

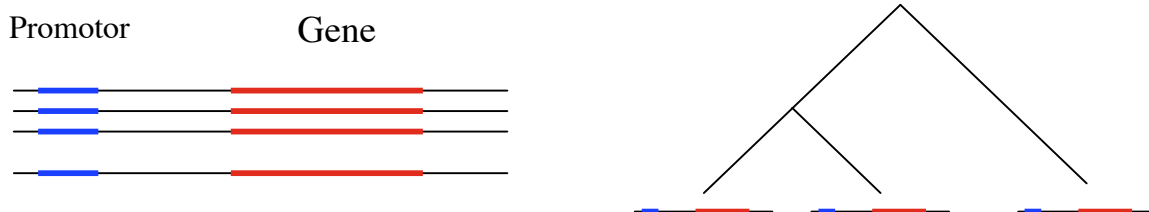
A Unified Approach to Signal Detection

Combining Footprinting with Motif Search

15.5.08

Background and Motivation. Two major classes of methods for finding regulatory signals are homologous and non-homologous analyses. Homologous will find signals by their modified or lower rate of evolution relative to a background sequence. Non-homologous will find signals because they occur with a high frequency, than in a background sequence. These two approaches have been combined before (Wang and Stormo, 2003; Zhou and Wong, 2007). Combining the approaches should have several advantages. Homologous detection is increasingly powerful due to the increase in know genomes. Non-homologous analysis is powerful in making functional statements since it can compare genes with different properties or that are co-regulated. This project proposes to unify the two approaches within the framework of statistical alignment.

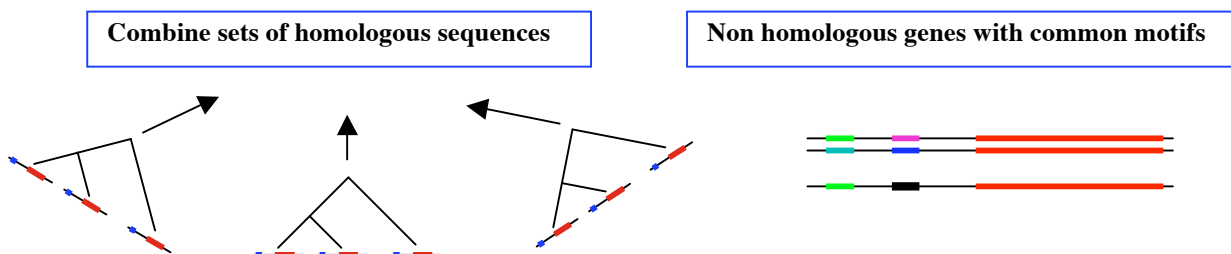
Acquiring knowledge of a gene is central in interpretation of what it does. A gene is close to the “the atom” of molecular biology (Keller, 2002) and must be annotated in terms of structure, selection profile, constraints, regulatory signals and its relation to other genes. In its extreme form, this necessitates understanding of the complete organism as a gene is part of a large interacting network of genes and signals. So making restrictions of this problem to define a worthwhile and still challenging problem is necessary. This project focuses on the identification and the characterisation of regulatory elements, which is likely to be a major challenge for quite some time to come and this project would easily scale to a complete PhD. However, it is also possible to formulate a pilot version, that is of interest in itself.



On the left we see a set of non-homologous genes and we search for a common motif in all of them. This can be generalized to zero multiple copies of the motif or a battery of motifs. The key word is non-homologous!! On the right the key word is homologous and motifs will be detected by a lowering of evolutionary rates or a change in the evolutionary process.

It is still a significant challenge to recognise and categorise the wide range of *cis*-acting regulatory elements by computational analysis alone (e.g. promoters, enhancers, locus control regions, boundary elements and silencers) that control gene expression (Maston et al., 2006).

Homologous analysis has recently become more powerful due to the availability of many orthologous, non-coding sequences from a variety of species separated by a wide evolutionary time scale (~50-500 MYRs, Prakash and Tompa, 2005). It has emerged that many multi-species conserved regulatory elements contain highly conserved transcription factor binding sites. Recently, phylogenetic footprinting (Blanchette and Tompa 2003; Taylor *et al.* 2006) has been very popular and it is clear that this can be put on a better statistical basis by using the framework of statistical alignment (Hein *et al.* 2000) that could be combined with phylogenetic models of regulatory signals (Jensen *et al.* 2005). Satija, Pachter and Hein (2008) have recently combined footprinting and statistical alignment into a combined method, that has substantial advantages, but can only analyze up to 5 sequences. One major advantage of this method is that it not dependent on a single alignment and a consequent ability to incorporate more distant genomes in the analysis. Satija, Miklos, Novak and Hein (2008) have used MCMC to extend the capability to 10-13 sequences. We are presently extending this further considerably further to include refined description of the molecular evolution of the regulatory signal and allow the analysis of hundreds of sequences.



Left: Now we have a set of sets of genes/sequences. Within a set of sequences the sequences are homologous, but between the sets they are not. Within a set statistical alignment and footprinting can be applied. However, one can't just use the methods used in the first illustration to find motifs in non-homologous as now we don't have sequences, but sets of sequences. Wang and Stormo solved this by using profiles that essentially reduced a set of sequences to one sequence. It would clearly be a major advantage if statistical alignment and footprinting could be generalized to cover this situation.

Right: This shows realisations of genes from a model where genes are independent, but motifs can be identical illustrated by little boxes of same color.

Non-homologous analysis provides a very different framework, since tools such as alignment and evolutionary model cannot be used. However, such comparison provides important additional information since signals can be common for functional reasons. This can only be revealed by comparing non-homologous genes. Lawrence et al. (1993) provided a non-homologous method that found a single overrepresented motif in a series of unrelated strings. The method used a Gibbs sampler that both found and defined the motif. This work has since been followed up by a series of useful generalisation by for instance Jun Liu. For instance the description of the motif can use a first order Markov chain instead of assuming independence between positions and could allow gaps in the signal. The assumption that 1 motif exactly must be present can be relaxed to an arbitrary number, could be of different kinds or the order of the motifs could itself be determined by a Markov chain.

Combining *homologous and non-homologous* analyses could provide major advantages. Wang et al. (2003) combined homologous with non-homologous methods, but in a non-statistical fashion. Zhou and Wong (2008) combined alignment HMM (not evolutionary models) by tying together description of common motifs.

Clearly the main challenge is how to handle the non-homologous situation, but there clearly are models where this can be done. If the footprinting model has a probabilistic descriptions of motifs, then once could choose to have common description of motifs between sequences or motifs for each set of sequences. This would tie sequences analyses together: Within sets we have evolutionary models and between the models are only tied together by the motif descriptions.

Data example. Start with the gene ORMDL3 and the 5 references as mentioned in the project “Fine Scale Regulatory Annotation of a Gene” and expand this with the most obvious other Asthma genes.

Project Plan

- Weeks 1-2 Read the literature, collect data and test key programs – SAPF, MultiModule, PhyloGibbs, Footprinter and PhastCons. Also read papers on ORMDL3
- Week 3 Create 3-5 data sets or increasing evolutionary time span with associated annotations – for instance Primates, Mammals, Vertebrates covering first many genes of low shallow time depth to fewer genes with longer time depth.
- Week 4 Run PhastCons (Siepel and Haussler, 2004) , SAPF (Satija, Pachter and Hein, 2008), MultiModule (Zhou and Wong, 2007), PhyloGibbs, and Footprinter on these.
- Week 5-6 Develop and implement algorithms that find segments of high probability common to a set of given HMMs.
- Week 7 Run the programs on HMMs created by SAPF.
- Week 8 finish report.

Comments. Clearly this project can be applied to any small set of genes of interest.

References

- Blanchette and Tompa (2003) FootPrinter: a program designed for phylogenetic footprinting. *Nuc. Acids Res.* 31: 3840-3842.
- Conlon, Liu, Lieb and Liu (2003) Integrating Motif Discovery and Expression Analysis *Proc.Natl.Acad.Sci.* 100.3339-44
- Evelyn Fox Keller (2002) “Century of the Gene” Harvard University Press
- Hein J., Wuof C., Moller M.B., Knudsen B., Wibling G. (2000) Statistical alignment: computational properties, homology testing and goodness-of-fit. *J Mol Biol* 302: 265-279.
- Janky and van Helden (2008) Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution *BMC Bioinformatics.* 2008; 9: 37-
- Jensen S., Shen L. and Liu J. (2005) Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics* 21: 3832-3839.
- Lawrence, Altshul, Boguski, Liu, Neuwald and Wooton (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 8:262(5131):208-14
- Maston, GA et al. (2006) Transcriptional regulatory elements in the human genome. *Ann. Rev. Genomics and Hum. Genet.* 7: 29-59.
- Prakash A. and Tompa M. (2005) Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.* 23:1249-1256.
- Ptashne (2002) “Signals and Genes” CSHL Press
- Satija R, L. Pachter and J. Hein (2008) “Statistical Alignment and Footprinting” (in press *Bioinformatics*)
- Satija R, I.Miklos, A. Novak and J. Hein (2008) “MCMC Statistical Alignment Footprinting” (submitted to *Bioinformatics*)
- Siddharthan, Siggia and van Nimwegen (2005) *PhyloGibbs PLOS Biology*
- Siepel, A. and Haussler D. (2004) “Combining phylogenetic and Hidden Markov Models in biosequence analysis.” *J. Comput. Biol.* 2004 11:413-428.
- Siepel, A. (2005) “Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes” *Genome Res.* 15:1034-1050
- Sinha and He (2007). MORPH: Probabilistic alignment combined with Hidden Markov Models of cis-regulatory modules. *PLoS Computational Biology.* 3(11):e216.
- Taylor, M.S. et al. (2006) Heterotachy in mammalian promoter evolution. *PLoS Genetics* 2: 30-39.
- Wingende et al. (2000) “TRANSFAC: an integrated system for gene expression regulation” *Nucleic Acids Research.* 2000, Vol. 28, No. 1 316-319
- Zhou, Q. and WH Wong (2007) “Coupling Hidden Markov Models for the Discovery of Cis-regulatory Modules in Multiple Species” *Annals of Applied Statistics* 1.1:36-65.
- Wang and Stormo (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs *Bioinformatics* Vol. 19 no. 18 2369-2380
- Wasserman WW and Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics.* 2004;5:276-287.

Comments. i. The main difficulty in the project stems from the fact that statistical alignment is so hard – writing the SAPF corresponds to more than a years work for Rahul Satija – so this project can only be done by efficient collaboration with for instance Rahul Satija. This will need a detailed work plan for writing the software. ii. Clearly if regulatory motifs could be found efficiently, it would be a major achievement and of great value in genomic analysis. One should not expect to “solve” this problem. Firstly, motifs evolve very quickly, so the background model of conserved motifs is probably naïve. Secondly, there are other factors than a local sequence motif that determines transcription factor binding.