**Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006**

FrD03.2

Combinatorial model for sequence and spatial motif discovery in short sequence fragments: Examples from β-barrel membrane proteins

Ronald Jackups, Jr. and Jie Liang

Abstract— **Motifs are over-represented sequence or spatial patterns appearing in proteins. They often play important roles in maintaining protein stability and in facilitating protein functions. When motifs are located in short sequence fragments, as in transmembrane domains that are only 10-20 residues in length, and when there is only very limited data, it is difficult to identify motifs. In this study, we develop combinatorial models for assessing statistically significant sequence and spatial patterns. We show our method can uncover previously unknown sequence and spatial motifs in β-barrel membrane proteins.**

I. INTRODUCTION

The identification of sequence motifs plays an important role in understanding protein stability and function. In many cases, these motifs are embedded in short sequence fragments, as those in the transmembrane domains of membrane proteins, which are usually only 10-20 residues in length. In studies of α-helical membrane proteins, sequence motifs were found to play important roles in the folding and assembly of TM helices. Examples include the well-known GxxxG motifs that promote the dimerization of Glycophorin A [4], and other Small-xxx-Small motifs [3]. Similarly, spatial motifs also play important roles.

Discovery of these motifs is a challenging task, as these sequence fragments are short and the amount of available data is limited. The statistics of sequence motifs in short fragments cannot be approximated by a binomial-type distribution, as in [1]. For spatial motifs, the χ^2 distribution, as used in [5], is also inappropriate.

In [4], Senes *et al* discovered a large number of sequence motifs in transmembrane helices based on exhaustive computer enumeration. Here we develop combinatorial models that give analytical forms of propensity value and p-value calculations for sequence motifs and additionally spatial interaction motifs in β-barrel membrane proteins.

II. MODEL AND METHODS

A. Propensity of intrastrand two-residue sequence patterns.

We use β-strands as our examples of sequence fragments for motif discovery. We introduce the propensity $P(X,Y|k)$ for two ordered intrastrand residues of type X and type Y that are k positions away on the same strand. We define the propensity as:

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

where $f(X,Y|k)$ is the observed frequency of XYk patterns, and $\mathbb{E}[f'(X,Y|k)]$ is the expected frequency of XYk patterns.

In our null model, the residues within each strand are permuted exhaustively and independently, and each permutation occurs with equal probability. An XYk pattern forms if in a permuted strand an X residue happens to be followed by a Y residue at the k-th position down the strand in the N-C direction. $\mathbb{E}[f'(X,Y)]$ is the expected number of XYk patterns over the entire dataset:

$$\mathbb{E}[f'(X,Y|k)] = \frac{xy(l-k)}{l(l-1)}, \qquad (1)$$

where l is the length of the strand, x is the number of residues of type X, and y is the number of residues of type Y.

We can represent $f'(X,Y|k)$ as the sum of identical Bernoulli variables $f'(1,Y|k)$, each of which equals 1 if one of the y residues of type Y is in the k-th position past a specific residue of type X when the strand is randomly permuted, or 0 if the k-th position is not a Y residue. The probability that the residue of type X is placed in one of the first $l-k$ positions is $\frac{l-k}{l}$. If it were placed in one of the last k positions, there would not be enough space for an XYk motif to form. The probability that one of the y residues of type Y is placed in the k-th position past the residue of type X once the latter has been placed is $\frac{y}{l-1}$. Thus,

$$\mathbb{E}[f'(1,Y|k)] = \mathbb{P}_{1,Y|k}(1) = \frac{(l-k)}{l}\cdot\frac{y}{(l-1)}.$$

There are x such identical variables (one for each residue of type X), and the expectation of their sum is the sum of their expectations, leading to Equation (1). For XXk motifs, *i.e.* two residues of the same type displaced by k residues, the expectation is calculated as

$$\mathbb{E}[f'(X,X|k)] = \frac{x(x-1)(l-k)}{l(l-1)},$$

as there will be $x-1$ residues available to form an XXk motif with a specific residue of type X. Although these Bernoulli random variables are dependent (*i.e.* the placement of one XYk motif will affect the probability of another XYk motif), the expectation of their sum is the sum of their expectations, because expectation is a linear operator. However, in order to calculate statistical significance in terms of p-values, special formulae must be derived to determine $\mathbb{P}_{X,Y|k}(i)$, the probability of the occurrence of $i = f'(X,Y|k)$ XYk motifs.

This work is supported by grants from NSF (CAREER DBI0133856), NIH (GM68958), and ONR (N000140310329)

Program in Bioinformatics, Department of Bioengineering, University of Illinois at Chicago, 851 S. Morgan Street, Room 218, Chicago, IL 60607–7052, U.S.A. Corresponding author: Jie Liang,`jliang@uic.edu`

Motifs for residues of different types if $k = 1$. To determine $\mathbb{P}_{X,Y;k}(i)$, we first consider the case where $X \neq Y$ and $k = 1$. Without loss of generality, we take a Y-centric view, and examine the probability of forming i number of X-Y patterns when there are x number of X residues, y number of Y residues, and l residues total in a strand. Altogether, there are $\binom{l}{y}$ different ways to place the y number of Y residues in the l available positions on the β-strand.

Among the y number of Y residues to form i number of $XY1$ motifs, i number of them will be immediately following an X residue. This depends on which i of the x number of X residues we choose to follow, and there are $\binom{x}{i}$ different ways to do so.

After placing the $l-y$ non-Y residues, there are $l-y+1$ positions ("slots" between each of the non-Y residues and at both ends of the strand) to place the Y residues with replacement (*i.e.* multiple Y residues can be placed in one slot). Since $x - i$ of those positions are preceded by the X residues which are not followed immediately by the i Y residues, and since there are exactly i $XY1$ motifs, we have only $(l - y + 1) - (x - i)$ slots in which to place the $y - i$ residues, although we can place multiple Y residues in one slot (*i.e.* with replacement). The number of different ways to achieve this is:

$$\binom{[(l-y+1)-(x-i)]+(y-i)-1}{y-i} = \binom{l-x}{y-i},$$

following the standard formula for choosing objects with replacement but without regard to order.

Combining these three terms, the probability of i $XY1$ motifs in one strand follows a hypergeometric distribution:

$$\mathbb{P}_{X,Y;1}(i) = \frac{\binom{x}{i}\binom{l-x}{y-i}}{\binom{l}{y}},$$

where x is the number of residues of type X in the strand, y is the number of residues of type Y, and l is the length of the strand.

Motifs for residues of the same type if $k = 1$. When $X = Y$ and $k = 1$, the probability of i X-X patterns in one strand follows a different hypergeometric distribution:

$$\mathbb{P}_{X,X;1}(i) = \frac{\binom{l-x+1}{x-i}\binom{x-1}{i}}{\binom{l}{x}} \tag{2}$$

with the conventional notation that $\binom{n}{r} = 0$ if $n < r$.

To illustrate this, first place $x - i$ of the x residues of type X in the $l - x + 1$ slots between the residues that are not of type X *without* replacement, so that no two residues of type X are adjacent. There are$\binom{l-x+1}{x-i}$ ways to do this. Then place the remaining i residues of type X in the $x - i$ slots following the already placed residues of type X *with* replacement so that there are i $XX1$ motifs. Here multiple X residues can be placed at any of the $x-i$ slots. There are $\binom{(x-i)+i-1}{i} = \binom{x-1}{i}$ ways to do this. Since there are $\binom{l}{x}$ total ways to arrange x residues of type X on the strand, the probability is defined by Equation (2).

Motifs for residues of different types if $x \leq 2$ or $y \leq 2$. If either $x = 1$ or $y = 1$, then

$$\mathbb{P}_{X,Y;k}(1) = \mathbb{E}[f'(X,Y;k)] = \frac{xy(l-k)}{l(l-1)},$$

since the maximum possible number i of XYk motifs is 1. This is the same as Equation (1). We also note that it does not matter whether $x = 1$ or $y = 1$, since $\mathbb{P}_{X,Y;k}(i) = \mathbb{P}_{Y,X;k}(i)$. To show this, we can simply reverse the residue sequence of the strand. As a result, it is possible to determine $\mathbb{P}_{X,Y;k}(i)$ for all values of k if the number count of either one of the residue types is 1. When $i = 0$, we have:

$$\mathbb{P}_{X,Y;k}(0) = 1 - \mathbb{P}_{X,Y;k}(1).$$

If $x = 2$ or $y = 2$, the probability of two XYk motifs is:

$$\mathbb{P}_{X,Y;k}(2) = [\binom{l-k}{2} - (l-2k)] \Big/ \frac{l(l-1)(l-2)(l-3)}{x(x-1)y(y-1)}. \tag{3}$$

There are $\binom{l-k}{2}$ positions in which to place two XYk motifs. However, the terminal residue of type Y in the first motif forbids the placement of the initial residue of type X in the second motif in $l - 2k$ cases, in which the initial residue of type X in the first motif is placed in one of the first $l - 2k$ positions of the strand. Thus, there are $\binom{l-k}{2} - (l - 2k)$ possible ways to place two XYk motifs. Since there are $\frac{l(l-1)(l-2)(l-3)}{x(x-1)y(y-1)}$ possible ways to place two residues of type X and two resides of type Y, the probability of exactly two XYk residues is as shown in Equation (3).

Since there can only be a maximum of two XYk patterns when $x = 2$ or $y = 2$, it is possible to determine the probability of exactly one XYk motif or zero motifs using the definition of expectation. Because $\mathbb{E}[f'(X,Y;k)] = \sum_{i=0}^{2} i \cdot \mathbb{P}_{X,Y;k}(i) = 0 \cdot \mathbb{P}_{X,Y;k}(0) + 1 \cdot \mathbb{P}_{X,Y;k}(1) + 2 \cdot \mathbb{P}_{X,Y;k}(2)$ and $\mathbb{P}_{X,Y;k}(0) + \mathbb{P}_{X,Y;k}(1) + \mathbb{P}_{X,Y;k}(2) = 1$, we have:

$$\mathbb{P}_{X,Y;k}(1) = \mathbb{E}[f'(X,Y;k)] - 2\mathbb{P}_{X,Y;k}(2) \tag{4}$$

$$\mathbb{P}_{X,Y;k}(0) = 1 - [\mathbb{P}_{X,Y;k}(1) + \mathbb{P}_{X,Y;k}(2)] \tag{5}$$

Motifs for residues of the same type if $x \leq 3$. If $x = 2$, then the probability of one XXk motif is:

$$\mathbb{P}_{X,X;k}(1) = \mathbb{E}[f'(X,X;k)] = \frac{x(x-1)(l-k)}{l(l-1)},$$

since it is only possible to have one XXk motif. Then:

$$\mathbb{P}_{X,X;k}(0) = 1 - \mathbb{P}_{X,X;k}(1).$$

If $x = 3$, then the probability of exactly two XXk motifs is:

$$\mathbb{P}_{X,X;k}(2) = \frac{l-2k}{\binom{l}{x}},$$

since there are only $l - 2k$ positions in which to place an $X \cdots X \cdots X$ motif (*i.e.* the only way to obtain two XXk motifs if $x = 3$), and $\binom{l}{x}$ ways to place x residues of type X in a strand of length l. It is then possible to determine the remaining probabilities using expectation, as was done in Equations (4) and (5), since at most only two XXk motifs are possible when $x = 3$ (*i.e.* an $X \cdots X \cdots X$ motif, where "$\cdots$" corresponds to $k-1$ residues):

Motifs for residues if $k > 1$, $x > 2$, and $y > 2$. When $k > 1$, $x > 2$, and $y > 2$, the analytical formulae for $\mathbb{P}_{X,Y;k}(i)$ become very complicated. However, because the strands in the dataset are short, it is possible to fully enumerate all permutations of a strand and calculate $\mathbb{P}_{X,Y;k}(i)$ and p-values exactly, as shown by Senes *et al.* [4].

B. Interstrand spatial contact propensities.

To identify spatial motifs, we calculate the interstrand spatial propensity $P(X,Y)$ for contacting pairs of residue types X and Y:

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

where $f(X,Y)$ is the observed frequency of X-Y contacts of a specific type in the strand pair dataset, and $\mathbb{E}[f'(X,Y)]$ is the expected frequency of X-Y contacts in a null model.

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. 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.

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

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

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. For statistical significance, two-tailed p-values can be calculated using the hypergeometric distribution.

Contacts between residues of different types. If the two contacting residues are not of the same type, *i.e.* $X \neq Y$, 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 $sp \in \mathcal{SP}$:

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

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. 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 detailed 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 $n_1 = l - x_1 - y_1$ count of type "neither". If we exhaustively permute the l residues, we have the trinomial coefficient number of different permutations. We denote this as: $T(l, x_1, y_1) \equiv (x_1, y_1, n_1)!$.

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 n_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: $T(l, x_2, y_2) = (x_2, y_2, n_2)!$.

Consider the residues on strand 2 that match to the x_1 number of residues of type X on strand 1. 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 $T(n_1,\ x_2 - h - j,\ y_2 - i - k)$ number of permutations to match the remaining n_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 n_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:

$$\begin{aligned}\mathbb{P}(h, i, j, k) = {} & T(x_1, h, i) \cdot T(y_1, j, k) \\ & \cdot T(n_1, x_2 - h - j, y_2 - i - k)/T(l, x_2, y_2).\end{aligned} \tag{6}$$

This can be illustrated as follows. When randomly picking x_2 of type X residues, y_2 of type Y residues, and the remaining n_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 n_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 $(n_1) - (x_2 - h - j) - (y_2 - i - k)$ 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:

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

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, 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-values for this null model of $f'(X, Y)$ number of observed X-Y contacts.

III. Results

We use the structures of 19 β-barrel membrane proteins with resolution of 2.6 Å or better as in [2] as out dataset, with a total of 262 β-strands. All proteins share no more than 26% pairwise sequence identity. In Table I, we report the pairwise intrastrand sequence motifs we discovered with calculated propensities and p-values. Only motifs with $k = 2$ or $k = 4$ that are significant at the threshold p-value of 0.05 are listed. Detailed biological implications of these motifs, including the AY2 dichotomy, will be published elsewhere.

$k = 2$			$k = 4$		
Pair	Odds	p-Value	Pair	Odds	p-Value
GR	2.14	5.6×10^{-6}	LY	1.90	8.2×10^{-5}
AY	1.75	5.6×10^{-4}	WV	2.79	1.1×10^{-3}
LG	1.63	1.0×10^{-3}	TY	1.93	9.3×10^{-3}
LA	1.61	1.1×10^{-3}	TG	1.68	1.0×10^{-2}
AA	1.47	2.3×10^{-2}	GD	1.86	1.3×10^{-2}
IL	1.73	2.7×10^{-2}	HR	5.27	2.0×10^{-2}
ND	2.16	3.0×10^{-2}	GN	1.88	2.9×10^{-2}
VY	1.43	3.2×10^{-2}	VG	1.60	3.3×10^{-2}
IA	1.60	3.8×10^{-2}	FA	1.75	3.6×10^{-2}
KW	3.31	4.0×10^{-2}	IG	1.75	4.5×10^{-2}
VP	2.24	4.2×10^{-2}			
YQ	1.74	4.9×10^{-2}			

TABLE I

Table II lists pairwise interstrand spatial motifs we discovered, divided into H-bonded and non-H-bonded pairs (see [2] for definitions). Only motifs significant at the threshold p-value of 0.05 are listed. Detailed biological implications of these motifs are described in [2], including the newly discovered "positive-outside" rule and aromatic rescue.

Strong H-Bonds			Non-H-Bonded Interactions		
Pair	Odds	p-Value	Pair	Odds	p-Value
GY	39	8.0×10^{-4}	WY	2.71	4.1×10^{-7}
ND	10	1.2×10^{-3}	GI	1.77	1.2×10^{-2}
GF	18	4.7×10^{-3}	RE	1.87	1.3×10^{-2}
IY	18	4.8×10^{-3}	GV	1.60	1.6×10^{-2}
KS	12	1.2×10^{-2}	QG	1.57	2.3×10^{-2}
LW	10	2.5×10^{-2}	LL	1.44	2.7×10^{-2}
LY	34	2.9×10^{-2}	AV	1.39	3.7×10^{-2}
RP	3	3.1×10^{-2}	LP	2.07	3.7×10^{-2}
AA	16	4.8×10^{-2}			
HK	3	5.0×10^{-2}			

TABLE II

IV. Conclusions

In this study, we have developed exact models for the discovery of sequence motifs from fragments of residues of very short length (about 10), as well as spatial interaction motifs when these fragments form β-strand pairs as in β-barrel membrane proteins. These models are essential for discovery of biologically important motifs when only very limited data is available, as in β-barrel membrane protein structures. Our results show that a number of significant motifs can be successfully uncovered, and the results can be used to understand the mechanisms of membrane protein folding and to predict membrane protein structures.

References

[1] R. Hart, *et al.*, *J. Comput. Biol.*, vol. 7, pp. 585–600, 2000.

[2] R. Jackups and J. Liang, *J Mol Biol*, vol. 354(4), pp. 979–93, 2005.

[3] A. Senes, *et al.*, *Curr Opin Struct Biol.*, vol. 14, pp. 465–479, 2004.

[4] A. Senes, *et al.*, *J Mol Biol.*, vol. 296, pp. 921–936, 2000.

[5] M. Wouters and P. Curmi, *Proteins*, vol. 22(2), pp. 119–31, 1995.