

CUBIC: Search for Binding Sites

Victor Olman¹, Jizhu Lu¹, PhuongAn Dam², Zhengchang Su², Ying Xu^{1,2}

¹Department of Biochemistry and Molecular Biology, University of Georgia, and ²Institute of Computational Biology, Oak Ridge National Laboratory
olman@csbl.bmb.uga.edu

Abstract

The regulation of gene transcription is achieved through specific interactions between transcription factors and their binding sites in the upstream region of the gene being regulated. Correct identification of these binding sites represents a key challenging problem in computational biology. Our approach to the problem is to find a “clear” cluster in the space of all k-mers from the upstream regulatory regions of a set of genes that potentially share similar binding sites. The cluster identification is performed by using Minimal Spanning Tree (MST) technique with a special distance between k-mers based on the chosen Profile. It’s shown that widely used “conservation” characteristic in position is a result of a “common sense” requirement for “conservation”. The local convergence of algorithm for “conservation” maximization of Profile has been proved and the method for statistical significance evaluation of results is presented. All ideas have been implemented in a form of software CUBIC.

1. Introduction

A key challenging problem in computational biology is to identify the regulatory transcription factor binding-site (BS) in the upstream region of genes. BSs are short (6-30 bp) DNA motifs, and generally similar to each other but with great variability. The short varying length and great variability in sequence make identification of BSs in the promoter region of a gene extremely difficult. Currently methods for BSs identification relies on a set of genes whose promoter regions are highly likely to contain similar binding sites.

The idea behind the software CUBIC is that similar BSs should look like a cluster in the space of all k-mers from the upstream region. Comparing to known programs [2-4] of binding site identification, such as GIBBS, MEME and AlignACE, multiple alignment of upstream regions is only the first step of

the CUBIC algorithm. The alignment is used to compute a distance between k-mers. All k-mers from the upstream regions is then represented as a Minimal Spanning Tree. CUBIC was successfully applied to the search of binding sites in many data sets. Here we present the example of binding site identification for orthologous genes from closely related species of cyanobacteria.

2. Mathematical Model

The pivotal point in search of binding sites through cluster identification is a distance measure between k-mers. The trivial mismatch function (a number of positional mismatches for two k-mers) can not create the picture of cluster in the fancy space of k-mers. So we define a distance between 2 arbitrary k-mers S_1, S_2 based upon a specific profile PF as follows:

$$\sum_{j=1, \dots, k} w_j * (C_0 * (1 - U_j) + C_1 * V_j),$$

where C_0, C_1 are positive weights for 2 contributions

to the distance: positional average conservation

$$U_j = .5 * (freq_j(S_{1j} | PF) + freq_j(S_{2j} | PF)),$$

and a relative difference of two substrings

$$V_j = |freq_j(S_{1j} | PF) - freq_j(S_{2j} | PF)| / U_j$$

where $freq_j(l | PF), j = 1, \dots, k, l = a, c, g, t$ form

a frequency matrix of profile PF with $\sum_{l=a,c,g,t} freq_j(l) = 1$ for $j=1, 2, k$,

Using this particular distance we build MST, and a set of binding sites can be easily found from a deep valley of Linear Representation of Data (LPD) from Prim algorithm [1].

The only parameter of a distance is the profile PF .

First, we define the similarity Q of a particular k-mer M to profile PF as

$$Q(M, w | PF) = \sum_{j=1}^{j=k} w_j * F(freq_j(M_j | PF))$$

where M_j is a nucleotide in M at j -th position, $w = (w_1, \dots, w_k) \in R^k$ is a normalized vector, and $F(t)$ is a monotonically increasing function for $t > 0$. By definition the consensus of profile PF has a maximal similarity to PF for any weight vector w . Let a conservation of substrings $(S_1, \dots, S_d) = S$, $d > 1$, in respect to profile PF , at j -th position be

$$Cons_j(S | PF) = \sum_{i=1}^{i=d} F(freq_j(S_{ij} | PF))$$

The natural requirement for a function F is

$$\max_{PF} (Cons_j(S | PF) = Cons_j(S | PF(S))) \quad (1),$$

or in other words, the highest conservation for the set S should be achieved through the profile built from the same set S .

Lemma 1. The **only** differentiable function that satisfies (1) is $F(t) = \log(t)$, $t > 0$.

The total similarity of a set S of k -mers S_1, S_2, \dots, S_d to the profile PF is

$$\sum_{i=1}^{i=d} \sum_{j=1}^{j=k} (w_j * F(freq_j(S_{ij} | PF))) \quad (2)$$

The higher total similarity the more conservative profile and “deeper” valley in LRD is. So our goal is to maximize the total similarity over possible profiles PF , sets S of k -mers and weight w .

Lemma 2. Maximization of (2) is equivalent to

$$\max_S \left(\sum_{j=1, \dots, d} entr_j^2(S) \right)$$

where $entr_j(S)$ is defined as

$$\sum_{i=1, \dots, d} freq(S_{ij} | S) * F(freq(S_{ij} | S)).$$

Algorithm of the best profile construction is based on both Lemmas.

3. Statistical Significance of results

For the evaluation of statistical significance we accept two assumptions related to the absence of binding site:

- All k -mers are scattered “uniformly” in the space, or in other words there is no concentration (based on distance above) of k -mers.
- Values of LRD form a random vector with Dirichle distribution.

This particular model is an extrapolation of 1-dimensional Euclidean case for uniform distribution to our problem. Calculations are based on results presented in [1].

4. Application.

For the set of upstream regions of orthologs of gene *cbbR* in 10 cyanobacteria : Nostoc sp. PCC 7120, Nostoc punctiforme, Prochlorococcus marinus subsp. Marinus, Prochlorococcus marinus subsp. Pastoris, Prochlorococcus marinus str. MIT, Synechococcus sp. WH 8102, Synechococcus sp. PCC7942, Synechocystis sp. PCC 6803, Thermosynechococcus elongatus BP-1, Trichodesmium erythraeum we got potential BSs with $P\text{-value} = 1.38 * 10^{-4}$.

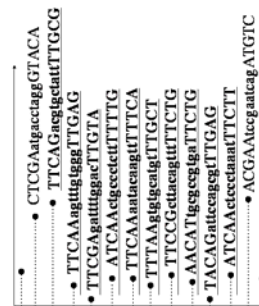


Figure. A part of LRD with a deep valley for a set of potential BSs. Y-axis is an index and X-axis is an edge length in Prim algorithm. Low-case letters in BS means a low informational content in a position.

5. Acknowledgment.

The work was supported by DOE office of Biological/Environmental Research, Genome to Life Project, "Carbon Sequestration in Synechococcus sp.: From Molecular Machines to Hierarchical Modeling."

4. References

- [1] V. Olman, D. Xu, Y. Xu, "CUBIC: Identification of regulatory binding sites through data clustering", Journal of Bioinformatics and Computational Biology, Vol. 1, No. 1, (2003), pp. 21-40.
- [2] C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, J.C. Wootton, "Detecting subtle sequence signals: Gibbs sampling strategy for multiple alignment", Science 262, (1993), pp. 208-214.
- [3] G.Z. Hertz, G.D. Stormo, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences", Bioinformatics 15, (1999), pp. 563-577.
- [4] T.L. Bailey, M. Gribskov, "Methods and statistics for combining motif match scores", Journal of Computational Biology 5, (1998), pp.211-221.

