Local multiple alignment

Sep. 27, 2011

Local multiple sequence alignment involves the discovery, modeling, and recognition of conserved patterns or motifs in multiple (and potentially very many) DNA or protein sequences.

- In *discovery*, we are given *unlabeled* sequences. The task is to identify one or more shared, conserved motifs in these sequences. In machine learning terms, this is equivalent to *labeling* the sequences. For example, each symbol in a sequence might be labeled "1", if it is in the conserved pattern, and "0", if it is not. More complex labeling schemes representing more than one motif or substructure within a motif are also possible.

- In *modeling*, we are given a local multiple alignment, or labeled sequences, as input. The task is to construct a probabilistic model that represents the properties of each column in the alignment (i.e., the symbols we are likely to observe at that position) in an efficient manner and that can be used for searching for new instances of the pattern.

- In *recognition*, we are given a new unlabeled sequence containing zero or more instances of the motif of interest. A probabilistic model of the motif is used to search the unlabeled sequence. The location and extent of each motif identified is reported.

Last Thursday, we discussed *Position Specific Scoring Matrices (PSSM's)*, a formalism for *modeling* ungapped local alignments. Given an ungapped local alignment, A, representing a motif of width w in k sequences, the frequency of symbol $i \in \Sigma$ at position j in the alignment is

$$q[i,j] = \frac{c[i.j] + b}{k + b \cdot |\Sigma|}, \tag{1}$$

where $c[i,j]$ is the number of i's at position j and b is a pseudocount. Note that A is a $k \times w$ matrix and $q[i,j]$ is a $|\Sigma| \times w$ matrix. From this, we obtain the propensity matrix

$$P[i,j] = \frac{q[i,j]}{p[i]}, \tag{2}$$

where $p[i]$ is the background distribution of symbol i. The log odds scoring matrix is

$$S[i,j] = \log_2 P[i,j]. \tag{3}$$

Today, we will discuss the *Gibbs sampler*, an algorithm for *discovery* of ungapped local alignments that uses the PSSM formalism as its basic data structure. The application of the Gibbs sampler for motif finding in biomolecular sequences was proposed first by Chip Lawrence and his colleagues in 1993 (Lawrence et al., Science. 1993 262(5131):208-14.) The Gibbs sampler is a general method for estimating a joint probability distribution by repeated calculations of a conditional distribution. For those interested in a general introduction to the Gibbs sampler in a statistical context, *Explaining the Gibbs sampler*, G. Casella & E. I. George, The American Statistician, 46:167-174, 1992, is listed under "optional readings" on the syllabus page. This is not required for the course.

PSSM's and the Gibbs sampler are suitable for ungapped motifs only. The *Hidden Markov Model (HMM)* is a formalism that can be used for both modeling and discovery of patterns that contain gaps. We will discuss HMM's immediately following the Gibbs sampler.

Gibbs sampler for motif discovery

The Gibbs sampler takes as input k sequences, $t_1 \ldots t_k$, of lengths $n_1 \ldots n_k$. The output is a set of k subsequences, one in each input sequence, that are "most similar" to each other. Here, our measure of "most similar" is a likelihood function derived from the propensity matrix, P, defined in equation 2. Note that the Gibbs sampler assumes that the sequences share an ungapped pattern of length w and that each sequence contains exactly one instance of this pattern. The length of the pattern, w, must either be supplied by the user or determined during the discovery process.

A brute force approach to identifying such a pattern is exhaustive enumeration: Consider all possible sets of indices $\{i_1 \ldots i_k\}$, where $1 < i_j < n_j - w + 1$, and score the local alignment consisting of the k subsequences of length w, $t_1[i_1 \cdots (i_1 + w - 1)]$, $t_2[i_2 \cdots (i_2 + w - 1)]$, and so on. The alignment of the highest scoring pattern is then reported. The computational cost of this approach is prohibitive for all but the smallest problem instances. The Gibbs sampler is a more efficient approach to searching the space of all possible motifs, which does not require that all possible alignments be considered. Another probabilistic search procedure called *Expectation maximization (EM)* can also be used to identify conserved, ungapped motifs. We will discuss EM briefly in the context of HMM's later in the course. EM is discussed in detail in 03-712.

Algorithm: Gibbs

Input:

Sequences $t_1, \ldots, t_k$ of lengths $n_1, \ldots, n_k$.

Initialization:

$t^* = t_1, n^* = n_1$ # t1 is the special sequence.

for ($x = 2$ to k) $index[x\text{–}1] = x$ # index of non-special sequences

for ($y = 1$ to $k - 1$) {

$x = index[y]$

$i_x = rand(1, n_x\text{–}w\text{+}1)$ # Guess starting positions

$A'[y, 1 \cdots w] = t_x[i_x \cdots (i_x\text{+}w\text{–}1)]$

}

Calculate $P[i, j]$, the propensity matrix of A' with pseudocounts

Search for pattern:

Repeat {

for ($i = 0$ to $(n^*\text{–}w)$)

$$pdf[i+1] = \frac{\prod_{j=1}^{w} P[t^*[i\text{+}j], j]}{\sum_{l=0}^{n^*\text{—}w} \prod_{j=1}^{w} P[t^*[l\text{+}j], j]}$$

}

With probability $pdf[i]$, $i^* = i$ # Select starting position in t^*

$y = rand(1, k\text{–}1)$ # Select new special sequence

$r = index[y]$

$A'[y, 1 \cdots w] = t^*[i^* \cdots (i^*\text{+}w\text{–}1)]$

$t^* = t_r$; $n^* = n_r$

Calculate $P[i, j]$, the propensity matrix of A' with pseudocounts

} until($P[\cdot, \cdot]$ stops changing)

Obtain A by adding $t^*[i^* \cdots (i^*\text{+}w\text{–}1)]$ to A'

Compute the log odds scoring matrix, S, from A.

Output:

Local multiple sequence alignment A with scoring matrix S.

In the above algorithm, the matrices P and S are the propensity and log odds matrices defined in equations 2 and 3. The notation $t[i \cdots j]$ is used as shorthand for the substring of a given string, t, starting at position i, up to and including position j. Note that A' and P are $(k\text{–}1) \times w$ matrices, whereas the output matrices A and S are $k \times w$ matrices. The use of pseudocounts when calculating P and S is recommended to ensure all characters are represented.

Comments

There are various potential pitfalls associated with the Gibbs sampler, as with any algorithmic attempt to discover biological "truth". Problems can arise if a sequence has no copy of the pattern or has more than one copy. Alternatively, you could find a statistically or biologically meaningful pattern that is not the pattern you are looking for.

Using this algorithm to obtain meaningful solutions requires a number of decisions that are not programmatically determined and require *ad hoc* solutions, possibly guided by the user's "biological intuition":

- Selecting the window size, w
- Selecting the starting configuration
- Selecting values for pseudocounts
- Termination condition: how should the algorithm decide when to stop?

These issues are discussed in greater detail in Lawrence et al. (1993), which is available via the "optional readings" column of the syllabus.

Convergence: The Gibbs sampler models the search for an optimal local alignment as a Markov Chain, in which each state is a set of k subsequences of length w. It can be shown that this Markov Chain has a stationary distribution and that the state corresponding to the most likely pattern has high probability in that distribution. In theory, this process is guaranteed to converge to the optimal solution, given "enough time". In practice, the sampler can get stuck in local optima. An approach to avoiding this problem is to run the procedure several times with different starting configurations. This is discussed in greater detail in the materials listed under "optional reading".