Structure-based statistical analysis of transmembrane helices.

Carlos Baeza[1], Marc A. Marti-Renom[2,*] and Ismael Mingarro[1,*]

1. Departament de Bioquímica i Biologia Molecular. Universitat de València. Burjassot, Spain.

2. Structural Genomics Group. National Center for Genomic Analysis (CNAG). Barcelona, Spain.

* Corresponding authors:

Marc A. Marti-Renom
Phone: +34 934 033 743
Fax: +34 934 037 279
E-mail: mmarti@cpb.ub.cat

Ismael Mingarro
Phone: +34 963 543 796
Fax: +34 963 544 635
E-mail: Ismael.Mingarro@uv.es

Version: 3/21/2012

Abstract

Recent advances in high-resolution structure determination of membrane proteins enable now the analysis of the main features of amino acids in transmembrane (TM) segments in comparison with amino acids in water-soluble helices. In this work, we introduced a large-scale analysis of amino acid propensities using a data set of 170 structures of integral membrane proteins obtained from MPTopo database and 930 structures of water-soluble helical proteins obtained from the Protein Data Bank. Large hydrophobic residues (Leu, Val, Ile and Phe) plus Gly had a clear preference for TM helices, while polar residues (Glu, Lys, Asp, Arg and Gln) were less frequent in this type of helices. The distribution of residues along the TM helices was also examined. As expected, hydrophobic and slightly polar amino acids are commonly found in the hydrophobic core of the membrane, while aromatic (Trp and Tyr) and Pro together with hydrophilic (Asn, His, and Gln) residues are frequent in the interface regions. Charged residues also have statistically preferred locations avoiding the hydrophobic core of the membrane, but while acidic residues are frequently found at both the cytoplasmic and extra-cytoplasmic interfaces, basic residues cluster at the cytoplasmic interface. These results strongly support the experimentally demonstrated biased distribution of positively charged residues (that is, the so-called the positive-inside rule) with structural data.

Keywords

Membrane protein; transmembrane helices; amino acid distribution; statistical analysis.

Introduction

Although helical membrane proteins represent about one fourth of all proteins in living organisms (Wallin & Heijne, 1998), the rules governing its folding are still not completely established. The hydrophobic effect is a dominant driving force to the folding of water-soluble proteins, but its contribution to the folding of membrane proteins is further more complex given that these proteins "live" in a biophysical environment –the membrane–, which is clearly different from the aqueous media. The cell membrane is a very heterogeneous media, composed mainly of phospholipids that are self-organized in two leaflets giving rise to the formation of a bilayer. The hydrocarbon core is the hydrophobic part of the membrane, covering approximately 30 Å. The polar head groups of the phospholipids define the lipid/water interphase and add approximately 15 Å to the thickness of each leaflet (White & Wimley, 1999). It is in this complex environment in which membrane proteins have to fold into their native conformations.

The hydrocarbon core of the biological membranes and the interior of folded water-soluble proteins are hydrophobic. In such a hydrophobic environment, the polarity of the polypeptide backbone is energetically unfavorable. Thus, in protein structures, nearly all the polar groups of the peptide bond (carbonyl and amide groups) tend to hydrogen bond with one another, leading to secondary structure that stabilizes the folded state. Alpha-helices are the commonest secondary structural elements found in water-soluble as well as in membrane protein structures. However, the distribution of the helices in these two groups of proteins is very different. While helices in water-soluble proteins can be exposed to both the hydrophobic core and the water-

accessible surface, transmembrane (TM) helices in membrane proteins are surrounded by a hydrophobic lipid phase where water is essentially absent. Therefore, for the structural stabilization of helical membrane proteins that reside in this apolar (low dielectric) environment, hydrogen bonding and van der Waals packing forces have an increased importance.

Although the great majority of membrane proteins integrate into biological membranes through the translocon (see for a recent review (Martínez-Gil et al, 2011)), our current biophysical understanding of its folding and function is hampered by the scarcity of structural information. Fortunately, the number of high-resolution structures of membrane proteins has increased exponentially in the last years (White, 2004; 2009). Consequently, a new statistical survey of TM helices properties is timely.

In this paper, we revisit the differences between helices from water-soluble proteins and TM helices in terms of length and amino acid composition. In addition, we analyze the distribution of amino acid residues in TM segments, which have to energetically accommodate into the highly heterogeneous media of biological membranes by interacting favorably with its local environment. The present study involved 170 helical membrane proteins with known three-dimensional structure and topology, containing a total of 792 TM segments and compared with 7,348 helices from 930 water-soluble protein structures. About half of all amino acids are randomly distributed when allocated into the membrane, but the rest show a strong correlation for residue positions along the TM regions.

Methods

Helix data sets

Two data sets of water-soluble and TM helices were obtained from the Protein Data Bank (PDB) (Berman et al, 2000) and the MPTOPO databases (Jayasinghe et al, 2001b), respectively.

First, a total of 4,405 structural chains deposited in the PDB (as of November 17th, 2011) that passed the following criteria were selected: (i) their total secondary structure had more than 60% of α-helices and no β-strands; (ii) their crystallographic resolution was 2.0 Å or higher; and (iii) the word *MEMBRANE* did not appear in the "TITLE" nor the "DESCRIPTION" fields of the PDB file. Furthermore, to remove redundancy, the 4,405 chain sequences were compared to each other with the *cd-hit* program (Huang et al, 2010) and pairs resulting in sequence alignments with 80% or higher identity were discarded. The final set of 930 non-redundant PDB chains was parsed to identify a total of 7,348 helices from "HELIX" fields of each PDB chain entry. Thus, the data set of water-soluble helices contained 930 non-redundant and high-resolution protein structures, 7,348 α-helices and 108,277 amino acids.

Second, all α-helical membrane proteins deposited in the MPTOPO database (last updated on January 19th, 2010) (Jayasinghe et al, 2001b), and thus with known membrane insertion topology, were selected. The initial set was further filtered by: (i) removing any entry of unknown structure as based on the MPTOPO entry classification (*i.e.*, keeping only entries described as "3D_helix" and "1D_helix"); (ii) removing redundant pairs at 80% sequence identity by applying the *cd-hit* program (Huang et al, 2010). The final data set

of TM helices contained 170 non-redundant structures, 837 TM helices, and 20,079 amino acids. Furthermore, to properly analyze the amino acid propensities in single membrane spanning TM helices, we discarded any helix shorter than 17 amino acids or larger than 38 amino acids. The resulting TM data subset contained 792 TM helices, and 19,356 amino acids.

Amino acid propensity measures

We calculated three different amino acid measures: (i) probability and percent, (i) Odds, and (iii) LogOdds. The probability (p_i) of an amino acid i is defined as:

$$p_i = \frac{n_i}{N}$$

where i is the amino acid type (one of the 20 amino acids), n_i is the observation count of the amino acid i, and N is all amino acids in the data set. Similarly, the percent of a given amino acid i is defined as its probability multiplied by 100. The Odds (O_i) of an amino acid i is defined as:

$$O_i = \frac{p_{i,c}}{(1 \quad p_{i,c})} \Big/ \frac{p_{i,r}}{(1 \quad p_{i,r})}$$

where $p_{i,c}$ is the probability of the amino acid i in the class c (for example, TM helix) and $p_{i,r}$ is the probability of the amino acid i in the class r (for example, water-soluble helix). Similarly, the LogOdds of a given amino acid i is defined as the logarithm in base 10 of its Odds. Briefly, Odds higher than 1 (or positive LogOdds) indicate over-occurrence of the amino acid type in the class. Odds smaller than 1 (or negative LogOdds) indicate under-representation of the amino acid type in the class.

Results and Discussion

Helix length in membrane and water-soluble proteins

Length distributions for helices found in high-resolution structures deposited in PDB (Berman et al, 2000) are very different for TM and water-soluble proteins (Fig. 1).

Helices in TM proteins are in average 24.0 (± 5.6) amino acid residues long, this result slightly differs from previous data obtained using databases with 45 (Bowie, 1997) and 129 (Ulmschneider & Sansom, 2001) TM helices, where average helix length was 26.4 and 27.1 amino acid residues, respectively. As the translation per residue in a canonical helix is 1.5Å, a stretch of about 20 consecutive hydrophobic residues can span the 30 Å of the hydrocarbon core of biological membranes. Indeed, the more prevalent (~12%) length for TM helices in our data set was 21 residues (Fig. 1). Longer helices can span the bilayer with a concomitant tilting of the helix axis respect to the membrane plane. Other options are also feasible ranging from lipid accommodation till polypeptide backbone deformation (Holt & Killian, 2009).

Helices from water-soluble proteins have an average length of 14.7 (±8.7) residues, which agrees with previous studies where the more prevalent helix length was 10-11 residues long (Engel & DeGrado, 2004; Pal et al, 2003). The reduced length for helices from water-soluble proteins is due to the absence of the restrictions imposed by the low dielectric constant at the hydrocarbon core of biological membranes, which forces the polypeptide backbone to adopt on average larger secondary structures.

Amino acid composition of α-helices

The amino acid composition for both, TM and water-soluble helices, have been examined (Fig. 2). TM helices of lengths between 17 and 38 residues were selected from the MPTOPO database (Jayasinghe et al, 2001b), which included helical segments that do completely span the hydrophobic core of the membrane. TM helices shorter than 17 residues as well as larger than 38 residues were excluded since they may not cross entirely the membrane (Fig. 1 inset a) or may contain segments parallel to the membrane (Fig. 1 inset b). Note that in the case of water-soluble helices all lengths were included in our analysis because no restrictions in terms of length can be assumed for water-soluble proteins in an aqueous milieu.

As expected, hydrophobic residues Leu, Ala, Val and Ile constitute the bulk of the amino acids in the TM region accounting for almost half (47.0%) of all residues. Similarly, these residues are also frequently found in helices of water-soluble proteins (34.1%). However, there are, as noted previously using smaller datasets (Bywater et al, 2001), differences in composition of the two types of helices. Despite sharing the same structural features, the differences between the two types of helices are reflected by their preferential occurrences measured by the logarithm of the Odds of finding a given amino acid in a TM helix with respect to its frequency in a water-soluble helix (Fig. 2 bottom panel). For example, while charged and polar residues are much more frequently found in helices from water-soluble proteins, Trp, Gly and Phe have higher propensities in TM helices. Interestingly, in contrast to their conformational preferences in water, the helical propensities of residues such as Val, Ile, Phe and Met are notably increased in the membrane environment,

where it has been suggested that their helical proclivity is primarily governed by their side chain hydrophobicity and by the hydropathy of the local polypeptide region in which the residues reside span the membrane (Li & Deber, 1994). Significantly, Gly and Pro are more frequent in TM helices relative to water-soluble helices. Although commonly considered as 'helix breakers' it has been reported that Gly residues occur frequently in TM helix-helix interactions, especially in association with β-branched residues at neighboring positions (Senes et al, 2000), and that Pro, in addition to its role in signal transduction and gating across the membrane, may also play a significant role in these processes (Orzáez et al, 2004).

A comparison of the amino acid frequency between TM and water-soluble helices confirmed that strongly polar residues (Glu, Lys, Asp, Arg, and Gln) are more prevalent in water-soluble helices (Fig. 3). These residues constitute only 8.2 % of the residues within TM helices compared to 30.9 % in water-soluble helices. Despite their lower presence, polar residues are evolutionary conserved in TM proteins, which has been partially explained by their tendency to be buried in the protein interior and also in many cases due to their direct involvement in the function of the protein (Illergård et al, 2011). Conversely, hydrophobic amino acids (Leu, Val, Ile, Gly, and Phe) are over-represented in TM helices (Fig. 3). Interestingly, Ala although being the second more abundant residue in TM helices (Fig. 2), it is not over-represented in this type of helices likely because its higher helical propensity in aqueous (Blaber et al, 1993) compared to membrane-mimetic environments (Li & Deber, 1994). In fact, both biological (Nilsson et al, 2003; Hessa et al, 2005) and biophysical (Jayasinghe et al, 2001a) measurements

have poised Ala at the threshold between those amino acids that promote membrane integration of TM helices and those residues that preclude membrane insertion.

Position dependent distribution of amino acid residues in TM helices

A comparison of the amino acid frequency at different positions in a TM segment, taking as reference the TM center, confirmed that about half of the natural amino acid residues have similar distributions at positive positions (towards inside the cell) than at negative positions (towards outside the cell) (Fig. 4). It was found that not only the strongly hydrophobic residues but also Gly and the hydroxylated residues Ser and Thr are equally distributed along the hydrophobic core of the membrane. It is important to note that Gly is a residue type that is normally regarded as being conductive to turn (Williams et al, 1987), yet it is a common residue in TM helices (Fig. 2). There are important folding reasons for incorporating Gly into TM helices. The absence of side-chain of the Gly allows for bulkier groups to be accommodated close to the polypeptide backbone of the TM helices. This might be important for intramolecular helix-helix packing, for homo-oligomerization, or for recognition of other membrane proteins, among other factors. Indeed, it has been observed that Gly has the highest overall packing value in membrane proteins (Eilers et al, 2002). Ser or Thr residues within TM helices participate in hydrogen-bonding networks through hydrogen bond linking of the side chain oxygen atom to acceptor side chain or peptide bond groups. These effects, intimate packing (Gly) and hydrogen bonding (Ser and Thr), can be relevant at any position along the TM region, which would explain the absence of position preference for these residues in TM helices. Met or Cys are also

frequent at different locations within the hydrophobic core, but a relative prevalence can be observed in a region that would correspond with the initial portion of the polar headgroups of the phospholipids, consistent with the slightly amphipathic nature of these residues and in agreement with its distribution in the lipid bilayer recently obtained from molecular dynamics simulation (MacCallum et al, 2008).

While Phe has a flat distribution in TM helices, behaving as a hydrophobic residue, Trp, Tyr and Pro residues are distributed in a biased manner: they are found preferentially at the ends of the bilayer (*i.e.* at the interface between the hydrophobic core of the bilayer and the bulk water). At this location, aromatic residues may serve as anchors for the TM helices into the membrane. In fact, Trp and Tyr positioned 7 to 9 residues away from the center of a TM segments result in a reduction in free energy (Hessa et al, 2007), which nicely correlates with the present statistical distribution from three-dimensional structures (Fig. 4). The biophysical reason for the observed distribution of Trp and Tyr residues could rely on the relatively amphipathic nature of their side chains, which can form hydrogen bonds as well as exhibit hydrophobic character. Actually, this preferred location has previously been observed not only for α-helical but also β-barrel membrane proteins (Ulmschneider & Sansom, 2001). A similar distribution is observed for Pro residues, although an increased presence is detectable towards the center of the bilayer, which can be associated with the fundamental and subtle role that Pro residues play in the dynamics, structure and function of many membrane proteins by inducing the formation of molecular hinges (Cordes et al, 2002). Indeed, thirteen TM helices with known structure have a

Pro residue at the 0 position, which in all cases results in a kink in the helix. Nevertheless, it should be noted that the interfacial preference of these three residues is somehow more pronounced at the non-cytoplasmic interface. This was also observed in the case of aromatic residues (Trp and Tyr) in a membrane protein prediction analysis using sequence information from 107 genomes (Nilsson et al, 2005).

The distribution pattern for Asn, His and Gln, corresponds to an interfacial preference close to the end of the TM regions, which is consistent with the amphipathic nature of these molecules. This pattern was previously reported for His residues (Ulmschneider & Sansom, 2001), which is in good agreement with our data. Interestingly, in more recent studies using computer simulations, it has been noted that small molecule analogs of Asn (MacCallum et al, 2008) and Asn, His, and Gln (Johansson & Lindahl, 2007) result in an energetic minimum for partition into model lipid bilayers.

Since the energetic cost of inserting an ionizable group in the hydrophobic environment of the membrane is very high (White & Wimley, 1999), charged amino acids should generally be excluded from the hydrophobic core of the TM helices. Interestingly, nearly all membrane proteins with six or more predicted TM helices contain at least one ionizable residue (Arkin & Brunger, 1998). However, charged amino acids consistently clustered at the TM flanking regions (Fig. 4). For example, acidic (Asp and Glu) residues result in an increased distribution at both cytoplasmic and extra-cytoplasmic side of the membrane, although with some prevalence for the cytoplasmic region. Positively charged (Arg and Lys) residues distribution is even more strongly asymmetric between opposite sides of the membrane, in good agreement

with the positive-inside rule (Heijne, 1992). Moreover, it has been experimentally demonstrated that basic residues act as stronger topological signals compared to acidic residues (Nilsson & Heijne, 1990; Saurí et al, 2009), which is reflected by their different statistical preferences on either end of the TM segments. Nevertheless, when considered globally, charged residues cluster preferentially near the cytoplasmic end of the TM segments (Fig. 5, orange line). This effect was already noted in a previous structure-based analysis that included a lower number of structures available at the time (Ulmschneider et al, 2005). On the contrary, although polar residues (Gln, His, and Asn) mimic the distribution pattern of charged residues avoiding the more hydrophobic region of the bilayer, they show a preference for the extra-cytoplasmic region (Fig. 5). Trp, Tyr and Pro are more abundant about 8 to 9 residue positions away from the center of the membrane, that is, within the interface region, but with some bias toward the extra-cytoplasmic interface. The rest of natural amino acids are more abundant at the center of the bilayer, within 7 amino acid positions on both sides of the membrane normal, but they are also very frequently found beyond this boundary as noted by their overall proximity to the Odd value of 1 for positions >10 on both sides of the center of the membrane (Fig. 5). Interestingly, the amino acid distribution patterns at both interface regions are slightly different. There is a sharper transition from mainly hydrophobic to charged, polar and aromatic residues at the cytoplasmic side of the membrane (positions 6 to 8) compared to that at the extra-cytoplasmic side (positions -5 to -9). The different lipid composition between the two lipid leaflets in biological membranes and the strong electrochemical potential over the prokaryotic inner cell membranes can exert

an important effect, which may be reflected by this difference. For instance, it has been recently reported an asymmetry in the distribution of amino acid residues within TM segments from plasma membrane proteins (Sharpe et al, 2010), which has been attributed to an asymmetry in the state of lipid order in the membrane. Such an asymmetry is likely due to the enrichment of lipids such as sterols and sphingolipids in the extra-cytoplasmic leaflet, where a more gradual amino acid distribution can be expected.

Finally, we analyzed and plotted the odd ratio for each amino acid in three regions in a membrane, that is, taking the hydrophobic TM region as the central 19 positions (~30Å) and 9 residue positions (~15Å) on both sides as the extra-cytoplasmic (from -10 to -18 residues) and cytoplasmic (from 10 to 18) flanking regions (Fig. 6). Hydrophobic amino acids (blue colored) populated preferentially the hydrophobic center. However, this trend is not observed for the more prevalent residues in TM segments (for example Leu, Fig. 2), which are also frequently found at the flanking regions. Trp, Tyr, and Pro (green) have a minor increase for the extra-cytoplasmic flanking region. The absence of higher differences for the distribution of these residues is probably due to their precise location at the interface between the hydrophobic core and the flanking hydrophilic environment. Polar (orange) residues (Gln, His, and Asn) have a preference for both flanking regions since they are energetically unfavorable within the membrane core. These residues do not ionize at the physiological pH and are able to donate and accept hydrogen bonds simultaneously. Such an effect translates into a higher preference of Gln, His and Asn for the rich hydrogen bond network environment of the interface. Charged residues (red) are underrepresented at

the hydrophobic core and resulted in preferences for the cytoplasmic flanking region being acidic residues more prevalent at the extra-cytoplasmic flanking region. Furthermore, basic residues are strong topological determinants that heavily populate the cytoplasmic flanking region. The effect of positively charged residues located near the cytoplasmic end of hydrophobic segments has been in fact estimated to be approximately -0.5 kcal/mol to the apparent free energy of membrane insertion (Lerch-Bader et al, 2008). This energetic contribution can be extremely relevant to precisely anchor hydrophobic regions into biological membranes.

Concluding remarks

We have compared the length and the amino acid composition of helices in TM and water-soluble proteins. Overall, significant differences are present in both types of proteins, which may be attributed to the biophysical differences between the two environments in which they fold. First, TM helices adapt their length to the dimensions and constraints of biological membranes, while water-soluble helices are statistically shorter since they do not have to satisfy the demanding restrictions imposed by the complexity of the membrane environment. Second, the observed differences highlight that in the lipid bilayer, which environment forces secondary structure formation, amino acid side chain hydrophobicity prevails to helicity. Accordingly, aliphatic residues with a reduced helical propensity (Val, Ile, Gly, and Phe) are abundant in TM helices, while polar residues (Glu, Lys, and Arg) with high helical propensity are consistently less frequent in TM helices. Third, half of the natural amino acid residues are equally distributed along the TM helices, whilst aromatic,

polar and charged residues plus Pro are biased toward the ends of the TM helices. Fourth, as previously observed, the distribution of charged residues was asymmetric occurring more frequently on the cytoplasmic side of the membrane, which causes a net charge unevenness on both sides of the membrane. In addition to this asymmetry, Trp, Tyr and Pro residues were found to be more frequent at the extra-cytoplasmic interface of the membrane and the polar residues (Gln, His, and Asn) at the extra-cytoplasmic flanking region of the TM helices. Fifth, transitions between the different types of residues at the ends of the hydrophobic core occur in a more defined region at the cytoplasmic side than at the extra-cytoplasmic face, likely reflecting the differences in lipids composition on both leaflets of biological membranes.

The conclusions on TM helix architecture described here should prove useful for constructing models of membrane proteins with desired properties, which could help filling in some of the many gaps in the field.

Acknowledgments

This work was supported by grants BFU2009-08401 (to I.M.) and BFU2010-19310 (to M.A.M-R.) from the Spanish Ministry of Science and Innovation (MICINN, ERDF supported by the European Union), as well as by PROMETEO/2010/005 and ACOMP/2012/226 (to I.M.) and ACOMP/2011/048 (to M.A.M-R.) from the Generalitat Valenciana. C.B. was recipient of a predoctoral FPI fellowship from the MICINN.

References

Arkin IT & Brunger AT (1998) Statistical analysis of predicted transmembrane alpha-helices. *Biochim Biophys Acta* **1429:** 113–128

Berman HM, Berman HM, Westbrook J, Westbrook J, Feng Z, Feng Z, Gilliland G, Gilliland G, Bhat TN, Bhat TN, Weissig H, Weissig H, Shindyalov IN, Shindyalov IN, Bourne PE & Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res.* **28:** 235–242

Blaber M, Zhang XJ & Matthews BW (1993) Structural basis of amino acid alpha helix propensity. *Science* **260:** 1637–1640

Bowie JU (1997) Helix packing in membrane proteins. *J Mol Biol* **272:** 780–789

Bywater RP, Thomas D & Vriend G (2001) A sequence and structural study of transmembrane helices. *J. Comput. Aided Mol. Des.* **15:** 533–552

Cordes FS, Bright JN & Sansom MSP (2002) Proline-induced distortions of transmembrane helices. *J Mol Biol* **323:** 951–960

Eilers M, Patel AB, Liu W & Smith SO (2002) Comparison of helix interactions in membrane and soluble alpha-bundle proteins. *Biophys J* **82:** 2720–2736

Engel DE & DeGrado WF (2004) Amino acid propensities are position-dependent throughout the length of alpha-helices. *J Mol Biol* **337:** 1195–1205

Heijne von G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J Mol Biol* **225:** 487–494

Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH & Heijne von G (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433:** 377–381

Hessa T, Meindl-Beinker NM, Bernsel A, Kim H, Sato Y, Lerch-Bader M, Nilsson I, White SH & Heijne von G (2007) Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* **450:** 1026–1030

Holt A & Killian JA (2009) Orientation and dynamics of transmembrane peptides: the power of simple models. *Eur Biophys J* **39:** 609–621

Huang Y, Huang Y, Niu B, Niu B, Gao Y, Gao Y, Fu L, Fu L, Li W & Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26:** 680–682 Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20053844&retmode=ref&cmd=prlinks

Illergård K, Kauko A & Elofsson A (2011) Why are polar residues within the

membrane core evolutionary conserved? *Proteins* **79:** 79–91

Jayasinghe S, Hristova K & White SH (2001a) Energetics, stability, and prediction of transmembrane helices. *J Mol Biol* **312:** 927–934

Jayasinghe S, Jayasinghe S, Hristova K, Hristova K, White SH & White SH (2001b) MPtopo: A database of membrane protein topology. *Protein Sci* **10:** 455–458

Johansson ACV & Lindahl E (2007) Position-resolved free energy of solvation for amino acids in lipid membranes from molecular dynamics simulations. *Proteins* **70:** 1332–1344

Lerch-Bader M, Lundin C, Kim H, Nilsson I & Heijne von G (2008) Contribution of positively charged flanking residues to the insertion of transmembrane helices into the endoplasmic reticulum. *Proc Natl Acad Sci USA* **105:** 4127–4132

Li SC & Deber CM (1994) A measure of helical propensity for amino acids in membrane environments. *Nat Struct Biol* **1:** 558

Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL (2012) OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res. **40:** 370-376

MacCallum JL, Bennett WFD & Tieleman DP (2008) Distribution of amino acids in a lipid bilayer from computer simulations. Biophys J **94:** 3393–3404

Martínez-Gil L, Saurí A, Marti-Renom MA & Mingarro I (2011) Membrane protein integration into the endoplasmic reticulum. *FEBS J* **278:** 3846–3858

Nilsson I & Heijne von G (1990) Fine-tuning the topology of a polytopic membrane protein: role of positively and negatively charged amino acids. *Cell* **62:** 1135–1141

Nilsson I, Johnson AE & Heijne von G (2003) How hydrophobic is alanine? *J Biol Chem* **278:** 29389–29393

Nilsson J, Persson B & Heijne von G (2005) Comparative analysis of amino acid distributions in integral membrane proteins from 107 genomes. *Proteins* **60:** 606–616

Orzáez M, Salgado J, Giménez-Giner A, Pérez-Payá E & Mingarro I (2004) Influence of proline residues in transmembrane helix packing. *J Mol Biol* **335:** 631–640

Pal L, Chakrabarti P & Basu G (2003) Sequence and structure patterns in proteins from an analysis of the shortest helices: implications for helix nucleation. *J Mol Biol* **326:** 273–291

Saurí A, Tamborero S, Martínez-Gil L, Johnson AE & Mingarro I (2009) Viral membrane protein topology is dictated by multiple determinants in its sequence. *J Mol Biol* **387:** 113–128

Senes A, Gerstein M & Engelman DM (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J Mol Biol* **296:** 921–936

Sharpe HJ, Stevens TJ, Munro S (2010) A comprehensive comparison of transmembrane domains reveals organelle-specific properties. *Cell* **142**, 158-169.

Ulmschneider MB & Sansom MS (2001) Amino acid distributions in integral membrane protein structures. *Biochim Biophys Acta* **1512:** 1–14

Ulmschneider MB, Sansom MSP & Di Nola A (2005) Properties of integral membrane protein structures: derivation of an implicit membrane potential. *Proteins* **59:** 252–265

Wallin E & Heijne von G (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci* **7:** 1029–1038

White SH (2004) The progress of membrane protein structure determination. *Protein Sci* **13:** 1948–1949

White SH (2009) Biophysical dissection of membrane proteins. *Nature* **459:** 344–346

White SH & Wimley WC (1999) Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct* **28:** 319–365

Williams RW, Chang A, Juretić D & Loughran S (1987) Secondary structure predictions and medium range interactions. *Biochim Biophys Acta* **916:** 200–204

Figure Legends

Figure 1. Length distributions for 837 TM and 7,348 water-soluble helices from a set of non-redundant proteins of known structure (see Methods). Transmembrane helices in blue (pale blue correspond to discarded lengths) and water-soluble helices in orange. (a) Example of a short 9 amino acid length helix in the ClC chloride channel from *E. coli* (1KPK entry in PDB). Membrane boundaries were obtained from the PPM Server (Lomize et al, 2012). The selected membrane is shown in rainbow coloring from N- (blue) to C-terminal (red) ends. (b) Example of a large 43 amino acid length helix in the chicken cytochrome BC1 complex (1BCC entry in PDB), which N-terminus of the helix (blue) lies at the membrane/water interface. Representation as in inset (a).

Figure 2. Amino acid type distribution from 792 TM and 7,348 water-soluble helices from a set of non-redundant proteins of known structure (see Methods). (Upper plot) Amino acid type distribution for TM helices in blue and for water-soluble helices in orange. (Lower plot) LogOdds values for comparing the relative abundance of each amino acid type in TM and water-soluble helices. Amino acid types are ordered by its LogOdds

Figure 3. Amino acid type percentage comparison between TM and water-soluble helices. Blue colored amino acids are over represented (difference > 3 % points) in TM helices compared to water-soluble helices. Orange colored amino acids are over represented (difference > 3 % points) in water-soluble

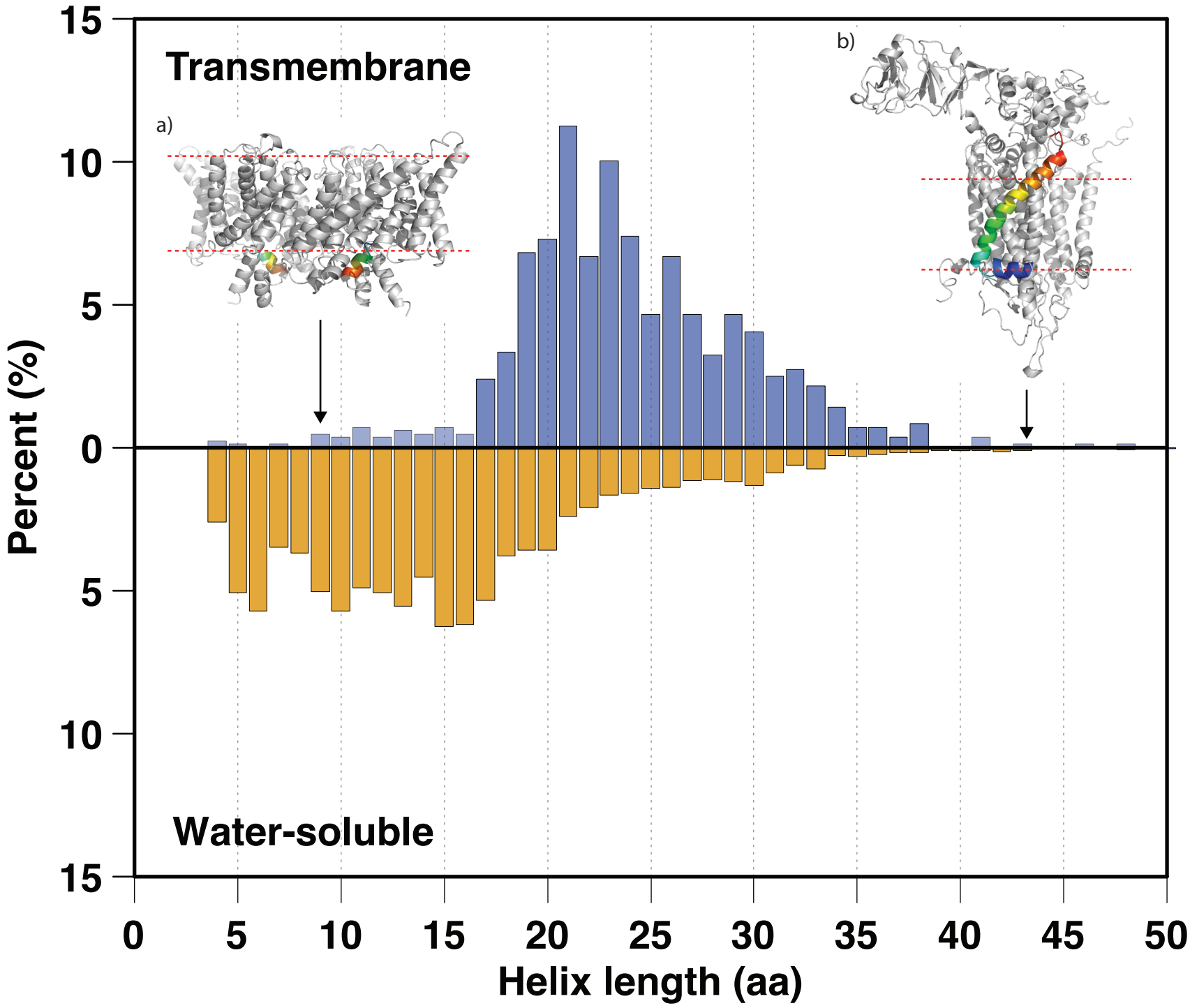
helices compared to TM helices. Dashed grey lines indicate a cut-off of 3 % difference points.

Figure 4. Amino acid type and position distribution in TM helices. Each amino acid type and their positioning in the TM helix is represented by their positional normalized Odds (that is, for each column the Odds are normalized to an average of zero and standard deviation of one). The amino acids are clustered based on their positional normalized Odds within the helices. Positively labeled positions refer to the cytoplasmic side of the membrane and its flanking region whilst negatively labeled positions refer to extra-cytoplasmic regions.

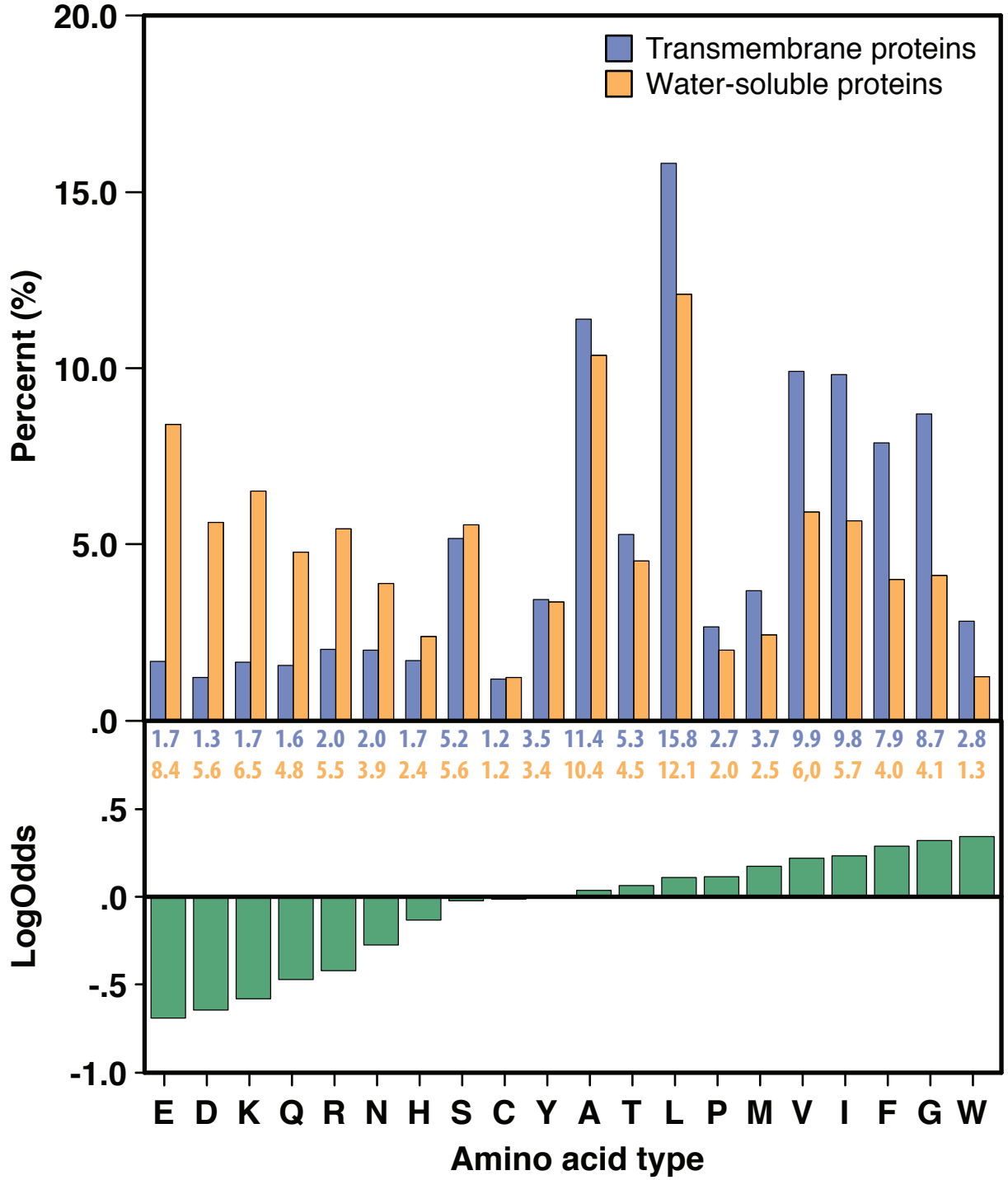
Figure 5. Amino acid groups positional preferences in a membrane. Thin lines represent the positional Odds for each amino acid individually, whilst thick lines represent the averaged positional Odds for each group of amino acids obtained from Figure 4. Amino acid types are grouped as in the dendogram in Fig. 4. That is, charged residues (red, KRED), polar residues (orange, QHN), aromatic residues plus Pro (green, PYW), and the rest of residues (blue, CMTSGVFAIL).

Figure 6. Amino acid location preferences in a membrane. Letter size is proportional to the odds (relative preference) of finding a given amino acid in the three regions in a membrane (*i.e.*, from top to bottom outer, membrane, and inner regions). Amino acids colored as in figure 5.

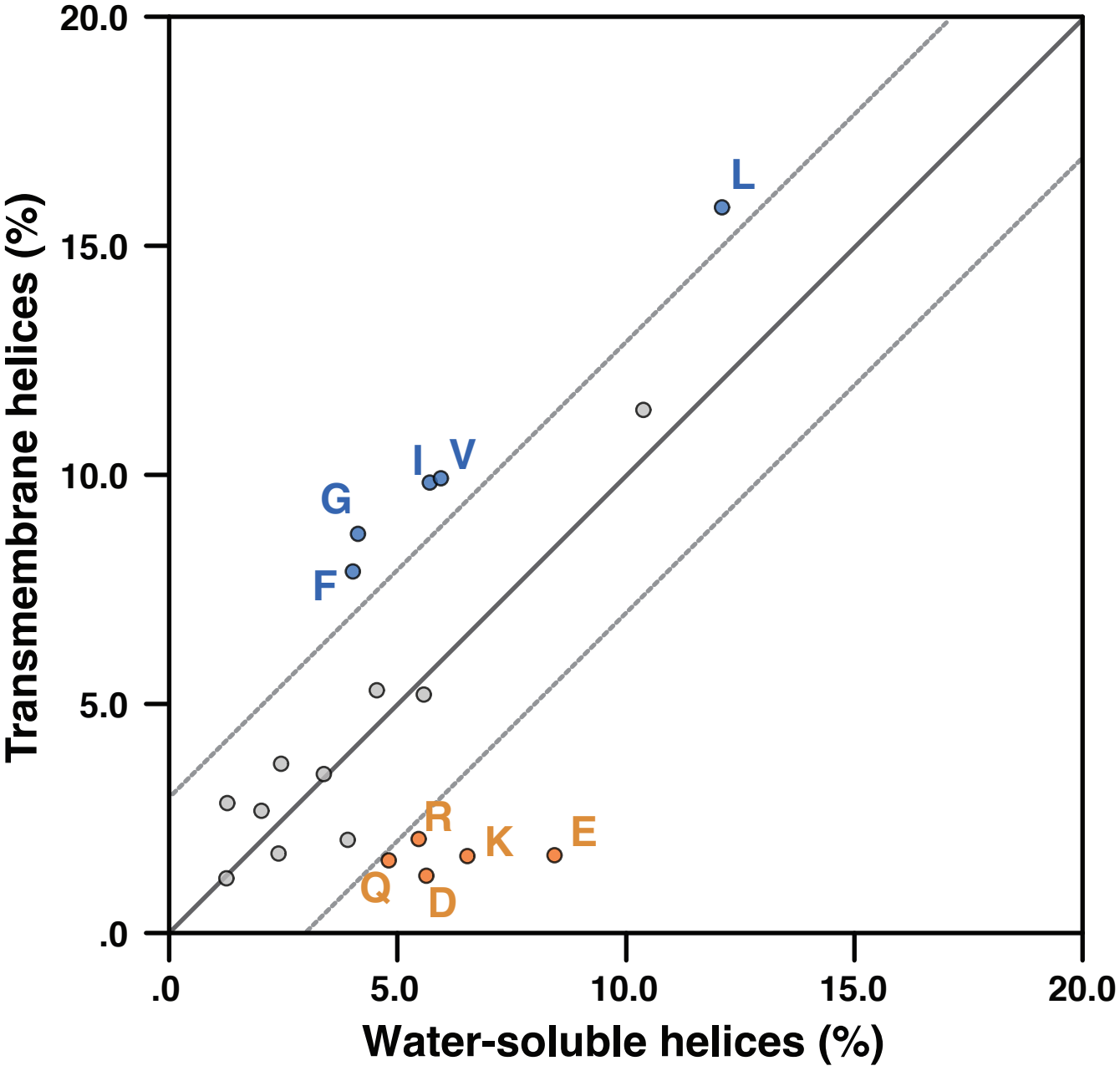
Images


Transmembrane
Water-soluble
a)
b)
Percent (%)
Helix length (aa)
15
10
5
0
5
10
15
0
5
10
15
20
25
30
35
40
45
50

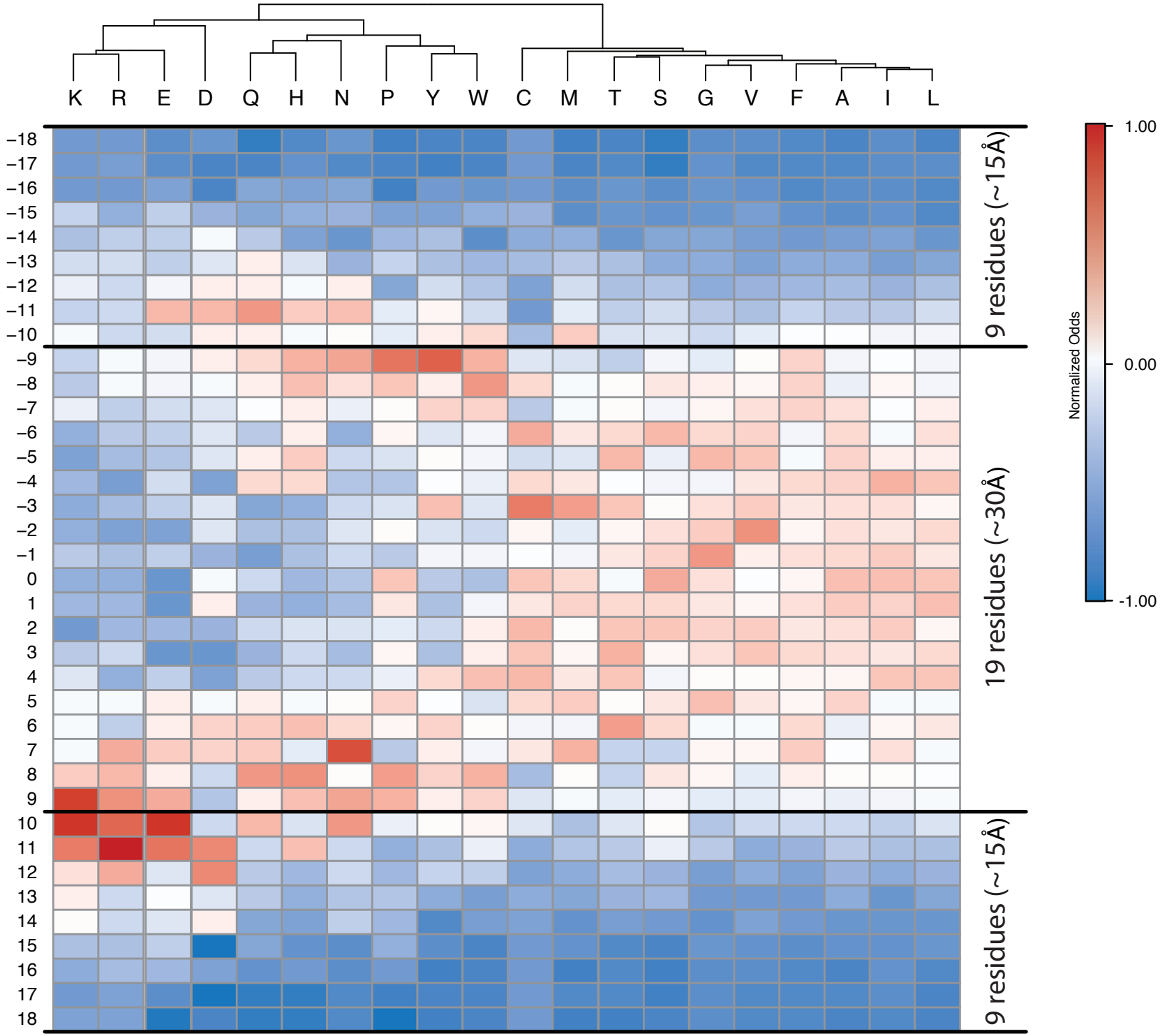
Images


Transmembrane proteins
Water-soluble proteins
Percernt (%)
20.0
15.0
10.0
5.0
.0
1.7 1.3 1.7 1.6 2.0 2.0 1.7 5.2 1.2 3.5 11.4 5.3 15.8 2.7 3.7 9.9 9.8 7.9 8.7 2.8
8.4 5.6 6.5 4.8 5.5 3.9 2.4 5.6 1.2 3.4 10.4 4.5 12.1 2.0 2.5 6,0 5.7 4.0 4.1 1.3
LogOdds
.5
.0
-.5
-1.0
E D K Q R N H S C Y A T L P M V I F G W
Amino acid type

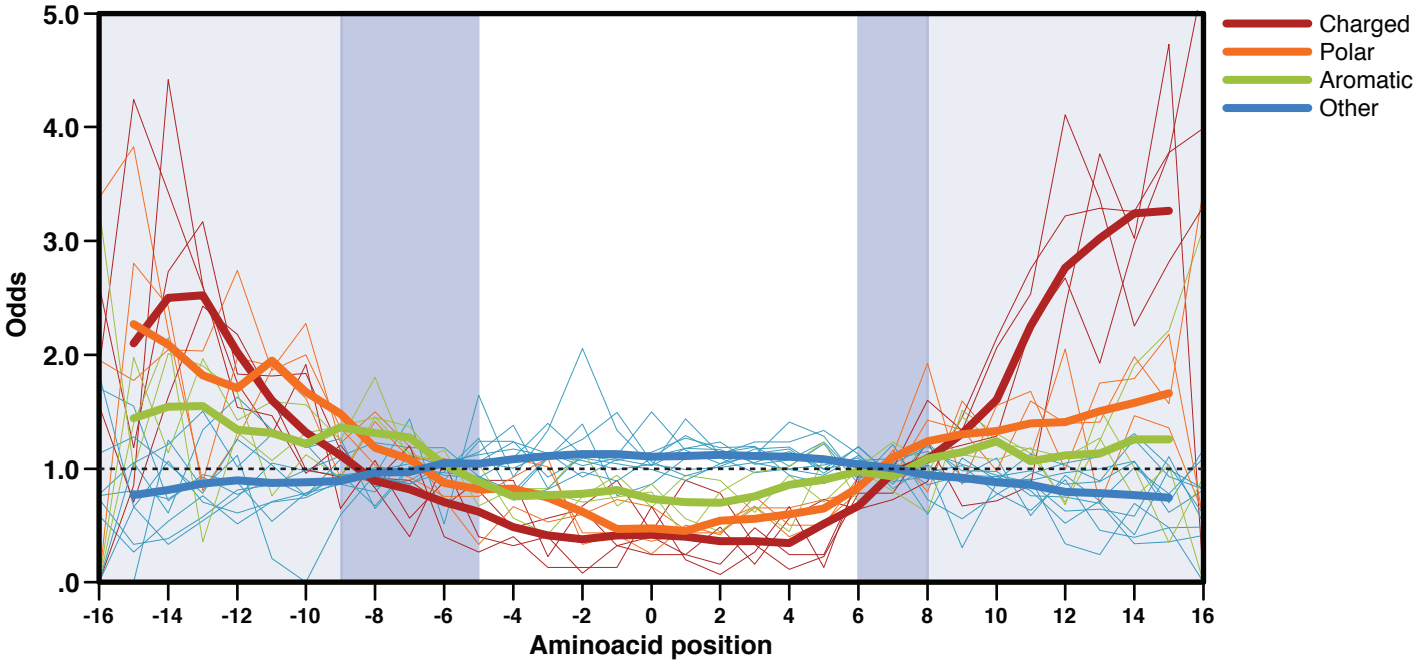
Images


20.0
15.0
10.0
5.0
.0
Transmembrane helices (%)
.0
5.0
10.0
15.0
20.0
Water-soluble helices (%)
L
I
V
G
F
R
K
E
Q
D

Images


K R E D Q H N P Y W C M T S G V F A I L
−18 −17 −16 −15 −14 −13 −12 −11 −10 −9 −8 −7 −6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
9 residues (~15Å)
19 residues (~30Å)
9 residues (~15Å)
Normalized Odds
1.00
0.00
-1.00

Images


5.0
4.0
3.0
2.0
1.0
.0
Odds
-16
-14
-12
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
14
16
Aminoacid position
Charged
Polar
Aromatic
Other

Images