Title

The Automatic Selection of TFBS Score Threshold in Comparative Genomics Approach

Authors

Stavrovskaya, Elena D.
Rodionov, Dmitry A.
Mironov, Andrey A.
et al.

Publication Date

2009-12-01

The automatic selection of TFBS score threshold in comparative genomics approach

Elena D. Stavrovskaya[1,2,*], Dmitry A. Rodionov[2,4] , Andrey A. Mironov[1,2], Inna Dubchak[3,5], Pavel S. Novichkov[3,5,*]

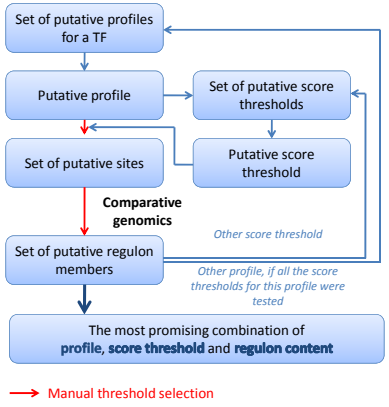
[1]Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119992, Russia;
[2]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russia;
[3]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA;
[4]Burnham Institute for Medical Research, La Jolla, CA 92037,USA;
[5]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA
* stavrovskaya@gmail.com, psnovichkov@lbl.gov

Overview

Reconstruction of transcriptional regulatory networks is one of the major challenges facing the bioinformatics community in view of constantly growing number of complete genomes. The comparative genomics approach has been successfully used for the analysis of the transcriptional regulation of many metabolic systems in various bacterial taxa. The key step in this approach is, given a position weight matrix, find an optimal threshold for the search of potential binding sites in genomes. Here we demonstrate that this problem is tightly bound to a problem of discovering the optimal content of regulon and suggest an approach to solve both problems simultaneously.

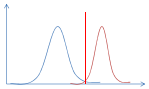
Manual analysis

The main idea is to model the approach used in manual inference.



Threshold selection problem

Bernoulli Estimator



Background distribution; **known**

"Signal" distribution; **unknown**

Consider a sample of $\{v_i\}$ of size n which is a mixture from **background** and **signal** distributions

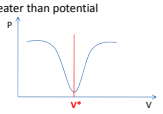
Task: select the threshold V*, which would maximize probability that all vi ≥ V* are from the **signal** distribution and at the same time that all vi < V* are from **background** one

- Go through all *vi* and consider each *vi* as a potential threshold *V*
- Calculate the number *k* of values *vi* greater than selected threshold *V*
- Supposing all *{vi}* were sampled from the **background** distribution **only**, calculate probability to observe *k* or more values in a sample to be equal or greater than potential threshold *V*

$$P_{BE}(V) = \sum_{i=k}^{n} C_n^i p^i (v \geq V) p^{n-i} (v < V)$$

- Select V*=V which delivers the minimum for P(V)

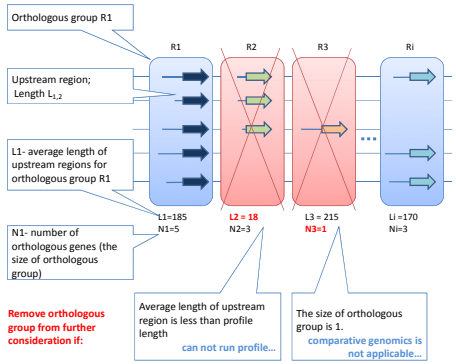
$$V^* = \underset{V}{\operatorname{argmin}}(p(V))$$



Input: { vi , p(v ≥vi) }
Output: V*, p(v≥V*)

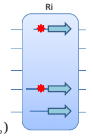
Regulatory potential

Procedure input: set of genomes, predefined groups of orthologous genes, fixed parameters for gene upstream region selection, and profile



Regulatory potential of orthologous group

- Run profile to search potential binding sites.
- Fix some threshold value S* for the score of the binding site.



$$P(s \geq S^* \mid L) = 1 - P^{L-L_p}(s < S^* \mid L_p)$$

- probability to find at least one binding site with score s ≥ S* in ***random sequence*** of length L, where Lp is a length of profile.

For a given orthologous group Ri:

- Calculate the number of genes Ki which have binding site with score ≥ S*
- Calculate the regulatory potential of orthologous group Z(S*)

$$Z_i(S^*) = -\log P(k \geq K_i \mid N_i, L_i, S^*) =$$

$$= -\log\left(\sum_{K=K_i}^{N_i} C_{N_i}^{K} P^{K}(s \geq S^* \mid L_i) P^{N_i - K}(s < S^* \mid L_i)\right)$$

P(K ≥ Ki | Ni,Li,S*) - probability to find at least Ki genes with site having score ≥ S* in a given orthologous group Ri, where the upstream regions where substituted by ***random sequences of legth Li***

Score threshold selection

Selection the set of significant orthologous groups

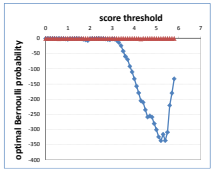
- Calculate quality Zi(S*) for each orthologous group
- Use **Bernoulli Estimator** to set threshold Z*(S*) for regulatory potential of orthologous groups Zi(S*) which would separate significant and non-significant groups
- Besides the threshold for regulatory potential Z*(S*) Bernoulli estimator returns optimal Bernoulli probability $P_{BE}(Z^*)$

Bernoulli Estimator requires the background probability

Selection the optimal threshold S* for the score of the binding sites

- Consider the score of each of the found binding sites as a potential threshold S*, and calculate the optimal threshold for regulatory potential Z*(S*) and minimal Bernoulli probability $P_{BE}(Z^*)$
- Calculate the optimal threshold for the score of the binding sites as

$$S^* = \arg\min_{S}(P_{BE}(Z^*))$$



Score threshold selection for lexA profile matrix applied to real (blue line) and simulated (red line) upstream regions. Score threshold corresponds to the global minimum of optimal Bernoulli Probability obtained using Bernoulli estimator.

Background distribution for regulatory potential

- For a given orthologous group Ri

$$P(z \geq Z_i \mid N_i, L_i, S^*) = P(k \geq K_i \mid N_i, L_i, S^*)$$

- For an arbitrary value Z and orthologous group Rj

$$P(z \geq Z \mid N_j, L_j, S^*) = P(k \geq K(Z) \mid N_j, L_j, S^*)$$

where $K(Z) = \min_{Z_i : -\log P(K \geq K_i \mid N_j, L_j, S^*) \geq Z} (K)$

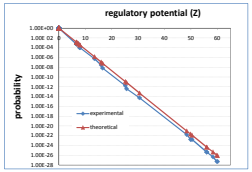
- The probability, that randomly selected orthologous group will have quality Z or better:

$$P(z \geq Z \mid S^*) = \sum_{N_i, L_i} P(z \geq Z \mid N_i, L_i, S^*) P(N_i, L_i) = \frac{1}{M} \sum_{i=1}^{M} P(z \geq Z \mid N_i, L_i, S^*)$$

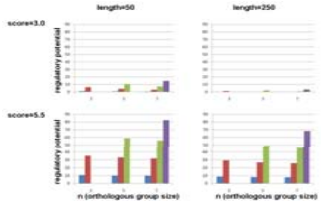
where M is a number of orthologous groups

This cumulative distribution can be well approximated by theoretical function:

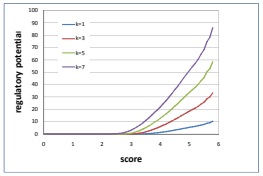
$$\exp(-Z^*)$$



Regulatory potential properties



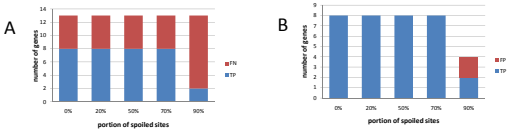
Diagrams of regulatory potential value (Z). The diagrams are presented for two typical upstream region lengths: 50 and 300 bp; and for two score values: 3.0 (weak sites) and 5.5 (strong sites). On each diagram the values for three orthologous group sizes n=3, 5 and 7 are shown. Column corresponds to *k* – number of genes having site with significant score. *k* can be less or equal to *n*. Blue columns correspond to *k*=1, red to *k*=3, green to *k*=5, purple to *k*=7



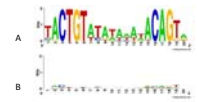
Behavior of regulatory potential depending on score by different *k* – number of genes with sites having significant score. Orthologous group size and upstream length are fixed: *n*=7, *l*=300.

Performance

Performance of regulon members prediction given different motif purity. Initially motif for *lexA* TF consisted of 78 sites. Then the motif was spoiled by substituting portion of sites (x-axis) for random ones. The performance is shown by number of correctly predicted regulon members (true positives) in comparison with: (A) number of unpredicted true regulon members (false negative); (B) number of overpredictions (false positives).



Logo diagrams for initial *lexA* motif (A) and motif spoiled by substituting 70% sites for random ones (B). The initial motif has information content 1.06 for a position. The spoiled motif has information content 0.46 for a position, but significant positions are still conservative and the *lexA* motif still remains.



Acknowledgments

This work was part of the Virtual Institute for Microbial Stress and Survival (http://VIMSS.lbl.gov) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Program: GTL through contractDE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy, Howard Hughes medical institute (55005610), RAS program "Molecular and cellular biology", Russian Foundation for Basic Research (08-04-01000-a).