Interstrand pairing patterns in β-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction

Ronald Jackups, Jr. and Jie Liang*
Department of Bioengineering, SEO, MC-063
University of Illinois at Chicago
851 S. Morgan Street, Room 218
Chicago, IL 60607–7052, U.S.A.

September 23, 2005

1 Supplementary Material

Database. The dataset used to derive the statistical models comprises 19 β-barrel membrane proteins found in the Protein Data Bank (Table 5), totaling 262 β-strands. All proteins share no more than 26% pairwise sequence identity. All structures have a resolution of 2.6 Å or better.

The dataset of soluble barrel and barrel-like β-sheets was compiled by searching structural homologs using the Combinatorial Extension method [2] with each of the 19 proteins in the TM dataset. Because the resulting dataset was small, CE was used again on the two closed barrels obtained from the first attempt, streptavidin (pdb `1stp`) and retinol binding protein (pdb `1brp`). As with the TM dataset, all proteins share no more than 26% pairwise sequence identity. All structures have a resolution of 3.3 Å or better. The dataset comprises 28 soluble β-sheets: `1avg`, `1ayr`, `1bbp`, `1bj7`, `1bpo`, `1brp`, `1dfv`, `1dmm`, `1d2u`, `1eg9`, `1ei5`, `1epa`, `1em2`, `1ewf`, `1fsk`, `1fx3`, `1h91`, `1jkg`, `1lkf`, `1m6p`, `1qfv`, `1std`, `1stp`, `1t27`, `1una`, `2a2u`, `3blg`, `4bcl`.

*Corresponding author. Phone: (312)355–1789, fax: (312)996–5921, email: `jliang@uic.edu`

Spatial Regions and Strand Model. In order to classify residues in β-barrel membrane proteins by specific spatial regions (core, headgroup, and polar caps), the coordinates in the protein's PDB file were translated and rotated so that the xy-plane was perpendicular to the vertical axis of the barrel and equidistant to the observed aromatic girdles presumed to be at the membrane interfaces. An example of such a transformation is illustrated in Figure 1. Each residue in the protein was assigned a region based on the z-coordinate of its associated α-carbon.

In addition, TM residues (those in the core or headgroup regions) were assigned as *internal* (*i.e.* their side-chains face into the center of the barrel) or *external* (*i.e.* their side-chains face away from the center of the barrel), depending on the angle between the vector from the barrel central axis to the α-carbon and the vector from the α-carbon to the β-carbon of the residue. If the angle is less than 90°, it is classified as internal; if the angle is greater, it is classified as external. If the residue is glycine, its orientation is extrapolated as the opposite of the residue previous to it on the β-strand.

In total, each barrel is divided into 8 distinct regions: periplasmic cap with $z \in (-20.5\text{Å}, -13.5\text{Å})$, periplasmic headgroup (internal and external) with $z \in (-13.5\text{Å}, -6.5\text{Å})$, core (internal and external) with $z \in (-6.5\text{Å}, 6.5\text{Å})$, extracellular headgroup (internal and external) with $z \in (6.5\text{Å}, 13.5\text{Å})$, and extracellular cap with $z \in (13.5\text{Å}, 20.5\text{Å})$. Residues that are not assigned to be in β-strands are automatically excluded from the TM (headgroup and core) regions. Unless noted, any residue in the protein not assigned to any of these regions was excluded from the calculation.

Single-body propensities. We define the single-body propensity $P_r(X)$ of residue type X in region r as the odds ratio comparing the frequency of a residue type in one region to its expected frequency when all eight regions are combined:

$$P_r(X) = \frac{f(X|r)}{\mathbb{E}[f'(X|r)]},$$

where $f(X|r)$ is the observed frequency (number count) of residue type X in region r, and $\mathbb{E}[f'(X|r)]$ is the expected frequency of residue type X in region r.

In order to calculate $\mathbb{E}[f'(X|r)]$, we need a random null model. Here we chose as our model

exhaustive permutation of all residues in all eight regions, such that each permutation occurs with equal probability. That is, the residues in all eight regions of all proteins in the dataset are permuted exhaustively, without replacement, and assigned regions based on their new positions. For each permutation, $f'(X|r)$ records the number of occurrences of residue type X in region r. Under this model, the probability $\mathbb{P}_{X|r}(i)$ of $i = f'(X|r)$ residues of type X being assigned to region r follows a hypergeometric distribution:

$$\mathbb{P}_{X|r}(i) = h(i|n, n_r, n_x) = \frac{\binom{n_x}{i}\binom{n-n_x}{n_r-i}}{\binom{n}{n_r}},$$

where n is the number of residues of all types in all eight regions in the entire dataset, n_r is the number of residues of all types in region r, and n_x is the number of residues of type X in all regions. Analogously, we can think of $\mathbb{P}_{X|r}(i)$ as the probability of selecting without replacement n_r residues out of a total of n residues contained in an urn, such that i of the residues are of type X. The expected frequency of residue type X occurring in region r is the mean of the hypergeometric distribution $h(i|n, n_r, n_x)$:

$$\mathbb{E}[f'(X|r)] = \sum_{i=0}^{n_x} i \cdot \mathbb{P}_{X|r}(i) = \frac{n_r \cdot n_x}{n}. \tag{1}$$

Therefore, the propensity $P_r(X)$ is:

$$P_r(X) = \frac{f(X|r)}{\mathbb{E}[f'(X|r)]} = \frac{\frac{f(X|r)}{n_r}}{\frac{n_x}{n}}.$$

That is, $P_r(X)$ is the ratio of the proportion of residue type X in region r to the proportion of residue type X in all eight regions combined.

Because the frequency of occurrences of residues of type X in region r in the null model follows a hypergeometric distribution, we can calculate an exact p-value for an observed $f(X|r)$ to assess its statistical significance. We calculate a two-tailed p-value based on the null hypothesis that $P_r(X) = 1.0$. If the observed $P_r(X) < 1.0$, then:

$$p = 2 \cdot \sum_{i=0}^{f(X|r)} \mathbb{P}_{X|r}(i). \tag{2}$$

If the observed $P_r(X) > 1.0$, then:

$$p = 2 \cdot \sum_{i=f(X|r)}^{x_n} \mathbb{P}_{X|r}(i). \tag{3}$$

That is, the p-value is the probability that the propensity calculated from the dataset would deviate as much or more from the observed propensity, higher or lower, assuming that the actual propensity is 1.0.

Determination of pairwise contacts. The three types of pairwise contacts (strong H-bond, side-chain, and weak H-bond) were assigned for interacting residues based on contact types defined by the DSSP program (Definition of Secondary Structure of Proteins [1]) based on the atomic coordinates from PDB files. For β-sheets, DSSP defines *bridge partners* as residues across from each other on adjacent β-strands, and also determines whether bridge partners interact via backbone N-O H-bonds. Two TM residues on adjacent β-strands contribute to a backbone *strong H-bond* interaction if they are listed as bridge partners that contribute to backbone H-bonds. They contribute to a *non-H-bonded interaction* if they are listed as bridge partners but do not contribute to backbone H-bonds. They contribute to a *weak H-bond* if the residue with the smaller residue number (*i.e.* closer to the N-terminus of the protein) is a bridge partner to the residue immediately following the other residue of the pair. This is because a weak C_α-O H-bond extends by one residue in the N-C direction, as seen in Figure 2. Residues j and y contribute to a weak H-bond in Figure 2, for example, because the residue with the lower number (y, since it is closer to the N-terminus), is a bridge partner to i, the residue immediately following j. For a strand pair between the first and last strands of a protein, the larger and smaller numbered residues must be switched, to account for the fully circular nature of β-barrels. Residues outside the TM regions (core and headgroup) were not considered in our calculations of pairwise contacts.

Interstrand two-body spatial contact propensities. The interstrand pairwise propensity (TMSIP) $P(X, Y)$ of residue types X and Y for each of the three types of pairwise contacts is given

by:

$$P(X,Y) = \frac{f(X,Y)}{\mathbb{E}[f'(X,Y)]},$$

where $f(X,Y)$ is the observed frequency of X-Y contacts of a specific type in the TM regions (core and headgroup), and $\mathbb{E}[f'(X,Y)]$ is the expected frequency of X-Y contacts in a null model. In Tables 2 and 3, $f(X,Y)$ is listed under "Obs.", and $\mathbb{E}[f'(X,Y)]$ is listed under "Exp."

In order to calculate $\mathbb{E}[f'(X,Y)]$, we choose a null model in which residues within each of the two adjacent strands in a strand pair are permuted exhaustively and independently, and each permutation occurs with equal probability. In this null model, an X-Y contact forms if in a permuted strand pair two contacting residues happen to be type X and type Y. $\mathbb{E}[f'(x,y)]$ is then the expected number of X-Y contacts over the entire dataset.

Contacts between residues of the same type. When X is the same as Y, the probability $\mathbb{P}_{X,X}(i)$ of $i = f'(X,X)$ number of X-X contacts in a strand pair follows a hypergeometric distribution:

$$\mathbb{P}_{X,X}(i) = \frac{\binom{x_1}{i}\binom{l-x_1}{x_2-i}}{\binom{l}{x_2}},$$

where x_1 is the number of residues of type X on the first strand, x_2 is the number of residues of type X on the second strand, and l is the length of the strand pair (the lengths of the two strands must be equal). This mimics the random selection of residues from one strand to pair up with residues from the other strand.

This is the same as picking x_2 balls from an urn of l balls, i of which are of type X and $x_2 - i$ of which are not type X. In this case, the urn represents the first strand, containing x_1 residues of type X and $l - x_1$ residues not of type X. The x_2 balls picked from the urn represent the residues in the first strand selected to be placed adjacent to the x_2 residues of type X in the second strand, i of which are of type X, and $x_2 - i$ of which are not of type X.

$\mathbb{E}_{\text{all}}[f'(X,X)]$ is then the sum of the expected values of $f'(X,X)$ for the set $\mathcal{SP}$ of all strand

pairs in the dataset.

$$\mathbb{E}_{\text{all}}[f'(X,X)] = \sum_{sp \in \mathcal{SP}} \mathbb{E}_{sp}[f'(X,X)] = \sum_{sp \in \mathcal{SP}} \frac{x_1(sp) \cdot x_2(sp)}{l(sp)},$$

where $x_1(sp)$ and $x_2(sp)$ are the numbers of residues of type X in the first and second strand of strand pair $sp \in \mathcal{SP}$, respectively, and $l(sp)$ is the length of strand pair sp. The right-hand side is determined by using the expectation of the hypergeometric distribution, analogous to Equation (1). For statistical significance, two-tailed p-values can be calculated using formulae similar to Equations (2) and (3).

Contacts between residues of different types. If the two contacting residues are not of the same type, *i.e.* $X \neq Y$, then the number of X-Y contacts in the random model for one strand pair is the sum of two dependent hypergeometric variables, one variable for type X residues in the first strand and type Y in the second strand, and another variable for type Y residues in the first strand and type X in the second strand. The expected frequency of X-Y contacts $\mathbb{E}[f'(X,Y)]$ is the sum of the two expected values over all strand pairs $sp \in \mathcal{SP}$:

$$\mathbb{E}[f'(X,Y)] = \sum_{sp \in \mathcal{SP}} \{\mathbb{E}[f'_{sp}(X,Y)] + \mathbb{E}[f'_{sp}(Y,X)]\} = \sum_{sp \in \mathcal{SP}} \{\frac{x_1(sp) \cdot y_2(sp)}{l(sp)} + \frac{y_1(sp) \cdot x_2(sp)}{l(sp)}\},$$

where $x_1(sp)$ and $x_2(sp)$ are the numbers of residues of type X in the first and second strand, $y_1(sp)$ and $y_2(sp)$ are the numbers of residues of type Y in the first and second strand, and $l(sp)$ is the length of strand pair sp. The right-hand side is determined by using the expectation of the hypergeometric distribution, analogous to Equation (1). Despite the fact that the variables $f'_{sp}(X,Y)$ and $f'_{sp}(Y,X)$ are dependent (*i.e.* the placement of an X-Y pair may affect the probability of a Y-X pair in the same strand pair), their expectations may be summed directly, because expectation is a linear operator.

Generalized hypergeometric model. However, because $f'_{sp}(X,Y)$ and $f'_{sp}(Y,X)$ are dependent, to determine p-values for a specific number of observed X-Y contacts, a more complex hypergeometric formula for the null model must be established. The probability of a specific number of X-Y contacts occurring in one strand pair does not follow a simple hypergeometric distribution. Here

we develop a generalized hypergeometric model based on the trinomial coefficient to characterize such a probability. First, we have a 3-element trinomial function $(a, b, c)!$ defined as:

$$(a, b, c)! \equiv \frac{(a + b + c)!}{a!b!c!}.$$

It represents the number of distinct permutations in a multiset of three different types of elements, with number count a, b, and c for each of the three element types. Consider residues in the first strand of length l of a strand pair. These l residues are of three types: x_1 count of type X residues, y_1 of type Y residues , and $l - x_1 - y_1$ count of type "neither". If we exhaustively permute the l residues, we have the trinomial coefficient number of different permutations. We denote this as:

$$T(l, x_1, y_1) \equiv (x_1, y_1, l - x_1 - y_1)!.$$

We now first fix the positions of residues on strand 1, and permute exhaustively all matching l residues on strand 2. Let x_2, y_2, and $l - x_2 - y_2$ be the numbers of residue of type X, Y, and "neither" on strand 2, respectively. The total number of permutations for strand 2 is:

$$T(l, x_2, y_2) = (x_2, y_2, l - x_2 - y_2)!.$$

Consider the residues on strand 2 that match to the x_1 number of residues of type X on strand 1. (This and all further descriptions are illustrated in Figure 1 for clarification.) These x_1 residues on strand 2 consist of h number of type X residues, i number of type Y residues, and $x_1 - h - i$ number of type "neither" residues. They can be permuted in

$$T(x_1, h, i) = (h, i, x_1 - h - i)!$$

different ways. By analogy, the y_1 residues on strand 2 that match type Y residues in strand 1 consist of j number of type X residues, k number of type Y residues, and $y_1 - j - k$ of type "neither"

residues, and thus the total number of permutations for these y_1 residues is:

$$T(y_1, j, k) = (j, k, y_1 - j - k)!.$$

Similarly, there are $T(l - x_1 - y_1,\, x_2 - h - j,\, y_2 - i - k)$ number of permutations to match the remaining $l - x_1 - y_1$ of type "neither" residues on strand 1.

We characterize the probability $\mathbb{P}(h, i, j, k)$ of interstrand matches: a) the x_1 type X residues on strand 1 with h type X residues, i type Y residues, and $x_1 - h - i$ type "neither" residues on strand 2; b) the y_1 type Y residues on strand 1 with j type X residues, k type Y residues, and $y_1 - j - k$ type "neither" residues on strand 2; and c) the remaining $l - x_1 - y_1$ type "neither" residues on strand 1 with $x_2 - h - j$ type X residues, $y_2 - i - k$ type Y residues, and the remaining type "neither" residues from strand 2. Equivalently, $\mathbb{P}(h, i, j, k)$ is the probability of h X-X contacts, i X-Y contacts, j Y-X contacts, and k Y-Y contacts occurring in a random permutation.

We introduce a higher order hypergeometric distribution for $\mathbb{P}(h, i, j, k)$ as follows:

$$\mathbb{P}(h, i, j, k) = \frac{T(x_1, h, i) \cdot T(y_1, j, k) \cdot T(l - x_1 - y_1, x_2 - h - j, y_2 - i - k)}{T(l, x_2, y_2)}.$$

This can be illustrated as follows. When randomly picking x_2 of type X residues, y_2 of type Y residues, and the remaining $l - x_2 - y_2$ type "neither" residues from an urn for strand 2, we have: (1) those matching the x_1 residues of type X on strand 1 are of h number of type X, i number of type Y, and $x_1 - h - i$ of type "neither"; (2) those matching the y_1 residues of type Y on strand 1 are of j number of type X, k number of type Y, and $x_2 - j - k$ of type "neither"; and (3) those matching the $l - x_1 - y_1$ residues of type "neither" on strand 1 are of $x_2 - h - j$ number of type X, $y_2 - i - k$ number of type Y, and $(l_1 - x_1 - y_1) - (x_2 - h - j) - (y_2 - i - k)$ of type "neither".

The marginal probability $\mathbb{P}_{X,Y}(m)$ that there are a total of $i + j = m$ X-Y contacts in the random model, namely, the pairings where a residue of type X in the first strand is paired with a residue of type Y in the second strand, summed with the pairings in which a residue of type Y in

the first strand is paired with a residue of type X in the second strand is:

$$\mathbb{P}_{X,Y}(m) = \sum_{h=0}^{x_1} \sum_{i=0}^{x_1-h} \sum_{k=0}^{y_1-(m-i)} \mathbb{P}(h, i, m-i, k),$$

where h is the number of matched X-X contacts, i the number of matched X-Y contacts, $m-i$ the number of matched Y-X contacts(j in Figure 1), and k the number of matched Y-Y contacts. The remaining contacts involving residues of type "neither" will then automatically be assigned, since all matches involving X and Y have been accounted for. There are x_1 possible values for h, one for each residue of type X on strand 1; $x_1 - h$ possible values for i, once h has been determined; and $y_1 - j = y - (m - i)$ possible values for k, once i has been determined. The i number of X-Y contacts plus the $m-i$ number of Y-X contacts will sum to the m number of contacts desired. This closed-form formula allows us to calculate analytically the two-tailed p-value for this null model of $f'(X, Y)$ number of observed X-Y contacts using formulae similar to Equations (2) and (3).

Confounding between single-body propensity and interstrand two-body propensity. Because single-body propensities can vary significantly, it is possible that differences in two-body propensities may simply be reflections of differences in single-body propensities, *e.g.* two polar residues might have high strong H-bond pairwise propensities simply because both residues have an independent preference for the same side of the TM barrel (internal-facing), and not because of any direct significant propensity between the two. This artifact can be eliminated by dividing each strand into two "substrands," each of which contains only residues facing the same direction. This correction is automatic for strong H-bonds and non-H-bonded interactions, as all of the residues participating in each of these interactions in a single strand pair face the same direction (*e.g.* the residues in a particular strand pair participating in a strong H-bond must either all be internal or all be external). For weak H-bond interactions, in which one residue is internal and one is external, each strand pair must be divided into two substrand pairs: one pair in which the first substrand is internal and the second is external, and another pair in which the first substrand is external and the second is internal. In this way, the often dominating effects of single-residue orientation are removed from two-body propensity calculation. Results reported in Tables 2 and 3 are obtained after these cor-

rections. For the analysis performed on the soluble β-sheet dataset, there is no strong distinction between internal and external residues, since only some of the proteins are closed barrels. Thus, no correction was used for weak H-bonds.

Pairwise propensities for a reduced alphabet. To obtain an objective reduced alphabet of amino acids for studying membrane β-barrel proteins, we cluster amino acids by their location preference and strand pairing propensities. We represent each amino acid as a vector and use hierarchical clustering to define residue groups. Each vector consists of 68 z-scores: one for each of the 20 pairwise contact propensities including the residue for each of the 3 interaction types (strong H-bonds, non-H-bonded interactions, and weak H-bonds), and one for each single-body regional propensity of the residue in each of the 8 regions. The z-scores for pairwise propensities are calculated as

$$z(X,Y) = \frac{f(X,Y) - \mathbb{E}[f'(X,Y)]}{\sqrt{\text{var}\,[f'(X,Y)]}}, \tag{4}$$

and the z-scores for single-body propensities are calculated as

$$z(X|r) = \frac{f(X|r) - \mathbb{E}[f'(X|r)]}{\sqrt{\text{var}\,[f'(X|r)]}}.$$

We use hierarchical clustering by average linkage with a Euclidean distance function to obtain the clustering shown in Figure 3. We place the distance threshold so an alphabet of 5 residues is formed.

Strand register prediction. Our algorithm consists of two steps, the prediction of exact strand starts and the prediction of strand register. We use two sequence models for these two tasks, namely, a 16-residue single strand model and a 9-residue TM strand pair model (Figure 2). The regional designations in the 16-residue canonical strand model are based on known physical attributes of TM β-strands: an average length of 9-10 residues in the headgroup and core regions, and an alternating internal-external pattern. The size of each region (the cap regions, headgroup regions, and core region) is determined by dividing the total number of residues in a particular region by the number of strands in the dataset. We have 4 residues for the extracellular cap region, 3 for the periplasmic

cap region, 2 each for the two headgroup regions, and 5 for the core region.

The 9-residue strand pair model is derived from the canonical 16-residue models for two adjacent strands. These 9 residues are those designated to be in the transmembrane region, *i.e.*, in the headgroup or core regions. We exclude the 7 cap residues because the cap regions do not contribute to strand pairing. We also incorporate the physical properties of the 3 types of strand interactions in the model: The strong H-bond and non-H-bonded interactions must alternate, and the weak H-bonds must extend one residue in the N-C direction. Due to chirality constraints in antiparallel β-sheets (*i.e.* all amino acids in biological proteins are L-amino acids), the backbone H-bonding pattern is fixed once the N-C direction of the strands and the internal-external pattern are determined by step 1. We describe steps 1 and 2 in more detail:

1 *Predict exact starts.* For a 16-residue window fitted to our strand model (Figure 2a), we calculate the *single strand energy* $E(s;i)$ in kT units for strand i:

$$E(s;i) = -\sum_{k=1}^{16} \ln P(k;i,s),$$

where s is the displacement of the window position and $P(k;i,s)$ is the single body propensity for the k-th residue in this window of strand i, where the residues are alternatively assigned as internal and external. We calculate the energy score for a total of 11 possible windows: the window starting at the given approximated strand start $s = 0$ (either provided beforehand or calculated by a strand predictor), and all windows 5 residues up ($s \leq +5$) or down ($s \geq -5$) in the amino acid sequence. Because there are two possibilities for the internal-external pattern of TM residues in our model, we calculated strand energy scores for both possibilities for each window. We take the window with the lowest energy as the exact strand start that we use in step 2. The prediction also identifies which residues are internal and which are external.

2 *Predict strand register.* After excluding the 7 cap residues, we fit the 9-residue windows for two adjacent strands to the strand pair model (Figure 2b), and calculate the *strand pairing*

energy $E(s; i, i+1)$ in kT units for the strand pair $(i, i+1)$ with *strand shear* s as

$$E(s; i, i+1) = -\sum \ln P_{SHB}(k; i, i+1, s) - \sum \ln P_{NB}(k; i, i+1, s) - \sum \ln P_{WHB}(k; i, i+1, s) + \alpha|\frac{N+2}{N} - s|.$$

Here i and $i+1$ are the labels of the two adjacent strands, $P(k; i, i+1, s)$ is the propensity of pairing between the k-th residues in strand i and strand $i+1$, and the strand shear s is the residue displacement between the starts of the two strands. The first 3 terms refer to each of the 3 interaction types: strong H-bonds (SHB), non-H-bonded interactions (NB), and weak H-bonds (WHB). The last term represents a penalty to the score when the strand shear s deviates from the average strand shear, approximated as the average shearing number $N+2$ of known TM β-barrels divided by the number of strands N. α is a coefficient determined computationally. We find that $\alpha = 2$ works well.

We calculate the strand pairing score for a total of 11 possible windows by sliding one strand in a pair against the other strand: the windows with a strand shear of $s = -4$ residues up to a strand shear of $s = +6$. We also exclude strand shears that place internal residues next to external residues, as this would violate the H-bonding patterns of β-sheets. This reduces the search space by half. The strand shear s with the lowest strand pairing energy is taken as the true strand register.

To approximate the strand starts of the β-barrel membrane proteins in our dataset, we use the hidden Markov model predictor of Bigelow *et al.* [3], one of the most successful predictors for strand starts. This predictor uses only the amino acid sequence as input, and outputs a designation for each residue from a list of 4 possibilities: "up-strand" (referring to an odd-numbered strand), "down-strand" (even-numbered strand), loop between an up- and down-strand, and loop between a down- and up-strand. This predictor will therefore also predict the number of strands in the protein *ab initio*. We use the first instances of the up- and down-strand designations for each strand as approximate strand starts.

We exclude from our prediction analysis two proteins in our dataset (6 strand pairs) for which

the strand start predictor failed: TolC and α-HL. For TolC, the predictor did not detect any TM strands. For α-HL, the predictor detected far too many strands (8 instead of the actual 2 strands). For the remaining 17 of the 19 proteins, the hidden Markov model correctly predicted 252 of the 256 strands, with only 4 false positives. The false positive strands were all short (7 residues) and had irregular composition. Our register prediction was also incorrect for all strand pairs involving these 4 false positives. We apply our prediction to the 17 β-barrel membrane proteins in leave-one-out fashion: we calculate single and pairwise propensities using 16 of the proteins, and use them to predict strand registers in the 17th protein. Since the pairwise propensities are derived from a very small dataset, we introduce a pseudocount of 1 to the observed and expected numbers of each pair when calculating propensities.

References

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

[2] I. N. Shindyalov, P. E. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path., Protein Eng. 9 (1998) 739–747.

[3] H. R. Bigelow, D. S. Petrey, J. Liu, D. Przybylski, B. Rost, Predicting transmembrane β-barrels in proteomes, Nucleic Acids Res. 32 (2004) 2566–2577.

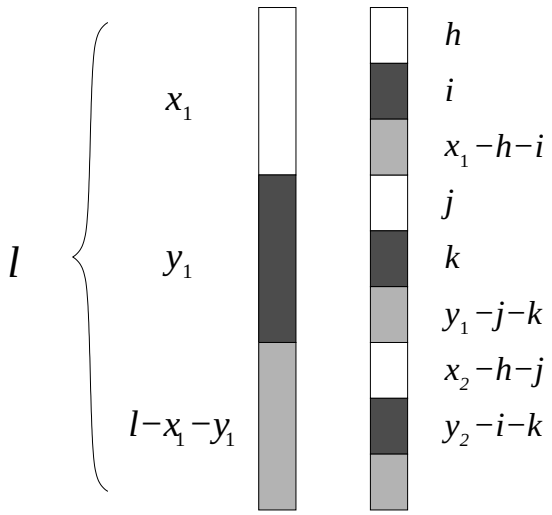

Figure 1: Illustration of the null model for interstrand spatial motifs when $X \neq Y$. White represents X residues, black Y residues, and grey "neither" residues (*i.e.* neither X nor Y). X-Y motifs are represented by the i residue pairs in which there is an X residue in the first strand and a Y residue in the second, and by the j residue pairs in which there is a Y residue in the first strand and an X residue in the second.

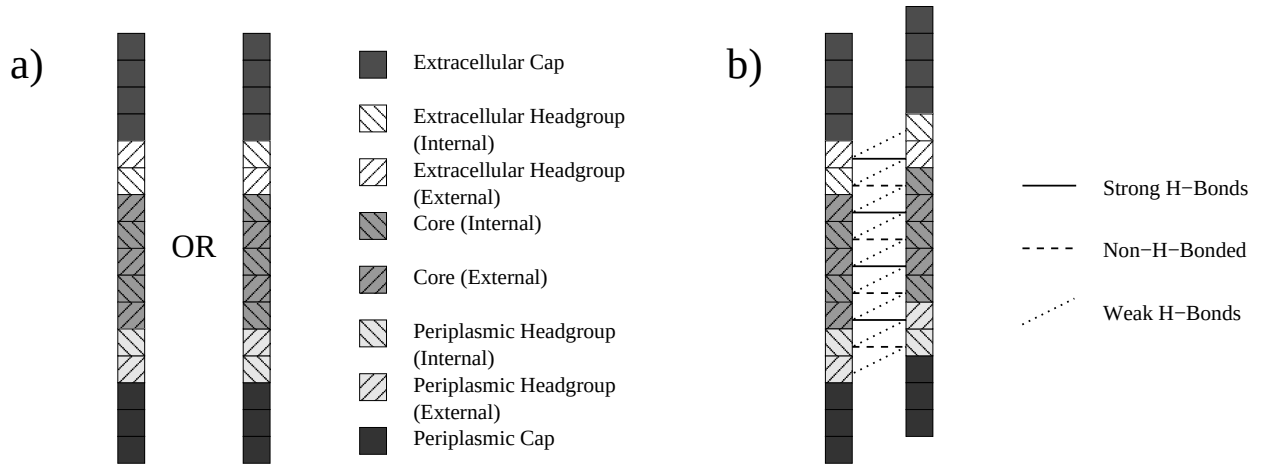

Figure 2: Illustrations of the models for the strand register prediction algorithm. a) 16-residue TM β-strand model. Two models are scored for each window with different internal-external designations (hatch marks). b) 9-residue strand pair model. The example shown has a strand shear of +1. One strand is shifted up or down against the other (fixed) strand to score different strand shears.