

# Statistical Analysis of Amino Acid Patterns in Transmembrane Helices: The GxxxG Motif Occurs Frequently and in Association with $\beta$ -branched Residues at Neighboring Positions

Alessandro Senes, Mark Gerstein and Donald M. Engelman\*

Department of Molecular  
Biophysics & Biochemistry,  
Yale University, P.O. Box  
208114, New Haven  
CT 06520-8114, USA

To find motifs that mediate helix-helix interactions in membrane proteins, we have analyzed frequently occurring combinations of residues in a database of transmembrane domains. Our analysis was performed with a novel formalism, which we call TMSTAT, for exactly calculating the expectancies of all pairs and triplets of residues in individual sequences, taking into account differential sequence composition and the substantial effect of finite length in short segments. We found that the number of significantly over and under-represented pairs and triplets was much greater than the random expectation. Isoleucine, glycine and valine were the most common residues in these extreme cases. The main theme observed is patterns of small residues (Gly, Ala and Ser) at  $i$  and  $i + 4$  found in association with large aliphatic residues (Ile, Val and Leu) at neighboring positions (i.e.  $i \pm 1$  and  $i \pm 2$ ). The most over-represented pair is formed by two glycine residues at  $i$  and  $i + 4$  (GxxxG, 31.6% above expectation,  $p < 1 \times 10^{-33}$ ) and it is strongly associated with the neighboring  $\beta$ -branched residues Ile and Val. In fact, the GxxxG pair has been described as part of the strong interaction motif in the glycoporphin A transmembrane dimer, in which the pair is associated with two Val residues (GVxxGV). GxxxG is also the major motif identified using TOXCAT, an *in vivo* selection system for transmembrane oligomerization motifs. In conjunction with these experimental observations, our results highlight the importance of the GxxxG +  $\beta$ -branched motif in transmembrane helix-helix interactions. In addition, the special role for the  $\beta$ -branched residues Ile and Val suggested here is consistent with the hypothesis that residues with constrained rotameric freedom in helical conformation might reduce the entropic cost of folding in transmembrane proteins. Additional material is available at <http://engelman.csb.yale.edu>/tmstat and <http://bioinfo.mbb.yale.edu>/tmstat.

© 2000 Academic Press

**Keywords:** membrane proteins; protein folding; glycine;  $\beta$ -branched; sequence analysis

\*Corresponding author

## Introduction

The two dozen high-resolution structures of integral membrane proteins available so far have revealed only two simple folds, the helical bundle and the closed beta barrel. These folds are the sim-

plest solutions to satisfying the hydrogen bonding potential of the polypeptide backbone amide groups in the lipid bilayer. In the helical family, the membrane-spanning domains are generally composed of very hydrophobic stretches of 20-30 amino acid residues. Algorithms based on hydrophobicity scales (Boyd *et al.*, 1998; Engelman *et al.*, 1986; Kyte & Doolittle, 1982; von Heijne, 1992) reliably identify these domains from primary sequences. As a consequence, a large database of predicted helical transmembrane (TM) domains (TMD) exists.

Abbreviations used: TM, transmembrane; TMD, transmembrane domain; GpA, glycoporphin A; CD, circular dichroism.

E-mail address of the corresponding author:  
[don@paradigm.csb.yale.edu](mailto:don@paradigm.csb.yale.edu)

Thus, structural information in a membrane protein sequence can be statistically interpreted. Elements of the structural simplicity of these proteins suggest the existence of commonly used patterns in transmembrane helix-helix interactions. First, the space that natural selection can sample in search of favorable combinations seems to be limited by the low complexity of the sequences, since two-thirds of transmembrane residues comprise, on average, only six amino acids (Leu, Ile, Val, Phe, Ala, and Gly), as schematized in Figure 1(a). The helices tend to adopt perpendicular orientation in order to span the bilayer (Bowie, 1997), and helix packing theories suggest that only a subset of the relative inter-helical orientations are optimal for interaction (Bowie, 1997; Chothia *et al.*, 1981; Richmond & Richards, 1978; Walther *et al.*, 1996). Moreover, the need for a detailed fit to maximize weak van der Waals interactions and the preference for preformed interfaces to minimize entropy lost upon packing, as postulated by MacKenzie & Engelman (1998), could also limit the number of conformations suitable for interaction.

The existence of correlations between residues has been suggested in previous statistical studies on predicted TM sequence databases (Arkin & Brunger, 1998; Landolt-Marticorena *et al.*, 1993; Samatey *et al.*, 1995). Here, we present a rigorous analysis of the frequency of occurrence of all pairs and triplets of amino acids in a large non-homologous set of sequences, compared with their theoretical expectancies.

Expectancy can be calculated trivially by the product of the frequency of amino acids in the database. However, this method requires the assumption that in terms of composition, the sequences belong to a homogeneous population. We have formulated a procedure for calculating analytically expectancy distributions of occurrence

based on the composition of the individual sequences. Our method takes into account the finite-length effect that for short sequences, such as transmembrane domains, becomes very important. The ratios between the observed occurrences and their relative mean expectancy value (odds ratio) allowed for the identification of the over-represented and under-represented pairs and triplets. The exact expectancy distribution permitted the calculation of a precise statistical significance for the observed differences from expectation. Our results show that a large number of significant cases exist, suggesting structural themes in helix-helix interaction and a special role for glycine and the  $\beta$ -branched Ile and Val in transmembrane domains.

## Results

### Characteristics of the database

The collection of sequences used in this analysis was obtained from the 49,946 transmembrane domains annotated in the Swiss-Prot database. In 94% of the cases, the annotations were marked as potential, possible or probable, indicating that they were identified by hydrophobicity algorithms (Table 1). In order to remove homology, TM sequences with high similarity scores to others were excluded, as described in Methods. The procedure yielded a database with 13,606 TMDs, an adequate size for the proposed analysis.

Helical transmembrane domains generally vary in length from 20 to 30 residues. The sequences annotated in Swiss-Prot are mainly in that range, as shown in the histogram in Figure 1(b). Instead of using the complete annotations, however, we performed the analysis on fixed-length windows of 18 residues selected for maximum hydrophobicity.

**Table 1.** Transmembrane annotations in the Swiss-Prot database (rel. 37 and updates until March 17, 1999)

	All proteins			Bitopic proteins <sup>a</sup>			Polytopic proteins <sup>b</sup>		
	Protein records <sup>c</sup>	TMs <sup>d</sup>	Potential (%) <sup>e</sup>	Protein records	TMs <sup>d</sup>	Potential (%) <sup>e</sup>	Protein records <sup>c</sup>	TMs <sup>d</sup>	Potential (%) <sup>e</sup>
Complete database <sup>f</sup>	10,769	46,946	94.3	3863	3863	78.5	6906	43,053	95.8
Eukaryota	6587	27,288	92.5	2630	2630	74.4	3957	24,658	94.5
Bacteria	3156	16,881	97.2	706	706	93.6	2450	16,175	97.4
Archaea	297	1341	98.2	71	71	94.4	226	1270	98.4
Viruses	729	1436	90.6	456	456	76.3	273	980	97.2
Non-homologous database <sup>g</sup>	5309	13,606	96.4	1174	1174	84.2	4135	12,432	97.6
Eukaryota	2510	5619	95.2	749	749	79.0	1761	4870	97.6
Bacteria	2333	6963	97.4	294	294	95.2	2039	6669	97.5
Archaea	244	666	98.6	43	43	97.7	201	623	98.7
Viruses	222	358	93.6	88	88	84.1	134	270	96.7

<sup>a</sup> Proteins containing a single transmembrane domains (single-span).

<sup>b</sup> Proteins containing multiple transmembrane domains (multi-span).

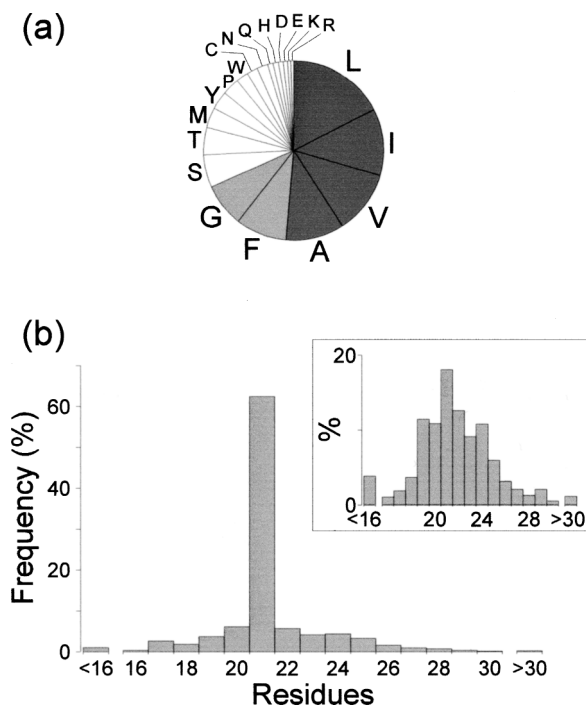
<sup>c</sup> Number of proteins containing transmembrane annotations.

<sup>d</sup> Annotated transmembrane domains (TRANSMEM entries in the FT field).

<sup>e</sup> Transmembrane annotations marked as POTENTIAL or POSSIBLE or PROBABLE.

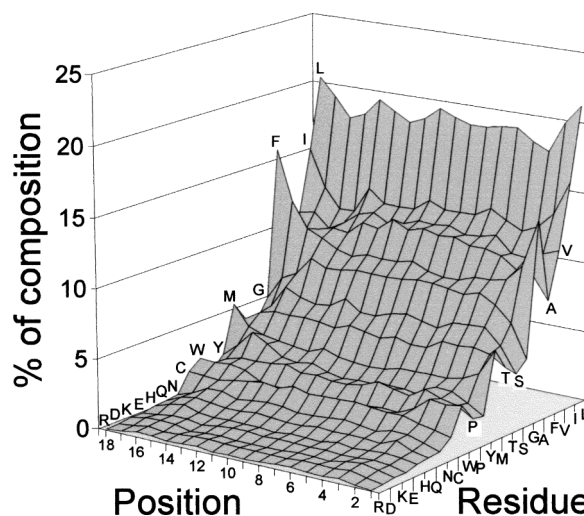
<sup>f</sup> Database containing all Swiss-Prot annotated transmembrane domains.

<sup>g</sup> Database used for the statistical analysis, obtained from the complete database after homology removal.



**Figure 1.** Average composition and length of the transmembrane annotations in Swiss-Prot. (a) Composition; the shaded areas emphasize the fact that, on average, one half of transmembrane residues comprise only four amino acids and two-thirds of the total only six amino acids. (b) Length distribution of all transmembrane annotations. Inset: distribution of the sequences not labeled as POTENTIAL, PROBABLE or POSSIBLE.

The composition of transmembrane sequences varies along sections exposed to different environments (water interface, lipid head-group and hydrocarbon regions). To avoid highlighting amino acid correlations that are due to these variations, it is important to limit the analysis to the portion of the sequence likely to be exposed to the hydrocarbon region. Moreover, the exact definition of the boundaries of a transmembrane domain is a non-trivial problem, even among solved structures. The majority of the sequences are putative, and incorrect assignment of the boundaries would result in contamination from flanking regions; using shorter sequences selected for hydrophobicity should minimize this risk. The exact boundaries of the transmembrane regions are generally not well established in Swiss-Prot, as demonstrated by the peak at 21 residues in the length distribution of the annotations (Figure 1(b)). Such a sharp demarcation is probably an artifact of the algorithms commonly used to identify transmembrane domains. The distribution is much more widespread when the putative annotations are excluded (inset). The amino acid composition of each of the 18 positions of the analyzed sequences is shown in Figure 2. The major differences in com-



**Figure 2.** Variation in the amino acid composition at different positions in the 18 amino acid residue transmembrane sequences used for the pair and triplet correlation analysis. Each 18 residue sequence corresponds to the most hydrophobic window of the 30 residue span centered on each Swiss-Prot annotation.

position are limited to the extremities, and the variations are significantly reduced compared to those of the original annotations (data not shown).

### Analytical procedures

The one-letter code of the two amino acids followed by their separation  $k$  will be used to indicate pairs of residues at distance  $i, i+k$  ( $k$  is also referred to as the register). For example, the pair in which Ala and Leu are at  $i, i+3$  (AxxL) is indicated by AL3. The occurrences in the TM sequence database of all 4000 pairs formed by all combinations of the 20 amino acids at registers 1 to 10 were counted. Raw counts are not very informative, since the main factor determining the gross number of occurrences of a pair in a set of sequences is the relative frequency of its residues. The Leu-Leu pairs at all separations (LL $k$ ), for instance, are the most numerous among the pairs in the database, since Leu is by far the most frequent residue. Therefore, to identify specific relationships that might be clues to helix interactions, it is necessary to refer to an expectation of the occurrence of each amino acid pair and triplet that permits distinguishing over-represented and under-represented pairs while accounting for the relative frequency of the amino acids. We calculated the expectation with a novel formalism named TMSTAT that incorporates both composition and length of every individual sequence and thus does not require the assumption that all sequences belong to a homogeneous composition distribution and accounts for finite-length effects. Based on formally derived probability distri-

butions, a statistical significance ( $p$ ) was assigned to any observed difference from expectation, i.e. the probability that a difference equal to or larger than that observed could occur by chance if the residues were actually randomly distributed. The TMSTAT method is presented in detail in Appendix and Figure 9.

### GGxxG (GG4) is the most significant pair

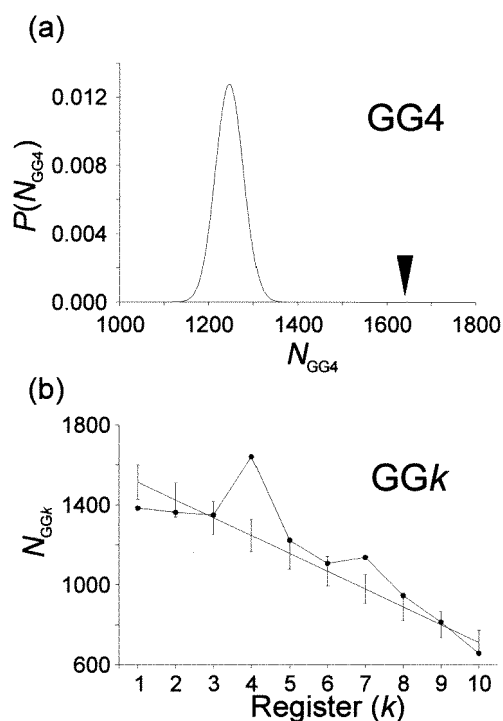
In analyzing 4000 randomly distributed variables, one would expect to observe by chance one instance of a difference from expectation with significance  $p < 0.00025$  ( $1/4000$ ); in the present analysis, 117 pairs deviate from the expected occurrences with at least that significance. The data for the most significant outliers found are shown in Table 2A (over-represented pairs) and Table 2B (under-represented pairs). At least one of the ten most significant over-represented pairs is found in 76% of the sequences in the database. (A table containing observed and expected occurrences of all 4000 pairs is available at <http://engelman.csb.yale.edu/tmstat> and <http://bioinfo.mbb.yale.edu/tmstat>).

GG4 (GxxxG) is the pair with the strongest positive correlation. In the database, GG4 occurs 1641 times, 32% more than its expectation of 1247 (odds ratio observed/expected = 1.32). The occurrence of the GG4 pair is compared with its expectation curve in Figure 3(a). The probability of observing an equal or greater difference from expectation by chance is unrealistic ( $p = 6.4 \times 10^{-34}$ ). The high frequency of GG4 in a database of predicted TM sequences was reported by Arkin & Brunger (1998); however, the expectation and significance for the occurrences of the pair were not calculated.

The observed occurrences of all Gly-Gly pairs as a function of separations and their expectation values are plotted in Figure 3(b). In the profile, the pair GG4 peaks between two positions that are relatively "unbiased." A significant negative bias is observed for the interaction of two adjacent Gly residues (GG1, 9% below expectation,  $p = 1.0 \times 10^{-4}$ ). A second positive, corresponding to GG7, is observed 16% more often than its expected occurrences ( $p = 2.9 \times 10^{-8}$ ). Clearly, the observed correlations must derive from specific position-dependent selection of residue properties.

### Pairs containing isoleucine, glycine and valine are the most biased

The 30 most significant over-represented pairs, shown in Table 2A, frequently contain isoleucine, glycine and valine residues. Ile is present in 14 pairs (five of the top eight cases). Gly exists in 11 pairs (five of the top six cases). Val is found in eight pairs (four of the top ten cases). Leu, the most common residue in transmembrane domains, is found in only five pairs and never in the top ten cases.



**Figure 3.** Occurrences of Gly-Gly pairs and their expectation. (a) Probability function  $P(N_{GG4})$  associated with any possible number of occurrences ( $N_{GG4}$ ) of the pair GG4 (GxxxG) in the database of TM sequences. The arrow marks the actual occurrences observed (observed 1641; expected 1246.8; statistical significance  $p = 6.4 \times 10^{-34}$ ). (b) Observed occurrences  $N_{GGk}$  of all GGk pairs (●) as a function of distance  $i, i+k$ . The straight line represents the expectation. The line is sloped due to end effects: in a sequence with finite length more pairs are possible at short register than at longer register. The error bars define the 99% interval of confidence around the expectations (the range outside which a value has significance  $p < 0.01$ ).

### Pair correlation results reflect helical periodicity

The pie diagram shown in Figure 4(a) depicts the proportion of the most significant pairs of Table 2A grouped by register. Pairs with registers 1 to 4 comprise 90% of the cases. Together, registers 1 and 4 are found in almost two-thirds of the cases. A stronger tendency of pairs at  $i, i+1$  and  $i, i+4$  to deviate from expectation is found in the entire set of 4000 pairs. The tendency is evidenced by the  $\chi^2$  scores of pairs as a function of register, shown in Figure 4(b), that clearly peak at these two positions. These results are strongly consistent with helical geometry. Four residues comprise about one helical turn in regular  $\alpha$ -helix conformation (3.6 residues per turn); at  $i, i+1$  and  $i, i+4$ , both residues of the pair are presented on the same face of a regular  $\alpha$ -helix, as schematized in the wheel diagram of Figure 4(c).

**Table 2.** The 30 most significant over-represented and under-represented pairs sorted by significance

Pair	Occurrences <sup>a</sup>	Expectation <sup>b</sup>	Standard deviation <sup>c</sup>	Significance ( $p$ ) <sup>d</sup>	Odds ratio <sup>e</sup>
<i>A. Over-represented pairs</i>					
GG4	1641	1246.8	31.3	$6.4 \cdot 10^{-34}$	1.32
II4	3782	3289.2	48.2	$8.4 \cdot 10^{-24}$	1.15
GA4	2057	1698.4	37.0	$3.6 \cdot 10^{-21}$	1.21
IG1	2721	2318.4	41.9	$4.8 \cdot 10^{-21}$	1.17
IG2	2528	2182.1	41.1	$1.3 \cdot 10^{-16}$	1.16
VG2	2268	1945.2	39.1	$5.7 \cdot 10^{-16}$	1.17
IV4	3003	2636.3	45.5	$2.1 \cdot 10^{-15}$	1.14
IP1	992	788.8	25.2	$4.5 \cdot 10^{-15}$	1.26
VV4	2770	2443.2	42.5	$3.8 \cdot 10^{-14}$	1.13
VI4	2965	2636.3	45.5	$1.1 \cdot 10^{-12}$	1.12
AV1	3149	2823.2	45.8	$2.2 \cdot 10^{-12}$	1.12
GL3	3392	3062.7	47.7	$9.7 \cdot 10^{-12}$	1.11
AG4	1929	1698.4	37.0	$9.1 \cdot 10^{-10}$	1.14
WQ1	88	45.8	6.5	$3.9 \cdot 10^{-9}$	1.92
IL4	4784	4446.3	57.3	$4.9 \cdot 10^{-9}$	1.08
AA3	2719	2477.0	42.2	$1.3 \cdot 10^{-8}$	1.10
VG1	2295	2066.7	40.0	$1.8 \cdot 10^{-8}$	1.11
GG7	1138	979.6	28.0	$2.9 \cdot 10^{-8}$	1.16
VL4	4362	4064.3	54.9	$7.7 \cdot 10^{-8}$	1.07
IS2	1916	1717.0	36.7	$9.0 \cdot 10^{-8}$	1.12
SI2	1912	1717.0	36.7	$1.5 \cdot 10^{-7}$	1.11
GI1	2536	2318.4	41.9	$2.9 \cdot 10^{-7}$	1.09
IY10	496	397.0	19.0	$4.5 \cdot 10^{-7}$	1.25
YY3	245	180.9	12.4	$6.3 \cdot 10^{-7}$	1.35
IF10	1617	1443.1	35.7	$1.6 \cdot 10^{-6}$	1.12
GI2	2375	2182.1	41.1	$3.3 \cdot 10^{-6}$	1.09
PI3	809	696.0	24.0	$4.0 \cdot 10^{-6}$	1.16
PV1	777	667.8	23.4	$5.0 \cdot 10^{-6}$	1.16
PL1	1342	1203.8	30.0	$5.4 \cdot 10^{-6}$	1.11
LP1	1342	1203.8	30.0	$5.4 \cdot 10^{-6}$	1.11
<i>B. Under-represented pairs</i>					
II2	3223	3759.1	50.6	$5.1 \cdot 10^{-27}$	0.86
GI4	1564	1909.3	39.1	$1.4 \cdot 10^{-19}$	0.82
IL1	4906	5399.1	60.7	$2.5 \cdot 10^{-16}$	0.91
FL1	3954	4394.8	55.4	$9.4 \cdot 10^{-16}$	0.90
FI4	2182	2525.4	44.5	$4.1 \cdot 10^{-15}$	0.86
IG4	1620	1909.3	39.1	$4.8 \cdot 10^{-14}$	0.85
LW4	611	786.7	25.0	$5.2 \cdot 10^{-13}$	0.78
IV2	2683	3013.0	47.6	$2.3 \cdot 10^{-12}$	0.89
YL4	788	974.5	27.9	$7.3 \cdot 10^{-12}$	0.81
PG1	311	434.2	19.3	$2.8 \cdot 10^{-11}$	0.72
CP1	56	113.1	10.1	$9.0 \cdot 10^{-10}$	0.50
FV3	1991	2244.3	42.1	$1.1 \cdot 10^{-9}$	0.89
AP1	508	642.2	22.9	$1.8 \cdot 10^{-9}$	0.79
IW4	376	493.2	20.4	$2.9 \cdot 10^{-9}$	0.76
IM1	922	1091.4	29.5	$4.7 \cdot 10^{-9}$	0.84
FL3	3575	3877.7	53.3	$1.1 \cdot 10^{-8}$	0.92
FV4	1869	2094.6	41.0	$2.5 \cdot 10^{-8}$	0.89
FI3	2462	2705.7	45.6	$6.7 \cdot 10^{-8}$	0.91
LW3	707	842.8	25.7	$7.5 \cdot 10^{-8}$	0.84
V12	2759	3013.0	47.6	$7.7 \cdot 10^{-8}$	0.92
GP1	335	434.2	19.3	$1.2 \cdot 10^{-7}$	0.77
YI4	575	694.8	24.0	$3.7 \cdot 10^{-7}$	0.83
FL2	3862	4136.3	54.4	$3.9 \cdot 10^{-7}$	0.93
VG4	1517	1702.0	37.2	$5.0 \cdot 10^{-7}$	0.89
FF2	2244	2450.3	41.9	$7.1 \cdot 10^{-7}$	0.92
FM1	743	872.3	26.7	$8.4 \cdot 10^{-7}$	0.85
FL6	2861	3102.2	49.4	$88 \cdot 10^{-7}$	0.92
II1	3744	3994.0	51.6	$1.1 \cdot 10^{-6}$	0.94
WV1	454	549.1	21.3	$5.1 \cdot 10^{-6}$	0.83
LL2	7509	7821.3	69.4	$6.5 \cdot 10^{-6}$	0.96

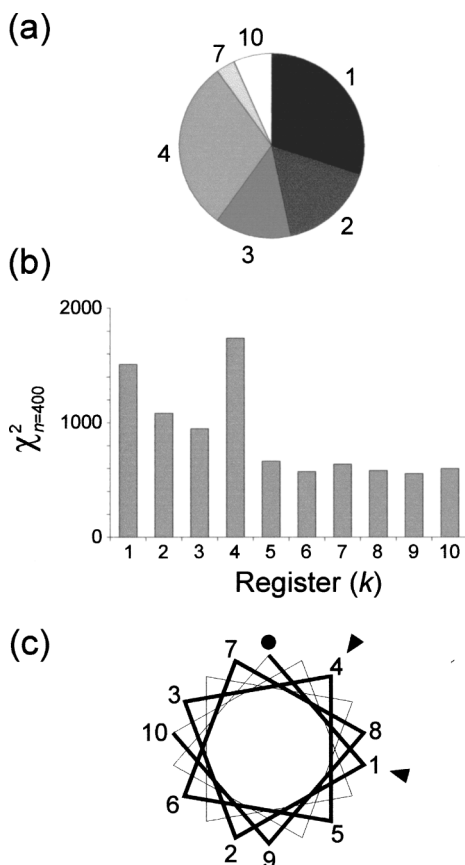
<sup>a</sup> Number of observed occurrences of the pair in the database.

<sup>b</sup> Average expected number of occurrences.

<sup>c</sup> Standard deviation of the expectation distribution.

<sup>d</sup> Calculated as two-tailed integral of the expectation distribution.

<sup>e</sup> Occurrences/Expectation ratio.



**Figure 4.** (a) Relative frequency of the 30 most significant pairs of Table 2A grouped by register. (b) Overall deviation from expectation at different registers, calculated as  $\chi^2$  score on the entire set of pairs. Pairs grouped by register (group size  $n = 400$ ):

$$\chi^2 = \sum_{\text{pairs}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

(c) Relative angular position along the helical axis of residues in pairs at different registers. The filled circle at the top of the helical wheel diagram (3.6 residues per turn) indicates the residue at  $i$  (●). The position of the residue at  $i + k$  is indicated by the respective number  $k$ . The arrows mark the registers with the highest overall tendency to diverge from expectation, as observed in (a) and (b).

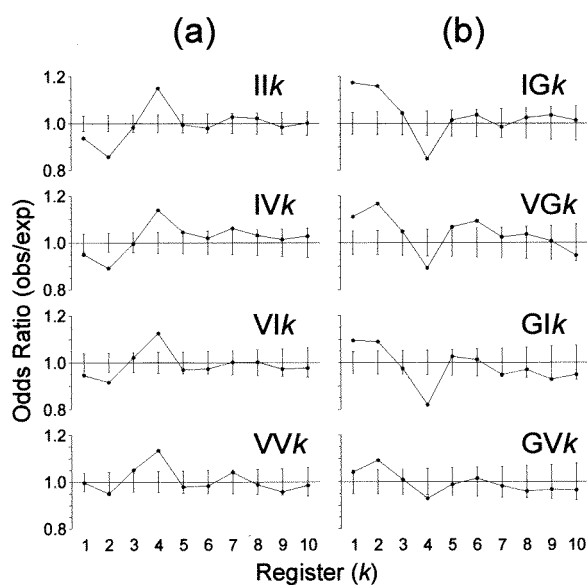
### Similar biases are found with pairs of residues with similar structure

A remarkable feature of the results shown in Table 2 is that most of the pairs can be grouped by register and side-chain chemistry into a few categories. For example, GG4, GA4 and AG4 are all observed among them. Similarly, all combinations of the  $\beta$ -branched aliphatic residues at  $i$ ,  $i + 4$  (II4, IV4, VI4 and VV4) are extremely significant. The pairs IL4 and VL4 are also among the most significant pairs (Leu is isomeric to Ile but  $\gamma$ -branched). There are many pairs formed

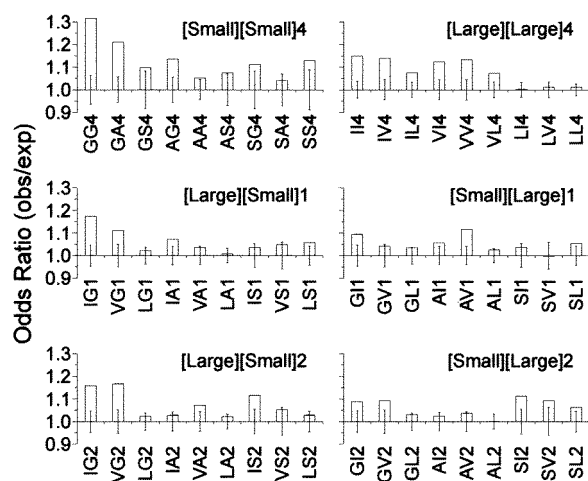
by one small residue (Ala, Gly and Ser) and a  $\beta$ -branched aliphatic residue at  $i$ ,  $i + 1$  (IG1 and VG1; GI1 and AVI) and  $i$ ,  $i + 2$  (IG2, VG2 and IS2; SI2 and GI2). Finally, a number of pairs are formed by Pro and large aliphatic residues (Ile, Val and Leu) at register 1 (IP1 and LP1; PV1 and PL1). In the list of the significant under-represented pairs, combinations of  $\beta$ -branched residues and glycine are very disfavored at  $i$ ,  $i + 4$  (GI4, IG4 and VG4) and neighboring Pro and Gly are also disfavored (PG1 and GP1).

The correspondence between the observed biases and side-chain chemistry also appears in the comparison of pairs profiles at all registers. In Figure 5, the occurrences of pairs normalized to their expectancy (odds ratios, observed/expected) are plotted as a function of register. In Figure 5(a), all pairs formed by combinations of Ile and Val have very similar profiles with a strong positive correlation at  $i$ ,  $i + 4$  and a negative peak at  $i + 2$ . Striking similarity is also evident in Figure 5(b), where the profiles of pairs formed by Gly and Ile or Val are shown.

These results suggest a general tendency for two large aliphatic residues (in particular the  $\beta$ -branched ones) to correlate at  $i$ ,  $i + 4$  when they are on the same face of the helix and to anti-correlate when they are on opposite faces. Pairs of smaller residues (in particular Gly) on the same face of the helix are also favored. Lastly, pairs formed by one small and one large residue correlate positively on adjacent ( $i$ ,  $i + 1$ ) or opposite faces ( $i + 2$ ) and, conversely, are strongly disfavored on the same face ( $i + 4$ ).



**Figure 5.** Normalized occurrences of pairs formed by combinations of the  $\beta$ -branched residues Ile, Val and Gly at all registers. Odds ratio = (observed occurrences/expected). The bars represent the 99% confidence interval around the expectation. (a) Pairs formed by Ile and Val. (b) Pairs formed by Gly with Ile or Val.



**Figure 6.** Odds ratios of pairs formed by similar residues. [Small], small residues, Gly, Ala and Ser; [Large], large aliphatic residues, Ile, Val and Leu. The error bars mark the 99% confidence interval around the expectation.

These general themes can be appreciated in Figure 6, where sets of pairs at the same register are grouped by side-chain size and compared. All pairs formed by two small residues (Gly, Ala and Ser) at register 4 are positively biased with a significance of at least  $p < 0.01$  (except the case of AS4). The  $\beta$ -branched residues Ile and Val correlate very strongly at register 4 (II4, IV4, VI4, VV4). Interestingly, Leu seems to be part of the trend of positively correlating [Large][Large] pairs at  $i, i + 4$  only when it is occupying the C-terminal position (IL4, VL4); all pairs in which Leu precedes a second large residue (LI4, LV4, LL4) are unbiased. The majority of the combinations of large and small residues at registers 1 and 2 have a positive bias with a significance of at least  $p < 0.05$ . Not all deviations from expectation are large or very significant. However, the observed trends can be taken with more confidence than the individual deviations, as it is less probable for a series of random deviations to occur all in the same direction.

### Analysis of triplets shows that residue correlations extend beyond the pair level

The relationships between pairs of larger and smaller residues suggested that the correlations were not limited to the pairs, since positively correlating pairs can be consistently combined to form higher-order patterns. This was confirmed by extending the analysis to triplets. The occurrences of 200,000 amino acid triplets were counted and compared to an expectation computed with the same method used for the pairs. The reference was therefore calculated on the

frequency of the single residue in the sequences and not relative to the pairs. The 30 triplets with the strongest positive correlation are listed in Table 3. The most significant triplets were indeed composed of combinations of strongly biased pairs. For example, the most significant case IG1L3 (IGxxL) is composed of IG1, IL4 and GL3, all observed in the 15 most biased pairs. The significance of IGxxL ( $p = 1.8 \times 10^{-20}$ ) is slightly lower than that of IG1 but higher than those of IL4 and GL3. However, it is incorrect to compare the  $p$  values, since, on average, the triplets have a smaller number of occurrences than the pairs, and  $p$  values strongly depend on "sample size" (for example, when a coin is tossed once, 100% "heads" is not a significant result, but in one million tosses 51% heads undoubtedly indicates a defective coin). A more appropriate value for comparison is the odds ratio (observed/expected occurrences). In the most significant triplets, the observed odds ratios always exceed those of the corresponding pairs.

Triplets containing the pair GG4 are present many times in Table 3, mostly in conjunction with Ile, Val or Leu at registers  $\pm 1$  and  $\pm 2$  with respect to the Gly residues. The interactions of the GG4 pair with Ile and Val at these distances is evident in Figure 7, which illustrates the effect of a third residue at positions relative to the GG4 pair. In addition, many strongly correlating triplets in Table 3 contain two large aliphatic residues interacting with one Gly or another small residue at position  $\pm 1$  and  $\pm 2$  (IG1L3, IG2I2, VG2I2, IG1I3, IS2I2, IA3V1, etc.). Together, these correlations define the main theme of the analysis, i.e. patterns of larger and smaller residues that are strongly favored to coexist at neighboring helical faces.

## Discussion

Many of the amino acid correlations that were found in the present analysis are readily interpretable in terms of helix-helix interaction patterns. Most of the positively correlating pairs occur at separations  $i, i + 1$  and  $i + 4$ , i.e. on the same face in  $\alpha$ -helical conformation. At register  $i, i + 4$  there is a marked preference for pairs of residues with similar size, while combinations of a small and a large residue are strongly disfavored. Furthermore, the GG4 pair and its relationship with  $\beta$ -branched residues at  $i \pm 1$  relative to the glycine residues has been observed in two important membrane oligomerizing systems: in the interface of glycoporphin A (GpA) transmembrane dimer (Lemmon *et al.*, 1994; MacKenzie *et al.*, 1997), and by an *in vivo* selection system for transmembrane helix-helix association (Russ & Engelman, 2000). The other strong correlation of GG4 with  $\beta$ -branched residues at  $i \pm 2$  is more difficult to explain in terms of helix-helix interaction, because these patterns in an  $\alpha$ -helical conformation would

**Table 3.** The 30 most significant over-represented triplets sorted by significance

Triplet	Occurrences <sup>a</sup>	Expectation <sup>b</sup>	Standard deviation <sup>c</sup>	Significance ( $p$ ) <sup>d</sup>	Odds ratio <sup>e</sup>
IG1L3	535	353.4	18.3	$1.8 \cdot 10^{-20}$	1.51
IG2I2	399	258.1	15.6	$3.8 \cdot 10^{-18}$	1.55
IG2G4	244	137.6	11.5	$6.1 \cdot 10^{-18}$	1.77
IG1A4	309	191.7	13.6	$1.6 \cdot 10^{-15}$	1.61
GV2G2	244	143.3	11.7	$3.6 \cdot 10^{-15}$	1.70
VG2I2	331	211.0	14.3	$6.7 \cdot 10^{-15}$	1.57
IG1I3	382	258.1	15.6	$7.4 \cdot 10^{-14}$	1.48
GG4G4	146	75.9	8.8	$1.4 \cdot 10^{-13}$	1.92
IV4L4	488	348.1	18.3	$5.1 \cdot 10^{-13}$	1.40
IP1I3	162	88.7	9.2	$5.9 \cdot 10^{-13}$	1.83
IS2I2	319	211.4	14.1	$1.1 \cdot 10^{-12}$	1.51
GI2G2	255	160.6	12.4	$1.6 \cdot 10^{-12}$	1.59
IG1G4	236	149.1	11.9	$4.7 \cdot 10^{-12}$	1.58
IA3V1	388	274.0	16.2	$7.7 \cdot 10^{-12}$	1.42
IG2L2	485	353.4	18.4	$1.1 \cdot 10^{-11}$	1.37
VG2G4	201	122.9	10.9	$2.6 \cdot 10^{-11}$	1.64
II4L4	555	419.6	19.7	$2.7 \cdot 10^{-11}$	1.32
VV4G2	257	169.5	12.7	$1.2 \cdot 10^{-10}$	1.52
PI3G2	90	43.1	6.5	$2.2 \cdot 10^{-10}$	2.09
VG5L3	334	234.8	15.1	$4.4 \cdot 10^{-10}$	1.42
IA2I2	428	316.7	17.2	$7.1 \cdot 10^{-10}$	1.35
GG4I2	213	137.6	11.5	$9.6 \cdot 10^{-10}$	1.55
AC3A4	71	32.3	5.6	$1.6 \cdot 10^{-9}$	2.20
VG2L3	413	305.3	17.1	$1.7 \cdot 10^{-9}$	1.35
VG1G4	206	133.1	11.3	$1.9 \cdot 10^{-9}$	1.55
IG2L3	439	328.2	17.7	$2.2 \cdot 10^{-9}$	1.34
GG4L3	274	189.6	13.5	$3.3 \cdot 10^{-9}$	1.45
VG2L2	438	328.8	17.7	$4.0 \cdot 10^{-9}$	1.33
IV4V4	298	210.7	14.1	$4.5 \cdot 10^{-9}$	1.41
GL3G1	334	241.3	15.1	$4.9 \cdot 10^{-9}$	1.38

<sup>a</sup> Number of observed occurrences of the triplet in the database.

<sup>b</sup> Average expected number of occurrences.

<sup>c</sup> Standard deviation of the expectation distribution curve.

<sup>d</sup> Calculated as two-tailed integral of the expectation distribution.

<sup>e</sup> Occurrences/Expectation ratio.

place the residues on opposite sides of the helix. We propose a possible explanation for this pattern in terms of helix flexibility modulation.

### Comparison with GpA transmembrane dimer

GG4 is the key feature of the dimerization interface of glycophorin A, the best characterized transmembrane helix-helix interaction. The single TMD of GpA forms a symmetric right-handed homodimer based on the seven residue motif LIxxGVxxGVxxT (Lemmon *et al.*, 1992, 1994). The glycine residues allow the backbones to reach close proximity and the larger side-chains pack in a "ridges into grooves" fashion (MacKenzie *et al.*, 1997).

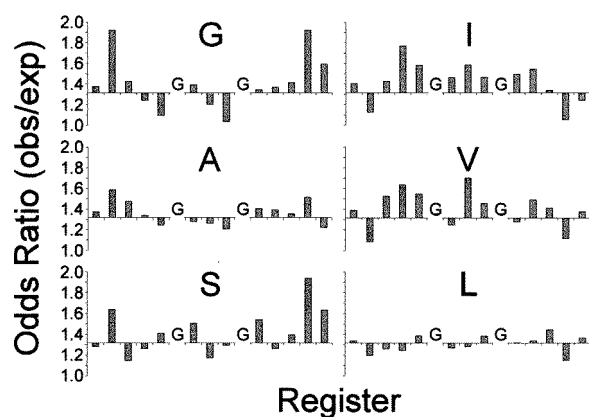
Many other features of the GpA interaction motif are found among the most significant results of the present analysis: IV4 and VV4, for instance, are two of the most strongly correlating amino acid pairs. In addition, the majority of the amino acid triplets of the motif correlate positively in this analysis, as shown in Table 4.

### Comparison with the TOXCAT *in vivo* selection system for helix-helix interaction

The GG4 pair is almost invariably present in the transmembrane oligomerization motifs identified from randomized sequences by the TOXCAT *in vivo* selection system, presented in the accompanying paper (Russ & Engelman, 2000). Seven positions with the periodicity of the GpA motif were randomized to a set of possibilities at each motif position in the context of a poly-leucine (Leu library) or poly-alanine (Ala library) background. The results (refer to Figure 4 in the accompanying paper) often contained the theme of large residues (Ile, Val or Leu) associated with the GG4 pair at positions  $\pm 1$ , in excellent agreement with the present statistical analysis.

In the TOXCAT library with a Leu context, the larger residues occurred at positions  $i + 1$  relative to the two glycine residues (G[IVL]xxG[IVL]). The  $\beta$ -branched residues were prevalent, especially in the first position. In addition, Thr was often found in the selection system at position  $i + 4$  from the second glycine residue. In the present statistical analysis, we find that the GG4T4 triplet, which is observed also in the GpA motif, is strongly over-





**Figure 7.** Triplet analysis: interaction of a third residue with the GG4 pair. The Figure represents the odds ratios of triplets containing the pair GG4 in conjunction with either a small residue (Gly, Ala and Ser, left panels) or a large aliphatic residue (Ile, Val and Leu, right panels). The position of the bars along the  $x$ -axis reflects the actual position of the residue relative to the pair GG4. The baseline is set at 1.316, the odds ratio observed for the pair GG4.

represented (+58%,  $p = 3.4 \times 10^{-4}$ ). In the Ala library, the two large residues occurred at position  $i - 1$ , on the N-terminal side of the GG4 pair ([IVL]Gxx[IVL]G).  $\beta$ -Branched residues were again prevalent. A schematic comparison of our results with the TOXCAT selection can also be found in Table 2 of Russ & Engelman (2000).

The convergence of the results obtained with such dissimilar approaches is remarkable,

especially if one considers that the TOXCAT system reports the oligomerization events of bitopic (single-span) transmembrane domains, while the correlation analysis is based mostly on polytopic (multi-span) proteins (Table 1). The frequent finding of GG4 with large flanking residues by TOXCAT, which selects for strong transmembrane interactions, probably reflects the excellent opportunity provided by the deep groove and ridge of the motif for bringing two helices in extensive contact, as observed in the GpA structure. If strong interactions are important in polytopic proteins, they are essential in oligomerizing helices, as more energy is required to compensate for the larger entropy cost of association of helices that are not covalently joined by extra-membranous loops.

Following this line of reasoning, one could expect the GG4 pair to be more frequent in the TMDs of single-span transmembrane proteins. To address this question, we analyzed bitopic and polytopic sequences separately (data not shown). In a raw count, the pair is indeed found more frequently in bitopic sequences (on average, in 12.5% of bitopic domains and in 12.1% of polytopic transmembrane domains). The GG4 pair is the most significant outlier in both databases, but it is more over-represented relative to its expected occurrences in the bitopic (+37.8%) than in the polytopic set (+30.8%). However, caution should be exercised when inferring the relative importance of the motif in the two different topologies from these results. In polytopic proteins, weak helix-helix interactions embedded in a bundle might be tolerable and extra-membranous loops might sometimes direct the folds. On the other hand, the fraction of transmembrane anchors in the single-span database that are not engaged in interactions

**Table 4.** Results of triplet analysis for all triplets present in the dimerization motif of glycoporphin A (LlxxGVxxGVxxT), sorted by decreasing odds ratio

Triplet	Significance ( $p$ )	Odds ratio
GG4T4	$3.4 \cdot 10^{-4}$	1.58
IG3G4	$1.8 \cdot 10^{-6}$	1.43
IV4V4	$4.5 \cdot 10^{-9}$	1.41
GG4V1	$1.6 \cdot 10^{-3}$	1.28
GV1G3	$3.1 \cdot 10^{-3}$	1.25
LG4G4	$9.6 \cdot 10^{-3}$	1.20
LV5V4	$2.8 \cdot 10^{-3}$	1.17
IG3V1	$1.9 \cdot 10^{-2}$	1.16
IV4G3	$4.9 \cdot 10^{-2}$	1.15
IG3V5	$5.7 \cdot 10^{-2}$	1.15
VG3V1	$3.4 \cdot 10^{-2}$	1.15
LG4V1	$6.5 \cdot 10^{-2}$	1.10
GV1V4	$2.1 \cdot 10^{-1}$	1.09
LI1G3	$2.6 \cdot 10^{-1}$	1.06
LI1V4	$2.8 \cdot 10^{-1}$	1.05
VV4T3	$8.8 \cdot 10^{-1}$	1.01
VG3T4	$1.0 \cdot 10^0$	1.00
LG4V5	$7.0 \cdot 10^{-1}$	0.97
GV5T3	$8.5 \cdot 10^{-1}$	0.96
LV5G3	$6.0 \cdot 10^{-1}$	0.96
GV1T3	$4.1 \cdot 10^{-1}$	0.90

is also unknown. "Passive" sequences with nearly randomly distributed residues would provide only an increase in the background noise and a decrease in the significance of the results. Thus, the only conclusion supported by the data is that the GG4 pair is very important in both bitopic and polytopic membrane proteins.

### **$\beta$ -Branched residues could minimize entropy loss upon packing**

Upon solution of the NMR structure of the GpA transmembrane dimer, MacKenzie *et al.* (1997) proposed that the association of the monomers might occur between two largely preformed interfaces. The idea was based on a fundamental implication of the two-stage model for membrane protein folding (Popot & Engelman, 1990). The first stage of the model involves the partitioning of largely hydrophobic TM segments in the lipid bilayer, which is strongly favored by the hydrophobic effect. The backbone adopts a helical conformation to satisfy its strong hydrogen bonding potential in the low-dielectric environment. Sequence specificity comes into play only in stage 2, when the equilibrium of associations of the preformed helices is established. Given the two-stage model, it is possible to have a notion of the structure of the unassociated state (helical) that is generally not available with the unfolded state of soluble proteins. This information is crucial to relating observed structural features of the native state to the energetics of folding, since stability depends on the differential between the energies of folded and unfolded states.

In the GpA dimer, many interfacial side-chains (Ile, Val, and Thr) have only one populated rotamer as a consequence of being in a helix (Dunbrack & Karplus, 1993; Schrauber *et al.*, 1993). Under the assumption that the GpA TM is helical in the monomeric state (recently confirmed experimentally by Fisher *et al.*, 1999), MacKenzie and colleagues pointed out that minimal loss of rotameric freedom upon dimerization was therefore expected. Later, a theoretical model based on a large number of GpA mutants indicated loss of side-chain entropy as one of the major factors destabilizing dimerization (MacKenzie & Engelman, 1998), supporting further the hypothesis that rotamerically constrained interfaces could provide a significant contribution to the stability of association.

In our results, there is a significant dichotomy in the role of the three larger aliphatic residues Ile, Val and Leu. The  $\beta$ -branched Ile and Val are, with Gly, the residues involved in the strongest correlations. Conversely, Leu, the most frequent residue in transmembrane domains, has only a secondary role. As a  $\gamma$ -branched side-chain, Leu can sample more conformations in helical secondary structure. Our results are therefore consistent with the hypothesized importance of a "preformed interface" and the possibility that the use of residues

with constrained side-chains in helical conformation might have general significance in limiting the entropic cost of association in a large set of membrane proteins.

### **Interaction of $\beta$ -branched residues at $i, i+4$ might modulate helix flexibility in TMs**

A combination of theoretical arguments and experimental evidence suggests the hypothesis that pairs of Ile and Val at  $i, i+4$ , which we find all strongly over-represented in this analysis, might influence flexibility in TM helices. Helical conformation prevents the  $\chi_1$  dihedral from positioning a heavy  $\gamma$ -substitute in *gauche*<sup>-</sup> orientation due to the steric clashes with the backbone carbonyl oxygen atom at  $i-3$  (McGregor *et al.*, 1987). In an analysis of intrahelical side-chain/side-chain interactions in soluble proteins, Walther & Argos (1996) reported that the majority of the contacts occurred between pairs of residues with spacing  $i, i+4$ . As they pointed out, interactions can occur at this separation, since they are promoted by  $\chi_1$  rotamers that involve a combination of a *trans* (at  $i$  position) and a *gauche*<sup>+</sup> (at  $i+4$ ) dihedral. Conversely,  $i, i+1$  and  $i, i+3$  interactions require the unfavorable *g*<sup>-</sup> conformation (*g*<sup>-</sup>/*g*<sup>+</sup> and *t/g*<sup>-</sup>, respectively). The two C $\gamma$  atoms of  $\beta$ -branched residues are forced to occupy simultaneously *g*<sup>+</sup> and *t* positions to avoid the *g*<sup>-</sup> dihedral (Schrauber *et al.*, 1993). For this reason,  $\beta$ -branched residues are good candidates for intrahelical interactions at  $i, i+4$ . This is consistent with the high scores of Ile and Val in the  $i, i+4$  contact propensity calculated by Walther & Argos (1996), a scale in which Leu scored only slightly above average.

Padmanabhan & Baldwin (1994) used circular dichroism (CD) to measure the interactions of L[IVL] and [IVL]L pairs at  $i, i+3$  and  $i, i+4$  in soluble peptides, and observed stronger helix stabilization in  $i+4$  pairs. The energy of interaction of pairs of hydrophobic residues at different registers in an  $\alpha$ -helix has been calculated by Creamer & Rose (1995) using an exhaustive Boltzmann-weighted conformational search. The interactions of pairs formed by Ile, Val and Leu at  $i, i+4$  were more stabilizing than those of  $i, i+3$  pairs. The energy ranking observed for these pairs at  $i, i+4$  agrees with our data (summarized in the [Large][Large]4 panel in Figure 6) in the fact that the smallest effects are observed when there is a Leu residue on the N-terminal side in the pair (LL4, LI4, LV4). The calculations made by Creamer & Rose (1995) were in only partial agreement with the experimental results reported by Padmanabhan & Baldwin (1994), who, conversely, observed higher helix content in L[IVL]4 than in [IVL]L4 pairs. However, Creamer & Rose (1995) calculated the interaction energies relative to the same pair at  $i, i+2$  (on opposite faces in helical conformation) while the CD data reflects the position of a helix-coil/strand equilibrium.

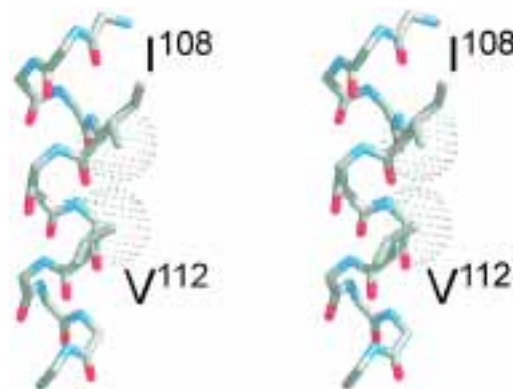
These three studies relate intrahelical side-chain interactions to helical stability in aqueous solution and they concur on the importance of  $i, i + 4$  contacts. In the membrane, the helix is already stabilized by the environment, but side-chain interactions might additionally affect the flexibility of the helix. This might be especially true for pairs of  $\beta$ -branched residues at  $i, i + 4$ , as their only favorable  $\chi_1$  rotamer conformation locks them in close proximity. In bacteriorhodopsin, the only helical membrane protein structure available at better than 2 Å resolution (Luecke *et al.*, 1999), the four [IV][IV]4 pairs found in regular  $\alpha$ -helical conformation have an average minimal distance (center to center of the closest  $C^\gamma$  or  $C^\delta$  atoms) of only  $4.2(\pm 0.3)$  Å ( $\pm$ SD). An example (residues I108 and V112 on the fourth transmembrane segment) is shown in Figure 8. Whether the strongly correlating pairs of  $\beta$ -branched residues at  $i, i + 4$  are important to diminishing transmembrane helix flexibility is an interesting question. If validated experimentally, it could provide further support to the hypothesis that a reduction of entropy in the helical unassociated state (in turn a destabilization of the unfolded state, if independent helices are stable in the bilayer) could be a significant factor in the transmembrane association equilibrium.

On the other hand, glycine is frequently observed in membrane helices and induces flexibility. Glycine is compatible with helical conformation in membrane proteins, as evident in GpA, which is largely helical in both the monomeric and dimeric states despite three glycine residues in its TM sequence (Fisher *et al.*, 1999). However, extensive studies in host peptides by Deber and colleagues have shown that, while Gly has a considerable tendency to form  $\alpha$ -helices in membrane mimetic environments, it is somewhat destabilizing compared to the more hydrophobic side chains (Li & Deber, 1992a,b; Liu & Deber, 1998). This is consistent with the observation by Ri *et al.* (1999) using a Monte Carlo simulation of a single TM. The ranking observed for increased flexibility (Gly > Ala > Val) correlated well with the severity of voltage-dependent gating phenotypes when these three residues were substituted for the wild-type Pro residue in connexin32.

Thus, a pair of  $\beta$ -branched residues  $i, i + 4$  and a pair of glycine residues at  $i, i + 4$  might lie at opposite sides of a hypothetical flexibility scale in TM helices. The favorable role of Gly in helix interactions might require the presence of additional stability from the  $\beta$ -branched residues. This argument provides a speculative but plausible explanation for the strong correlations between the GG4 pair and [IV][IV]4 pairs observed in opposite faces of the helix at  $i + 2$ , which could perhaps have a compensatory role in modulating helix flexibility.

### Final remarks

Many instances of the "GG4 +  $\beta$ -branched" motif and its variations can be found in the avail-



**Figure 8.** Example of a pair of  $\beta$ -branched aliphatic residues at  $i, i + 4$  in the fourth transmembrane segment of bacteriorhodopsin (RSCB PDB code *1c3w*). Both I108 and V112 are in their standard helical rotamer in which the  $\gamma$ -carbon atoms are positioned away from the disfavored *gauche*<sup>-</sup> orientation. According to the IUPAC nomenclature rules, the rotamers are designated respectively as *trans* and *gauche*<sup>+</sup>. The van der Waal sphere of the carbon atoms of closest approach is represented by dots (1.9 Å). The center-to-center distance between I108- $C^{\gamma 2}$  and V112- $C^{\gamma 2}$  is 4.1 Å.

able X-ray structures of helical transmembrane proteins. An in-depth comparison of the results of our analysis with the structural models has not been completed at this stage. This comparison could offer further insights into the physical role of this motif and of other observed correlations. For example, it would be interesting to put the strong association of Ile, Val and Leu with neighboring Pro residues in relation to the geometry of the kink.

We have shown that the inherent simplicity of helical membrane proteins structure results in correlations between residues that are detectable with simple statistical methods and that suggest interpretations in terms of protein chemistry. In turn, our results also support the validity of TMD prediction techniques. With the growth of primary data provided by the genome projects, these results are an indication of the important role that sequence analysis will assume in the near future in the membrane protein field as a complement to the interpretation of experimental and structural data.

## Methods

### Database

The source of transmembrane sequences for this work was the annotated database Swiss-Prot, release 37 and updates to March 17, 1999 (Bairoch & Apweiler, 1999). All sequence fragments corresponding to a TRANSMEM annotation in the FT field were extracted and a database of 46,946 transmembrane domains was compiled (Table 1).

Phase One: sequence-level distributions,  $P(N_{ALk} | I, k, N_A, N_L)$ 

Example of calculation with the following parameters

$I$  sequence length = 5  
 $k$  register = 3  
 $N_A$  Ala residues = 2  
 $N_L$  Leu residues = 2

$N_{AL3}$	Permutations	$P(N_{AL3} 5,3,2,2)$
0	19	0.63
1	10	0.33
2	1	0.03
TOTAL	30	1.00

Permutation	$N_{AL3}$	Permutation	$N_{AL3}$	Permutation	$N_{AL3}$
1	AALL-	1	11	AL-LA	1
2	AAL-L	1	12	A-LLA	1
3	AA-LL	2	13	LAAL-	0
4	ALAL-	1	14	LAA-L	1
5	ALA-L	0	15	-AALL	1
6	A-ALL	1	16	LALA-	0
7	ALLA-	0	17	LA-AL	1
8	AL-AL	0	18	-ALAL	1
9	A-LAL	0	19	LAL-A	0
10	ALL-A	0	20	LA-LA	0
			21	-ALLA	0
			22	LLAA-	0
			23	L-AAL	0
			24	-LAAL	0
			25	LLA-A	0
			26	L-ALA	0
			27	-LALA	0
			28	LL-AA	0
			29	L-LAA	0
			30	-LLAA	0

Phase Two: database-level distributions,  $P_{DB}(N_{ALk})$ Example of calculations with the pair AL3 ( $I = 18, k = 3$  for all sequences).**Seq. 1:** LLIVVATIIIVLIVIVWI $N_A = 1, N_L = 3$ 

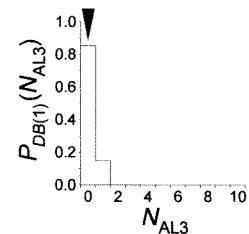
Observed occurrences of AL3 in Seq. 1: 0

$N_{AL3}$	$P_1 = P(N_{AL3} 18,3,1,3)$
0	0.853
1	0.147

**DB(1):**

Observed occurrences of AL3 in DB(1): 0

$$P_{DB(1)}(N_{AL3}) = P_1(N_{AL3})$$

**Seq. 2:** FGVGVALVAVLGAALLAL $N_A = 5, N_L = 5$ 

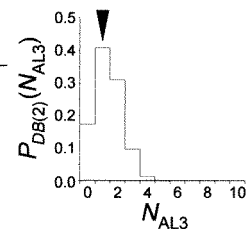
Observed occurrences of AL3 in Seq. 2: 1

$N_{AL3}$	$P_2 = P(N_{AL3} 18,3,5,5)$
0	0.203
1	0.442
2	0.286
3	0.064
4	$4.5 \cdot 10^{-3}$
5	$5.9 \cdot 10^{-5}$

**DB(2):**

Observed occurrences of AL3 in DB(2): 0+1 = 1

$N_{AL3}$	$P_{DB(2)}(N_{AL3})$	Calculated as
0	0.173	$P_2(0) \cdot P_{DB(1)}(0)$
1	0.407	$P_2(1) \cdot P_{DB(1)}(0) + P_2(0) \cdot P_{DB(1)}(1)$
...	...	...
5	$7.1 \cdot 10^{-4}$	$P_2(5) \cdot P_{DB(1)}(0) + P_2(4) \cdot P_{DB(1)}(1)$
6	$8.7 \cdot 10^{-6}$	$P_2(5) \cdot P_{DB(1)}(1)$

**Seq. 3:** IVMFVAVLLIAAFVAGIL $N_A = 4, N_L = 3$ 

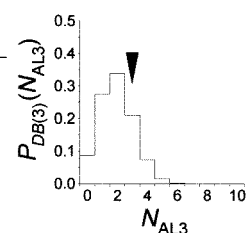
Observed occurrences of AL3 in Seq. 3: 2

$N_{AL3}$	$P_3 = P(N_{AL3} 18,3,4,3)$
0	0.500
1	0.416
2	0.081
3	$3.3 \cdot 10^{-3}$

**DB(3):**

Observed occurrences of AL3 in DB(3): 1+2 = 3

$N_{AL3}$	$P_{DB(3)}(N_{AL3})$	Calculated as
0	0.086	$P_3(0) \cdot P_{DB(2)}(0)$
1	0.275	$P_3(1) \cdot P_{DB(2)}(0) + P_3(0) \cdot P_{DB(2)}(1)$
2	0.338	$P_3(2) \cdot P_{DB(2)}(0) + P_3(1) \cdot P_{DB(2)}(1) + P_3(0) \cdot P_{DB(2)}(2)$
...	...	...
9	$2.9 \cdot 10^{-8}$	$P_3(6) \cdot P_{DB(2)}(3)$

**Final result****DB(13,606):**

Observed occurrences of AL3 in DB: 4,140  
 Expected occurrences: 4,043.1  
 Odds Ratio (Observed/Expected): 1.02  
 Significance ( $p$ ):  $7.5 \cdot 10^{-2}$

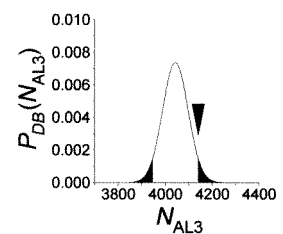


Figure 9. Calculation of probability distributions of pair occurrence with the TMSTAT method. The Figure is explained fully in the Appendix.

## Homology cleanup

Homology removal was performed at the level of the TM sequences by eliminating each sequence that was exceedingly similar to another sequence. Given the large number of proteins in the database, homology elimination at the level of the TMDs was a practical and effective alternative to more complex and intensive clustering procedures at the protein level (Boberg *et al.*, 1992; Brenner *et al.*, 1998; Gerstein, 1998; Hobohm & Sander, 1994; Hobohm *et al.*, 1992). In addition, the TMD-level procedure takes care of the "internal homology" sometimes present within a given protein while preserving any non-homologous TMDs of otherwise homologous proteins. The annotated sequences were extended (or occasionally shortened) to a length of 30 residues using the flanking regions. Two sequences were compared in all possible frame shifts using a 100 PAM matrix derived from the Mutation Probability Matrix of Jones *et al.* (1994) and the maximum score was recorded as the similarity score of the pair.

Sequences were eliminated according to the following process. First, all pairs with similarity scores of 50 or higher were ranked by score, from highest to lowest. Then, beginning with the pair with the highest score, one member of each pair was marked for removal. The particular sequence in a pair chosen for removal was determined by its priority number. Priorities, assigned according to the description of the annotation in the Swiss-Prot database, gave preference to non-potential transmembrane domains:

- 0, transmembrane sequences of potential protein (ORFs identified in Swiss-Prot with IDs starting with the letter Y);
- 1, transmembrane domains marked as POTENTIAL, PROBABLE or POSSIBLE;
- 2, annotations that included the words BY SIMILARITY;
- 3, remaining annotations.

Sequences with larger priority numbers were kept in the database, and when members of the pair shared the same priority number, one was randomly chosen for removal. The cleanup proceeded down the list of pairs so that when a pair in which neither sequence had been marked for removal was encountered, priority numbers were assigned and only one sequence was subsequently kept.

## Pair and triplets definition

The analysis of the pairs correlation was performed on all combinations of amino acids separated by one to ten residues ( $20 \times 20 \times 10 = 4000$  pair correlations analyzed). Pairs at  $i, i+k$  are indicated using the one-letter code of the two residues followed by the separation  $k$  (register): for example, AL1 corresponds to the sequence AL and AL3 to AxxL.

The triplets analyzed were formed by all combinations of residues at separations ranging from 1 to 5 ( $20 \times 20 \times 20 \times 5 \times 5 = 200,000$  triplet correlations). Triplets are represented by VI2P3 (corresponding to VxIxxP).

## Input sequences

The analysis was performed on sequences of fixed length instead of the entire annotation, in order to limit the analysis to the hydrophobic core of the sequences. The most hydrophobic window of 18 amino acid residues in a span of 30 residues centered on each annotation was selected using the GES scale (Engelman *et al.*, 1986). Occasionally, the selected window included residues outside the original annotations.

Exceedingly hydrophilic sequences with a hydrophobicity score below 15 were excluded from the analysis (4.9% of all sequences). Low-complexity sequences (when a single residue represented more than half of the composition of the sequence or two residues more than two-thirds of the composition of the sequence) were also excluded (0.5%).

## Pair and triplet correlation analysis with TMSTAT

The occurrences in the database of all pairs and triplets of residues were counted. The probability distributions associated with any possible number of occurrences of each pair and triplet were calculated from the composition of the individual sequences, as explained in Appendix and in the scheme in Figure 9. The statistical significance of the observed deviations of each occurrence from its respective average expected value was calculated by the two-tailed integral of their probability distributions.

## Acknowledgments

We thank Mark Bowen, Zimei Bu, Lilian Fisher, Karen Ho, Yuval Kluger, Albert Lee, Huiming Li, Maura Mezzetti, Gigi Riva, William Russ, Koji Sonoda, Iban Ubarretxena, Fang Zhou and other members of the Engelman group for helpful discussion and critical reading of the manuscript. This work was supported by grants from the NIH and NSF.

## References

- Arkin, I. T. & Brunger, A. T. (1998). Statistical analysis of predicted transmembrane alpha-helices. *Biochim. Biophys. Acta*, **1429**, 113-128.
- Bairoch, A. & Apweiler, R. (1999). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucl. Acids Res.* **27**, 49-54.
- Boberg, J., Salakoski, T. & Vihinen, M. (1992). Selection of a representative set of structures from Brookhaven Protein Data Bank. *Proteins: Struct. Funct. Genet.* **14**, 265-276.
- Bowie, J. U. (1997). Helix packing in membrane proteins. *J. Mol. Biol.* **272**, 780-789.
- Boyd, D., Schierle, C. & Beckwith, J. (1998). How many membrane proteins are there? *Protein Sci.* **7**, 201-205.
- Brenner, S. E., Chothia, C. & Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073-6078.
- Chothia, C., Levitt, M. & Richardson, D. (1981). Helix to helix packing in proteins. *J. Mol. Biol.* **145**, 215-250.

- Creamer, T. P. & Rose, G. D. (1995). Interactions between hydrophobic side chains within alpha-helices. *Protein Sci.* **4**, 1305-1314.
- Dunbrack, R. L., Jr & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **230**, 543-574.
- Engelman, D. M., Steitz, T. A. & Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.* **15**, 321-353.
- Fisher, L. E., Engelman, D. M. & Sturgis, J. N. (1999). Detergents modulate dimerization, but not helicity, of the glycoprotein A transmembrane domain. *J. Mol. Biol.* **293**, 639-651.
- Gerstein, M. (1998). Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins: Struct. Funct. Genet.* **33**, 518-534.
- Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522-524.
- Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409-417.
- Jones, D. T., Taylor, W. R. & Thornton, J. M. (1994). A mutation data matrix for transmembrane proteins. *FEBS Letters*, **339**, 269-275.
- Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* **157**, 105-132.
- Landolt-Marticorena, C., Williams, K. A., Deber, C. M. & Reithmeier, R. A. (1993). Non-random distribution of amino acids in the transmembrane segments of human type I single span membrane proteins. *J. Mol. Biol.* **229**, 602-608.
- Lemmon, M. A., Flanagan, J. M., Treutlein, H. R., Zhang, J. & Engelman, D. M. (1992). Sequence specificity in the dimerization of transmembrane alpha-helices. *Biochemistry*, **31**, 12719-12725.
- Lemmon, M. A., Treutlein, H. R., Adams, P. D., Brunger, A. T. & Engelman, D. M. (1994). A dimerization motif for transmembrane alpha-helices. *Nature Struct. Biol.* **1**, 157-163.
- Li, S. C. & Deber, C. M. (1992a). Glycine and beta-branched residues support and modulate peptide helicity in membrane environments. *FEBS Letters*, **311**, 217-220.
- Li, S. C. & Deber, C. M. (1992b). Influence of glycine residues on peptide conformation in membrane environments. *Int. J. Pept. Protein Res.* **40**, 243-248.
- Liu, L. P. & Deber, C. M. (1998). Uncoupling hydrophobicity and helicity in transmembrane segments. Alpha-helical propensities of the amino acids in non-polar environments. *J. Biol. Chem.* **273**, 23645-23648.
- Luecke, H., Schobert, B., Richter, H. T., Cartailler, J. P. & Lanyi, J. K. (1999). Structure of bacteriorhodopsin at 1.55 Å resolution. *J. Mol. Biol.* **291**, 899-911.
- MacKenzie, K. R. & Engelman, D. M. (1998). Structure-based prediction of the stability of transmembrane helix-helix interactions: the sequence dependence of glycoprotein A dimerization. *Proc. Natl Acad. Sci. USA*, **95**, 3583-3590.
- MacKenzie, K. R., Prestegard, J. H. & Engelman, D. M. (1997). A transmembrane helix dimer: structure and implications. *Science*, **276**, 131-133.
- McGregor, M. J., Islam, S. A. & Sternberg, M. J. (1987). Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J. Mol. Biol.* **198**, 295-310.
- Padmanabhan, S. & Baldwin, R. L. (1994). Tests for helix-stabilizing interactions between various non-polar side chains in alanine-based peptides. *Protein Sci.* **3**, 1992-1997.
- Popot, J. L. & Engelman, D. M. (1990). Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*, **29**, 4031-4037.
- Ri, Y., Ballesteros, J. A., Abrams, C. K., Oh, S., Verselis, V. K., Weinstein, H. & Bargiello, T. A. (1999). The role of a conserved proline residue in mediating conformational changes associated with voltage gating of Cx32 gap junctions. *Biophys. J.* **76**, 2887-2898.
- Richmond, T. J. & Richards, F. M. (1978). Packing of alpha-helices: geometrical constraints and contact areas. *J. Mol. Biol.* **119**, 537-555.
- Russ, W. P. & Engelman, D. M. (2000). The GxxxG motif: a framework for transmembrane helix-helix association. *J. Mol. Biol.* **296**, 911-919.
- Samatey, F. A., Xu, C. & Popot, J. L. (1995). On the distribution of amino acid residues in transmembrane alpha-helix bundles. *Proc. Natl Acad. Sci. USA*, **92**, 4577-4581.
- Schrauber, H., Eisenhaber, F. & Argos, P. (1993). Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J. Mol. Biol.* **230**, 592-612.
- von Heijne, G. (1992). Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* **225**, 487-494.
- Walther, D. & Argos, P. (1996). Intrahelical side chain-side chain contacts: the consequences of restricted rotameric states and implications for helix engineering and design. *Protein Eng.* **9**, 471-478.
- Walther, D., Eisenhaber, F. & Argos, P. (1996). Principles of helix-helix packing in proteins: the helical lattice superposition model. *J. Mol. Biol.* **255**, 536-553.

### Appendix I: Calculation of Expectation Distributions for the Occurrence of Pairs and Triplets of Amino Acids in a Database of Short Sequences with the TMSTAT Method

The aim of the present analysis is to survey frequently occurring patterns of residues (pairs and triplets) in transmembrane sequences. For this, we need some measure of the expectation of occurrence of the patterns. The simplest way to calculate this is from the average composition (i.e. the probability of finding a particular residue is constant at all positions in all sequences and corresponds to its frequency in the database). However, this approach requires the assumption that, in terms of composition, all sequences derive from a homogeneous population and that residues do not co-segregate or anti-segregate in different sequences.

This assumption is not required if the expectation is based on the composition of each individual sequence instead of the overall composition of amino acids in the database (i.e. the probability of finding a particular residue is constant at all positions within a sequence and corresponds to its frequency in the sequence). However, finite sequence length effects also need to be accounted for, since they are quite important for short sequences (18 residues in our case). A solution is to base the

calculation on all theoretically possible internal permutations of the sequences, that is, to take into account the length and composition of each sequence once internal positional information has been removed. A way to conceptualize this is to ask: What would be the probability of finding a certain number of occurrences of a pair in the database after all sequences have been randomly permuted? Considering the entire theoretical set of different databases that can be obtained from the original when the sequences are allowed to independently assume any possible internal permutation, the probability corresponds to the fraction of all permuted databases that contain that exact number of occurrences of the pair.

The expectancy distribution of a pair based on all theoretical permutations of all sequences could be approximated by cycles of random shuffling of the sequences and sampling of the occurrences. However, a sampling algorithm would produce estimates with errors that are higher at the tails of the distribution, i.e. where greater precision would be desirable. To completely avoid errors, we have calculated analytically the exact theoretical distributions of expectancy of any pair. The TMSTAT method is schematized in Figure 9 of the main text. The calculation is divided into two phases: in phase 1, the probability distributions for occurrences of pairs in single sequences were calculated and stored in a matrix table for later use. Consider the pair  $ALk$ , A and L as examples of any two non-identical residues at positions  $i, i+k$ : the probability that pair  $ALk$  will occur  $N_{ALk}$  times in a particular sequence is:

$$P(N_{ALk}|l, k, N_A, N_L)$$

which depends on four parameters; the length of the sequence  $l$ , the register  $k$  and how many Ala ( $N_A$ ) and Leu residues ( $N_L$ ) are in the sequence. It is defined as the fraction of all possible permutations of the sequence containing exactly  $N_{ALk}$  occurrences of the  $ALk$  pair. An example of the calculation is shown explicitly in the scheme for a short five residue sequence with two Ala and two Leu residues and at register 3. The box shows all 30 possible permutations of the short sequence (the non-A and non-L residue is symbolized by a dash): of the 30 possible permutations, 19 (63.3%) have no occurrences of the  $AL3$  pair. The pair occurs once in ten (33.3%) and twice in one (3.3%) of the permutations. All sequence probability distributions for all relevant combinations of the four parameters ( $l = 18$ ;  $k = 1$  to  $10$ ;  $N_A = 1$  to  $9$ ;  $N_L = 1$  to  $9$ ) were calculated and tabulated for later use. Pairs formed by two identical residues, as for example  $LLk$ , obey different distributions,  $P(N_{LLk} | l, k, N_L)$ , that were analogously calculated and tabulated.

The specific database is considered only in phase 2, when actual occurrences of the pairs are counted and the database probabilities are computed. The overall probability distribution of occurrence of the

pair  $ALk$  in the database,  $P_{DB}$ , was calculated by iteratively convoluting the specific single-sequence  $P_j(N_{ALk})$  distributions tabulated in phase 1 relative to the  $[l, k, N_A, j, N_L, j]$  parameters provided by each  $j$  sequence of the database considered. The probability of observing  $N_{ALk}$  occurrences of the pair  $ALk$  in a database of  $n$  sequences can be calculated according to:

$$P_{DB(n)}(N_{ALk}) = \sum_{i=0}^{N_{ALk}} P_{DB(n-1)}(i)P_n(N_{ALk} - i | l, k, N_A, n, N_L, n)$$

defined recursively, with initial  $P_{DB(0)}(0) = 1$ .  $N_{A,n}$  and  $N_{L,n}$  are the number of Ala and Leu residues in sequence  $n$ .

An example of the process is shown in the scheme where the first three steps and the final result are illustrated for the analysis of the occurrences of the pair  $AL3$ . All sequences in the database analyzed have fixed length  $l$  of 18 residues (this restriction is not necessary in general and the method applies to mixed-length sequence databases). The first sequence of the database contains two Ala and three Leu residues. No occurrence of  $AL3$  is observed in this sequence (black arrow at zero occurrence in chart) In the first step of the procedure only one sequence has been considered and the probability distribution of the database,  $P_{DB(1)}(N_{AL3})$  (bar chart) corresponds to the probability distribution of sequence 1,  $P_1 = P(N_{AL3} | l = 18, k = 3, N_A = 2, N_L = 3)$ .

The second sequence of the database contains five Ala and five Leu residues, and in this case one occurrence of  $AL3$  is observed.  $P_2$  is thus  $P(N_{AL3} | l = 18, k = 3, N_A = 5, N_L = 5)$  and the cumulative  $P_{DB(2)}$  distribution is then obtained from  $P_2$  and  $P_{DB(1)}$ , as shown in the example. Two occurrences of  $AL3$  are found in the third sequence, bringing the total to three for the database at this stage, and  $P_{DB(2)}$  is then obtained from  $P_3$  and  $P_{DB(2)}$ . The calculation becomes more complex as more combinations are available and the curve assumes a more bell-shaped character.

Once all 13,606 sequences had been analyzed, the  $P_{DB}$  distribution has converged to a bell curve. Average expected values and standard deviations were calculated from the probability distribution curves according to:

$$\bar{N}_{ALk} = \sum_{N_{ALk}} [N_{ALk} P_{DB}(N_{ALk})]$$

$$SD_{ALk} =$$

$$\sqrt{\sum_{N_{ALk}} [N_{ALk}^2 P_{DB}(N_{ALk})] - \left[ \sum_{N_{ALk}} N_{ALk} P_{DB}(N_{ALk}) \right]^2}$$

The observed 4140 occurrences of  $AL3$  in the database are slightly above the average expectation value of 4043.1. The two-tailed integral of

the  $P_{DB}(N_{AL3})$  function provided a significance for the observed occurrences of a pair. The integration was computed on formally derived curves; therefore, no assumption regarding the nature of the distributions was necessary. Two-tailed integrals were used, since both above and below-expectation values were considered significant. The significance of the occurrences of the AL3 pair is low ( $p = 0.075$ ), that is, if the residues were actually randomly distributed there would be a realistic possibility of observing an equal or greater number of occurrences by random chance.

The analysis of the triplets was performed with an analogous method. The single-sequence probability distributions were calculated for the triplet  $ALk_1Vk_2$  as:

$$P(N_{ALk_1Vk_2} | l, k_1, k_2, N_A, N_L, N_V)$$

based on all possible sequence permutations and tabulated for the relevant ranges of  $l$ ,  $k_1$ ,  $k_2$ ,  $N_A$ ,  $N_L$  and  $N_V$  (Ala, Leu and Val representing any three non-identical residues at relative spacing  $k_1$  and  $k_2$ ). Probability distributions were also calculated for triplets in which residues are repeated ( $AAk_1Lk_2$ ,  $ALk_1Ak_2$ ,  $ALk_1Lk_2$ ,  $AAk_1Ak_2$ ). The cumulative probability distribution,  $P_{DB}$ , for the occurrence of each triplet in the database was calculated with the same recursive formula of the pairs. The TMSTAT method is, in principle, applicable to quadruplets and higher-order multiplets, although the increased number of combinations can limit the feasibility.

*Edited by G. von Heijne*

*(Received 4 November 1999; received in revised form 29 December 1999; accepted 29 December 1999)*