

Regulatory signals

Project Assignment – Bioinformatics module 2009

March 23, 2009

1 Introduction

Transcriptional regulation plays a vital role in the development and functioning of each cell of any organism. It is the key link between components of regulatory networks that ultimately determine the expression level of each individual gene in a cell. Several genetic diseases have been shown to be caused by a defect in gene regulation [1].

There are a number of different factors involved in the regulation of transcription. Some of these work by binding to the RNA polymerase itself and some do bind to the DNA chain, to specific sites upstream, downstream or at introns within the gene's coding sequence. In this project we are not addressing protein-protein type interactions but rather concentrate on detecting binding sites that could potentially play a role in the regulation of a target gene (i.e. cis-regulatory elements), including binding sites of transcription factors, activators, repressors and – within the promoter – of the RNA polymerase itself.

One approach to identifying putative cis-regulatory elements is through methods that look for known motifs. These motifs are a few base-pair long consensus sequences of experimentally verified binding sites usually given in the form of position-specific weight matrices (*PSWMs*). A number of tools exist that can perform such a search.

However, we take the comparative approach that has the advantage of being able to predict novel binding sites [3, 2]. The key observation that lies behind this family of methods is that sites involved in regulation are significantly more conserved across closely related species than other non-coding regions lacking biological function. By detecting highly conserved sites in a homologous set of sequences it is possible to identify most of the regulatory elements. Of course, binding sites can appear and disappear over time, so the success of the analysis largely depends on the choice of the species involved.

2 Your task

Given a selected gene, your task is to identify all of the conserved motifs in the upstream region that could potentially correspond to a binding site and thus take part in the regulation of the gene.

1. Select a gene of interest. This could either be a human gene (e.g. one of *CHRNA5*, *CHRNA3* or *CHRNA4*) or you can make your life easier by analysing a bacterial gene, such as *motA*.

2. Download the upstream region of the selected gene and its homologs from 4-5 species that are relatively closely related to the target one. The length of the upstream sequences should be at least 1000 for a human and 300 for a bacterial gene. Collect all your sequences in a FASTA file.
3. Download *BigFoot*, a recently developed tool for TFBS detection from <http://www.stats.ox.ac.uk/~satija/BigFoot/>
4. Build a phylogenetic tree from your sequences using either *ClustalW* or *FSA server*. These tools usually generate an unrooted tree that you will need to transform to a rooted one (choosing the position of the root arbitrarily).
5. Run *BigFoot* on your data, making sure you add your tree beforehand. The default number of MCMC cycles is 1 million which is usually needed to get optimal results. Note that several hours of computation will be required so leave enough time the analysis to finish.
6. Use the MPD view to identify highly conserved motifs: the red curve will show the level conservation at each site and the candidate TFBS motifs will be written in capitals.
7. Ideally you will get short conserved regions embedded into long variable regions. However, if too much or too little of the sequences is conserved you chose too closely/distantly related species, so you will have to replace some of your sequences and re-run the analysis.

When you have successfully identified a few putative regulatory motifs try to compare your results to experimentally verified data. Find regulatory databases for your chosen target species – for human sequences, *CisRED* is a good place to start, for bacterial genes *EcoCyc* is recommended.

If interested, you can try *PSWM*-based tools such as *MotifScanner* to search for known binding sites in your sequences or other methods that detect statistically overrepresented oligonucleotides [4].

References

- [1] M. F. Moffatt et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, July 2007.
- [2] R. Satija, I. Miklós, Á. Novák, R. Lyngsø, and J. Hein. BigFoot: Bayesian alignment and phylogenetic footprinting with MCMC. *BMC Evolutionary Biology*, 2008. (submitted).
- [3] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050, August 2005.
- [4] S. Sinha and M. Tompa. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, 30(24):5549–5560, 2002.