Title
Regulon inference without arbitrary thresholds: three levels of sensitivity

Permalink
https://escholarship.org/uc/item/5c17b3wg

Author
Dubchak, Pavel Novichkov, Elena Stavrovskaya, Dmitry Rodionov, Andrey Mironov, Inna

Publication Date
2011-02-15
Regulon inference without arbitrary thresholds: three levels of sensitivity

Elena D. Stavrovskaya1,2, Dmitry A. Rodionov3,4, Andrey A. Mironov5,6, Irina Dubchak6,5, Pavel S. Novichkov1,3,6

1Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119992, Russia; 2Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russia; 3Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; 4Lumen Institute for Medical Research, La Jolla, CA 92037, USA; 5Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

* stavrovskaya@gmail.com, pavel.novichkov@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

Score threshold selection
Bernoulli Estimator
- Background distribution, known
- "Signal" distribution, unknown

Task: where the threshold \( T \), which would maximize probability that all \( N \) \( x \)'s are from the \( \text{signal} \) distribution and at the same time that all \( y \)'s \( \neq 0 \) are from background one
- Go through \( x \)'s and \( y \)'s
  - For all \( x \) in \( x \)
    - Calculate the number of \( y \) values or greater than selected threshold \( y \)
    - Calculate probability to observe \( y \) or more

Performance
7 genomes
- Shewanella oneidensis MR-1
- Shewanella baltica OS150
- Shewanella densitivitica OCSY
- Shewanella putrefaciens NCPPB 405
- Shewanella amarae ISB8
- Shewanella alvei IAM4178
- Shewanella putrefaciens ATCC 70054

62 regions:
- Local regions 29
- Middle-size regions 15
- Global regions 17

Iterative approach
Calculate the regulatory potential for orthologous group \( i \) for a particular threshold \( T \)

Three levels of sensitivity
1. Common threshold for all orthologous groups
- Calculate regulatory potential \( P(T) \) for each orthologous group
- Use Bernoulli Estimator to calculate threshold for regulatory potential of orthologous group \( i \) \( P(T)_i \) and corresponding Bernoulli probability \( \pi_i \)
- Iterate through each \( T \) to find the optimal threshold \( \hat{T} \) delivering minimum to \( \pi_i(T) \)

2. Threshold individual for each orthologous group
- Calculate the optimal threshold \( \hat{T}_i \) for each individual orthologous group delivering the maximum to the regulatory potential \( P(T_i) \)
- Use Bernoulli Estimator to calculate \( \hat{T}_i \) for regulatory potential of orthologous group

3. No score threshold, all putative TFBSs are considered
- Calculate the regulatory potential for each individual orthologous group \( \pi_i \)
- Optimal threshold for regulatory potential of orthologous group

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, Generation Program (G1) through contract number DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy; Howard Hughes medical institute (50009824), NIH program "Molecular and cellular biology" Russian foundation for Basic Research (15-04-00041-a).