Transcription factor centric discovery of regulatory elements in mammalian genomes using alignment-independent conservation maps

Nilanjana Banerjee and Andrea Califano

Department of Biomedical Informatics, Columbia University, New York, New York 10032

Introduction

The computational identification of DNA binding sites that have high affinity for a specific transcription factor is an important problem that has only been partially addressed in prokaryotes and lower eukaryotes. Given the higher length of regulatory regions and the relative low complexity of DNA binding signature, however, methods to address this problem in higher order eukaryotes are lacking. Thus, the specific interactions between transcription factors (TFs) and their cognate cis-regulatory elements – ultimately responsible for transcriptional network function – remain largely unmapped in a mammalian context.

Method

Our method assumes that a candidate pool of targets of a specific human transcription factor is available. In the context of this paper these were generated either by mining the literature data (only for the set of biochemically validated targets of *MYC*) or by network reverse engineering using the ARACNE algorithm [Basso et al., 2005, Margolin et al., 2005].. Targets were inferred from a phenotypically rich collection of 254 microarray expression profiles of normal and tumor related human B lymphocytes, also described in [Margolin et al., 2006].

Targets that are positively and negatively correlated with the transcription factor are assembled in separate sets and analyzed separately by our method (details shown in Figure 1).

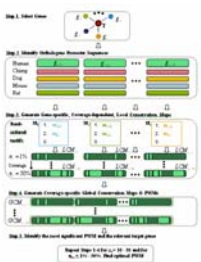


Fig. 3. Schematic diagram of our TF-centric approach to discover regulatory elements using alignment-independent conservation maps. Step 1 identifies candidate targets of a TF by ARACNE bootstrapping. In Step 2, the orthologous promoter sequences are extracted. Step 3 generates gene-specific conservation maps varying over percent coverage. Step 4 compiles a global conservation map over all targets with varying percent coverage. The process is repeated for varying number of top-ranked targets. Finally, Step 5, identifies the 2 most significant PWMs for a TF – the relevant target genes and the associated percent coverage.

Results

We report the optimal PWMs identified by our approach for eight transcription factors. We chose six TFs whose binding sites have been well-characterized (*MYC, E2F1, TFDP1, IRF7, FOSL1* and *NFkB2*) and two TFs which do not have well-characterized binding sites (*BCL6* and *HOXD13*).

Transcription Factors with a previously known PWM

We first tested our approach on the positively correlated (activated) and negatively correlated (repressed) putative targets of the *MYC* proto-oncogene. Analysis of the positively-correlated targets yields the well-known binding site (E-box) as the optimal PWM for both tested motif lengths ($w = 6$ and $w = 8$). The optimal statistical significance was obtained for top-ranked putative targets by ARACNE and n_c=12%). Figure 2a shows the statistical significance of the PWM enrichment as a function of both n_t and n_c.

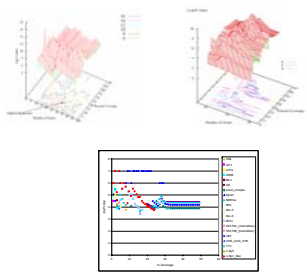
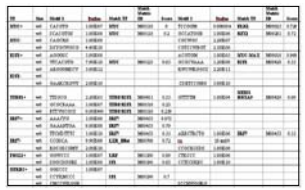


Fig. 2. (a) Analysis of ARACNE-predicted positively-correlated MYC targets showing the statistical significance of PWM enrichment as a function of both percent coverage (n_c) and number of genes (n_t). The E-box was the optimal PWM (w=6) with n_c=12% and n_t= 30. **(b)** Analysis of validated MYC targets. E-box is the most significant PWM. The significance value is high even for n_t= 10 target genes and increases modestly as more validated genes are included. **(c)** All PWMs generated from n_t= 30 and n_c= 1 to 40 were matched against the TRANSFAC database. TFs MYC, RFX1 and ELK1 have close match (K-L divergence score) with the some of the most significant PWMs.

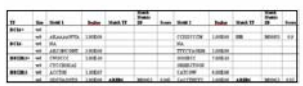
Table 1. Statistically significant PWMs discovered for TFs with a previously known PWM



The complete set of discovered PWMs (for previously characterized TFs), matched matrices and significance values are shown in Table 1. We find that four of the six tested TFs (*MYC, E2F1, TFDP1* and *IRF7*) the discovered PWM matched the known ones at high specificity in the expected set of targets. Most of the other discovered PWMs match the profile of some other well-characterized binding sites in TRANSFAC or JASPAR, revealing possible combinatorial mechanisms. For *NFkB2* and *FOSL1* none of the discovered PWMS matched the known one.

Transcription Factors with a previously unknown PWM

Table 2. Statistically significant PWMs discovered for TFs with a previously unknown PWM



BCL6 is a transcriptional repressor whose activity is important in the creation of the Germinal Center and in formation non-Hodgkin's B cell lymphoma. The PWMs discovered in the positively-correlated targets do not match any known TRANSFAC or JASPAR motif. Analysis of the negatively-correlated targets, reveals that the second most significant PWM of length w=9 (TTYCYAGRM), with n_t = 28 and n_c = 29, closely matches a known BCL6 binding site discovered from in vitro binding assays.

HOXD13 belongs to the homeobox family of transcription factors. Analysis of the negatively correlated targets allowed the identification of a PWM (GDSVAGGTG) which is very similar to that of AREB6, a zinc-finger homeodomain enhancer-binding protein.

Further analysis of these discovered PWMs can possibly help understand the details of BCL6 and HOXD13 protein function.

Conclusions

We report a systems biology approach that combines network reverse engineering, comparative genomics, and bioinformatics analysis which is successful in identifying the correct binding site for a significant number of the tested transcription factors. We show that for 4 of the 6 previously characterized TFs, our approach is successful in discovering PWMs that closely match the previously known profiles.

For *BCL6* we show that the inferred PWM matches a previously known binding site instance. In addition, for HOXD13 negatively-correlated targets we identify a PWM that matches (albeit with some variability) with the known motif for another homeodomain protein.

The new approach innovates on previous methods in several ways: (1) while sets of co-regulated genes were used before to identify binding sites, we rely on a network reverse engineering algorithm (ARACNe) to identify putative targets of a known transcription factor for further analysis. This allows to causally link the discovered profile with a specific transcription factor; (2) rather than relying on sequence alignment, the method uses pattern discovery to identify regions that are highly conserved across orthologous promoter regions in several mammalian genomes; (3) rather than using ad-hoc parameters, the method systematically explores the parameter space to discover the most statistically significant PWMs.

Literature cited

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R. & Califano, A. 2005. Reverse engineering of regulatory networks in human B cells. *Nature Genetics* **37,** 382-90.

Califano, A. 2000. SPLASH: structural pattern localization analysis by sequential histograms. *Bioinformatics* **16**(4):341-57.

Margolin A.A., Nemenman I., Basso K., Wiggins C., Stolovitzky G., Dalla Favera R., and Califano A., 2006. *ARACNE*: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context, *BMC Bioinformatics*, **7**(Suppl 1):S7.

Acknowledgments

We thank H. Erikson for computer assistance and P. Morozov, M. Kustagi, K. Basso, R. Dalla-Favera for helpful discussions. This study was supported in part by NCI grant R01 CA109755-01(PI:AC).