Computational Promoter Analysis of Metazoan α-Clusters

Background

Although it is now possible to identify and annotate coding sequences from primary DNA sequence relatively efficiently, 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. 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 (MRS-R) contain highly conserved transcription factor (TF) binding sites. However, it is still not clear why these sequences (often as few as 6-10 bases) bind TFs in their chromosomal context, while other apparently identical, and yet non-conserved binding sites do not. It suggests that there is more to these MCS-R elements than meets the eye and indeed closer inspection shows that sequences surrounding the conserved TF binding sites may also be relatively well conserved although not to the same degree as seen for the TF binding sites themselves.

To address these issues we have studied the structure, evolution and function of the α-globin gene cluster which encodes one subunit of haemoglobin and when mutated gives rise to the serious human disease thalassaemia. A variety of approaches characterising the genetic and epigenetic features of this region have defined the regulatory domain containing the α-globin genes. We define this as the segment of the chromosome which contains all of the sequences required to direct fully regulated, tissue- and developmental stage- specific expression. In the case of the human α-cluster this region spans ~ 120kb. From a survey of 22 species Prof Higgs' group has identified 7 *cis*-acting regulatory elements within the domain which have been conserved throughout 150MY of evolution. These and other data from their studies on this locus suggest that these elements nucleate multiprotein complexes which are responsible, in turn, for regulating expression of the α-globin genes during erythropoiesis.

This collection of conserved *cis*-elements will provide a model to understand, in general, how *cis*-elements evolve and will form the basis for developing computational tools to distinguish functional *cis*-acting regulatory elements from other potential binding sites which clearly do not bind TFs *in vivo*.

The Higgs Group

The Higgs group has applied comparative genomics techniques to map conserved sequences (MCSs) across 22 species (Hughes *et al.* 2005). In contrast to most earlier attempts to do this, attempts that typically found a very large number of false positive (Prakash and Tompa 2005), the Higgs group has been very successful in finding conserved segments with major overlap in what has been experimentally verified. Motivated by this success, it is natural to bring this approach much further, both in terms of larger data sets, more refined statistical methods, but also in attempting to make more specific predictions about function of the MCSs.

Biological Issues to be addressed

- How detailed an annotation can be achieved by comparative methods? Finding multi-species conserved segments (MSCs) 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 that 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.

- The role of RNA structure and periodicities in the regulation of the α-globin complex: The search for RNA structural constraints (Knudsen and Hein 2003; Pedersen *et al.* 2004a,b) could reveal positions were base pairing or complementarity had a key role. Not all information necessarily resides in the local sequence structure, but potentially in chromatin structure. It would be very natural to investigate if rules can be defined for when a regulatory signal has the potential to be active in the context of chromatin and when not. This last question is difficult, but essential if regulation is to be understood.

- How can additional information be brought in to annotate and generate testable hypotheses even further? "Additional information" is available at many levels: a. Detailed information in the regulatory proteins and complexes in some species. b. Expression data. c. Protein interaction data can give some direction concerning which proteins could form complexes, when full information isn't available.

- Success in 1 and 2 would tell us much about the evolution of regulation over the last 500 MY of the α-globins, both in terms of what has been conserved and what has been adapted.

Data

- Genomic Data: Genomic sequence of up to 300kb of 22 animal species with a root about 500 MY ago. The number of species analysed will continue to grow over the coming years and more examples may be sequenced in the Higgs' laboratory. The genomic regions have been extensively characterised for genetic (transgenics, homologous recombination, stable transfectants, interspecific hybrids) and epigenetic (chromatin accessibility, chromatin modification, nuclear localisation, timing of replication and DNA methylation) characters.

- Structural Data: An increasing number of protein structures and DNA/protein structures are available that could be used to formulate criteria describing DNA binding motifs. There is a limited number of DNA binding proteins in the human genome (Maston *et al*. 2006), so making predictions about specific proteins will be increasingly realistic.

- Expression Data are available for the transcription programmes that accompany erythropoiesis as the globin genes are activated.

Methods

Routes forward to predicting protein sequences and linking to experimental inputs include:

- Comparative **Gene Finding** and **RNA structure** prediction methods (to identify regulatory RNAs) which are based on stochastic context free grammars - SCFGs. Hidden Markov Models scan sequences and generate symbols, where the probability only depends on the last state (Pedersen and Hein 2003; Siepel and Haussler 2004). SCFGs are a generalization of Hidden Markov Models that allows the simultaneous generation of non-neighbouring pairs of symbols in a nested fashion (Knudsen and Hein 1999), thereby allowing base pairings, loops and other structural elements to be modelled.

- **Regulatory Signals** are much more heterogeneous and comparative approaches are the main way forward. 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 HMM. 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).

- **Integration of Models** is needed for both description of structure and the structure dependent evolution. Integration of models is needed since a position can be subject to multiple constraints involving reading frames, RNA structure and regulatory signals (Pedersen *et al*. 2004).

- **Annotation restricted by given biological knowledge**. It is increasingly important that, when applying bioinformatics techniques, a variety of levels and data types are taken into account. The full project will focus on a set of 50-300 kb genomic regions that we will trace in a large number of genomes, allowing very detailed annotation.

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.
- Hughes *et al.* (2005) Annotation of cis-regulatory elements by identification, sub-classification, and functional assessment of multispecies conserved sequences. *PNAS* 102: 9830-9835.
- 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.
- Maston, GA et al. (2006) Transcriptional regulatory elements in the human genome. *Ann. Rev. Genomics and Hum. Genet.* To appear.
- 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.
- Prakash A. and Tompa M. (2005) Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.* 23:1249-1256.
- Siepel, A. and Haussler D. (2004) Combining phylogenetic and Hidden Markov Models in biosequence analysis. *J. Comput. Biol.* 2004 11:413-428.
- Taylor, M.S. *et al.* (2006) Heterotachy in mammalian promoter evolution. *PLoS Genetics* 2: 30-39.