

## *Extending the Domain of Comparative Genomics*

The comparative approach to Genomics is a major success at present and due to the continued accumulation of genomes, the demand for such methods will only increase. As data increases and methods of analysis are refined, increasingly ambitious questions can be addressed. RNA structure has been a front runner, because a single sequence here carries little structural information, but the substitution process is highly dependent on the structural context (Rivas and Eddy, (2001) and Knudsen and Hein (1999)). Gene prediction uses a variety of principles, but comparatively the distinction between coding and non-coding regions is the dominant principle (Hein and Pedersen (2003) and Siepel and Haussler (2003)). Gene prediction is relatively more successful on a single sequence, than RNA prediction because features such as the absence of stop codons and triple periodicities strongly distinguish genes from non-genes. But the comparative approach will eventually dominate as the number of genomes become large. The characterization of regulatory signals is highly challenging and will be of tremendous importance in coming years (Boffelli, D. et al. (2003), Kells et al. (2003)). The proposed research will focus on gene finding, but will also involve regulatory signals as these are important characteristics used in descriptions of gene structure.

Comparative methods have 2 obvious limitations at present that easily can be identified: Their dependence on an alignment and the assumption of a common structure shared by the compared genomes. These limitations have created a “Scylla and Carybdis” of comparative genomics: If the sequences are too close, the genomes are easy to align and the gene structure probably hasn’t changed, but there are too few evolutionary events for the comparative principle to be exploited maximally. If the genomes are too distant, the alignment becomes unreliable and the gene structure could have changed, but there will be plenty of evolution that could distinguish hypotheses about gene structure. To be more specific in relation to present and coming genomes: For (human, chimp) the alignment and common structure assumption will hold, it can still be usefully assumed for (human, mouse), but for (human, chicken) these assumptions are seriously violated. In data analysis, this will be seen as an increased amount of noise, because regions have been misaligned and even when correctly aligned, human reinterpretation of the annotation is necessary because the method has chosen an “average” gene structure, because the structures are different in the compared genomes.

We will here propose a model that could alleviate both these issues in the case of gene finding and improve comparison of more distant genomes. Apart from better annotation, an analysis of the evolution of gene structure, not only its sequence content, will be possible. As with any application of models of evolution, this will have two parts: First, a better understanding of the process of gene evolution. And second, the ability to make statements about the evolution of specific genes (at times called ancestral analysis).

Modeling gene structure evolution will at first be done by a simple, but tractable model and we will then move towards more complicated models in response to data analysis. There is such a simple starting point that also could be made more realistic. This is a natural extension of the model used for statistical alignment - the insertion-deletion process first described in 1991 by Thorne, Kishino and Felsenstein (TKF91).

The TKF91 process describes the insertion and deletion of single elements. In the resulting string all elements are equivalent, making it unsuitable for gene structure description. However, it is possible to make a two-layer version of this process, where the top layer deletes and inserts intron-exons. The lower layer is the old TKF91 model describing the expansion/shrinkage of individual introns-exons. The two-layer TKF91 will be more complicated than the simple TKF91 model, but such calculations are based on the combination of well explored recursions from the TKF91 model with recursions describing gene structure as in Hein and Pedersen (2003). The overall outcome of such a method will be: First, an estimation of all involved parameters – insertion/deletion rates at nucleotide and gene structure level and substitution rates and relative branch lengths in the phylogeny relating the genomes. Second, probability assignments to all alignments at the two levels (nucleotide and gene structure) and determination of the most probably alignment, gene structure annotation and gene structure alignment.

Our group has already contributed significantly to the methodologies behind the proposed project. Knudsen and Hein (1999) were the first to combine a structure model with structure dependent evolution to predict RNA secondary structure. Pedersen and Hein (2003) have developed the analogous methodologies in the context of comparative gene prediction. The Pedersen and Hein

approach has independently been developed by a series of researchers, including the large genome annotation group in Santa Cruz around David Haussler. Since 2000 our group has published 10 papers on statistical alignment extending the original TKF91 model to more sequences, faster algorithms and more realistic models and additionally a number of papers have been submitted or are in preparation. All our methods are made into practical programs, so development of the proposed methods should have much experience to build on and applications should be possible quite early in the process. In summary, a successful applicant will do the research in a group working on all the involved issues in completing this project successfully.

Both the development and final test of the proposed methods is in the application to real data and at the OCGF this will be done in collaboration with Prof. Kay Davies' group. They are interested the transcriptional control of genes responsible for muscle stem cell development and muscle cell differentiation. We will therefore focus on the annotation of muscle protein genes using genomic data.

This is obviously an ambitious project, but there will be strong demand for progress on this problem in coming years.

### **References**

Boffelli, D. et al. (2003): Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome. *Science* vol. 299 1391-95

Kells et al. (2003): Sequencing and Comparison of Yeast Species to Identify Genes and Regulatory Elements. *Nature* vol 423.241-254.

Rivas,E. and S. Eddy (2001): Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2.8.1-19

Siepel and Haussler (2003): Combining Phylogenetic and Hidden Markov Model Biosequence Analysis. *RECOMB03* p277-86

Thorne, JL, Kishino, H. and J. Felsenstein (1991): An Evolutionary Model for Maximum Likelihood Alignment of DNA Sequences. *In Journal of Molecular Evolution*, 33:114-124

Knudsen, B. and J.J.Hein (1999) "Using stochastic context free grammars and molecular evolution to predict RNA secondary structure (*Bioinformatics* vol 15.5 15.6.446-454)

Hein,J., C.Wiuf, B.Knudsen, Møller, M., and G.Wibling (2000): Statistical Alignment: Computational Properties, Homology Testing and Goodness-of-Fit. (*J. Molecular Biology* 302.265-279)

Steel, M. & J.J.Hein (2001): A generalisation of the Thorne-Kishino-Felsenstein model of Statistical Alignment to k sequences related by a star tree. (*Letters in Applied Mathematics* vol. 14.679-684)

J.J.Hein (2001): A generalisation of the Thorne-Kishino-Felsenstein model of Statistical Alignment to k sequences related by Relevant a binary tree. (*Pac.Symp.Biocompu.* 2001 p179-190 (eds RB Altman et al.)

Pedersen, J.S. and J.J. Hein (2003) "Gene finding with as hidden Markov model of genome structure and evolution" *Bioinformatics* 19.2.219-227.

Knudsen,B. and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.* 2003 Jul 1; 31(13): 3423-8.

Lunter, Song, Miklos & Hein (2003) "A one-state recursion for multiple statistical alignment" (in press, *J.Compu.Biol.*)

Lunter, Miklos, Drummond, Jensen & Hein (2003) "Bayesian Phylogenetic Inference under a statistical insertion-deletion model" (in press, *WABI03, Hungary.*)