I. Appendix

This is the appendix to *Ronald Jackups, Jr and Jie Liang.(2008) Combinatorial analysis for sequence and spatial motif discovery in short sequence fragments. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Accepted*

A. Positional null model

The other motif analyses we have developed are based on an *internally random* null model in which the residues within each sequence are permuted, and each permutation is equally likely. This assumption can be problematic in certain cases where there are biases of residue types for certain positions in a sequence known *a priori*. For instance, aromatic residues tend to be favored at either end of a transmembrane α-helix or β-strand [1–3]. These single-residue biases may confound two-residue propensities without providing additional information into the preferences of these patterns. When such biases are known, it may be helpful instead to consider a null model that accounts for them.

We therefore introduce a *positional null model*. Instead of permuting residues across all positions within individual sequences, we permute residues across all sequences in a dataset

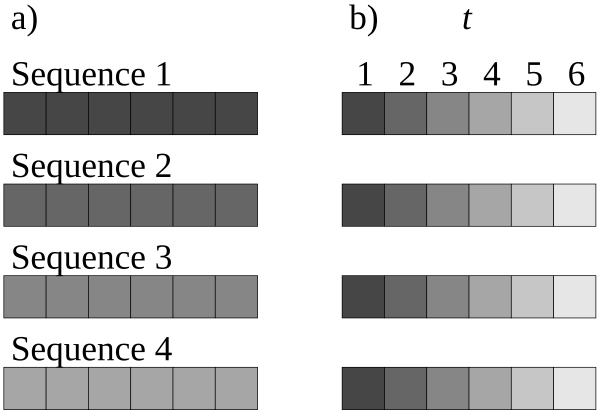

Fig. 1. Difference between a) an *internally random* null model for sequence motif analysis and b) a *position-dependent* null model. In both cases, only residues of the same shade are permuted with each other. In a), residues are permuted only within each sequence individually, while in b), residues are permuted across sequences but only within their specified position t.

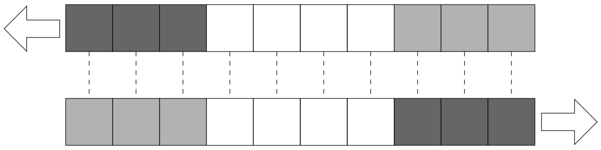

Fig. 2. An example of interacting antiparallel β-strands for use with a positional null model, as described in the text. The three regions are *N-terminal* (gray), *C-terminal* (black), and *core* (white). Arrows represent the N-to-C direction. The two spatial pairing types are N-terminal with C-terminal and core with core. The N-terminal with C-terminal pairing type is an example of paired residues from different regions ($r \neq s$), while the core with core pairing type is an example of paired residues from the same region ($r = s$). These pairing types do not overlap: N-terminal residues may only pair with C-terminal residues, and core residues may only pair with other core residues.

within specific positions (Figure 1). We have adapted this null model for both spatial and sequence motifs.

1) Positional null model for spatial interaction pairs: When calculating position-dependent propensities for spatial motifs, we need a meaningful definition of position. Here, we allocate residues into regions. Although these regions do not have to be the same length along a sequence, interacting regions within a sequence pair must have equal length, and regions may not overlap. In other words, if a residue in region r interacts with a residue in region s on a spatially adjacent sequence fragment, all residues in region r in the dataset must only interact with residues in region s. It is possible for residues in the same region to interact with each other, as long as no residue in that region interacts with any other region in the dataset. For example, for interacting antiparallel β-strands, we may divide each strand into three regions, the N-terminal, central core, and C-terminal regions, and all interacting strand pairs into two spatial pair types, N-terminal with C-terminal and core with core (Figure 2). This would require that no core residue interact with an N-terminal or C-terminal residue.

The null model for position-dependent spatial motifs differs depending on whether paired residues are from the same region ($r = s$) or different regions ($r \neq s$), and whether the residue types in the pair are the same ($X = Y$) or different ($X \neq Y$).

Null model when $r = s$ and $X = Y$.

Let n_r be the number of residues in region r. Because residues in region r may only interact with other residues in region r, there will be $\frac{n_r}{2}$ residue pairs in region r. The probability that

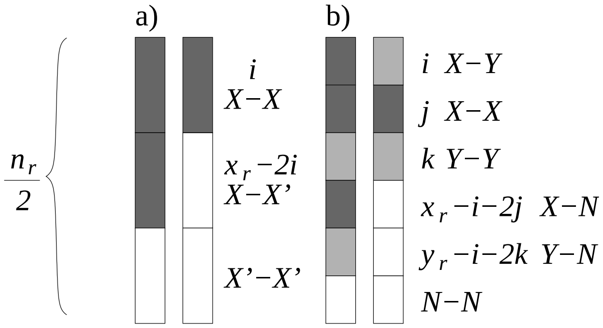

Fig. 3. Division of residue pair types in the position-dependent spatial motif null model when $r = s$ and a) $X = Y$ or b) $X \neq Y$. Black = X, gray = Y, white = X' (not X) or N (neither X nor Y).

an arbitrary pair in region r will be of type X-X is the number of ways to choose 2 of the x_r residues of type X in region r divided by the number of ways to choose 2 of the n_r residues of all types in region r. Thus, the expected value of X-X pairs in region r is:

$$\mathbb{E}(X, X|rr) = \frac{\binom{x_r}{2}}{\binom{n_r}{2}} \cdot \frac{n_r}{2} = \frac{x_r(x_r-1)}{2(n_r-1)}.$$

To determine $\mathbb{P}_{XX|rr}(i)$, the probability of i X-X pairs in region r in the dataset, from which p-values may be calculated, we first introduce the 3-element multinomial function:

$$M(a, b, c) \equiv \frac{a!}{b!c!(a-b-c)!},$$

where $M(a, b, c) = 0$ if $a - b - c < 0$. We visualize region r in the dataset as two interacting columns of length $\frac{n_r}{2}$. When $X = Y$, we must assign the x_r number of X residues in r to each column. There are $\binom{n_r}{x_r}$ ways to do this. In order to obtain exactly i X-X pairs, the $\frac{n_r}{2}$ residue pairs must be divided into 3 distinct groups: i X-X pairs, $x_r - 2i$ X-X' pairs, and $\frac{n_r}{2} - x_r + i$ X'-X' pairs, where X' is any residue other than X (Figure 3a). There are $M(\frac{n_r}{2}, i, x_r - 2i)$ ways to do this. Finally, there are two ways to place each of the $x_r - 2i$ X-X' pairs, depending on which column contains the X residue. Multiplying these factors together, we obtain:

$$\mathbb{P}_{XX|rr}(i) = \frac{M(\frac{n_r}{2}, i, x_r - 2i) \cdot 2^{x_r-2i}}{\binom{n_r}{x_r}}.$$

Null model when $r = s$ and $X \neq Y$.

The expected value when $X \neq Y$ is similar to that when $X = Y$, except that there are $x_r y_r$ ways to choose an X-Y pair:

$$\mathbb{E}(XY|rr) = \frac{x_r y_r}{\binom{n_r}{2}} \cdot \frac{n_r}{2} = \frac{x_r y_r}{n_r - 1}.$$

However, $\mathbb{P}_{XY|rr}(i)$, the probability of i X-Y pairs in the dataset, is more complicated. We must first introduce a six-variable multinomial function:

$$M(a,b,c,d,e,f) \equiv \frac{a!}{b!c!d!e!f!(a-b-c-d-e-f)!}$$

Again, we visualize region r as two interacting columns of length $\frac{n_r}{2}$. We must assign the x_r number of X residues and the y_r number of Y residues in region r to each column. There are $M(n_r, x_r, y_r)$ ways to do this. In order to obtain exactly i X-Y pairs, the $\frac{n_r}{2}$ residue pairs must be divided into 6 distinct groups: i X-Y pairs, j X-X pairs, k Y-Y pairs, $x_r - i - 2j$ X-N pairs, $y_r - i - 2k$ Y-N pairs, and $\frac{n_r}{2} - x_r - y_r + i + j + k$ N-N pairs, where N, “neither,” represents any residue other than X or Y (Figure 3b). There are 3 degrees of freedom, i, j, and k, as long as none of the six quantities is negative. There are $M(\frac{n_r}{2}, i, j, k, x_r - i - 2j, y_r - i - 2k)$ ways to distribute the residues into these pairs, and two ways to place each pair of type X-Y, X-N, or Y-N, depending on which column contains which residue type, comprising a total of $x_r + y_r - i - 2j - 2k$ pairs. The probability of each combination of i, j, and k is then:

$$\mathbb{P}(i,j,k) = \frac{M(\frac{n_r}{2}, i, j, k, x_r - i - 2j, y_r - i - 2k) \cdot 2^{x_r + y_r - i - 2j - 2k}}{M(n_r, x_r, y_r)}$$

The marginal probability distribution function for i X-Y pairs, then, is the sum over all possible values of j and k:

$$\mathbb{P}_{XY|rr}(i) = \sum_{j=0}^{\frac{x_r - i}{2}} \sum_{k=0}^{\frac{y_r - i}{2}} \mathbb{P}(i,j,k).$$

The summation limits ensure that none of frequencies of the six pair types is negative.

Null model when $r \neq s$.

Whether $X = Y$ or $X \neq Y$, the null model when $r \neq s$ follows the same distribution. A distinction must be made between X_r, a residue of type X occurring in region r in one sequence, and X_s, a residue of type X occurring in region s in the other sequence. Thus, an X-Y pair, which we define as an X_r-Y_s pair, is different from a Y-X pair, which is Y_r-X_s. Because there is a one-to-one correspondence between residues in region r and region s, $n_r = n_s$ is the total number of r-s pairs.

In order for exactly i X-Y pairs to occur, i X_r residues must be drawn from a possible x_r residues of type X to match i Y_s residues drawn from a possible y_s residues of type Y. This situation can be modeled with a simple hypergeometric distribution:

$$\mathbb{P}_{XY|rs}(i) = \frac{\binom{x_r}{i}\binom{n_r - x_r}{y_s - i}}{\binom{n_r}{y_s}}.$$

The expected value can be calculated from this distribution:

$$\mathbb{E}(XY|rs) = \frac{x_r y_s}{n_r}.$$

2) Positional null model for sequence pairs: We first define the *positional residue frequency* x_t as the number of residues of type X occupying the t-th position of all sequences in the dataset. If all sequences in a dataset are the same length, then all positions t will have the same total number of residues of all types, which is also the number of sequences in the dataset. If different lengths are represented in the dataset, it is necessary to normalize t to be within an appropriate range $[1, l]$, to approximate an average or predetermined sequence length of l:

$$t = \lceil \frac{l(t_{obs} - 0.5)}{l_{obs}} \rceil,$$

where $t_{obs} \in \{1, 2, 3, \cdots, l_{obs}\}$ is the actual position of the residue within its sequence, l_{obs} is the actual length of the sequence, $\lceil x \rceil$ represents the ceiling function, equal to the lowest integer greater than or equal to x, and the 0.5 factor is a correction for continuity to round to the next integer. This ensures that $1 \leq t \leq l$, no residues are removed from the model by truncation, and each position t will be represented by nearly the same number of residues.

In order to calculate position-dependent sequence propensities, we use permutation within each position in a sequence *with replacement* across all sequences. Although all of the other null models in this study rely on permutation without replacement, such permutation would be complex for this null model. Since this model is based on datasets of multiple sequences instead of individual sequences, the approximation of sampling without replacement will not be problematic as long as a large enough sample of sequences is used.

The probability that an XYk pattern appears at position t is 0 if $t > l - k$, because an XYk pattern at that position would span across the end of a sequence of length l. Otherwise, the probability of an XYk pattern at position t is the probability of a residue of type X being placed at position t multiplied by the probability of a residue of type Y being placed at position $t + k$:

$$\mathbb{P}(X, Y|k, t) = \frac{x_t}{n_t} \cdot \frac{y_{t+k}}{n_{t+k}},$$

where x_t is the number of residues of type X in position t on all sequences, y_t is the number of residues of type Y in position t, and n_t is the number of all residues of all types in position t. This null model can be represented as a binomial distribution, such that the probability of i XYk patterns at position t in the dataset is:

$$\mathbb{P}_{XYk|t}(i) = \binom{n_t}{i} \mathbb{P}(X, Y|k, t)^i [1 - \mathbb{P}(X, Y|k, t)]^{n_t - i},$$

and the expected value is:

$$\mathbb{E}[f(X, Y|k, t)] = n_t \cdot \mathbb{P}(X, Y|k, t).$$

If instead we are interested in the dataset-wide probability of an XYk pattern at any arbitrary position of the sequence, we must calculate the average of $\mathbb{P}(X, Y|k, t)$ over all $l - k$ possible positions:

$$\mathbb{P}(X, Y|k) = \frac{1}{l - k} \sum_{t=1}^{l-k} \mathbb{P}(X, Y|k, t).$$

This null model can similarly be represented as a binomial distribution with probability distribution function:

$$\mathbb{P}_{XYk}(i) = \binom{n_k}{i} \mathbb{P}(X, Y|k)^i [1 - \mathbb{P}(X, Y|k)]^{n_k - i},$$

where n_k is the number of all pairs of all residue types k residues apart in the dataset. The expected value is then:

$$\mathbb{E}[f(X,Y|k)] = n_k \cdot \mathbb{P}(X,Y|k).$$

Unlike the situation where only one position t is concerned, this distribution represents the sum of dependent Bernoulli variables. Methods of accounting for this dependence can be found in Robin *et al.* [4].

REFERENCES

[1] W. C. Wimley, Toward genomic identification of β-barrel membrane proteins: composition and architecture of known structures, Protein Sci. 11 (2002) 301–312.

[2] G. von Heijne, Membrane proteins: from sequence to structure., Annu Rev Biophys Biomol Struct. 23 (1994) 167–192.

[3] R. Jackups Jr, J. Liang, Interstrand pairing patterns in beta-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction, J Mol Biol. 354 (2005) 979–993.

[4] S. Robin, F. Rodolphe, S. Schabth, DNA, words, and models: Statistics of exceptional words, Cambridge University Press, 2005.