

Comparative Genomics: Finding Regulatory Signals

Objective: To give a presentation of about 60 minutes at the end of the week covering the key aspects of the finding regulatory signals in genomes by computational means.

This project is devoted to finding regulatory signals in a large set of genomes. Acquiring knowledge of a gene is central in interpretation of what it does. A gene is close to the “the atom” of molecular biology (Keller, 2002) and must be annotated in terms of structure, selection profile, constraints, regulatory signals and its relation to other genes. In its extreme form, this necessitates understanding of the complete organism as a gene is part of a large interacting network of genes and signals. The questions and contents below are meant as motivators and need not be followed.

The Big Questions Are:

- What are the key classes of signals?
- Are signals fully defined in terms of local sequence?
- How well can signals be found computationally? (as function of signal, genomes,...)
- What is most efficient homologous or non-homologous annotation?
- Which additional data types are there and how helpful are they?
- How detailed can signal annotation become?
- How does signals evolve?

Maximal Contents of Presentation

The Data: Genomes,

Classes of Signals

Classes of Problems:

Known/unknown signals

Homologous/non-homologous sequences

Experiments

Knowledge available

Aligned/Unaligned sequences

Levels and Limits of Annotations

Recommended literature

Blanchette and Tompa (2003) FootPrinter: a program designed for phylogenetic footprinting. *Nuc. Acids Res.* 31: 3840-3842.

Evelyn Fox Keller (2002) “Century of the Gene” Harvard University Press

Janky and van Helden (2008) Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution *BMC Bioinformatics.* 2008; 9: 37-

Lawrence, Altshul, Boguski, Liu, Neuwald and Wootton (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*8;262(5131):208-14

Pashne (2002) “Signals and Genes” CSHL Press

Satija R, L. Pachter and J. Hein (2008) “Statistical Alignment and Footprinting” *Bioinformatics* 24(10): 1236-42.

Satija R, I.Miklos, A. Novak and J. Hein (2008) “MCMC Statistical Alignment Footprinting” *Bioinformatics*

Siddharthan, Siggia and van Nimwegen (2005) *PhyloGibbs PLOS Biology*

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

Sinha and He (2007). MORPH: Probabilistic alignment combined with Hidden Markov Models of cis-regulatory modules. *PLoS Computational Biology.* 3(11):e216.

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.

Wang and Stormo (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs *Bioinformatics* Vol. 19 no. 18 2369-2380

Wasserman WW and Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics.* 2004;5:276-287.