58. Alignment of Promoter Regions by Mapping Nucleotide Sequences into Arrays of Transcription Factors Binding Motifs

E. Blanco[1,2] eblanco@imim.es
X. Messeguer[2] peypoch@lsi.upc.es
R. Guigó[1,⋆] rguigo@imim.es

1: Grup de Recerca en Informàtica Biomèdica. Institut Municipal d'Investigació Medica / Universitat Pompeu Fabra / Centre de Regulació Genòmica. Pg. Maritim de la Barceloneta 37-49, 08003 Barcelona, Spain. **2:** Grup d'algorísmica i genètica. Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya. C/ Jordi Girona, 1-3, 08034 Barcelona, Spain. ⋆: Corresponding author.

Keywords: Sequence analysis, dynamic programming, comparative genomics, gene expression.

1 Introduction

We address the problem of comparing promoter regions from genes with similar expression patterns (e.g. homologous genes, [1]). Similarity in gene expression may not be reflected in sequence similarity in promoter regions which partially would explain the limited sucess of currently available computational methods for promoter prediction ([2]). In the approach described here, we attempt to overcome such a limitation representing a (potential) promoter region as a sequence in a new alphabet in which the different symbols denote different Transcription Factors. Thus, a promoter can be translated into a sequence of pairs containing the factor potentially binding to a motif and the associated position on the nucleotide sequence. Sequences in this new alphabet can be aligned. If the scoring model takes into account not only the presence/absence of a given symbol, but its relative position on the primary sequence as well, the optimal alignment between the promoter regions of two similarly expressed genes may reflect the underlying common configuration of binding motifs.

2 Algorithm

To obtain the optimal alignment between two sequences A and B in the new alphabet, we introduce a dynamic programming algorithm similar to a method initially developed to compare enzyme restriction maps, [3]. Elements $a_i =< r_i, p_i >$ in A and $b_j =< t_j, q_j >$ in B are said to match if and only if $r_i = t_j$ (the same TF). The optimal global alignment ending at match $a_i \leftrightarrow b_j$ is computed, then, according to the following recurrence:

if $r_i = t_j$ **then**

$$D(a_i, b_j) \equiv D_{ij} = \min_{\substack{0 < k \leq i \\ 0 < l \leq j}} \{D_{i-k,j-l} + \lambda(k + l - 2) + \mu(p_i - p_{i-k} - q_j + q_{j-l})\}$$

otherwise

$$D_{ij} = \infty$$

Initialization:

$$D_{00} = 0 \qquad D_{i0} = D_{i,m+1} = \infty \qquad D_{0j} = D_{n+1,j} = \infty$$

where λ and μ are two parameters to control the length and the compactness of the final alignment. If A and B contain n elements, the cost of the algorithm is $\theta(n^4)$. Because of the matrix is quite sparse, the final cost can be rewritten in terms of the number of matching positions between both lists $F << n$ as $\theta(F^2)$.

3 Figures

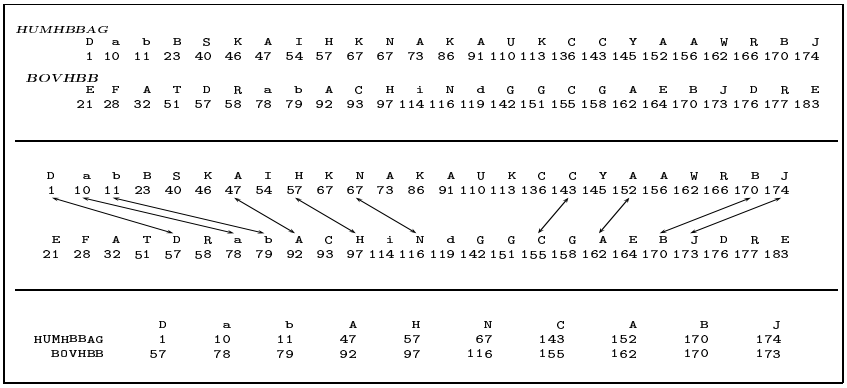

Figure 1: Alignment of the promoter regions from the homologues HUMHBBAG and BOVHBB of the *A-gamma-hemoglobin* gene.

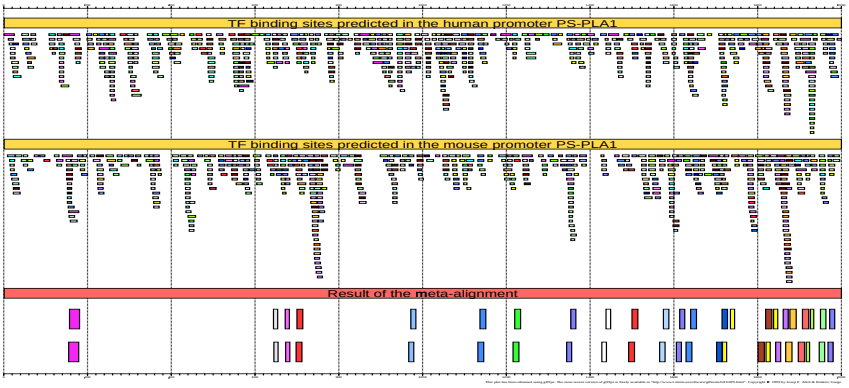

Figure 2: Graphical representation of the TF sites predicted by the program MatInspector ([4]) and the results of the alignment along the human and mouse promoters of the gene *PS-PLA1*.

References

[1] Dermitzakis, E.T. and Clark, A.G. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Molecular Biology and Evolution* 19:1114-1121.

[2] Fickett, J.W. and Wasserman, W.W. 2000. Discovery and modelling of transcription regulatory regions. *Current opinion in Biotechnology* 11:19-24.

[3] Waterman, M.S., Smith T.F. and Katcher, H.L. 1984. Algorithms for restriction map comparisons. *Nucleic Acids Research* 12:237-242.

[4] Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. 1995. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research* 23:4878-4884.