

Project title: Phylogenetic Analysis of “New” Homeobox in the lineage leading to Humans

The homeobox genes are extremely interesting and central in the mechanisms establishing form in metazoans, so tracing these genes evolutionary. Peter Holland group has found 4 new genes in humans, by comparison to the mouse genome (Booth et al., 2006). Since this work the number of genomes have increased significantly and it should now be possible to date the origin of these genes much more precisely.

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

- (i) Comparative **Gene Finding** methods which are based on Hidden Markov Models that describe what we would conceive as legal gene structures. The proposed genes can also be found very precisely by homology to known homeobox genes.
- (ii) Investigate the date of arisal of the four genes on the lineages leading to humans. This is done by finding homologous is closely related genomes chimp, orangutang, Consult this www-page <http://www.genome.gov/10002154> for a list of the available genomes.
- (iii) Test for rate of evolution, strength of selection and the molecular clock for these 4 genes and their orthologues.
- (iv) Map the position of these genes to genomes and investigate if they stay in the same synteny groups.

The project will be well defined, but can easily move into more open and challenging questions in case of fast initial progress. The core project will involve comparative genome annotation in a region with only 4 known genes. Examples of the basic idea behind comparative gene finding can be found in Pedersen and Hein (2003) and Siepel and Haussler (2004). There are countless methods for aligning sequences.

If this is efficiently solved, a series of problems will lie ahead of increasing difficulty.

- **Regulatory Signal** detection are much more heterogeneous and comparative approaches are the only 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.
- Statistical alignment of the genes involved using methods like Lunter et al., Hobolt et al.
- The analysis of multiple genes – hopefully a small number – will also involve **gene order** rearrangements, duplication, inversions,...

References:

- 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)
- Lunter, Miklos, Drummond, & Hein (2005) "Alignment, Statistics and Evolution" (in "Statistical Methods in Molecular Evolution" ed. Rasmus Nielsen,)
- [Hobolth A, Jensen JL](#).(2005) Applications of hidden Markov models for characterization of homologous

DNA sequences with a common gene. *J Comput Biol.* 12(2):186-203.

- Eddy SR.(2002) Computational genomics of noncoding RNA genes. *Cell.* 2002 Apr 19;109(2):137-40. Review.
- 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.
- Siepel, A. and D. Haussler (2004) “Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol.* 2004;11(2-3):413-28.
- Ruiz-Trillo, I. Burger, G., Holland, P.W.H., King, N., Lang, B.F., Roger, A.J. and Gray, M.W. (2007) The origins of multicellularity: a multi-taxon genome initiative. **Trends in Genetics.** In press.
- Booth, H.A.F. and Holland, P.W.H. (2007) Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. **Gene** 387, 7-14.