**6.096 Algorithms for Computational Biology**

# Problem Set 4

*Professor: Manolis Kellis*

## Small project proposal

All through this course, you have learned about algorithms for computational biology. Your last problem set will entail a small project. You will have the oportunity to explore the published literature in computational biology, writing a 5 page report on a published paper focussing on some of the following items:

1. Background in the biology problem related to the algorithm
2. Runtime analysis and if relevant space analysis
3. Intuition behind the some of the proofs provided and/or proofs of facts stated without proof
4. Implementation of the algorithm in python and application to a real or synthetic dataset or possible extensions of the algorithm
5. Literature search of other approaches to similar problem

You may choose to make your project more theoretical or more applied.

Please write one or two paragraphs choosing a paper for your project and discussing what you plan to explore. Send these two paragraphs by e-mail to the TA by Wednesday April 27, 2005. Note that the small project will be due on Friday May 6th, 2005.

Here is a list of papers that we find interesting and directly related to the class. If you would strongly prefer to work on a different paper on computational biology, please let us know in the e-mail due by Wednesday April 27th, 2005.

1. J. Buhler and M.Tompa. Finding Motifs using Random Projections. RECOMB 2001. (comments: random projections applied to motif discovery)
2. S. Altschul, et a l. Basic Local Alignment Search Tool. Journal of Mol Biology 1990. (comments: this is the original BLAST paper)
3. J. Buhler. Efficient large-scale sequence comparison by locality-sensitive hashing. Bioinformatics 2001 (comments: this is an extension of the second pset of random projections for rapid database search)
4. Kamvysselis et al, While-genome comparative annotation and regulatory motif discovery in multiple yeast species RECOMB 2003 (comments: a paper by the faculty teaching the class combining several yeast species)
5. G.Z. Hertz. Identifying DNA and protein patterns with statistical significan alignments of multiple sequences. Bioinformatics 1999. (comments: CONSENSUS)
6. P.A, Pevzner. Finding Motifs in the twilight zone. Bioinformatics 2002. (comments: MULTIPROFILER)
7. Lawrence, altschul et al. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science 1993.
8. Smith-Waterman. Identification of common molecular subsequences. Journal of Molecular Biology 1981.( comments: the smith-waterman algorithm is a seminal paper in computational biology. It uses local alignments)

9. Higgins, Thompson, Bibson. Using CLUSTAL for multiple sequence alignments. Methos in Enzymology. 1996.

10. Burge and Karlin. Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology 1997.(comments: applications of HMMs)

11. Bairoch et al. The prosite database, its status in 1997. Nucleic Acid Research 25. (comments: applications of regular grammars)

12. Dandekar et al. Finding the hairpin in the haystack: searching for RNA motifs. Trends in Genetics 11.1995

13. Baldi et al. Hidden Markov models of biological primary sequence information. PNAS of the USA 91. 1994

14. Haussler et al. Protein modeling using hidden markov models: Analysis of globins. Proceedings of the Hawaii International Conference on System Science pages 792-802.

15. Haussler et al. Hidden Markov models in computational biology: Applications to protein modeling. Journal of Molecular Biology, 235:1501-1531, 1994.

16. Ben-Dor et al. Clustering Gene Expression Patterns. Journal of Computational Biology, 6: 281-297, 1999.

17. Nei et al. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biological Evolution. 1987.