

Annotate 12 Drosophila genomes for regulatory signals.

By Jotun Hein and Vasile Palade

The fruit fly *Drosophila melanogaster* has been a key model organism for about a century and must be one of the organisms that are most thoroughly studied. Such studies spans from purely functional to evolutionary including population genetics, speciation and molecular evolution. Motivated by the great success of comparative genomics, 12 complete genomes of *Drosophila* species have been sequenced to provide a ideal test for the power of comparative approaches in prediction phenotype and organismal biology from sequence data. A series of researchers located in Oxford will focus on this data set starting in 2006 (Lior Pachter, Rahul Satija, Andreas Heger and probably others) This data set provides a long series of challenges. To mention a few:

- Whole Genome Alignment:
- Sequence Level Alignment
- Protein Gene Finding
- RNA Gene Finding
- **Regulatory Element Characterisation**
- Relating the above to the biology of the species.

It would be easy to expand this item list much further and these data are truly a resource for a variety of interesting biological problems. Given that we have a set of related sequences and the acknowledged strength of evolutionary approaches, it seems likely that evolution must be part of modelling approaches.

This project proposes a simple analysis of the 12 *Drosophila* genomes for regulatory signals. We will assume that our only data is the 12 genomes and thus we don't know anything about the expression levels.

Work Plan:

1. Read key *Drosophila* and regulatory signal finding articles with/without supervisors. Expand this page into a 3-5 page more detailed work plan.

2. Promoter recognition for *Drosophila* sequences using computational intelligence techniques Based on our previous experience on recognizing promoters in *E.Coli* and Human DNA sequences, this projects intend to apply computational intelligence techniques, like multiclassifiers, neural networks, genetic algorithms but not only, for recognizing promoters within the 12 *Drosophila* genomes. Multiple classifier systems (MCS) provides better recognition through the incorporation of diversity among a pool of individual classifiers. Each classifier could be trained to specialize on different aspects of the genome, and then combine their prediction results. More details on MCS and current approaches and results can be found here: <http://web.comlab.ox.ac.uk/oucl/work/romesh.ranawana/RP2005b.pdf>
<http://web.comlab.ox.ac.uk/oucl/work/romesh.ranawana/RP2004a.pdf>
<http://web.comlab.ox.ac.uk/oucl/work/romesh.ranawana/RP2005d.pdf>

3. Analysis of 12 *Drosophila* genomes properties using machine learning feature selection methods for data dimensionality reduction. The 300 bases long human DNA was dimensionally reduced to a 7 feature space using 7 property functions which numerically characterizes different aspects of human DNA sequences. This approach is very promising, as for human DNA produced an accuracy in excess of 89% using a 7 feature space only when training the classifiers.

Details are available on this paper:

<http://web.comlab.ox.ac.uk/oucl/work/romesh.ranawana/RP2005c.pdf>

4. The methods under 2 and 3 does not use phylogenetic information about the sequences. Could this be incorporated into these methods?

References:

- R. Ranawana, V. Palade (2005). "A Neural Network Based Multi-Classifer System for Gene Identification in DNA Sequences", *Neural Computing and Applications* (Springer-Verlag), vol. 14, no. 2, pp. 122-131.
C. Dewey, P. Huggins, K. Woods, B. Sturmfels and L. Pachter, Parametric alignment of *Drosophila* genomes, submitted.

K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin and I. Dubchak, VISTA: computational tools for comparative genomics, *Nucleic Acids Research* 32 (2004) p 273 -- 279.

N. Bray, I. Dubchak, L. Pachter, AVID: A global alignment program, *Genome Research*, 13 (2003) p 97--102.

Shane T. Jensen, Lei Shen, and Jun S. Liu

Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes

Bioinformatics Advance Access published on August 16, 2005

Kohler (1994) "Lord of the Flies" Chicago Press

Blanchette, M, B. Schwikowski and M. Tompa (2002) "Algorithms for Phylogenetic Footprinting" *J. Comp. Biol.* 9.2.211-

www-pages:

<http://rana.lbl.gov/drosophila/> (description and data of the 12 genomes)

<http://www.stats.ox.ac.uk/~hein/teaching.htm> (lecture on regulatory signals)

<http://www.people.fas.harvard.edu/~junliu/> (many articles, presentations and programs)