

# Integrated prediction of transcription factor binding sites and interactions in bacteria: a comparative genomics approach

Xin-Yi Chua<sup>1</sup>  
x.chua@qut.edu.au

James M. Hogan<sup>1</sup>  
j.hogan@qut.edu.au

<sup>1</sup> Faculty of Information Technology, Queensland University of Technology, 2 George St, Brisbane, Australia

## Abstract

A central goal in computational biology is to model accurately cellular processes such as transcriptional regulation. Transcription Regulatory Networks (TRNs) have proven to be invaluable scientific tools in bioinformatics, but regulatory interactions (RI) are inferred based only upon the identification of transcription factors (TFs) and target genes (TGs) orthologous to TFs and TGs for which an interaction is known [1-4]. Transcription factor binding sites (TFBSs) are omitted from the model. This work investigates the effect of TFBS inference on the TRNs identified, the goal being to filter putative interactions to improve the reliability of the networks. Unfortunately, the high false positive (FP) rate of TFBS predictions remains a hurdle. We describe a new approach to filter the FPs based on a set of independent features which significantly reduces the effect of this problem.

**Keywords:** transcriptional regulation, transcription factor binding sites, prediction, superfamily

## 1 Introduction

Transcriptional Regulatory Networks in bacteria are usually inferred through clustering of microarray expression data, and this approach has been used in the two most prominent model organisms – *E.coli* and *B.subtilis* [1,2]. While in time these approaches may be applied to a far broader set of organisms, it is common to rely at least initially on inference based on a model TRN via the Regulog assumption [1,2]. Under regulog, a regulatory relationship is assumed to exist if orthologs are found for both the TF and TG of a known interaction in the model. Such inferences may be used directly, or to inform the design of microarray experiment. In each case, identification of the associated TFBS may improve our confidence in the inferred relationship, and allow a more detailed understanding of the regulatory mechanism. However, such predictions are plagued by high FP rates, introducing an alternative form of uncertainty. Some studies have integrated prior knowledge to bias the prediction task, but these approaches are unsuitable for exploratory comparative genomics where biologists often do not know what they are expecting. Our focal point is to utilise a more extensive and general feature set to filter binding site candidates as a post-processing step. We present the results of two initial predictors.

## 2 Discussion and Results

### 2.1 Location Probability

Previous studies have shown TFBSs are generally located within specified regions with respect to the transcription start site (TSS) [2,4]. Repressors occur downstream of promoters while activators occur upstream. Using *E.coli* as our model organism, for each TFBS prediction we calculated a normalised *location probability* (LP) score based on prior distributions. Using LP score alone we were able to reduce the number of FP by 33% while sacrificing only 6% of TP. Our initial analysis revealed LP can be a good predictor but is limited to scenarios where the TSS is experimentally confirmed.

## 2.2 Superfamily share similar binding sites

Based on the assumption that structurally similar TFs share similar TFBSs we assessed how this would perform as a predictor. Previous groups have demonstrated the success of using *familial binding profiles* (FBPs) to bias TFBS predictions [6,7]; however, this approach is unsuitable for exploratory comparative genomics since biologists have no knowledge of the expected profile, and incorporating incorrect priors would overlook noteworthy motifs. We compared the TFBSs of 58 TFs against each other using T-Reg Comparator [5]. The mean *dissimilarity scores* (DSs) [5] for TFs within the same and those from different Superfamilies [8] are illustrated in Figure 1. This formed the basis of our predictor and a DS is calculated for each TFBS prediction. By combining LP and DS values, we reduced FP by 44% while sacrificing 7% of TP. The two predictors show potential in improving TFBS predictions without specific prior knowledge.

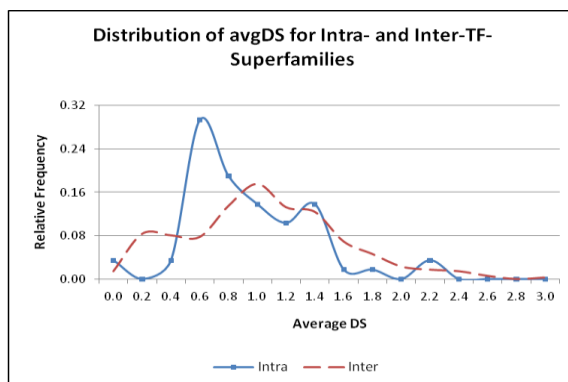


Figure 1: Comparison of TFBSs for TFs within the same (solid) and different (dashed) superfamilies. We assessed the performance of this feature as a potential predictor.

## 3 Conclusion

This study demonstrated both LP and DS show promise as predictors and are worthy of further investigation. We are interested in a set of other features including the ability to use pair-wise distributions of Superfamily binding sites to further reduce FPs, thereby increase TRN performance as aids for comparative genomics.

## References

- [1] Babu, M., Sarah T., and Aravind, L., Evolutionary Dynamics of Prokaryotic Transcriptional Regulatory Networks, *J. Mol. Biol.*, 358:614-633, 2006.
- [2] Espinosa, V., Gonzalez, A.D., Vasconcelos, A.T., *et al*, Comparative Studies of Transcriptional Regulation Mechanisms in a Group of Eight Gamma-proteobacterial Genomes. *J. Mol. Biol.*, 354:184-199, 2005.
- [3] Lozada-Chavez, I., Sarath C.J., and Collado-Vides, J., Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Research*, 34(12): 3434-3445, 2006.
- [4] Rodionov, D., Comparative Genomic Reconstruction of Transcriptional Regulatory Networks in Bacteria. *ChemInform*, 38(46): 3467-3497, 2007.
- [5] Roepcke, S., Grossman, S., Rahmann, S., and Vingron M., T-Reg Comparator: an analysis tool for the comparison of position weight matrices. *Nucleic Acids Research*, 33:W438-W441.
- [6] Sandelin, A., and Wasserman W., Constrained Binding Site Diversity with Families of Transcription Factors Enhances Pattern Discovery *Bioinformatics*. *J. Mol. Biol.*, 338: 207-215, 2004.
- [7] Tan, K., Lee A.M. and Stormo, G., Making connections between novel transcription factors and their DNA motifs. *Genome Research.*, 15:312-320, 2005.
- [8] <http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/>