BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 11 2005, pages 2604–2610
doi:10.1093/bioinformatics/bti369

Sequence analysis

A periodicity analysis of transmembrane helices

Hadas Leonov and Isaiah T. Arkin*

The Alexander Silberman Institute of Life Sciences, Department of Biological Chemistry, The Hebrew University, Givat-Ram, Jerusalem 91904, Israel

Received on December 2, 2004; revised on February 15, 2005; accepted on March 1, 2005
Advance Access publication March 3, 2005

ABSTRACT

Transmembrane helices and the helical bundles which they form are the major building blocks of membrane proteins. Since helices are characterized by a given periodicity, it is possible to search for patterns of traits which typify one side of the helix and not the other (e.g. amphipathic helices contain a polar and apolar sides). Using Fourier transformation we have analyzed solved membrane protein structures as well as sequences of membrane proteins from the Swiss-Prot database. The traits searched included aromaticity, volume and ionization. While a number of motifs were already recognized in the literature, many were not. One particular example involved helix VII of lactose permease which contains seven aromatic residues on six helical turns. Similarly six glycine residues in four consecutive helical turns were identified as forming a motif in the chloride channel. A tabulation of all the findings is presented as well as a possible rationalization of the function of the motif.

Contact: arkin@cc.huji.ac.il

INTRODUCTION

Proteins contain periodic structural elements such as α-helices, 3_{10} helices and β-strands. It is inevitable as such that some of these elements will be exposed to different environments, and therefore participate in different types of interactions. As an example, segments of protein secondary structure can be amphiphilic in the sense of having a hydrophobic side and a polar side. Accordingly methods have been developed utilizing Fourier transform based approaches along with hydrophobicity scales, in order to detect such amphiphilic α-helices (Eisenberg *et al.*, 1984; Cornette *et al.*, 1987; Phoenix and Harris, 2002).

Following the idea that different sides of a protein secondary structure element may be exposed to different surroundings (e.g. solvent, lipid, protein core and protein interface), it is possible that some structures contain a segment with a unique characteristic on one side of it. For instance:

(i) G-proteins undergoing palmitoylation were shown to have a patch of positive residues that is used to anchor them to the membrane (Kosloff *et al.*, 2002).

(ii) The GxxxG motif was experimentally found as a motif that is responsible for stabilizing helix–helix interactions in both membrane and water soluble proteins (Lemmon *et al.*, 1994; Arkin and Brunger, 1998; Russ and Engelman, 2000; Kleiger and Eisenberg, 2002). The two glycine residues are found at i and $i+4$ positions, and cover one helical turn. This sequence has been found 32% above expectations in transmembrane α-helices, and 41% above expectations in water soluble proteins (Senes *et al.*, 2000; Kleiger *et al.*, 2002). It is believed that oligomerization is enabled by the glycine's side chain that is the smallest of all amino acids, and allows a tighter packing of helices (Curran and Engelman, 2003; Russ and Engelman, 2000; Javadpour *et al.*, 1999). Furthermore the GxxxG motif was suggested to promote backbone-to-backbone contacts of the C_α-H$\cdots$O type, thereby stabilizing helix–helix interactions in soluble as well as membrane proteins (Kleiger *et al.*, 2002; Kleiger and Eisenberg, 2002; Arbely and Arkin, 2004). Domains in soluble proteins that adopt the Rossmann fold, and also bind FAD and NAD(P), were also discovered to contain GxxxG or GxxxA motifs (Kleiger and Eisenberg, 2002).

(iii) The occurrence of the AxxxA motif was also found above expectations in predicted α-helices, and is significantly enhanced in thermophiles. This suggests a mechanism for thermostability in protein structures, and is explained by the fact that alanine provides a complementary surface, allowing two helices to have a close contact without much loss of side-chain entropy (Kleiger *et al.*, 2002). In addition, a statistical analysis had found patterns of small residues (i.e. G, A and S) in i and $i+4$ positions, in various combinations such as AxxxG, SxxxG, SxxxS, etc. (Senes *et al.*, 2000).

(iv) The periodicity of serine and threonine residues was experimentally found using the TOXCAT method (Russ and Engelman, 1999). They are the polar residues that appear most frequently in transmembrane helices. These residues can drive oligomer formation through a network of inter-helical side-chain hydrogen bonds which do not require breaking the backbone hydrogen bond (MacKenzie and Engelman, 1998; Senes *et al.*, 2000; Dawson *et al.*, 2002).

In this study we have applied the Fourier transform method in order to detect different kinds of periodic patterns. This is not so much a search for 'amphiphilicy' in its broader term of having one side with a certain characteristic and an opposite side with another, but rather one side with a certain periodic property, and an opposite side without it.

*To whom correspondence should be addressed.

METHODS

Database 1: helices from solved membrane proteins

We have consulted the list given by White (http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html) in order to download a set of 35 non-homologous helical membrane proteins from the Brookhaven Protein Data Bank (PDB) (Bernstein *et al.*, 1977). These proteins comprised the solved membrane structures that were used in this study (Supplementary Table 1). A secondary structure state was assigned for each residue in the proteins using the program DSSP (Kabsch and Sander, 1983), which derives the secondary structure from the PDB 3D structure file. Helical sequences of length 7 or more residues were considered in the search for periodic motifs.

Database 2: SwissProt TM helices

A much larger database was used in order to check whether there are similar motifs in other membrane proteins whose structure was not solved, specifically in transmembrane helices. A total of 20 308 membrane proteins were obtained from SwissProt (release 42.10) and TrEMBL (release 25.10) (Boeckmann *et al.*, 2003). Their transmembrane regions were predicted using the TMHMM server (Sonnhammer *et al.*, 1998; Krogh *et al.*, 2001). Helices with 90% or more sequence identity were removed, thereby excluding transmembrane helices that approximately differ by <2 amino acids. It was done in order to avoid analyzing redundant helices, thereby reducing the computational power needed for the analysis. Finally, a database of 65 213 transmembrane helices was subjected to the periodicity search.

Properties and scales

All helices were scanned for the periodic appearance of various chemical properties. Each property is represented by assigning a 'property value', *F(aa)* to each amino acid.

- *Aromatic amino acids*: A binary value is used to represent aromatic residues: $F(F) = F(Y) = F(W) = 1$, while zero is assigned to the rest.
- *Positively charged amino acids*: $F(K) = F(R) = 1, F(H) = 2/3$, while zero is assigned to all other amino acids. Histidine is considered only partially charged due to its low side-chain p*K*a value.
- *Negatively charged amino acids*: $F(D) = F(E) = 1$, while zero is assigned to all other amino acids.
- *Small amino acids*: Repeats of small amino acids may point to a possible helix–helix interaction such as oligomerization or tighter packing that could not take place if larger amino acids were present due to steric interference. To that end, a volume scale was built by the formula

$$F(aa) = \overline{V} - V_{aa} \tag{1}$$

The formula assigns the difference of each amino acid volume (V_{aa}) from the average volume of all amino acids ($\overline{V}$), so that the volume score decreases as the amino acid's volume grows larger.

Mathematical methods: Fourier transform and the amphipathic index

The method for detecting the periodicity of some property in a protein sequence is similar to the way amphipathic α-helices are found (Eisenberg *et al.*, 1984; Cornette *et al.*, 1987; Phoenix and Harris, 2002). Such helices show a spatial segregation of polar and apolar amino acids. Polar residues are located on one side of the helix, while apolar residues are located on the opposite side. An ideal α-helix requires approximately 3.6 residues per turn. Therefore, a peptide that constitutes an amphipathic α-helix will have a periodic variation in its hydrophobic/hydrophilic amino acids: hydrophobic residues will appear every 3–4 residues in positions i and $i+4$, whereas hydrophilic residues will also appear at such a frequency on shifted positions that represent the opposite side of the helix, such as $i+2$ and $i+6$.

In general, the character moment of a protein can be estimated when the period of the sequence, m (the number of residues per turn), is known. When viewing the helix down its axis like in a helical wheel plot, $\theta = 360°/m$ is the angle in which successive side chains emerge from the backbone. Thus for an α-helix, θ is estimated in the range $85° \leq \theta \leq 110°$.

An analytical approach based on Fourier transformation had been developed to identify periodicity patterns in helices (Cornette *et al.*, 1987; Phoenix and Harris, 2002). The method is based on transforming an amino acid sequence into a numerical sequence, where each amino acid is represented by a character value. The Fourier power spectrum for an amino acid sequence of length n with character values $F_1, F_2, \ldots, F_n$ for each amino acid, is defined by

$$P(\theta) = \left[\sum_{k=1}^{n} F_k \cos(k\theta)\right]^2 + \left[\sum_{k=1}^{n} F_k \sin(k\theta)\right]^2 \tag{2}$$

$P(\theta)$ is investigated for the value $\theta = \hat{\theta}$ that maximizes it. Ideal helices with a periodic pattern will form a peak around $\theta = 100°$, and on average at around 97.5°, while other helices will contribute randomly to it (Cornette *et al.*, 1987).

Herein, patterns of periodicity were detected using a similar scoring system based on a measure previously termed as the amphipathic index (AI) (Cornette *et al.*, 1987). In short, it is a measure of how much of the power spectrum is concentrated around angles that represent a helical periodicity, as compared to the total area under the spectrum. It is given by

$$\mathrm{AI}[P(\theta)] = \frac{(1/25°)\int_{85°}^{110°} P(\theta)\,\mathrm{d}\theta}{(1/180°)\int_{0°}^{180°} P(\theta)\,\mathrm{d}\theta} \tag{3}$$

A large AI value will usually result from a helix that is amphipathic according to the property scale by which the power spectrum was computed. Furthermore, this measure overcomes the problem in which different helices of different lengths cannot be compared according to $P(\theta)$. This is due to the fact that a long amphipathic helix will exhibit a larger $P(\theta)$ value than a shorter amphipathic helix.

In this work, we have used different scales instead of the conventional hydrophobicity scales, in order to detect periodicity patterns of various properties. The sliding window method had been employed, using window sizes of 7–30 amino acids to slide over a helix sequence. The AI score has been computed for each window, where windows that passed a certain AI threshold (1.8 for the solved membrane proteins database), were considered as potential motif candidates. Overlapping windows with the same period were merged only if the merged sequence still passed the AI score threshold.

Evolutionary analysis and statistical significance

Potential motif candidates were chosen for further statistical and evolutionary analysis. Initially, a P-value[1] for observing a specific motif, given its amino acid composition, was calculated by shuffling its sequence 10 000 times and computing the AI score for each of these shuffled instances. Secondly, PSI-BLAST (Altschul *et al.*, 1990, 1997) was used in order to find homolog proteins and check whether the motif is conserved in them. A protein was considered as homologous if its sequence alignment with the original protein produced 40–85% sequence identity with an E-value ≤ 0.001. For each homologue, the following values were computed if the sequence that matched the original motif in it contained no gaps in the alignment:

- Fourier transform and AI scores.
- P-value[2] as described above.
- Relative conservation (RC) score. This is a measure for the conservation in the motif itself, compared to the conservation (% identity) of the whole protein, and is given by

$$\mathrm{RC} = \frac{\mathrm{MC}}{\mathrm{PC}} \tag{4}$$

where PC is the protein global conservation (identity percentage between original protein and the specific homolog protein) and MC is the Motif's

[1]This P-value is notated m-p value in Supplementary data.

[2]The P-value average of the homologue sequences is notated as h-p value in Supplementary data.

conservation that is computed by

$$\mathrm{MC} = \left\{ \begin{array}{l} \left\{ \sum_{i=1}^{k} b_{m[i]} * b_{h[i]} * \cos(i\theta) \right\} /m \\ \left\{ \sum_{i=1}^{k} b_{m[i]} * b_{h[i]} * \sin(i\theta) \right\} /m \end{array} \right. \quad (5)$$

where $b_{m[i]}$ is a binary value that receives the value of 1 when position i in the original motif sequence has a property table value $F[i] \geq \overline{F}$. $b_{h[i]}$ is computed in a similar fashion for the homologue motif sequence.

A motif was considered significant if three conditions were satisfied: (a) the AI average for all homolog motif sequences was >75% of the original AI; (b) ≥75% of the homolog motifs were statistically significant according to their P-value (≤0.05); and (c) the RC score was >1 for distant homologs.

In addition, we computed the probability of finding a helix with a motif that receives an AI score above a certain threshold. This was done by shuffling each SwissProt helix 20 times, thereby receiving a total of 1 304 260 random helices, and computing the AI score for each chemical characteristic.

Images were created with VMD 1.82 (Humphrey *et al.*, 1996) and PovRay 3.6 (Persistence of Vision Raytracer Pty. Ltd.; http://www.povray.org).

RESULTS AND DISCUSSION

Solved MP database

All motifs were validated and analyzed by visual inspection of the proteins' solved structure.

Aromatic motifs

Fourteen sequences with 3–6 repeats of aromatic amino acids were found and are listed in Supplementary Table 2. The most common type was of aromatic residues that appear in three sequential turns in a helix (1bcc, 1jb0-b, 1jb0-f, 1lgh and 1occ). Other types belong to sequences with 2–3 sequential appearances of an aromatic residue, a pause of 1–2 turns, and again 1–2 occurrences of an aromatic residue within the same period (1c3w, 1jb0-a, 1jb0-b and 1pv6). Some of the periodic segments that were found are discussed in detail below.

Bacteriorhodopsin (1c3w-A) The three aromatic residues, W182, Y185 and W189, constitute an aromatic motif.[3] The solved structure of this light driven H^+ pump and past studies show that the three aromatic residues located on consecutive turns are involved in affixing the retinal and stabilizing its pocket. W182 creates two types of interactions: (i) A hydrogen bond is formed between the indole of W182 (Nϵ) and A215 (O) via a water molecule 501. (ii) The plane of W86 and W182 contributes to the immobilization of the retinal. Y185 (OH) creates a hydrogen bond with D212 (Oδ1) which in turn forms a hydrogen bond with a water molecule near the retinal. W189 (Nϵ1) forms a hydrogen bond with Y83 (OH) on the opposite helix (Luecke *et al.*, 1999)

Light harvesting complex II (1lgh-B) Three aromatic residues (β-F20, β-F24 and β-F27) in three consecutive helical turns are found in the light harvesting complex II (LHII), a photosynthetic pigment–protein complex (Koepke *et al.*, 1996). Previous studies have shown that many carotenoids are surrounded by aromatic residues in known crystal structures of photosynthetic pigment–protein complexes, and specifically in LHII with the lycopene carotenoid. π–π interactions were discovered to be essential in binding and stabilization of carotenoids. In LHII, they are carried out through aromatic residues and the conjugated C=C bonds in the carotenoids (Wang and Hu, 2002).

[3]Note that F168 residue was found three helical turns away from W182, Y185 and W189, but is probably unrelated.

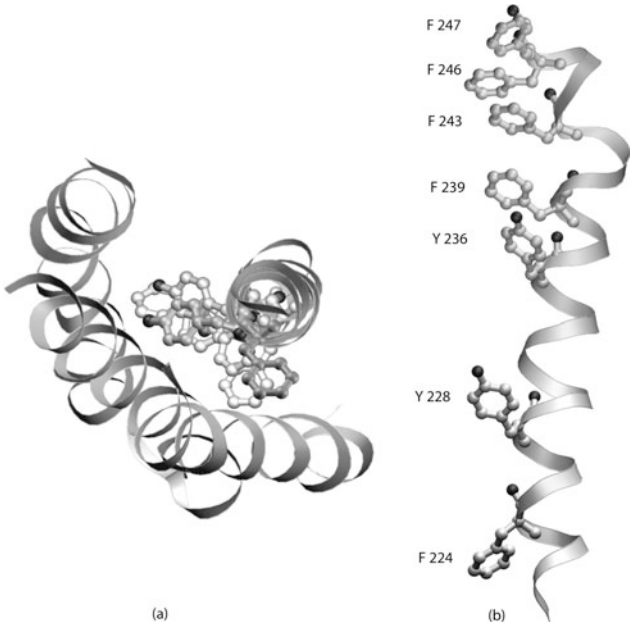

Fig. 1. Lactose permease (1pv6) (Abramson *et al.*, 2003): Orthogonal views of the seven aromatic residues on helix 7 are shown in a ball-and-stick model. **(a)** A view along the helix axis shows the tilt of the side chains, and the helices that pack against this helix. **(b)** Helix 7 is shown without its neighbors. Color legend: light grey—carbon; black—oxygen.

Lactose permease (1pv6-A) This is the largest aromatic motif that was found. The motif is located on helix VII of lactose permease, a polytopic membrane protein that catalyzes lactose/H^+ symport. Helix VII is located near the sugar translocation pathway, suggesting an active role of the helix in the sugar transport mechanism (Voss *et al.*, 1997). The helix contains seven aromatic residues on six helical turns: F224, Y228, Y236, F239, F243, F246 and F247 (Fig. 1). Cysteine-scanning mutagenesis of helix VII (Frillingos *et al.*, 1994) had shown that very few residues are directly involved in the actual transport mechanism of lactose. The critical residues on this helix are in fact D237 and D240: their replacement interferes with the charge-neutralizing interactions between D237 and K358 or D240 and K319. Cysteine scanning had measured the differences in lactose transport rates both in the initial rate of uptake and the steady-state accumulation. It had shown that the positions where Cys replacements exhibit high sensitivity of the permease to inactivation lie between Y236 and F247 where some of which (Y236, F243 and F247) lie on the same relatively hydrophilic face as D237 and D240. An exception to that is of Y228 in which a Cys replacement has a significantly lower initial transport rate (<20% of the wild-type rate). In addition, the mutant Y236C exhibits little transport activity which is abolished by Y236F, suggesting that Y236 is required for H-bonding. It was later revealed in the 3D-structure of lactose permease that Y236 participates in a hydrogen bond with H322 (Abramson *et al.*, 2003).

Furthermore, binding a lactose homologue (TDG) as a ligand is thought to cause a 'scissors-like' movement between helix VII and helix I or II. Such a movement allows conformational flexibility between the helices that is important for the conformational change

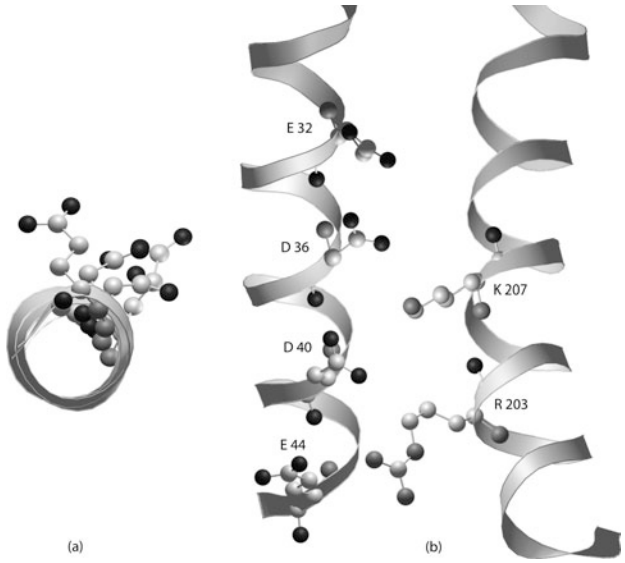

Fig. 2. Cytochrome bc1 (1bcc) (Zhang *et al*., 1998): **(a)** A view along the helix axis. **(b)** The negatively charged residues are on the left helix, whereas their possible partners for forming salt bridges, R203 and K207, are shown on the right helix. Color legend: light grey—carbon; grey—nitrogen; black—oxygen.

that results in the release of the proton and the substrate into the cell (Venkatesan *et al*., 2000).

Perhaps this conformational flexibility also allows the aromatic motif that we have found to create an aromatic lane. Such a lane offers both hydrophobic interactions and hydrogen bonds along which the lactose glides, much like the 'greasy slide' that was found to facilitate maltose translocation in the maltoporin channel (Dutzler *et al*., 2002b). If only one of the aromatic residues is replaced, it may only slow down the lactose uptake rate, rather than critically disrupt it. To the best of our knowledge, no research has been conducted up to this date that measures transport rates when mutating more than one of these aromatic residues.

On the other hand, site-directed spin labeling (Voss *et al*., 1997) demonstrated that the average accessibility of residues 233–246 is lower than that observed for helix XII (with one face towards the lipid). This suggests that helix VII is a part of the hydrophobic core, and might have a significant role in the protein's packing, folding and stability. In other words, this motif may be important in stabilizing the protein's core, with stacking interactions, hydrogen bonds and hydrophobic interactions that can be formed by these aromatic residues.

Aromatic repeats of 3–4 helical turns were also discovered in cytochrome bc1 (1bcc) and photosystem I (1jb0). Specifically in the photosystem I complex, four different aromatic repeats were found. This is not surprising with regard to the large amount of chlorophylls that surround each helix, which could interact with them.

Negatively charged motifs

Two sequences with amino acids that can potentially be negatively charged were found (Supplementary Table 3).

Cytochrome bc1 (1bcc-J) Four negatively charged amino acids are found in four consecutive helical turns (E32, D36, D40 and E44) in a part of the helix that is assumed to be outside the membrane or in the region of the lipid's polar head-groups (Fig. 2). These residues face chain D of the protein which has two positively charged amino acids (R203 and K207). The following distances are suggestive of possible salt bridges that may be used to stabilize the protein: R203[Nη] ↔ D40[Oδ1]: 4.03 Å; R203[Nη] ↔ E44[Oϵ2]: 3.53 Å; K207[Nζ] ↔ D36[Oδ1]: 3.3 Å; K207[Nζ] ↔ D40[Oδ1]: 4.105 Å; and K207[Nζ] ↔ D40[Oδ2]: 4.09 Å.

Calcium ATPase (1eul-A) Three glutamate residues in three helical turns (E109, E113 and E117) are found on the edge of a transmembrane helix which stretches out into a soluble medium. It faces the water, possibly forming hydrogen bonds. A previous study (Clarke *et al*., 1989) has shown that amino acid substitutions to E109, E113 and the segment 109-ERNAE-113 did not affect Ca^{2+} transport at all, thereby negating the importance of this acidic periodical region.

Positively charged motifs

Two sequences with potential positively charged amino acids were found (Supplementary Table 3).

Calcium ATPase (1eul-A) Three positive residues are found in four helical turns (R751, K758 and R762) on helix S5 of the sarcoplasmic reticulum ATPase. This is a helix that had been found to be important for mediating communication between the Ca^{2+} binding pocket and the catalytic domain. Past mutation studies (Sorensen *et al*., 1997; Sorensen and Andersen, 2000) revealed that they were all sensitive towards substitution: R751 appears to be crucial and therefore highly conserved in the P-type ATPase family. Mutations of R751 (except the conservative R751K mutant) were found to result in a completely non-functional or non-expressed protein. The 3D structure of the protein implies that R751 participates in hydrogen bonds with residues on the 'L-6-7' loop (Toyoshima *et al*., 2000). K758 had also been mutated and found to be important in regulating conformational equilibrium. A K758I mutant resulted in the increase of dephosphorylation of the E_2P conformation to E_2, and a decrease in the rate of Ca^{2+} binding to the E_2 state. The K758R mutant did not have any effect. In addition, K758 is also known to be hydrogen-bonded to a residue on the 'L6-7' loop. R762 also proved to be important and was assumed to interact with residues on helices M7 and M8. In the mutant R762I, the rate of Ca^{2+} binding transition was reduced by a factor of 3.5 at 25°C.

Thus, these residues are not involved in one specific role, but rather are responsible for stabilizing the overall structure of the protein and for performing a functional role in the Ca^{2+} binding structure.

Cytochrome C oxidase (1occ-E) Four positive residues are found in three non-consecutive helical turns (K46, R53, R56 and R57) on a helix that is in an aqueous environment. They are not mentioned in previous studies and are likely to be involved in hydrogen bonds with water molecules.

Small amino acids motifs

The repeats that were found (Supplementary Table 4) are mostly composed of glycine, alanine and serine residues, since these amino acids have the smallest volumes, 60.1, 88.6 and 89 $Å^3$ respectively.

Cytochrome C oxidase (1occ-A 451-461) Four serine residues are found in three helical turns, preceded by an aspargine residue. All are uncharged polar residues which create a network of possible

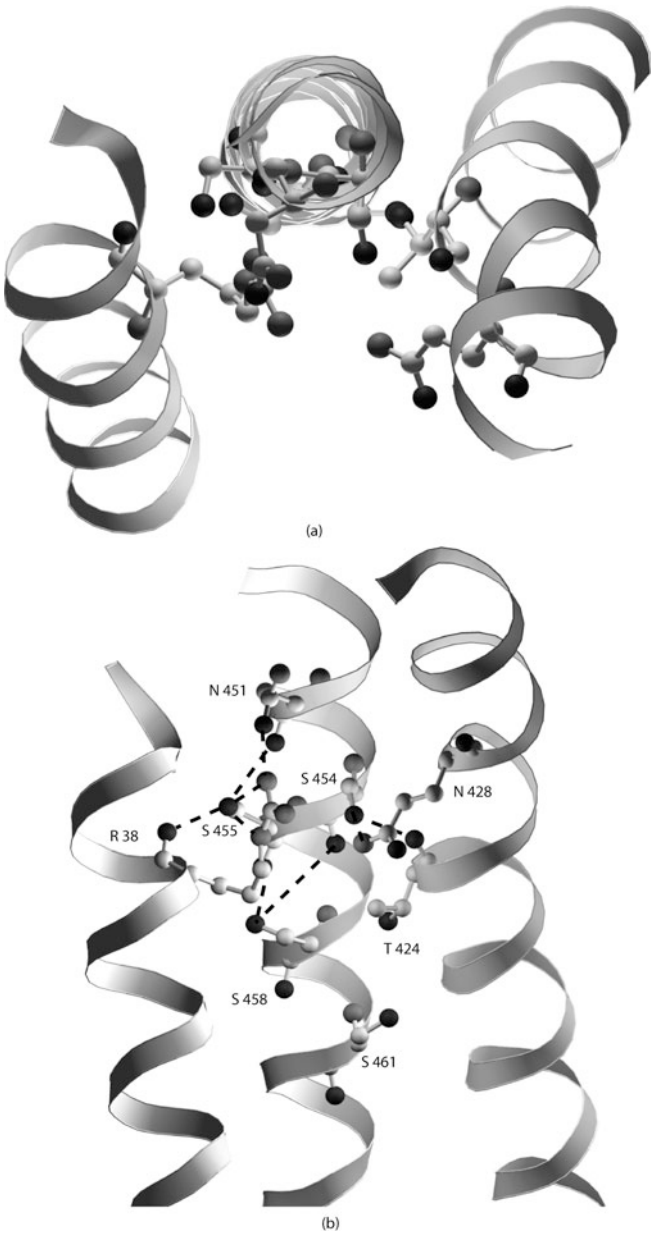

Fig. 3. Cytochrome C oxidase (1occ) (Tsukihara *et al*., 1996): Small and polar amino acids (Ser and Asn) that are able to form hydrogen bonds were found to be periodic in chain A 451–461. **(a)** The arrangement of the helices, where the Asn and Ser residues are located on the middle helix, are all shown on the same side. **(b)** The network of hydrogen bonds (dashed black) that is formed only between side chains of the middle helix (containing the motif), and neighboring residues. Color legend: light grey—carbon, black—oxygen.

hydrogen bonds (Fig. 3), some of which are bifurcated. Specifically, N451[Nδ2] $\leftrightarrow$ Y447[O]: 2.85 Å; N451[Oδ2] $\leftrightarrow$ R38[NH_2]: 3.13 Å; S454[Oγ] $\leftrightarrow$ N428[Nϵ2]; 3.04 Å; S454[Oγ] $\leftrightarrow$ T424[O]: 2.62 Å, S455[Oγ] $\leftrightarrow$ R38[O], [NE] and [NH_2]: 2.71, 2.85 and 3.03 Å; S455[Oγ] $\leftrightarrow$ N451[O]: 3.26 Å; S458[Oγ] $\leftrightarrow$ S455[O]: 2.76 Å; and S458[Oγ] $\leftrightarrow$ S454[O]: 4.6 Å. Moreover, the structure of this protein reveals a network of hydrogen bonds that takes place through water molecules. S461 and T424 are

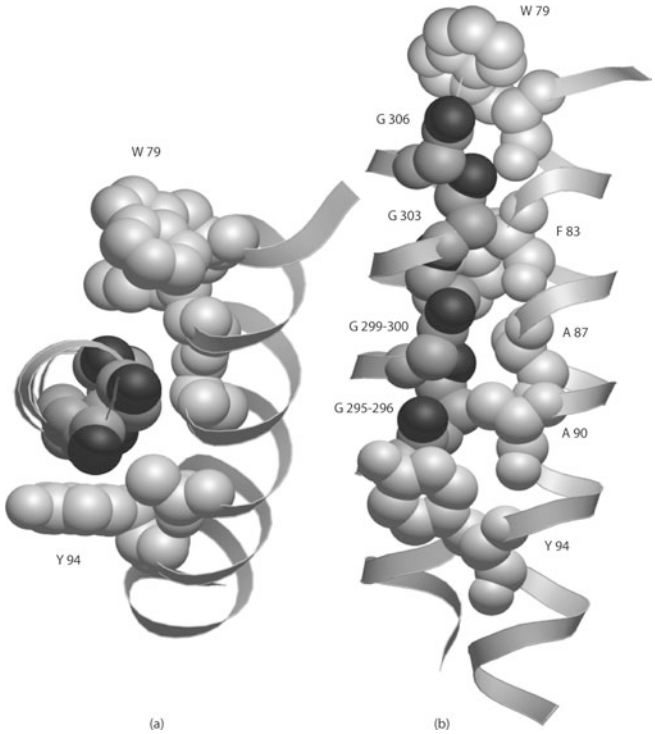

Fig. 4. Chloride channel (1kpl) (Dutzler *et al*., 2002a): The six glycine residues are shown in a CPK atom model in the left helix and colored by atom (dark grey—carbon; black—oxygen). Interface residues from a neighboring helix are shown and colored in light grey. Bulky aromatic amino acids reside in the helix ends. They may cause fixation of both helices. Alanines are found in the helix–helix interface with the glycines. **(a)** A view along the left helix axis, showing the 3D arrangement of the two helices. **(b)** The interface of the two helices.

hydrogen-bonded to a group on heme A via a water molecule. R38, which is also shown by our analysis to be stabilized by S455, is hydrogen-bonded to heme A as well (Tsukihara *et al*., 1996; Yoshikawa *et al*., 1998). Thus one can assume that these hydrogen bonds play a role in tighter packing of the helices and the heme.

Cytochrome C oxidase (1occ-A 14–40) Periodicity of four glycine residues and one serine in five helical turns. They are found to face another helix of chain A (56–82) where somewhat larger amino acids face them (M, H and I). The sequence of cytochrome C oxidase is very conserved. The homologues that were found were within the range of 74–85% identity; therefore the RC score for this motif does not show significance. This evolutionary factor is also related to the average P-value of the homologues which is only marginally significant.

Photosystem I (1jb0-B) The repeats that were found in this helix were in a 110° period. However the solved structure does not match such a period and only a part of the repeats reside on the same side of the helix (S426, S429, G433 and T436). A visual inspection of the structure finds them in a helix–helix interface.

Chloride channel (1kpl-B) Six glycine residues in four helical turns were found in this very conserved anion selective channel (Fig. 4). The RC score is relatively low because most homologues

except three have a total of ~80% sequence identity to the original (1kpl) sequence. The motif is highly conserved in these close homologues (only G295 is substituted twice by alanine). The more distant sequences preserve four of these glycines, where G295 is substituted by leucine, and G285 is substituted by serine or threonine which are also relatively small amino acids that also have hydrogen bonding possibilities.

SwissProt database

The motifs that were found from a scan of the SwissProt database are presented in Supplementary Tables 5–8. The AI threshold that was used here was higher than the threshold used in the database of solved membrane proteins, and different for each property, so that only the strongest and most significant motifs were considered. In order to evaluate these results we have computed the probabilities of finding a motif sequence with an AI that is larger than a certain threshold from 1 304 260 shuffled SwissProt helices. These results are depicted in Supplementary Tables 9–12.

(i) *Aromatic motifs*: Twenty-six motifs that passed an AI threshold of AI $\geq$ 2.5 and the thresholds of the evolutionary analysis were found. They contain periodic repeats of 4–6 aromatic acids and are depicted in Supplementary Table 5. The probabilities of finding such results are depicted in Supplementary Table 9. The probabilities for AI $\geq$ 2.5, 2.6, 2.7 and 2.8 are 0.002. 0.001, 0.00044 and 0.00023 respectively.

(ii) *Charged motifs*: Not surprisingly, very few repeats of charged amino acids were found. All of them have an AI $\geq$ 2. Three unique negative motifs and two positives were found (out of 16 motifs, some of which belong to similar proteins). They are depicted in Supplementary Tables 6 and 7. According to Supplementary Tables 10 and 11, it can be seen that the probability of finding a charged motif with AI $\geq$ 2.1 is in the order of 10^{-5}.

(iii) *Small amino acid motifs*: Twenty-two motifs that passed an AI threshold of AI $\geq$ 4 and the thresholds of the evolutionary analysis are presented in Supplementary Table 8. These motifs are mostly variations of G, A, S and T. A large part of these motifs had passed the *P*-value tests, but not the RC threshold. Furthermore, the probabilities of finding a small amino acid motif with AI $\geq$ 4, 4.2, 4.4 and 4.6 are 0.001, 0.00047, 0.00022 and 9.5×10^{-5} respectively.

Overall, the *P*-values are affected by the short sequence of the motifs and their relatively lower amino acid variation that results from their location within a membranous environment (Arkin and Brunger, 1998). However, the *P*-values obtained for the motifs are fairly reasonable with respect to these effects. Indeed, our raw results showed a higher number of motifs, some of which could have been found by random. Yet, the key to finding which of these results might be of actual biological importance is the evolutionary analysis that we conducted, which we used to filter out insignificant results, and which showed the statistical conservation of the motifs presented in this study.

GENERAL CONCLUSIONS

In this work we have utilized a mathematical method in order to detect structural motifs at the sequence level of membrane helices, given the assignment of their secondary structure elements. This method has been applied to a specific class of membrane proteins—those that are folded as α-helical bundles—but it could easily be applied to soluble proteins as well. This scan has the potential of locating important sites in a protein that could possess many possible properties and functions such as affinity for specific molecules [e.g. G-proteins with a positive patch that provide affinity for membrane anchorage (Kosloff *et al.*, 2002)], binding substrates, segments that are responsible for the protein's conformational stability, the folding process, membrane insertion, oligomerization of helical subunits and tight packing of the protein's core through hydrogen bonds or other hydrophobic interactions. Therefore such sites, when present, could point us to the crucial residues in a protein, thereby supplying information for mutations studies.

Aromatic repeats that were discovered usually contained 3–6 repeats of aromatic amino acids (which were not always consecutive). Most are assumed to have a role in the stability of the protein. For instance the aromatic residues found in bacteriorhodopsin have already been shown to be important in stabilizing the retinal binding pocket (Luecke *et al.*, 1999). Surprisingly, the longest motif that we found, in lactose permease, was not mentioned by previous studies and its importance remains unknown.

Polar charged residues are infrequent in membrane proteins because of their preference to reside in an aqueous environment. These motifs can contribute to the stability of the protein by forming salt bridges and hydrogen bonds. The charged amino acid repeats that were found in the database of solved membrane proteins were shown to be important mainly for stability reasons.

Repeats of small amino acids were found more frequently. The motifs that were found and had been structurally analyzed have been shown to lie within a close range of other helices with bulk residues, or within hydrogen-bonding range of neighboring polar residues.

The RC score has not always succeeded in verifying the results that were found (where RC $<$ 1), pointing that evolution had not preserved all of it or that the protein itself is highly conserved. Therefore it depends on the distribution of homologues—if there are more closer homologues than distant ones, the RC score will show a decrease. However, it does not always mean the motif or a part of it does not have an important biological function.

In general, periodicity patterns do not characterize transmembrane helices; this can be seen by the small proportion of motifs we have found in the SwissProt database (Supplementary Table 13). Thus when evolution chooses to keep a periodicity pattern in a protein, it is possible to suggest that the pattern is important for the protein's function or folding. The motifs that were found in this research both in the database of TM helices from SwissProt, and the database of solved membrane proteins are all theoretical suggestions for important (if not critical) sites. Experimental analyses shall prove if they are indeed significant.

ACKNOWLEDGEMENTS

This work was supported in part by a grant from the Israel Science Foundation (784/01) to I.T.A.

REFERENCES

Abramson,J. *et al.* (2003) Structure and mechanism of the lactose permease of *Escherichia coli*. *Science*, **301**, 610–615.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Arbely,E. and Arkin,I.T. (2004) Experimental measurement of the strength of a C alpha-H. . .O bond in a lipid bilayer. *J. Am. Chem. Soc*., **126**, 5362–5363.

Arkin,I.T. and Brunger,A.T. (1998) Statistical analysis of predicted transmembrane alpha-helices. *Biochim. Biophys. Acta.*, **1429**, 113–128.

Bernstein,F.C. *et al*. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Boeckmann,B. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Clarke,D.M. *et al*. (1989) Functional consequences of glutamate, aspartate, glutamine, and asparagine mutations in the stalk sector of the Ca^{2+}−ATPase of sarcoplasmic reticulum. *J. Biol. Chem.*, **264**, 11246–11251.

Cornette,J.L. *et al*. (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.*, **195**, 659–685.

Curran,A.R. and Engelman,D.M. (2003) Sequence motifs, polar interactions and conformational changes in helical membrane proteins. *Curr. Opin. Struct. Biol.*, **13**, 412–417.

Dawson,J.P. *et al*. (2002) Motifs of serine and threonine can drive association of transmembrane helices. *J. Mol. Biol.*, **316**, 799–805.

Dutzler,R. *et al*. (2002a) X-ray structure of a ClC chloride channel at 3.0 A reveals the molecular basis of anion selectivity. *Nature*, **415**, 287–294.

Dutzler,R. *et al*. (2002b) Translocation mechanism of long sugar chains across the maltoporin membrane channel. *Structure (Camb.)*, **10**, 1273–1284.

Eisenberg,D. *et al*. (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl Acad. Sci. USA*, **81**, 140–144.

Frillingos,S. *et al*. (1994) Cysteine-scanning mutagenesis of putative helix VII in the lactose permease of *Escherichia coli*. *Biochemistry*, **33**, 8074–8081.

Humphrey,W. *et al*. (1996) VMD: visual molecular dynamics. *J. Mol. Graph*, **14**, 33–38.

Javadpour,M.M. *et al*. (1999) Helix packing in polytopic membrane proteins: role of glycine in transmembrane helix association. *Biophys. J.*, **77**, 1609–1618.

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

Kleiger,G. and Eisenberg,D. (2002) GXXXG and GXXXA motifs stabilize FAD and NAD(P)-binding Rossmann folds through C(alpha)-H · · · O hydrogen bonds and van der waals interactions. *J. Mol. Biol.*, **323**, 69–76.

Kleiger,G. *et al*. (2002) GXXXG and AXXXA: common alpha-helical interaction motifs in proteins, particularly in extremophiles. *Biochemistry*, **41**, 5990–5997.

Koepke,J. *et al*. (1996) The crystal structure of the light-harvesting complex II (B800-850) from *Rhodospirillum molischianum*. *Structure*, **4**, 581–597.

Kosloff,M. *et al*. (2002) Structural homology discloses a bifunctional structural motif at the N-termini of G alpha proteins. *Biochemistry*, **41**, 14518–14523.

Krogh,A. *et al*. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.,* **305**, 567–580.

Lemmon,M.A. *et al*. (1994) A dimerization motif for transmembrane alpha-helices. *Nat. Struct. Biol.*, **1**, 157–163.

Luecke,H. *et al*. (1999) Structure of bacteriorhodopsin at 1.55 A resolution. *J. Mol. Biol.*, **291**, 899–911.

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.

Phoenix,D.A. and Harris,F. (2002) The hydrophobic moment and its use in the classification of amphiphilic structures (review). *Mol. Membr. Biol.*, **19**, 1–10.

Russ,W.P. and Engelman,D.M. (1999) TOXCAT: a measure of transmembrane helix association in a biological membrane. *Proc. Natl Acad. Sci. USA*, **96**, 863–868.

Russ,W.P. and Engelman,D.M. (2000) The GxxxG motif: a framework for transmembrane helix–helix association. *J. Mol. Biol.*, **296**, 911–919.

Senes,A. *et al*. (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.

Sonnhammer,E.L. *et al*. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175–182.

Sorensen,T.L. and Andersen,J.P. (2000) Importance of stalk segment S5 for intramolecular communication in the sarcoplasmic reticulum Ca^{2+}−ATPase. *J. Biol. Chem.*, **275**, 28954–28961.

Sorensen,T. *et al*. (1997) Mutation Lys758→Ile of the sarcoplasmic reticulum Ca^{2+}−ATPase enhances dephosphorylation of E2P and inhibits the E2 to E1Ca2 transition. *J. Biol. Chem.*, **272**, 30244–30253.

Toyoshima,C. *et al*. (2000) Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature*, **405**, 647–655.

Tsukihara,T. *et al*. (1996) The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å. *Science*, **272**, 1136–1144.

Venkatesan,P. *et al*. (2000) Site-directed sulfhydryl labeling of the lactose permease of Escherichia coli: helix VII. *Biochemistry*, **39**, 10641–10648.

Voss,J. *et al*. (1997) Site-directed spin-labeling of transmembrane domain VII and the 4B1 antibody epitope in the lactose permease of *Escherichia coli*. *Biochemistry*, **36**, 15055–15061.

Wang,Y. and Hu,X. (2002) A quantum chemistry study of binding carotenoids in the bacterial light-harvesting complexes. *J. Am. Chem. Soc*., **124**, 8445–8451.

Yoshikawa,S. *et al*. (1998) Redox-coupled crystal structural changes in bovine heart cytochrome c oxidase. *Science*, **280**, 1723–1729.

Zhang,Z. *et al*. (1998) Electron transfer by domain movement in cytochrome bc1. *Nature*, **392**, 677–684.