Lawrence Berkeley National Laboratory

LBL Publications

Title

Web-based Tool for Fast and Accurate de novo Inference of Regulons in the Sets of Closely Related Bacterial Genomes

Permalink

https://escholarship.org/uc/item/0ch5q4ph

Authors

Novichkov, Pavel S.
Stavrovskaya, Elena D.
Gelfand, Mikhail s.
et al.

Publication Date

2009-10-30

eScholarship.org Powered by the California Digital Library
University of California

Web-based tool for fast and accurate *de novo* inference of regulons in the sets of closely related bacterial genomes

Pavel S. Novichkov[1,6*], Elena D. Stavrovskaya[2,3], Mikhail S. Gelfand[2,3], Andrey A. Mironov[2,3], Inna Dubchak[1,5,6], Dmitry A. Rodionov[2,4,*]

*[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA; [2]Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow 127994, Russia; [3] Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119992, Russia; [4]Burnham Institute for Medical Research, La Jolla, CA 92037,USA; [5]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA; [6]*Virtual Institute for Microbial Stress and Survival, http://vimss.lbl.gov

* psnovichkov@lbl.gov, rodionov@burnham.org

Introduction

One of the major challenges for the bioinformatics community in view of constantly growing number of complete genomes is providing effective tools to enable high-quality reconstruction of transcriptional regulatory networks (TRN). Definition of a particular TRN includes specification of which transcription factors (TF) bind to TF-binding sites (TFBS) in the promoter regions of which genes and what is the integrated effect of all these TFs on the expression of al these genes. Reconstruction of TRNs helps to better understand the metabolism and functions of bacteria.

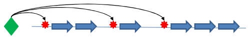
Among different approaches that are used for TRN reconstruction are an expression data-driven approach, and comparative genomic approaches that are either computing-driven, or subsystem (pathway) -driven . DNA microarrays, reporting gene expression, continue to be an important tool for high-throughput measurements on transcriptional levels, and machine-learning approaches were used to identify TRN (without a TFBS component) from a compendium of microarray expression profiles . However, in many cases the complexity of the interactions between regulons makes it difficult to distinguish between direct and indirect effects on transcription. Availability of a large number of complete genomes opens an opportunity to apply modern approaches of comparative genomics to expand the known regulons to yet uncharacterized organisms and to predict and describe new regulons with high precision .

Comparative genomics

Transcription factors ◆ (TFs)
proteins that recognize specific binding sites to either activate or repress transcription of genes.

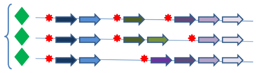
Binding sites ✸ (TFBSs)
(or operators) are *cis*-regulatory DNA sequences located in 5'-untranslated regions (UTRs) of genes.

Regulon (single genome)
is a set of genes in a particular genome directly regulated by the same transcription factor



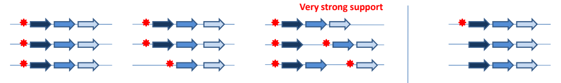
Regulog

is a group of regulons operated by orthologous TFs in closely-related genomes.
- involves set of genes with conserved TFBSs.
- is associated with a distinct TFBS-recognition profile.



Validation of predicted binding sites

Typical pros / *Typical cons*



Presence of binding sites in upstreams of orthologs

Birth of new binding site in upstream of ortholog

Operon rearrangement and birth of new binding sites

Presence of orthologs, **but absence** of binding sites

Clusters of regulated orthologous operons

Allow to

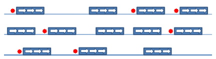
1.Separate all found potential TFBSs into subsets which can be analyzed independently

2.Collect pros and cons for decision making

Definition

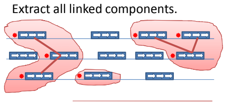
Two operons from two different genomes are called orthologous if they have one or more common orthologous genes

I. For a selected profile, search potential binding sites in upstream regions of genes. Consider each operon as a vertex in the graph



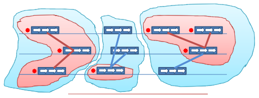
II. Join two vertices by edge if the correspondent operons:
i.are orthologous
ii.have potential binding sites

Extract all linked components.



Allows to collect pros…

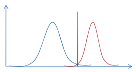
III. Build up each linked component by adding operons witout binding sites that are orthologous to one of operons from the linked component



Allows to collect cons…

Threshold selection problem

Bernoulli Estimator



Background distribution; known
"Signal" distribution; unknown

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

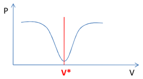
Task: select the threshold V*, which would maximize probability that all $v_i \geq V^*$ are from the signal distribution and at the same time that all $v_i < V^*$ are from background one

- Go through all v_i and consider each v_i as a potential threshold V
- Calculate the number k of values v_i greater than selected threshold V
- Supposing all $\{v_i\}$ 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

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

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

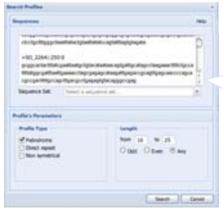
$$V^* = \arg\min_V (p(V))$$



Input: $\{ v_i , p(v \geq v_i) \}$
Output: $V^*, \ p(v \geq V^*)$

Web Based GUI

Search TFBS profiles



The types of sets of sequences to search profile in

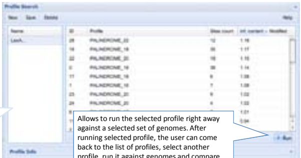
▪Sets of genes from the same metabolic pathway (based on SEED subsystems).
▪Sets of genes positionally linked to transcription factors
▪Sequences provided by user

Profile search parameters

The user can select a range of parameters (e.g. a 16-bp palindrome) and search all selected types of profiles at once.

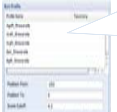
The result of the profile search procedure

The list of all found profiles. Each profile is supplied with a set of properties such as profile type, profile length, the size of the training set selected for a profile. Several types of scores allowing to estimate the profile quality are suggested to be implemented. Among them are information content, regulon conservation score, percent of significant positions in profile. The list of profiles can be sorted by any of the parameters of quality scores.



Allows to run the selected profile right away against a selected set of genomes. After running selected profile, the user can come back to the list of profiles, select another profile, run it against genomes and compare results to make a decision about which profile is the best

Run profile



The source of profiles

▪The library of profiles from RegPrecise database is provided. Each profile has link to RegPrecise record with complete description of corresponded of regulog.
▪Alignment of binding sites in fasta format provided by user

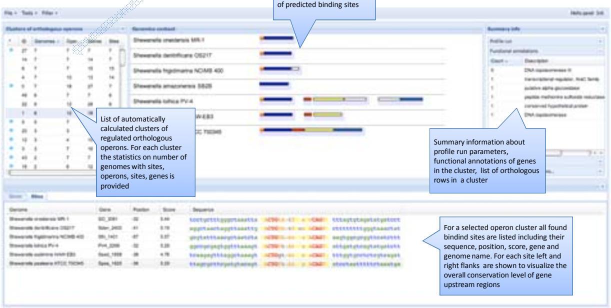
Parameters
▪Parameters for selection upstream regions
▪Threshold for the score of the potential binding site

RegPrecise database (http://regprecise.lbl.gov)

The main object in this database is regulog, which is a collection of inferred regulons of the same TF in a set of closely related bacterial genomes.

Each profile from library available to search was build on RegPrecise regulogs. For a given regulog the information provided about transcription factor, TF family, effector (if known), set of genomes under analysis, complete description of regulons in each genome, including regulated genes; sequence, position and score of the binding site.

Regulon annotation



Visual analysis for validation of predicted binding sites

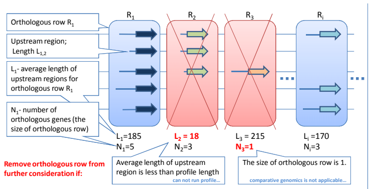
List of automatically calculated clusters of regulated orthologous operons. For each cluster the statistics on number of genomes with sites, operons, sites, genes is provided

Summary information about profile run parameters, functional annotations of genes in the cluster, list of orthologous rows in a cluster

For a selected operon cluster all found bindind sites are listed including their sequence, position, score, gene and genome name. For each site left and right flanks are shown to visualize the overall conservation level of gene upstream regions

Evolutionary regulon conservation score

Input for the procedure: set of genomes with predefined rows of orthologous genes and TFBS profile



Orthologous row R_1

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

L_1- average length of upstream regions for orthologous row R_1

N_1- number of orthologous genes (the size of orthologous row)

Remove orthologous row from further consideration if:

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

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

Quality of orthologous row

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

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

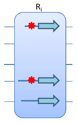
- probability to find at least one binding site with score s ≥ S* in ***random sequence*** of length L, where Lp is a length of profile.

For a given orthologous row Ri:
▪Calculate the number of genes Ki which have binding site with score ≥ S*

▪Calculate the quality of orthologous row $Z_i(S^*)$

$$Z_i(S^*) = 1 - P(k \geq K_i \mid N_i, L_i, S^*) = 1 - \sum_{K=K_i}^{N_i} C_{N_i}^K P^K(s \geq S^* \mid L_i) P^{N_i - K}(s < S^* \mid L_i)$$

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



L_i =170
N_i=3
K_i = 2

Selection the set of significant orthologous rows

- Calculate quality $Z_i(S^*)$ for each orthologous row
- Use **Bernoulli Estimator** to set threshold Z*(S*) for orthologous row quality values $Z_i(S^*)$ which would separate significant and non-significant rows

But, Bernoulli Estimator requires also the background probability to observe quality equal or better than given…

- For a given row R_i
$$P(z \geq Z_i \mid N_i, L_i, S^*) = P(k \geq K_i \mid N_i, L_i, S^*)$$
- For an arbitrary value Z and row Rj
$$P(z \geq Z \mid N_j, L_j, S^*) = P(k \geq K(Z) \mid N_j, L_j, S^*)$$
, where $K(Z) = \min_{Z > 1 - P(k \geq K \mid N_j, L_j, S_*)} (K)$
- The probability, that randomly selected orthologous row 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 rows

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 orthologous row quality Z*(S*)
- Calculate the optimal threshold for the score of the binding sites as
$$S^* = \arg\max_S (Z^*(S))$$

Evolutionary regulon conservation score $Z^* = \max_S (Z^*(S))$

Testing the platform for *de novo* regulon inference

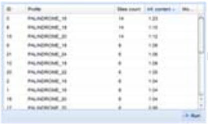
Input: genes from the same metabolic pathway

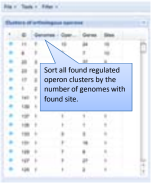
Histidine degradation (SEED subsystem)

Shewanella oneidensis MR-1	
SO_0095	imidazolonepropionase
SO_0096	histidine utilization repressor
SO_0097	urocanate hydratase
SO_0098	histidine ammonia-lyase
SO_4198	arginase family protein
SO_3164	conserved hypothetical protein

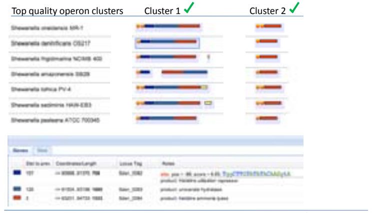
Select predefined set of genes

- Search profiles
- Sort profiles and run the most promising one

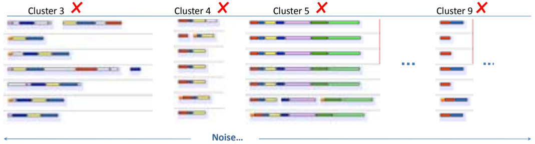




Sort all found regulated operon clusters by the number of genomes with found site.



Visual analysis for regulog annotation



Result

To test the platform for regulon inference we analyzed regulation of the histidine degradation in the group of 7 Shewanella genomes. For the training set of upstream regions, the procedure selected X palindromic profiles with length between 16 and 24 bp. The best scored profile (a 20-bp palindrom) was used to scan the genomes for binding sites resulting in identification of 143 clusters of candidate regulated operons. Cluster ranking and visual analysis allowed us to identify just two clusters with strong binding site conservation(clusters 1 and 2), whereas all other operon clusters appear to be linked to false positive sites that are fairly conserved across the genomes.

Acknowledgments

We are grateful to Andrey Osterman for useful discussions, Alexey Kazakov, Olga Laikova, and Anna Gerasimova for contributions to regulon reconstructions.

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.