

Strategies for computational transcription factor binding site discovery in humans

Geoff Macintyre¹

gmaci@csse.unimelb.edu.au

James Bailey¹

jbailey@csse.unimelb.edu.au

Adam Kowalczyk²

adam.kowalczyk@nicta.com.au

Izhak Haviv³

izhak.haviv@bakeridi.edu.au

¹ Department of Computer Science and Software Engineering,
University of Melbourne, Australia

² Victoria Research Laboratory, National ICT Australia, Melbourne, Australia

³ Baker IDI Heart and Diabetes Research Institute, Melbourne, Australia

Background

Computational approaches for discovery of Transcription Factor Binding Sites (TFBSs) are commonly focused around the Transcription Start Site (TSS) of a gene, typically 1kb to 10kb upstream. The promoter proximal search space is popular due to the fact that historically, TFBSs have been determined through experiments, both empirical and computational, on low order model organisms such as yeast or drosophila, where the majority of binding sites lie close to the TSS. This search space may not be suitable in humans, given the increase in complexity of the gene regulation system. In fact, a study into a cis-acting regulator causing preaxial polydactyly demonstrated a case where a TF was empirically determined to act over a distance of around 800kb [1].

Analysis

Recently, chromatin immunoprecipitation coupled with high-throughput sequencing technologies has provided the ability to empirically observe the behavior of particular Transcription Factors (TFs) and map their binding sites genome-wide, providing a valuable tool for the validation of TFBS discovery methods and elucidation of gene regulatory networks. We compiled a study of five different ChIP-PET and ChIP-SEQ datasets identifying TFBSs for ER [3], RELA [2] STAT1 [4], Myc[6] and p53 [5]. We matched TFBSs with carefully compiled lists of target genes highlighting typical genomic distances of TF action. Where possible genes upregulated in the presence of the protein synthesis inhibitor cycloheximide were selected as primary targets of a TF. See Figure 1 for a summary of results.

Strategies for TFBS discovery in humans

Figure 1 shows that average of only 28% of genes had a TFBS within 10000bp of their TSSs. In addition, all TFs showed relatively equal distributions of upstream vs downstream binding. Most importantly, on average, 57% of the binding site target gene pairs had genes residing between the TFBS and the TSS. These results suggest the following strategies may be adopted to improve future TFBSs discovery methods in humans:

- Identification of TFBSs must shift focus to genomic regions outside the proximity of the TSS.
- Careful consideration needs to be given to genes deemed as direct targets of a TF.
- Regulatory action of a TF should be considered beyond the boundary of the nearest gene.

Unfortunately, with large input sequences, the majority of current computational TFBS discovery approaches suffer from many false positive predictions. However with intelligent use of the increasing amount of high-throughput genomic data, we believe that current approaches can be improved to scale well when applied to the large genomic search spaces required in humans.

Category	STAT1	RELA	ER	p53	Myc	Average
TSS proximal binding	54.55	4.17	10.75	22.64	2.7	18.96
Intragenic binding	8.46	0	8.6	11.32	1.35	5.95
Upstream binding	18.81	45.83	44.09	28.3	54.05	38.22
Downstream binding	18.18	50	36.56	37.74	41.89	36.87
TFBS <5000bp	63.01	4.17	15.05	33.96	6.76	24.59
TFBS 5000-10000bp	4.7	1.39	3.23	7.55	1.35	3.64
TFBS 10000-50000bp	14.42	6.94	17.2	37.74	14.86	18.23
TFBS 50000-100000bp	6.27	4.17	11.83	9.43	18.92	10.12
TFBS 100000-200000bp	4.08	11.11	10.75	7.55	32.43	13.18
TFBS >200000bp	7.52	72.22	41.94	3.77	25.68	30.23
TFBS that skip over genes	19.75	87.5	56.99	37.74	83.78	57.15

Figure 1: This figure gives a summary of the frequencies of TFBS/target gene pairs under different categories. All values represent the percentage of pairs observed in a category for a particular gene list and TF. The last column is an average over all gene lists.

References

- [1] Lettice LA, *et al.*, Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proceedings of the National Academy of Sciences* 99:7548–7553, 2002.
- [2] Lim, C. *et al.*, Genome-wide Mapping of RELA(p65) Binding Identifies E2F1 as a Transcriptional Activator Recruited by NF-B upon TLR4 Activation. *Molecular Cell* 27, 622-635, 2007.
- [3] Lin CY, *et al.*, Whole-Genome Cartography of Estrogen Receptor α Binding Sites. *PLoS Genetics* 3:867–885, 2007.
- [4] Robertson G, *et al.*, Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Meth* 4:651–657, 2007.
- [5] Wei CL, *et al.*, A Global Map of p53 Transcription-Factor Binding Sites in the Human Genome. *Cell* 124:207–219, 2006.
- [6] Zeller KI, *et al.*, Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proceedings of the National Academy of Sciences* 103:17834–17839, 2006.