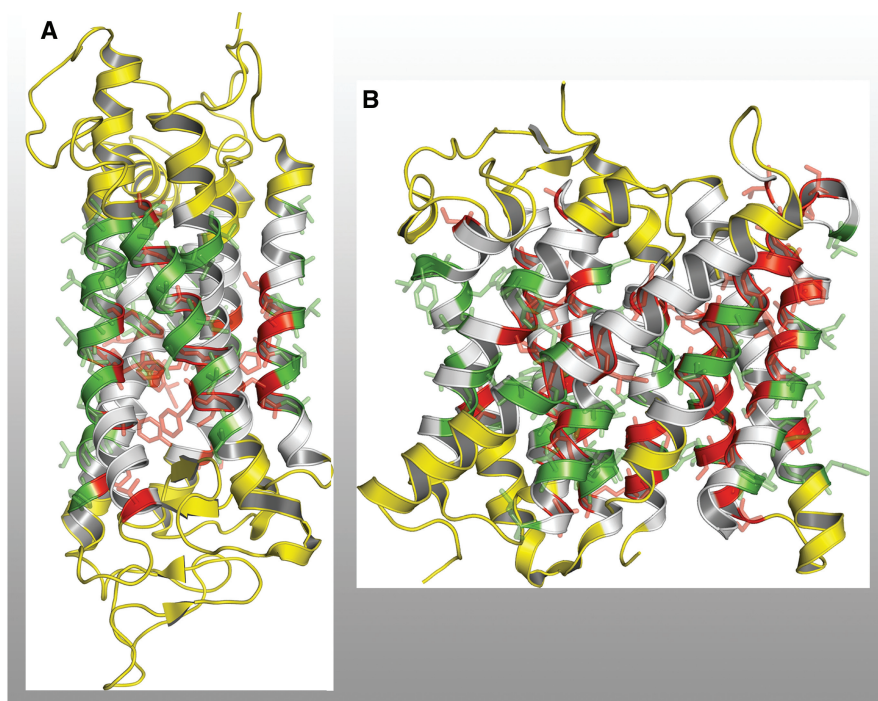
Figure 3. Two high-resolution crystal structures of (**A**) rhodopsin, PDB-entry: 1u19 and (**B**) the ammonium transporter, PDB-entry: 1xqf, were colored according to the predicted contact types (green = helix–membrane, red = helix–helix) using the downloadable PyMOL script from *RHYTHM*. Helical sections that are predicted by HMMTOP to protrude from the lipid bilayer are coloured yellow. The two structures represent two different architectures (23). Rhodopsin belongs to 'membrane-coils', where helix pairs are regularly arranged in small left-handed packing angles. The ammonium transporter belongs to 'channels' that compose of helix pairs packed at large right-handed angles. Different matrices are applied for the prediction of contact types of these two distinct packing modes.

REFERENCES

1. Becker,O.M., Marantz,Y., Shacham,S., Inbal,B., Heifetz,A., Kalid,O., Bar-Haim,S., Warshaviak,D., Fichman,M. and Noiman,S. (2004) G protein-coupled receptors: in silico drug discovery in 3D. *Proc. Natl Acad. Sci. USA*, **101**, 11304–11309.
2. Civelli,O. (2005) GPCR deorphanizations: the novel, the known and the unexpected transmitters. *Trends Pharmacol. Sci.*, **26**, 15–19.
3. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
4. Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2005) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.*, **33(Database Issue)**, D275–D278.
5. Bowie,J.U. (2005) Solving the membrane protein folding problem. *Nature*, **438**, 581–589.
6. Lehnert,U., Xia,Y., Royce,T., Goh,C., Liu,Y., Senes,A., Yu,H., Zhang,Z., Engelman,D. and Gerstein,M. (2005) Computational analysis of membrane proteins: genomic occurrence, structure prediction and helix interactions. *Quart. Rev. Biophys.*, **37**, 1–6.
7. Fleishman,S.J., Unger,V.M. and Ben-Tal,N. (2006) Transmembrane protein structures without X-rays. *Trends Biochem. Sci.*, **31**, 106–113.
8. Punta,M., Forrest,L.R., Bigelow,H., Kernytsky,A., Liu,J. and Rost,B. (2007) Membrane protein prediction methods. *Methods*, **41**, 460–474.
9. Forrest,L.R., Tang,C.L. and Honig,B. (2006) On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophys. J*, **91**, 508–517.
10. Moller,S., Croning,M.D. and Apweiler,R. (2001) Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**, 646–653.
11. Cuthbertson,J.M., Doyle,D.A. and Sansom,M.S. (2005) Transmembrane helix prediction: a comparative evaluation and analysis. *Protein Eng. Des. Sel.*, **18**, 295–308.
12. Bernsel,A., Viklund,H., Falk,J., Lindahl,E., von Heijne,G. and Elofsson,A. (2008) Prediction of membrane-protein topology from first principles. *Proc. Natl Acad. Sci. USA*, **105**, 7177–7181.
13. Granseth,E., Viklund,H. and Elofsson,A. (2006) ZPRED: predicting the distance to the membrane center for residues in alpha-helical membrane proteins. *Bioinformatics*, **22**, e191–e196.
14. Amico,M., Finelli,M., Rossi,I., Zauli,A., Elofsson,A., Viklund,H., von Heijne,G., Jones,D., Krogh,A., Fariselli,P. *et al.* (2006) PONGO: a web server for multiple predictions of all-alpha transmembrane proteins. *Nucleic Acids Res.*, **34**, W169–W172.
15. Adamian,L. and Liang,J. (2006) Prediction of buried helices in multispan alpha helical membrane proteins. *Proteins*, **63**, 1–5.
16. Adamian,L. and Liang,J. (2006) Prediction of transmembrane helix orientation in polytopic membrane proteins. *BMC Struct. Biol.*, **6**, 13.
17. Barth,P., Schonbrun,J. and Baker,D. (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. *Proc. Natl Acad. Sci. USA*, **104**, 15682–15687.
18. Beuming,T. and Weinstein,H. (2004) A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. *Bioinformatics*, **20**, 1822–1835.
19. Park,Y., Hayat,S. and Helms,V. (2007) Prediction of the burial status of transmembrane residues of helical membrane proteins. *BMC Bioinformatics*, **8**, 302.
20. Park,Y. and Helms,V. (2008) Prediction of the translocon-mediated membrane insertion free energies of protein sequences. *Bioinformatics*, **24**, 1271–1277.
21. Zhang,Y., Devries,M.E. and Skolnick,J. (2006) Structure modeling of all identified G protein-coupled receptors in the human genome. *PLoS Comput. Biol.*, **2**, e13.
22. Yarov-Yarovoy,V., Schonbrun,J. and Baker,D. (2006) Multipass membrane protein structure prediction using Rosetta. *Proteins*, **62**, 1010–1025.
23. Hildebrand,P.W., Lorenzen,S., Goede,A. and Preissner,R. (2006) Analysis and prediction of helix–helix interactions in membrane channels and transporters. *Proteins*, **64**, 253–262.
24. Park,Y. and Helms,V. (2006) How strongly do sequence conservation patterns and empirical scales correlate with exposure patterns of transmembrane helices of membrane proteins? *Biopolymers*, **83**, 389–399.
25. Walters,R.F. and DeGrado,W.F. (2006) Helix-packing motifs in membrane proteins. *Proc. Natl Acad. Sci. USA*, **103**, 13658–13663.
26. Eyre,T.A., Partridge,L. and Thornton,J.M. (2004) Computational analysis of {alpha}-helical membrane protein structure: implications for the prediction of 3D structural models. *Protein Eng. Des. Sel.*, **17**, 613–624.
27. Chamberlain,A.K. and Bowie,J.U. (2004) Analysis of side-chain rotamers in transmembrane proteins. *Biophys. J.*, **87**, 3460–3469.
28. Pellegrini-Calace,M., Carotti,A. and Jones,D.T. (2003) Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3D structures. *Proteins*, **50**, 537–545.
29. Eilers,M., Patel,A.B., Liu,W. and Smith,S.O. (2002) Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys. J.*, **82**, 2720–2736.
30. Fleishman,S.J. and Ben-Tal,N. (2002) A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices. *J. Mol. Biol.*, **321**, 363–378.
31. Pappu,R.V., Marshall,G.R. and Ponder,J.W. (1999) A potential smoothing algorithm accurately predicts transmembrane helix packing. *Nat. Struct. Biol.*, **6**, 50–55.
32. Pilpel,Y., Ben-Tal,N. and Lancet,D. (1999) kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J. Mol. Biol.*, **294**, 921–935.
33. Eilers,M., Shekar,S.C., Shieh,T., Smith,S.O. and Fleming,P.J. (2000) Internal packing of helical membrane proteins. *Proc. Natl Acad. Sci. USA*, **97**, 5796–5801.
34. Adamian,L. and Liang,J. (2001) Helix–helix packing and interfacial pairwise interactions of residues in membrane proteins. *J. Mol. Biol.*, **311**, 891–907.
35. Gimpelev,M., Forrest,L.R., Murray,D. and Honig,B. (2004) Helical packing patterns in membrane and soluble proteins. *Biophys. J.*, **87**, 4075–4086.
36. Hildebrand,P.W., Rother,K., Goede,A., Preissner,R. and Frommel,C. (2005) Molecular packing and packing defects in helical membrane proteins. *Biophys. J.*, **88**, 1970–1977.
37. Hildebrand,P.W., Gunther,S., Goede,A., Forrest,L., Frommel,C. and Preissner,R. (2008) Hydrogen-bonding and packing features of membrane proteins: functional implications. *Biophys. J.*, **94**, 1945–1953.
38. Walther,D., Eisenhaber,F. and Argos,P. (1996) Principles of helix-helix packing in proteins: the helical lattice superposition model. *J. Mol. Biol.*, **255**, 536–553.
39. Langosch,D. and Heringa,J. (1998) Interaction of transmembrane helices by a knobs-into-holes packing characteristic of soluble coiled coils. *Proteins*, **31**, 150–159.
40. Taylor,W.R., Jones,D.T. and Green,N.M. (1994) A method for alpha-helical integral membrane protein fold prediction. *Proteins*, **18**, 281–294.
41. Fuchs,A., Martin-Galiano,A.J., Kalman,M., Fleishman,S., Ben-Tal,N. and Frishman,D. (2007) Co-evolving residues in membrane proteins. *Bioinformatics*, **23**, 3312–3319.
42. Adamian,L., Nanda,V., DeGrado,W.F. and Liang,J. (2005) Empirical lipid propensities of amino acid residues in multispan alpha helical membrane proteins. *Proteins*, **59**, 496–509.
43. MacKenzie,K.R. and Engelman,D.M. (1998) Structure-based prediction of the stability of transmembrane helix–helix

interactions: the sequence dependence of glycophorin A dimerization. *Proc. Natl Acad. Sci. USA*, **95**, 3583–3590.

44. Park,Y., Elsner,M., Staritzbichler,R. and Helms,V. (2004) Novel scoring function for modeling structures of oligomers of transmembrane alpha-helices. *Proteins*, **57**, 577–585.

45. Yarov-Yarovoy,V., Baker,D. and Catterall,W.A. (2006) Voltage sensor conformations in the open and closed states in ROSETTA structural models of K(+) channels. *Proc. Natl Acad. Sci. USA*, **103**, 7292–7297.

46. Sammut,S.J., Finn,R.D. and Bateman,A. (2008) Pfam 10 years on: 10 000 families and still growing. *Brief Bioinform.*, **9**, 210–219.

47. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.

48. Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849–850.

49. Preissner,R., Goede,A. and Frommel,C. (1998) Dictionary of interfaces in proteins (DIP). Data bank of complementary molecular surface patches. *J. Mol. Biol.*, **280**, 535–550.

50. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

51. Tusnady,G.E., Dosztanyi,Z. and Simon,I. (2004) TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics*, **21**, 1276–1277.

52. Wistrand,M. and Sonnhammer,E.L. (2005) Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. *BMC Bioinformatics*, **6**, 99.

53. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**, 3038–3049.

54. Rother,K., Preissner,R., Goede,A. and Frömmel,C. (2003) Inhomogeneous molecular density: reference packing densities and distribution of cavities within proteins. *Bioinformatics*, **19**, 2112–2121.

55. Paci,E. and Marchi,M. (1996) Intrinsic compressibility and volume compression in solvated proteins by molecular dynamics simulation at high pressure. *Proc. Natl Acad. Sci. USA*, **93**, 11609–11614.

56. Cuff,A.L. and Martin,A.C. (2004) Analysis of void volumes in proteins and application to stability of the p53 tumour suppressor protein. *J. Mol. Biol.*, **344**, 1199–1209.

57. Rother,K., Hildebrand,P.W., Goede,A., Gruening,B. and Preissner,R. (2008) Voronoia: analyzing packing in protein structures. *Nucleic Acids Res*, **37**, D393–D395.

58. Adamian,L. and Liang,J. (2003) Interhelical hydrogen bonds in transmembrane region are important for function and stability of Ca^{2+}-transporting ATPase. *Cell Biochem. Biophys.*, **39**, 1–12.

59. Stevens,T.J. and Arkin,I.T. (2001) Substitution rates in alpha-helical transmembrane proteins. *Protein Sci.*, **10**, 2507–2517.