

Computational Promoter Analysis of non-Coding RNAs

17.7.07

One of the major recent scientific surprises has been the discovery of a variety of functional RNAs and that a large fraction of the genome is under purifying selection. It is clear that a large fraction of RNAs is involved in gene regulation and is especially important in brain development. Until recently only coding genes have been studied in detail, but it has been recently proposed that non-coding genes might play an essential role in development and disease (Ponting and Lunter, 2006; Mehler and Mattick, 2006; Pollard et al., 2006). The fact that these sequences are highly conserved during evolution is indicative of a presently unknown function. Moreover, there are indications that functional non-coding genes that are evolving under either negative or positive selection are particularly enriched in brain expression, hinting at hitherto unforeseen roles in controlling brain developmental programs.

Optimal use of comparative genomics in this context presents two challenging problems: RNA gene identification and the characterisation of regulatory elements. The project focuses on the latter, 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.

It is now possible to identify and annotate from RNA genes (Washielt et al., 2005), although it is harder than the corresponding protein gene problem. 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). This problem can be classified dependent on the nature of the data into: *Homologous*, *non-homologous* comparisons and *knowledge based* identification.

- *Homologous* analysis has recently this has become more tractable by comparing orthologous, non-coding sequences from a variety of species separated by a wide evolutionary time scale (~50-500 MYs, 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. 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.
- *Non-homologous* analysis provide 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 and Wang et al. (2003) combined homologous with non-homologous methods, but in a non-statistical fashion.
- *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.
- Additional data to pure genomic data can come in the form of population variation data, expression data, molecular evolution rate/selection or functional annotation. All these sources are of major value in increasing the power of analysis and providing the basis for biological interpretation.

The Davies/Ponting/Molnar Groups

Dorsal cortex of avian and reptilian brains contains only a component of the six layered isocortex of mammals. Although fundamental structures of the isocortex are similar in all mammals, there is a drastic increase in cortical size and complexity in mammals culminating with the human brain. The mechanisms and genes responsible for generating these variations can be understood by studying cortical development in various different species. We are focussing on the function of so-called "macro-RNA" genes that exhibit distinct signatures of purifying selection, suggestive of functionality. Using the tools of modern genetics and developmental neurobiology we are testing the function of selected genes in *in vitro* and *in vivo* experimental paradigms. We are analysing ncRNAs which show a specific developmental profile in the developing cortex as demonstrated by *in situ* hybridisation. We will then use microarray analysis of the cortical region when these RNA sequences have been knocked down by RNAi to determine their targets. We would also use bioinformatics approaches to find such targets in genomic sequence more broadly.

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.

Public Databases

- Rfam is the RNA equivalent of protein Pfam. It is a database of non-coding RNA (ncRNA) families, represented by multiple sequence alignments and covariance models (may be you want to define what covariance models are). It is available via the Web at <http://www.sanger.ac.uk/Software/Rfam/> and <http://rfam.wustl.edu/>. The 6.1 release includes 379 families annotating over 280 000 regions. The Rfam library of covariance models can be used to search sequences (including whole genomes) for homologues to known non-coding RNAs, in conjunction with the INFERNAL software. Indeed, Rfam makes available annotation of over 13,400 candidate ncRNA genes (plus 172 self-splicing introns and 1285 cis-regulatory RNA elements) belonging to 172 families in 224 completed chromosomes and genomes. Family-specific databases are also available; mirBase at <http://microrna.sanger.ac.uk/sequences/> for miRNAs, and snoRNA database at <http://lowelab.ucsc.edu/snoRNAdb/> for snoRNAs for example. The UCSC browser provides up-to-date genomic alignments of genomes at varying evolutionary distances from human to medaka, including chimp, rhesus, mouse, rat, cat, opossum, chicken, and fugu. It annotates these genomes with predicted ncRNAs using Evofold, and RNAz. These alignments can be used in conjunction with the covariance models provided by Rfam to obtain secondary structures.

Methodology: The pilot project will focus on *homologous* and *non-homologous* methods:

- Homologous methods have been developed by many, but it would be the easiest to use and extend SAPF as developed by Rahul Satija.
- Non-homologous methods can be employed by using existing programs, but it could also be good to use this to devise and implement a method to extract information from a set of HMMs describing non-homologous gene families.

Project Plan

- Weeks 1-2 Read the literature, collect data and test key programs – SAPF, MultiModule, PhyloGibbs, Footprinter and PhastCons.
- 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.

References

- Blanchette and Tompa (2003) FootPrinter: a program designed for phylogenetic footprinting. *Nuc. Acids Res.* 31: 3840-3842.
- 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.
- Jensen S., Shen L. and Liu J. (2005) Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics* 21: 3832-3839.
- Knudsen B. and Hein J. (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* 15: 446-454.
- Lawrence, C. et al. (1993) "Detecting Subtle Sequence Signals." A Gibbs Sampler approach to Multiple Alignment. *Science* 262:208-
- Maston, GA et al. (2006) Transcriptional regulatory elements in the human genome. *Ann. Rev. Genomics and Hum. Genet.* 7: 29-59.
- Mehler MF & JS Mattick (2007) "Non-coding RNAs in the nervous system" *J.Physiol.* 575.2.333-41
- Pedersen J.S. and Hein J.J. (2003) Gene finding with as Hidden Markov Model of genome structure and evolution. *Bioinformatics* 19: 219-227.
- Pedersen J.S., Meyer I.M., Forsberg R., Simmonds P. and Hein J. (2004) A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nuc. Acids Res.* 32: 4925-4936.
- Pedersen, JS et al. (2006) "Identification and Classification of Conserved RNA Secondary Structures in the Human Genome" *PLOS Computational Biology* 2.4.251-262
- Prakash A. and Tompa M. (2005) Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.* 23:1249-1256.
- Satija R, L. Pachter and J. Hein (2008) "Statistical Alignment and Footprinting" (in prep.)
- Siddharthan, Siggia and van Nimwegen (2005) "PhyloGibbs: A Gibbs Sampling Motif Finder that Incorporates Phylogeny" *PLOS Computational Biology* 17
- 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
- Taylor, M.S. et al. (2006) Heterotachy in mammalian promoter evolution. *PLoS Genetics* 2: 30-39.
- Wang and Stormo (2003) "Combining phylogenetic data with co-regulated genes to identify regulatory motifs" *Bioinformatics* 19.18.2369-80
- Washielt, S et al. (2005) "Mapping of conserved RNA secondary structures predict thousands of functional noncoding RNAs in the human genome" *Nat. Biotech.* 23.11.1383-90
- 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.