Assessment of Transcription Factor Binding Motif and Regulon Transfer Methods

Sefa Kilic and Ivan Erill

University of Maryland Baltimore County
Department of Biological Sciences
1000 Hilltop Circle, Baltimore, Maryland 21250
{sefa1,erill}@umbc.edu

Introduction. Comparative genomics has been leveraged in many studies to characterize transcriptional regulatory networks [1,2]. However, despite its fundamental importance in such studies, the effect of motif and regulon transfer methods remains largely unstudied. Thanks in large part to high-throughput experimental techniques, available experimental data has increased dramatically over the last few years and it has become possible for the first time to reliably assess methods used for regulatory network reconstruction. In this study, we describe three different transfer methods that define transcription factor (TF) binding motif in a target species given some regulatory activity information in a reference species. Motif-based transfer is performed using the reference binding motif to search for putative binding sites in the target genome on the assumption that, for a given TF, the binding relatively well conserved across closely related species. This method has been shown to perform well at inferring existing regulatory networks in previously uncharacterized genomes [3,4]. The alternative source of prior information is the regulatory network itself. The putative regulon is then constructed based on orthologous transfer of the reference regulon and *de novo* motif discovery is performed on the promoter regions of putatively regulated target genes.

Methods. We compiled binding site data from publicly available databases, mostly from CollecTF [5], a database of experimentally validate sites. The first method that we tested is the direct transfer using the collection of known binding sites from a model species to build a position-specific scoring matrix (PSSM) which is used then to scan the promoter regions of the target genome to identify putative sites. The second method defines the motif by performing motif discovery on pre-searched candidate sequences. After the PSSM search, promoters with high scoring sites are given as input to the motif discovery algorithm with the motivation of capturing motifs slightly different from the reference one and mitigating the effect of inaccurate PSSM score threshold. The final method that we tested, called network transfer, does not assume motif conservation unlike the other two methods which are expected to perform poorly if the motif is not conserved. The underlying hypothesis is that the regulon across two genomes might be functionally conserved to some degree even if the binding motif is not. To define the motif in target species through network transfer, the first step is to identify target regulon, the collection of genes that are orthologous to the ones in the reference regulon. In the next step, the promoters of operons in the

target regulon are used for motif discovery. To assess the performance of different transfer methods quantitatively, we measured both (a) Euclidean distance between the true motif and the inferred motif and (b) the area under ROC curve for the inferred motif. To assess the significance of performances, we computed the distance and area under ROC curve using a column-permuted version of the target motif as the inferred motif.

Results. We measured the performance of the transfer methods by applying them to all pairs of species with at least 10 binding sites for a particular TF, yielding 411 pairs of species where most of them belong to either Fur or LexA. Our results show that direct transfer and motif discovery on pre-searched promoters perform very similarly. Since these two methods are based on motif conservation, they perform well when the TF proteins in the reference and target species are highly similar. As the TF protein distance increases, their performances decrease dramatically. Although network transfer is capable of inferring non-conserved motifs for large protein distances in many cases, our permutation analysis showed that network transfer method does not perform significantly well for any level of reference-target TF distance overall. Our finding is consistent with previous studies reporting high plasticity in transcriptional regulatory networks [6]. Another reason for poor network transfer is the strictness of the orthology-based regulon transfer method. As future work, we intend to relax the network transfer method by using functional similarity (e.g., cluster of orthologous groups) for regulon transfer rather than direct orthology. Also, we plan to investigate whether combining the information from the extended network transfer with relaxed PSSM searches can enhance the performance of direct transfer as the similarity between reference and target motifs decays.

References

1. Ravcheev, D.A., Best, A.A., Sernova, N.V., Kazanov, M.D., Novichkov, P.S., Rodionov, D.A.: Genomic reconstruction of transcriptional regulatory networks in lactic acid bacteria. BMC Genomics **14** (2013) 94
2. Meireles-Filho, A.C., Stark, A.: Comparative genomics of gene regulation—conservation and divergence of cis-regulatory information. Current opinion in genetics & development **19**(6) (2009) 565–570
3. Leyn, S.A., Kazanov, M.D., Sernova, N.V., Ermakova, E.O., Novichkov, P.S., Rodionov, D.A.: Genomic reconstruction of the transcriptional regulatory network in Bacillus subtilis. J. Bacteriol. **195**(11) (Jun 2013) 2463–2473
4. Leyn, S.A., Suvorova, I.A., Kholina, T.D., Sherstneva, S.S., Novichkov, P.S., Gelfand, M.S., Rodionov, D.A.: Comparative genomics of transcriptional regulation of methionine metabolism in Proteobacteria. PLoS ONE **9**(11) (2014) e113714
5. Kılıç, S., White, E.R., Sagitova, D.M., Cornish, J.P., Erill, I.: CollecTF: a database of experimentally validated transcription factor-binding sites in Bacteria. Nucleic Acids Research (2013)
6. Price, M.N., Dehal, P.S., Arkin, A.P.: Orthologous transcription factors in bacteria have different functions and regulate different genes. PLoS Comput Biol **3**(9) (09 2007) e175