Ecosystems and Networks Integrated with Genes and Molecular Assemblies

Regulon inference without arbitrary thresholds: three levels of sensitivity

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

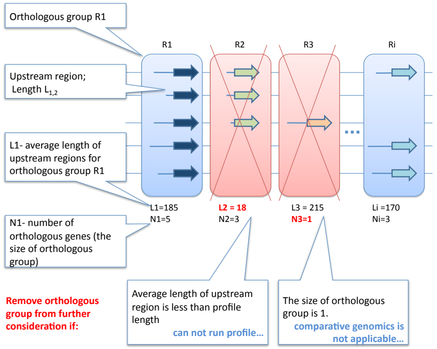
[1]*Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119992, Russia;*
[2]*Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russia;*
[3]*Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA;*
[4]*Burnham Institute for 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

Introduction

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. In our previous work we proposed an approach for automatic selection of TFBS score threshold coupled with inference of regulon content. In this study we developed two modifications of this approach providing two additional levels of sensitivity

Regulatory potential

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



Orthologous group R1

Upstream region; Length $L_{1,2}$

L1- average length of upstream regions for orthologous group R1

N1- number of orthologous genes (the size of orthologous group)

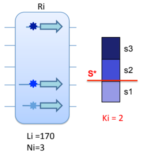
Remove orthologous group from further consideration if:

Average length of upstream region is less than profile length **can not run profile…**

The size of orthologous group is 1. **comparative genomics is not applicable…**

Regulatory potential of orthologous group

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



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

- 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_i (S*)

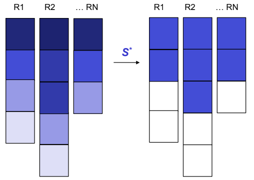
$$Z_i(S^*) = -\log P(k \geq K_i \mid N_i, L_i, S^*)$$

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

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***

Three levels of sensitivity

I. Common threshold for all orthologous groups

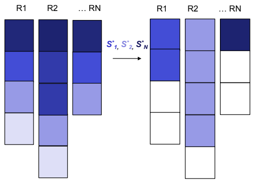


- ➢For a particular threshold S^*
- ➢Calculate regulatory potential $Z_i(S^*)$ for each orthologous group
- ➢Use **Bernoulli Estimator** to calculate threshold for regulatory potential of orthologous groups $\tilde{Z}(S^*)$ and corresponding Bernoulli probability $P_{BE}(S^*)$
- ➢Iterate through each S^* to find the optimal threshold $\tilde{S}$ delivering minimum to $P_{BE}(S^*)$

The outcome:

Optimal threshold for TFBS score $\tilde{S}$
Optimal threshold for regulatory potential of orthologous groups $\tilde{Z}$

II. Threshold individual for each orthologous group

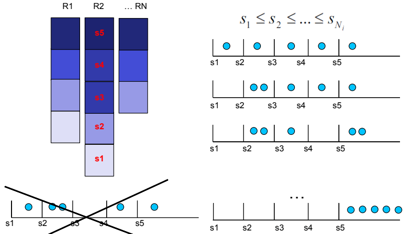


- ➢Calculate the optimal threshold $\tilde{S}_i$ for each individual orthologous group delivering the maximum to the regulatory potential $Z_i(S_i^*)$
- ➢Use **Bernoulli Estimator** to calculate threshold $\tilde{Z}$ for regulatory potential of orthologous groups

The outcome:

Optimal thresholds for TFBS score for each individual orthologous group $\tilde{S}_i$
Optimal threshold for regulatory potential of orthologous groups $\tilde{Z}$

III. No score threshold, all putative TFBSs are considered



$$s_1 \leq s_2 \leq \ldots \leq s_{N_i}$$

$$\forall i \quad \sum_{s \geq s_i} N_s \geq i$$

Probability to observe binding sites with the same scores or greater

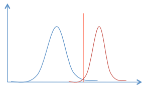
$$P_{mn} = \sum_{k=0}^{m-n+1} C_m^k p_n^k P_{m-k,n-1}$$

where p_i – probabiliy to observe TFBS score in the range $s_i < s < s_{i-1}$

- ➢Calculate the regualtory potential for each orthologous group $Z_i = -\log P_{mn}$
- ➢Use **Bernoulli Estimator** to calculate threshold $\tilde{Z}$ for regulatory potential of orthologous groups

Score threshold selection

Bernoulli Estimator



Background distribution; **known**

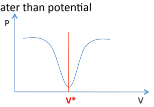
“Signal” distribution; **unknown**

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)

$$V^* = \arg\min_V (p(V))$$

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

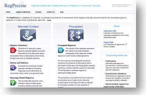
Performance

7 genomes

- *Shewanella oneidensis MR-1*
- *Shewanella baltica OS155*
- *Shewanella denitrificans OS217*
- *Shewanella frigidimarina NCIMB 400*
- *Shewanella amazonensis SB2B*
- *Shewanella sediminis HAW-EB3*
- *Shewanella pealeana ATCC 700345*

RegPrecise – manually curated database of regulons inferred by comparative genomics approach

NAR database issue, 2010



http://regprecise.lbl.gov

62 regulons

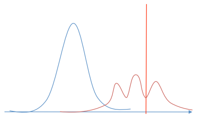
- Local regulons *39*
- Middle-size regulons *16*
- Global regulons *7*

	I. Common threshold		II. Individual treshold		III. No threshold, all sites considered	
	Sf	Sp	Sf	Sp	Sf	Sp
Gloabal	0,36	0,95	0,41	0,88	0,52	0,91
Middle-size	0,63	0,95	0,62	0,82	0,63	0,89
Local	0,74	0,74	0,92	0,72	0,86	0,81

Underprediction….

Iterative approach

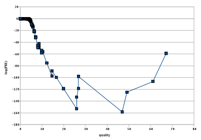
“Signal” distribution can be supperpositions of several distributions…



Background distribution; **known**

“Signal” distribution; **unknown**

Dependence of the Bernoulli probability on the regulatory potential threshold, NagR regulon



Second iteration

	I. Common threshold		II. Individual threshold		III. No threshold, all sites considered	
	Sf	Sp	Sf	Sp	Sf	Sp
Global	0,63	0,85	0,77	0,67	0,74	0,64
Middle-size	0,72	0,60	0,80	0,47	0,81	0,62

Acknowledgments

This work was part of ENIGMA, Scientific Focus Area Program 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).