Biophysics, Vol. 48, Suppl. 1, 2003, pp. S17–S20.

Clustering Regulatory Signals by Binary Trees

E. D. Stavrovskaya and A. A. Mironov

State Scientific Center GosNIIGenetika, 1 1st Dorozhny proezd, Moscow, 117545 Russia
Department of Bioengineering and Bioinformatics, E-mail: esta191@fromru.com
Moscow State University, 1-73 Vorobievy Gory, Moscow, 119992 Russia

Received November 3, 2003; in final form, November 19, 2003

Abstract—Application of the phylogenetic footprinting techniques to bacterial genomes generates a large number of potential regulatory sites identified upstream of orthologous genes. The next step of such analysis should be clustering of sites corresponding to one signal, that is, binding sites of one regulator. We describe an algorithm for clustering of regulatory sites and present the results of its testing on real data.

Key words: regulatory signal, regulon, cluster

INTRODUCTION

Study of molecular biology of prokaryotic cells is of considerable theoretical and practical value. In particular, one important aspect is regulation of metabolic and functional systems. Compared with an eukaryotic cell, a bacterial cell is relatively simple, and in most cases the regulation is performed on the transcription level.

Experimental analysis of gene regulation is very labor-intensive. Therefore, analysis of regulatory signals by comparative genomics techniques becomes increasingly popular, especially as more and more genomes are sequenced. Application of comparative techniques is essential for annotation of new genomes; besides, this helps development of the evolution theory of regulatory and other systems.

Like genes, regulatory signals are subject to natural selection. Unlike noncoding regions in general, they are conserved and thus can be identified by signal analysis techniques. There exist two strategies for identifying regulatory signals. One is to compile a set of experimentally determined sites and construct a recognition rule using statistical techniques. There are several problems with this approach. Firstly, in most cases it is impossible to create a reliable rule. Secondly, it is not possible to find sites for new, unstudied regulatory systems.

The second approach is based on comparative genomics. It relies on the assumption that a set of co-regulated genes (regulon) in one genome will constitute a regulon in a related genome (more exactly, the regulon will be constituted by orthologs of the genes from the first genome). Thus the regions upstream of orthologous genes will contain a signal recognized by one regulator. This allows one to describe regulatory signals without experimental data about regulation or gene function, as it is sufficient to have groups of orthologous genes from related genomes. This strategy was implemented in [1–3].

Analysis of regulatory networks is one of the main problems of contemporary molecular biology. The first step of such analysis is identification of regulons. Co-regulation of genes can be deduced from similarity of regulatory sites in upstream regions. Thus clustering of similar sites identified by comparison of upstream regions of orthologous genes leads to identification of candidate regulons.

The existing approaches to clustering of biological sequences can be divided into hierarchical ones [4, 5] and those based on statistical models [6, 7]. The former arrange the data basing on pairwise comparison and linking similar sequences. The latter make clusters using *a priori* probability distributions.

Here we present a clustering algorithm based on construction of a binary tree. It is a hierarchical

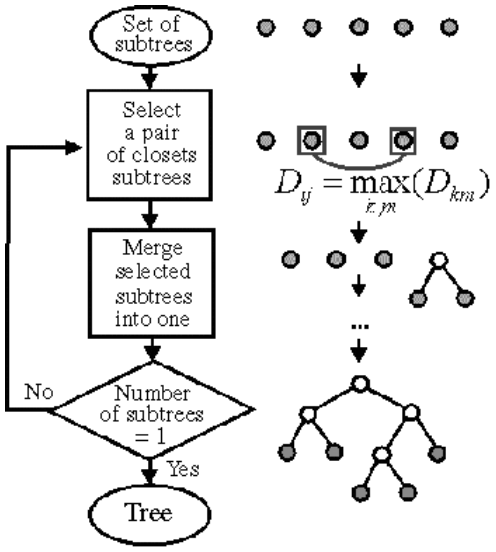

Fig. 1. Algorithm of tree construction.

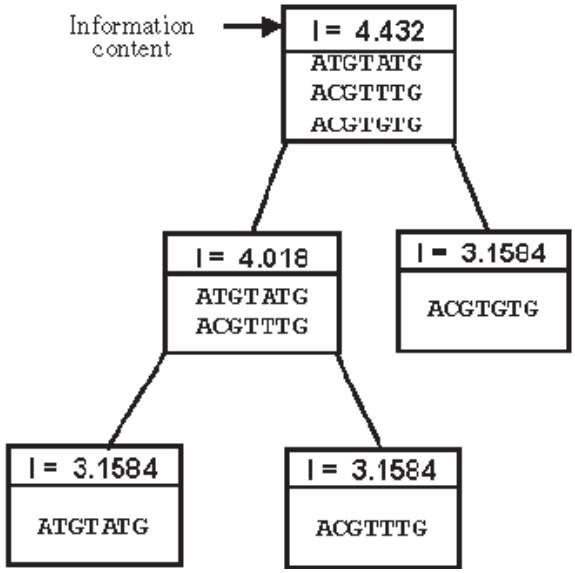

Fig. 2. Example of a tree.

algorithm similar to the one described in [5], but it differs from the latter in the method for identification of clusters in the constructed tree.

ALGORITHM

The algorithm has two main stages: construction of a tree, and tree analysis. Each node of the tree corresponds to a set of sites, and each tree corresponds to a site in the initial sample. Tree construction is done by the Simple Joining Algorithm. It is an iterative process, starting with a site set. For each pair of subtrees a distance is computed, defined as the correlation coefficient

$$D = \sum_{k=1}^{l} \sum_{i=A,C,G,T} f_1(i,k) f_2(i,k) \Big/ \sqrt{F_1 F_2}, \qquad (1)$$

where

$$F_j = \sum_{k=1}^{l} \sum_{i=A,C,G,T} f_j^2(i,k), \qquad (2)$$

Table 1. Clusters in IRSA data

Regulator	Gene	Position	Site	Number of known sites in the sample	Number of sites in the cluster	Number of known sites in the cluster
lexA Inf : 12.98	EC_umuD	161	ctactgtatataaaaacagtat	7	7	7
	EC_recN	67	ttactgtatataaaaccagttt			
	EC_lexA	112	ttgctgtatatactcacagcat			
	EC_dinP	86	tcactgtatactttaccagtgt			
	EC_ruvA	131	tcgctggatatctatccagcat			
	EC_recA	50	atactgtatgagcatacagtat			
	EC_ding	135	atattggctgtttatacagtat			
purR Inf : 13.28	EC_purM	119	gtctcgcaaacgtttgctttcc	6	6	6
	EC_purH	75	gttgcgcaaacgttttcgttac			
	EC_purE	78	gccacgcaaccgttttccttgc			
	EC_cvpA	126	cctacgcaaacgttttcttttt			
	EC_purR	138	taaaggcaaacgtttaccttgc			
	EC_purL	106	tccacgcaaacggtttcgtcag			

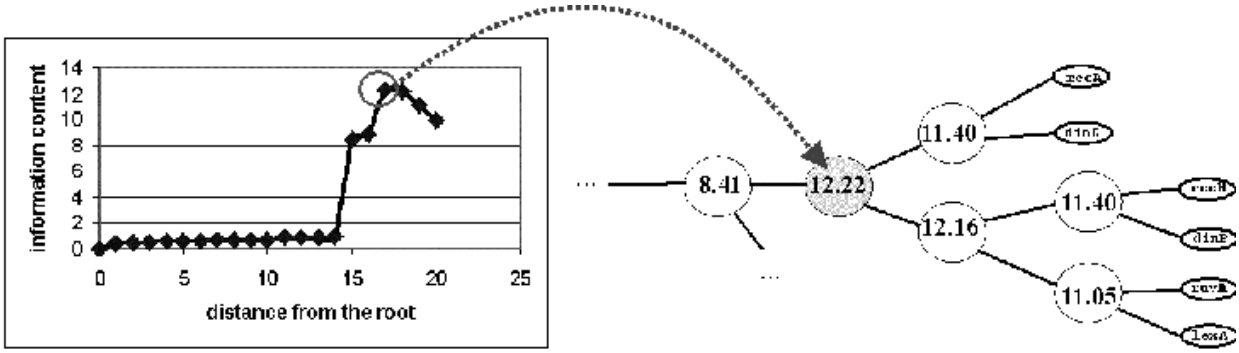

Fig. 3. Behavior of the information content.

$$f_j(i,k) = \frac{n_j(i,k) + 0.25\alpha\sqrt{N_j}}{N_j + \alpha\sqrt{N_j}}, \qquad (3)$$

$f_j(i, k)$ is the nucleotide frequency of nucleotide i in position k in the set of sites corresponding to subtree j; and are pseudocounts, N_j is the number of sites in sample j; α is a constant. Select a pair of closest subtrees and merge them into one subtree. The new subtree (parent) corresponds to the union of the site sets corresponding to the merged subtrees (children).

Thus at each iteration the number of subtrees decreases by 1. The procedure terminates at one tree corresponding to the total site set. The schematic representation of the algorithm is given in Fig. 1, and a simple example, in Fig. 2.

At the second step, all nodes are considered in order to determine the clusters. For each node the information content of the corresponding site set is computed [8]:

$$I = \sum_{k=1}^{l} \sum_{i=A,C,G,T} f(i,k) \log[f(i,k)/0.25]. \qquad (4)$$

Clusters are defined as nodes corresponding to local maxima of the information content. Thus the cluster node should correspond to the condition

$$(I > I_p) \text{ and } ((I > I_l) \text{ or } (I > I_r)), \qquad (5)$$

where I is the information content of the given node, I_p is the information content of the parent node, I_r and I_l are the information contents of the children nodes.

Thus we compute the information content of the signal corresponding to the node. If formula (4) is used without pseudocounts, single-site samples would have the maximal information content, that is, all constructed clusters would consist of only one site. However, the use of pseudocounts (3) allows one to construct non-trivial clusters.

Figure 3 shows the behavior of the information content at a route from the root to a leaf. As one can see, initially the information content is close to zero, and then it increases sharply. This means that the node corresponds to a group of similar sites that differ from the remaining sites in the sample. This could be one cluster or several clusters corresponding to similar signals. The maximum corresponds to the optimal division of the tree into clusters.

Table 2. Clusters in DPInteract data

Regulator	Number of known sites in the sample	Number of sites in the cluster	Number of known sites in the cluster
cpxR	12	10	10
crp	49	78	49
lparfadR	7	6	6
flhCD	3	3	3
fur	9	6	6
gcvA	4	4	4
hipB	4	3	3
lacI	3	3	3
lexA	19	15	15
metJ	15	6	6
metR	8	8	8
modE	3	3	3
nagC	6	4	4
ntrC	5	5	5
phoB	15	12	12
purR	22	16	16
trpR	4	4	3
tyrR	17	9	9

TESTING AND RESULTS

The algorithm was tested on regulatory signals of *E. coli*.

Table 1 lists the results of testing on candidate sites selected from regions upstream of orthologous genes of enteric bacteria using the IRSA algorithm [9]. Only *E. coli* sites were considered.

The program was also tested on sites from DPInteract [10]. The results are given in Table 2.

DISCUSSION

As one can see in Tables 1 and 2, the results are reasonable. However, Table 2 contains regulators whose clusters were not recovered completely. For regulators such as CRP and TyrR this could be expected, as the corresponding signals are rather weak. However, incompleteness of the PurR cluster is surprising. The cause of this is the fact that a higher-level cluster contains a subcluster of PurR sites as well as GalR sites that interferes with formation of a proper PurR cluster. Closer analysis showed that the reason for that is that PurR sites from DPInteract are longer than the signal, and nonsignificant positions at signal termini decrease the information content. We plan to implement a procedure for calculation of the statistical significance of positional information content, and to disregard nonsignificant columns.

ACKNOWLEDGMENTS

We are grateful to L. Danilova for the IRSA data, and to M. Gelfand and D. Ravcheev for useful discussions.

This study was partially supported by a grant from HHMI (55000309).

REFERENCES

1. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C., *Science*, 1993, vol. 262, pp. 208–214.
2. Bailey, T.L. and Elkan, C., *Machine Learning J*., 1995, vol. 21, pp. 51–83.
3. Pevzner, P.A. and Sze, S., *Proc Int Conf Intell Syst Mol Biol*., 2000, vol. 8, pp. 269–278.
4. Pietrokovski, S., *Nucleic Acids Res*., 1996, vol. 24, no. 19, pp. 3836–3845. Erratum in: Nucleic Acids Res., 1996, vol. 24, no. 21, p. 4372.
5. Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M., *J. Mol. Biol*., 2000, vol. 296, no. 5, pp. 1205–1214.
6. van Nimwegen, E., Zavolan, M., Rajewsky, N., and Siggia, E.D., *Proc. Natl. Acad. Sci. USA*, 2002, vol. 99, no. 11, pp. 7323–7328.
7. Qin, Z.S., McCue, L.A., Thompson, W., Mayerhofer, L., Lawrence, C.E., and Liu, J.S., *Nat. Biotechnol*., 2003, vol. 21, no. 4, pp. 435–439.
8. Gelfand, M.S., Koonin, E.V., and Mironov, A.A., *Nucleic Acids Res*., 2000, vol. 28, pp. 695–705.
9. Danilova, L.V., Gorbunov, K.Yu., Gelfand, M.S., and Lyubetskii, V.A., *Mol. Biol*., 2001, vol. 35, no. 6, pp. 987–995.
10. Robison, K., McGuire, A.M., and Church, G.M., *J. Mol. Biol*., 1998, vol. 284, no. 2, pp. 241–254.