COS 597c: Topics in Computational Molecular Biology
Lecture 10: October 20, 1999
Lecturer: Mona Singh
Scribes: Dawn Brooks[1]

Motifs and Profile Analysis

Broadly speaking, a sequence motif is a conserved element of a sequence alignment. Its function or structure may be known, or its significance may be unknown. Thus, one way to get functional or structural information about a sequence is to determine what motifs it contains. In a sense, we have already talked about this problem in the context of detecting sequence similarity using pairwise and multiple sequence alignments. Today we will discuss another method, where we build profiles (which also have been called, among other things, weight matrices, templates, and position specific scoring matrices) for motifs or sequence families. There are many websites that contain collections of amino acid sequence motifs that indicate particular structural of functional elements; searches of these websites with a new sequence allow inferences about its potential functional roles. Such websites include BLOCKS, PRINTS and Pfam.

Unlike the alignment methods we have talked about thus far, intutively, profile analysis uses the fact that certain positions in a family are more conserved than other positions, and allows substitutions less readily in these conserved positions. The latest approaches to building profiles are based on hidden Markov models, which we will talk about in a subsequent lecture.

We will restrict our discussion below to protein families, but the same techniques are also applicable to DNA sequences. (For example, profile analysis has also been applied to recognizing promoter sequences.)

Profiles for Aligned Sequences

The general framework for profile analysis is: given subsequences that are members of a family in interest, devise methods to determine whether a new sequences is a member of this family.

[1]Portions of the notes are adapted from lecture notes originally scribed by Xuxia Kuang in Fall 1998.

Here is an overview of the profile method:

- Align the sequences in the family.
- Use the alignment to create a profile.
- Test new sequences against the profile.

1. Initially, we will assume that there are no gaps in the alignment. We look at the alignment of N sequences of l positions as follows:

	Position					
Sequence	1	2	3	4	...	l
1	a_{11}	a_{12}	a_{13}	...	...	a_{1l}
2	a_{21}	a_{22}	a_{23}	...	...	a_{2l}
3	a_{31}					
.						
.						
.						
N	a_{N1}	a_{N2}	a_{N3}	...	...	a_{Nl}

   where a_{ij} denotes the amino acid from the ith sequence at the jth position.

2. We build the profile as follows. We compute:

$$\begin{aligned} f_{ij} &= \% \text{ of column } j \text{ that is amino acid } i \\ b_i &= \% \text{ of "background" which is amino acid } i \end{aligned}$$

   The "background" can be computed, for example, from a large sequence database, or from a genome, or from some particular protein family.

   Now compute the $20 \times l$ array P_{ij}, where

$$P_{ij} = \frac{f_{ij}}{b_i} \tag{1}$$

   Intuitively, P_{ij} is the "propensity" for amino acid i in the j position in the alignment.

   This gives us the following table:

	Position						
Sequence	1	2	3	4	5	...	l
L	P_{L1}	P_{L2}	P_{L3}	...	...		P_{Ll}
V	P_{V1}	P_{V2}	P_{V3}	...	...		P_{Vl}
F	P_{F1}						
.							
.							
.							

And we use this table to compute:

$$Score_{ij} = \log(P_{ij}) \tag{2}$$

For example, say we have the following alignment ($N = 4, l = 4$):

LEVK
LDIR
LEIK
LDVE

Assume for simplicity that all amino acids are equally likely in the in the background (e.g., the fraction of amino acid i in some large database is $\frac{1}{20}$). We have

$$P_{L1} = \frac{\frac{4}{4}}{\frac{1}{20}} = 20,\ P_{D2} = \frac{\frac{2}{4}}{\frac{1}{20}} = 10\ ...$$

3. To use the profile to score a new sequence, we do the following:
   (a) Slide a window of width l over the new sequence.
   (b) The score of the window equals the sum of the scores of each position in the window.

   For example, sequence $\boxed{LEVE}\,ER$,

   $$\text{Score of the first window} = Score_{L1} + Score_{E2} + Score_{V3} + Score_{E4}$$

   $$\text{Score of the second window} = Score_{E1} + Score_{V2} + Score_{E3} + Score_{E4}$$

   ...

   (c) If the score of the window is higher than the cutoff, which is determined empirically, we can conclude that the window is a member of the family. In addition, the higher the score, the more confident the prediction.

Simple probabilistic interpretation

Note that the profile method just described can be justified from a log-odds perspective. In particular, when scoring a subsequence, profile methods assume that each position is independent, and estimate:

$$\log\left(\frac{\Pr(\text{subsequence}|\text{family})}{\Pr(\text{subsequence}|\text{not family})}\right)$$

You can show this simply by repeated application of the definition of conditional probability, and by using our assumption of position independence. Probabilities from the numerator are estimated from frequencies for each amino acid in each position of the alignment, and probabilities in the denominator are estimated from the background frequencies.

Common Extensions to Profile Methods

There are several variations to the method we just described.

1. We can modify the profile method to incorporate gaps.
   - We now have 21 $\times l$ matrix, where l is the length of the alignment.
   - Gap costs vary at different positions. For example, we tend to see more gaps at certain positions in loops of protein structures.

2. The most important extension is how to handle the zero frequency case. In practice, handling the zero frequency case is essential for good performance.

   If an amino acid does not occur in a column, P_{ij} is zero and $Score_{ij} = \log(0)$, which is undefined. One common approach for dealing with this problem is to calculate P_{ij} as:

   $$P_{ij} = \frac{\frac{\#\text{ of amino acid } i \text{ in position } j + b}{N + 20b}}{\text{fraction of amino acid } i \text{ in GenBank}} \tag{3}$$

   where b is usually some small value (e.g. $b \leq 1$). This is often known as the "pseudo-count" method, and can be justified in a Bayesian framework. There are many other methods to deal with the zero-frequency case.

3. When several sequences are closely related, we do not want to overweigh these sequences; instead, we want to weigh more remote homologies equally. Thus, once common extension to the profile method is to weight the sequences used in building the profile.

   If sequence k is weighted $\mathcal{W}_k$ and $\sum_{k=1}^{N} \mathcal{W}_k = W$, then

   $$P_{ij} = \frac{\frac{\sum_{k=1}^{N} \mathcal{W}_k \cdot \delta_{k_{ij}}}{W}}{\text{fraction of amino acid } i \text{ in Genbank}} \tag{4}$$

   where $\delta_{k_{ij}} = 1$ if $a_{kj} =$ residue i, and 0 otherwise.

   (Sanity check: note that if each sequence is weighted 1, i.e. $\mathcal{W}_k = 1$, we get back the original formula.)

4. Sometimes profile methods incorporate known substitution matrices used for alignments. (The PAM250 matrix is a common example.) If δ is our substitution matrix, we compute:

   $$Score_{ij} = \sum_{d=1}^{20} \delta(i,d) \times \frac{\#\text{ of amino acid } d \text{ in position } j \text{ of alignment}}{N}, \tag{5}$$

   for $i \in \{L, V, ...\}$ and $1 \leq j \leq l$

   This variation is actually how profile methods were first described. Note that this formula gives an alternate way to handle the zero-frequency case; however, it does not fit into the log-odds perspective.

References

[1] R. Durbin, S. Eddy, A. Krogh and G. Mitchison. *Biological Sequence Analysis.* Cambridge University Press, 1998.

[2] M. Gribskov, A. D. McLachlan and D. Eisenberg. (1987) Profile Analysis: Dectection of distantly related proteins. *Proc. Natl. Acad. Sci., U S A* 84(13): 4355-8.