

Fine Scale Regulatory Annotation of a Gene

5.5.08

Background and Motivation. 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.

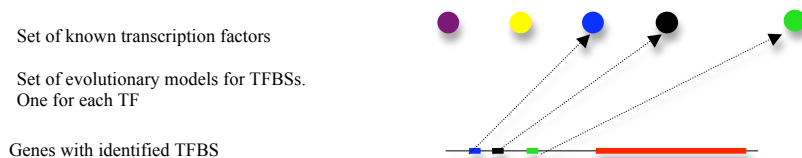


This is a cartoon of a standard gene. The red segments are exons, whose combined length must be a multiple of 3. The region between exons are called introns. Regulatory signals are shown in blue. Homologous variants of this gene is then observed in many species. Signals are often found by comparison and observing that they evolve slower than surrounding regions.

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).

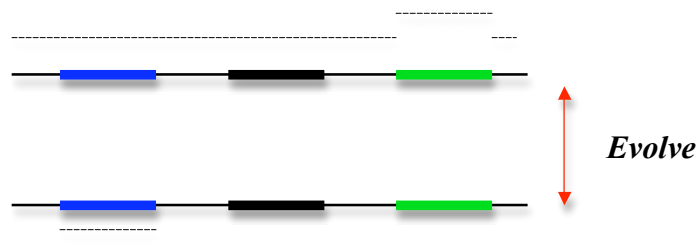
A ladder of ambition. The purpose of this section is both to state what is possible and what isn't.

1. Finding slowly evolving segments as indicator of regulatory signals. 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 (SAFP), that has substantial advantages, but can presently only analyze up to 5 sequences. One major advantage of this method is that it is 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 considerably further to include refined description of the molecular evolution of the regulatory signal and allow the analysis of hundreds of sequences. (see project description “k-restricted Steiner Trees as approximations to Phylogenies”)



2. Finding segments matching known transcription factors. (see illustration above). Knowledge based identification is relevant, if the regulatory signals are known then there are databases describing these and possible information concerning their interaction with regulatory molecules, and a probabilistic description of a signal is often done using a Hidden Markov Model. Knowledge based identification is quite limited and necessitates a degree of knowledge of regulation, that rarely is available. The database TRANSFAC (Wingende et al. (2000)) would be one example.

Additionally, a set of genes could be analyzed in conjunction with other data types such as expression levels and protein interaction data. Such models have been developed (for instance Conlon et al., 2002).



Here we have 2 homologous genes, but could have an arbitrary number. We only show their regulatory regions. One has been experimentally annotated for the blue TF and the other for the green, while no investigations have been made for the black TF. This knowledge can be entered as priors. Observing the genes and their evolution will define a posterior not only on the gene experimentally annotated, but this will be “transported” to the other genes by the evolutionary model.

3. *Homologous Knowledge Projection/Transfer among genes.* (see illustration above). The framework behind SAFP can provide basis for how to transfer knowledge of for instance transcription factors. An experiment relating a TF to a specific TFBS gives information specific for a given gene, specie and TF. This is most valid for a closely related species/gene. It is less likely to be valid if the relationship is more distant. An experiment can be represented as a prior on a gene for a TFBS at a given position and the evolutionary model determines how strong this prior is on other genes/species. See project description “Knowledge Transfer among Homologous Genes”

4. *Placing genes in a regulatory network (IG) and making predictive dynamic models of cells (SB).* The first of these topics is often associated with the field of Integrative Genomics (IG) and the second with Systems Biology (SB). Both fields needs data types beyond the sequence. IG is presently enjoying successes, but also failures. SB is extremely ambitious and is a more distant goal. See project description “Gaussian Processes and Gene Regulation”

Genes are also analysed by approaches that can go in at different places in this ladder:

A. *Non-homologous signal search.* It could be a major advantage to allow non-homologous analysis. This 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 and this can only be revealed by comparing non-homologous genes. Lawrence et al. (1993) provided a non-homologous method. Wang et al. (2003) combined homologous with non-homologous methods, but in a non-statistical fashion. For a suggestion of how to do this in a more statistical fashion, look at the project description “A Unified Approach to Signal Detection”.

B. *Functional analysis of a group of non-homologous genes.* Grouping genes as co-regulated, having the same function (for instance according to Gene Ontology) or according to some other criteria is important in formulated hypotheses about the genes.

1) can be done if the appropriate data is available (which it normally is in the form of many closely related genomes), but still with some error. It might seem hopefully unambitious, but is in reality very valuable due to the “needle in haystack” problem that genomic analysis present as what is functionally important for a given problem can be a very small fraction of the complete genome. No biologist believes methods promising to make predictive dynamic models from the genome). But even attempting to devise models that can do 4) can be valuable, since this is what a biologist will try to do manually. In several cases (2-4) it is possible to sketch a method, but there presently exists no programs that can be used.

Biological Issues to be addressed

How detailed an annotation can be achieved by comparative methods? Finding multi-species conserved segments has been done with great success in recent years and is now the standard starting point for any analysis relating to regulatory signals, but key questions remain unanswered: a. How detailed can the annotation go below the segment level toward assigning functional constraints on individual nucleotides? b. What is it that is conserved in regulatory signals? This is a much more difficult question than the analogous question for RNA structure or protein genes, and addressing the question might need much more data, in particular data from deeper species comparisons. The answer may involve physical-chemical parameters describing the potential of DNA segments to interact with regulatory proteins. Such descriptors are known and could directly be used to investigate, for instance, DNA-flexibility of a region relative to a random evolution model.

We will take a single gene as a test case and annotate it maximally by comparative methods. ORMDL3 is a recently identified susceptibility gene for asthma (Moffatt et al., 2007) It is member of a conserved new family of endoplasmic reticulum membrane proteins (Hjelmqvist et al.,2002). Its function is unknown. It appears to be differentially regulated in asthma. There are three human ORMDL genes, on different chromosomes. The genes themselves are high similar, but their regulator regions may vary. The project will annotate the 3 ORMDL regulatory regions in humans, and compare these to homologous regions in other species. The function of the gene is unknown, but has been unambiguously identified with Asthma in children and a third of afflicted children has a specific allele of this gene, so any information on regulation will be exciting and of great value.

Project Plan

- Weeks 1-2 Read the literature, collect data and test key programs – SAFP, 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) , SAFP (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 SAFP.
- Week 8 finish report.

Comments. Clearly this project can be applied to any gene of interest. Step one of the ladder is straight forward, but going to 2 and 3 will take some methodology development and programming.

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., Wiuf C., Møller 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.

ORMDL3:

- Galanter et al. (2008) "ORMDL3 Gene is Associated with Asthma in Three Ethnically Diverse Populations." *Am J Respir Crit Care Med*. 2008 Feb 2
- Hjeltnqvist, Tusaon, Marfany, Herrero, Balcells and Goncalves-Duarte (2002) "ORMDL proteins are a conserved new family of endoplasmic reticulum membrane proteins" *Genome Biology* 3.6.1-27
- Moffatt et al. (2007) "Genetic Variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 448.26.470-74.
- Tavendale et al.(2008) "A polymorphism controlling ORMDL3 expression is associated with asthma that is poorly controlled by current medications." *J Allergy Clin Immunol*. 2008 Apr;121(4):860-3.
- Zhang, Pare and Sandford (2008 Jan) "Recent Advances in Asthma Genetics" *Respir. Res.*