PROMOCO: a New Program for Prediction of *cis* Regulatory Elements:

From High-Information Content Analysis to Clique Identification

Guojun Li, Jizhu Lu, Victor Olman and Ying Xu
CSBL, Department of Biochemistry and Molecular Biology, the University of Georgia, GA30602, USA

{guojun, jlu, olman, xyn}@csbl.bmb.uga.edu

Abstract

We present a computational study for prediction of cis regulatory elements. We model the problem as follows. Each set of conserved binding motifs, evolved from one common ancestor, have a short (Hamming) distance from this ancestor. The problem is to identify a set of l-mers from a given set of promoter sequences which have at most k different positions from the to-be-identified ancestor. A number of papers published in the past attempt to solve this challenging problem. Although the putative ancestor is unknown, even it does not appear in whole background database, we may assume that an instance of it at hand since we can guess it. Our main contribution in this paper is to develop an algorithm, named PROMOCO (PROfile Motif Collection), to find a profile containing all the motifs and relatively small number of random l-mers so that the consensus of the profile would be the putative ancestor. The key idea of the PROMOCO algorithm lies in a new distance measure.

Two classes of computational approaches have been developed and widely used for prediction of cis regulatory elements. One class of methods essentially treat the identification problem of cis regulatory elements as identification of a group of l-mers that exhibit high-information content when aligned. Another class of methods solve the problem through identification of cliques in a graph representation of l-mers where a pair of l-mers are linked by an edge if and only if their (Hamming) distance is below some predefined threshold. While intuitively similar, the detailed relationship between these two classes of algorithms has not been carefully investigated.

We present a computational study for prediction of cis regulatory elements. We model the problem as follows. Each set of conserved binding motifs, evolved from one common ancestor, have a short (Hamming) distance from this ancestor. The problem is to identify a set of l-mers from a given set of promoter sequences which have at most different positions from the to-be-identified ancestor; (l,k)-motifs have been used to represent such a set of motifs. Rigorously solving (l,k)-motif problem has proved to be challenging - the general problem has been shown to be NP-hard. The difficulty comes mainly from the fact that the putative (to-be identified) ancestor is unknown, and the distances between this to-be-identified ancestor and the binding motifs in the provided promoter sequences are unknown. A number of papers have been published which attempt to solve this challenging problem. Based on an observation relating the distances of (l,k)-motifs to the unknown ancestor and the pairwise distances among the (l,k)-motifs, Pevzner and Sze showed that the problem of solving the (l,k)-motif problem is equivalent to solving a q-member clique problem (or q-clique) in a graph defined on the subsequences with length l, where q is the size of the set of motifs that are within distance k to the to-be-identified ancestor.

Consider a set of l-mers which are within distance k to the putative ancestor which we tend to find. We first establish the lower and upper bounds of information content of the profile of the l-mers.

Theorem 1 Given are n l-mers $s_1, s_2, \ldots, s_n$, and an integer $k \le \frac{3}{4}l.$ If there is an l-mer s_0 with $d_H(s_0, s_i) \le k$ for all $i \in [1,n],$ then

$$2l \ge IC(s_1, s_2, \ldots, s_n) \ge (l-k)\log_2 \frac{l-k}{l} + k\log_2 \frac{k}{3l} + 2l,$$

Where $d_H()$ represents the Hamming distance.

Let s_0 be a putative l-mer. $N_k(s_0, \mathcal{S}) = \{W \in \mathcal{S} \mid d_H(s_0, W) \le k\}.$ If we are lucky that the putative ancestor, say s_0 , of length l

disclose to us, $N_k(s_0, \mathcal{S})$ may be supposed to be a profile of motifs. Unfortunately, the putative ancestor could not be known even it does not appear in background at all. However, we may assume that an instance of it is at hand since we can guess it. Our main contribution in this paper is to develop an algorithm, named PROMOCO (PROfile MOtif COllection), to find a profile containing all the motifs and relatively small number of random l-mers so that the consensus of the profile would be the putative ancestor.

The key idea of the algorithm lies in a new distance measure. Let $L = \{1, 2, \ldots, l\}$ be the set of sequential positions of an l-mer, Q a subset of L , and $\overline{Q}$ the complement set of Q .

$$d_Q(W_1, W_2) = |\{i \in Q \mid W_1[i] = W_2[i]\}| + |\{i \in \overline{Q} \mid W_1[i] \neq W_2[i]\}|,$$

Where $W[i]$ represents the i^{th} letter of W . Then we define the (Q,k)-pseudo-neighborhood of a word W_0 as follows.

$$N_{Q,k}(W_0, \mathcal{S}) = \{W \in \mathcal{S} \mid d_Q(W_0, W) \leq k\}.$$

Theorem 2 Let P_1 be a motif to be identified, which has evolved from a putative ancestor P with t substitutions, Q be the set of t mutated positions, and k be an integer $\geq t$. Then

$$N_{Q,k}(P_1, \mathcal{S}) \supseteq N_k(P, \mathcal{S}).$$

It should be emphasized that the reason why our algorithm works well lies in the number of random words included $N_{Q,k}(P_1, \mathcal{S})$ is relatively small. Intuitively, $N_{Q,k}(P_1, \mathcal{S})$ and $N_k(P, \mathcal{S})$ have a same consensus which should be the to-be-identified ancestor. Theorem 2 lays the foundation of the PROMOCO algorithm.

Pseudo of PROMOCO algorithm:

Input: $s_1, s_2, \ldots, s_n$.

Output: all the profiles $\mathcal{S}$ for all l.

```
for $l$ : from 5 to 25 do
    $W^l = W_1, Q^l = \phi.$
    For $j$ : from 1 to $L - l + 1$ do
        If $\sigma_0^l(W_j)$ ¡Ý $\frac{l}{3n}$, then skip the current reference word $W_j$ and $\sigma^l(W_j) =$ å.
          for $q$: from 0 on do
            $\sigma^l(W_j) = \min\{\sigma_q^l(W_j), \sigma^l(W_j)\},$
            $Q_j^l, q = q + 1,$
            where $\sigma^l(W_j) = \sigma_{Q_j^l}(W_j),$
        if $W^l = W_{j\ 1}, Q^l = Q_{j\ 1}, j = 1, j + 1,$ then
          $W^l = W_{j\ 1}, Q^l = Q_{j\ 1}, j = j + 1,$
        else $W^l = W_j, Q^l = Q_j, j = j + 1,$
    $\mathcal{S}^l = \{s_i^l, i = 1,2,\ldots,n\},$
    where $d_{Q^l}^l(W^l, s_i) = d_{Q^l}^l(W^l, s_i^l).$
```

FEFERENCE:

[1] Baily, T. and Elkan, C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Machine Learning, 21, 51-80.

[2] Liu X, Brutlag DL, Liu JS. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomputing, pp. 127-38.

[3] Frances, M., and Litman, A. (1997) On covering problems of codes, Theoret. Comput. Syst., 30, 113-119.

[4] Hertz, G. and Stormo, G. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics, 15, 563-577.

[5] Keich, U., Pevzner, P. A. (2002) Finding motifs in the twilight zone, Bioinformatics, Vol.18, 1374-1381.

[6] Keich, U., Pevzner, P. A. (2002) Subtle motifs: defining the limits of motif finding algorithms, Bioinformatics, Vol.18,1382-1390.

[7] Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A. and Wootton, J. (1993) A Gibbs Sampling Strategy for Multiple Alignment, Science, 262, 208-214.