AUTOMATED REGULON CONTENT PREDICTION AND ESTIMATION OF PWM QUALITY

Elena Stavrovskaya[1,2], Andrey Mironov[1,2], Dmitry Rodionov[2,3] , Inna Dubchak[4] and Pavel Novichkov[4]

[1]*Department of Bioengineering and Bioinformatics, Moscow State University, Leninskiye Gory 1-73, Moscow, Russia*
[2]*Institute for Information Transmission Problems, Moscow, Russia*
[3]*Burnham Institute for Medical Research, La Jolla, CA, U.S.A.*
[4]*Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA, U.S.A.*

Keywords: Regulation of transcription, Bacteria, Regulon, PWM.

Abstract: Identification of genes regulated by the same transcription factor (TF) is a major problem in analysis of regulation. The key step in detection of a group of co-regulated genes (regulon) is prediction of TF binding sites (TFBS). This is what positional weight matrix (PWM) is for. This matrix is applied to upstream region of a gene, and high-scoring sites are considered as putative TFBSs. Choice of threshold for the scoring function is a separate complicated problem. Usually, the threshold is chosen manually. Some methods for automated threshold detection exist, but they are based on selection of threshold for different functions. In this paper, we present an approach for regulon prediction based on a probabilistic method of threshold detection. The optimal probability computed by this method can be used to estimate the quality of the PWM itself. It can be useful when the matrix is a result of regulatory motif prediction program.

1 INTRODUCTION

Reconstruction of transcriptional regulatory networks is one of the major challenges that the bioinformatics community is facing 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 selection of an optimal site score threshold for the search of potential transcription factor binding sites (TFBS), given a position weight matrix. Here we demonstrate that this problem is tightly coupled with a problem of discovering the optimal content of regulon and suggest an approach to solve both problems simultaneously.

In this study we developed an approach for selection of TFBS score threshold that allows automatic inference of the regulon content given known TFBS motif. We developed three modifications of this approach and extensively tested them on a set of manually curated regulons.

2 METHOD

2.1 Universal TFBS Score Threshold

At the first step we process all genomes under analysis and cluster all genes into orthologous groups (OGs). To enable comparative genomics analysis we consider only those groups that have at least two orthologous genes. Using the concept of OG, the problem of regulon inference can re-formulated as a problem of selecting a subset of OGs that most probably have a conservative TFBS.

Then, assuming S^* to be a potential optimal threshold, we calculate regulatory potential R_i for each orthologous group (OG). A particular OG is described by two parameters: the number of orthologous genes N_i, and the length of upstream region Li averaged by all genes in an OG. Applying the TFBS profile to upstream regions of all orthologs in a group, the number of genes K_i that have a potential TFBS with score $s \geq S^*$ can be calculated. We define the regulatory potential in terms of probability to observe the number of

Stavrovskaya E., Mironov A., Rodionov D., Dubchak I. and Novichkov P..
AUTOMATED REGULON CONTENT PREDICTION AND ESTIMATION OF PWM QUALITY.
DOI: 10.5220/0003787403220325
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2012), pages 322-325
ISBN: 978-989-8425-90-4

orthologous genes in a group, having at least one high scoring site, being K_i or greater, given that upstream regions are random sequences.

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

At the final step we utilize “Bernoulli Estimator” (BE) routine (Kalinina, 2004) which assumes that input values are a mixture from two distributions representing the noise and the signal. Only distribution that represents the noise is required for automatic inference of the optimal threshold distinguishing the signal from the noise. Applying BE to regulatory potentials calculated for all OGs given S^*, the most probable content of a regulon can be automatically identified. Considering all possible values of S^*, the optimal threshold delivering the minimum to BE probability, can be obtained.

As a result, the optimal threshold for TFBS score S^*, the optimal threshold for the regulatory potential, and the subset of OGs predicted to be members of the regulon can be calculated. It should be noted that in this case the same “universal” TFBS score threshold has been used for all OGs under consideration

We also implemented two additional modifications of the developed approach considering different levels of sensitivity.

2.2 Individual TFBS Score Threshold

In this minor modification, instead of one universal TFBS score threshold, we use individual threshold for each OG. It allows to take into account the possible difference in affinity of TF factor to DNA binding sites among different members of the same regulon. It was shown that such differences can be evolutionary conserved and thus have functional meaning. (Kotelnikova, 2005)

2.3 No TFBS Score Threshold

This modification is fundamentally different from the two previous ones. This version does not use threshold to filter out weak sites, but rather allows all putative binding sites to contribute to the regulatory potential of an OG. For a particular OG of size N we consider a set of N best scores $\{s_1, s_2 \ldots s_N\}$. The regulatory potential is calculated as a probability to observe OG with maximum scores $\{s_1, s_2 \ldots s_N\}$ or better by chance.

3 TESTING

3.1 Comparison with the Results of Manual Analysis

The developed algorithms have been extensively tested on 62 manually curated regulons from *Shewanella* collection retrieved from the RegPrecise database (http://regprecise.lbl.gov). All regulons were classified into three classes: i) local (1-2 operons), ii) medium (3-10 operons), and iii) global (more than 10 operons). As expected, all three versions performed similarly well on local regulons representing the most abundant class of regulons in microbial genomes. For 24 out of 39 local regulons the regulon content was predicted correctly.

For medium and large size regulons, the "universal TFBS score threshold" approach was able to predict 63% and 36% true members of regulons. The careful analysis of predicted regulon content revealed that in both cases it comprises the core of the true regulon with very high TFBS score and high level of TFBS conservation across all genomes under analysis. At the same time the specificity is very high in both cases (95%). Thus the approach can be used for automatic accurate reconstruction of the core of regulon, and provides a good starting point for the detailed manual curation.

3.2 PWM Quality

All the methods of regulatory motif prediction give some number of variants as output. This gives rise to the problem of motif quality estimation. Conservation of nucleotides in PWM positions does not always reflect the true quality of the motif found. The best test of motif quality is arrangement of sites recognized with PWM in the genome: if sites are found upstream of genes which can be included into one metabolic pathway, the motif is certainly found correctly. Unfortunately, we don't always have the information about gene function. On the other hand, by applying comparative genomics, we can select evolutionarily conservative sites. Our approach is based on comparative genomics; in addition, we compute the minimal probability of selecting these sites by chance. This probability could reflect the motif quality. To verify this assumption, we used the following test.

It is known that, as a rule, each TF regulates its own gene. Even when other regulated genes for a TF are unknown, one can search for motifs upstream of TF gene in a set of closely related genomes. Such a motif is a first approximation, and can be improved

Table 1: Average results of testing on 62 Shewanella regulons. Sf (sensitivity) is the fraction of the correctly predicted operons among all the operons under regulation; Sp (specificity) is the fraction of the correctly predicted operons among all the predicted operons.

	universal score threshold		individual score threshold		no score threshold	
	Sf	Sp	Sf	Sp	Sf	Sp
local	0.74	0.74	0.92	0.72	0.86	0.81
medium	0.63	0.95	0.62	0.82	0.63	0.89
global	0.36	0.95	0.41	0.88	0.52	0.91

Table 2: Logarithm of the Bernoulli probability produced by the test of “de novo” motif prediction. The values for the true motifs are on gray background.

TF	motif	method		
		universal score threshold	individual score threshold	no score threshold
LexA	*LexA_0*	**-165.22**	-117.98	-62.21
	LexA_1	-143.41	**-118.85**	**-83.89**
NtrC	*NtrC_0*	**-89.54**	**-69.19**	**-45.84**
	NtrC_1	-56.7	-39.25	-29.47
PdhR	*PdhR_0*	-80.77	-56.56	-35.19
	PdhR_1	-68.65	-59.24	-35.99
	PdhR_2	**-87.81**	**-62.27**	**-37.55**
TrpR	*TrpR_0*	**-59.66**	**-38.9**	**-27.68**
	TrpR_1	-23.88	-14.96	-14.36
	TrpR_2	-54.26	-38.39	-20
TyrR	*TyrR_0*	-65.86	**-58.24**	**-38.37**
	TyrR_1	**-68.27**	-48.05	-32.27

by the further analysis. We selected several known regulons and used this strategy. We applied the motif search algorithm SignalX (Mironov, 2000) to upstream regions of TF genes in a set of Shewanella genomes.

Next, we applied our approach of regulon inference to the 2-3 best hits. In almost all cases, the correct motif has the lowest minimal Bernoulli probability (see Table 2). In PdhR motifs, the false motif (PdhR_2) has lower Bernoulli probability in all the approaches. The reason for this is that this motif is gc-reach; such a motif will always have a lower Bernoulli probability, because Shewanella genomes are gc-poor.

4 CONCLUSIONS

We have developed an approach for an automatic reconstruction of regulons in microbial genomes without arbitrary thresholds given the known TFBS motif. The approach allows automatic selection of the optimal TFBS score threshold and uses it for prediction of the regulon content. The testing of the approach on the manually curated regulons showed that it performs best on local regulons. In the case of medium and large size regulons, the approach allows automatic detection of the core of the regulon and has low level of overprediction. We also showed that the same probabilistic schema can be used for selection of the best candidate PWM in a course of motif search procedure.

ACKNOWLEDGEMENTS

This work was partially supported by Federal Agency on Education (#P999), Russian Ministry of Education and Science (Program “Research and Development”, contract #07.514.11.4007), Russian Foundation of Basic Research (grant #09-04-92742).

REFERENCES

Novichkov P. S., Laikova O. N., Novichkova E. S., Gelfand M. S., Arkin A. P., Dubchak I., Rodionov D. A., 2010. RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Res.* Jan;38(Database issue).

Kalinina O. V., Mironov A. A., Gelfand M. S., Rakhmaninova A. B., 2004. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci*. Feb;13(2).

Kotelnikova E. A, Makeev V. Yu, Gelfand M.S., 2005 Evolution of transcription factor DNA binding sites. *Gene*. 347 (2).

Mironov A. A., Vinokurova N. P., Gel'falnd M. S., 2000. Software for analyzing bacterial genomes. *Mol Biol (Mosk)*. 34(2).