

## **BI6103: COMPUTATIONAL BIOLOGY**

Core course for M.Sc. (Bioinformatics), NTU

Spring Semester 2011

Academic Units: 3

Pre-requisites: BI6101 – Introductory Biology

BI6102 – Introductory Bioinformatics

Thursdays 6.30-9.30pm at SBS CR5

Lecturer: Professor Jagath Rajapakse

### Syllabus for 2009/10 Semester 2

1. INTRODUCTION AND MATHEMATICAL FOUNDATIONS (27/01, 10/02)  
Sets and sequences; functions and spaces; probability theory; Bayes' theorem; random variables; probability distributions; multidimensional density functions; information theory; sequence alignment by information minimization, mutual information (MI), characterization of splice sites with MI
2. PROBABILISTIC MODELS OF SEQUENCES (17/02, 24/02)  
Parameter estimation: ML, and MAP estimators; constrained optimization (Lagrange theory); prior models: maximum entropy principles, Gaussian priors, Dirichlet priors; dice models of sequences given data/counts; dice models for pairs/ multiple sequences; random and match models and log-odds ratios for alignment; Markov chains, Markov models of sequences, modeling CpG islands
3. HIDDEN MARKOV MODELS & GENE STRUCTURE PREDICTION (03/03, 10/03)  
Definition of hidden Markov models (HMM); dice models of sequences; likelihood of sequences; forward algorithm; backward algorithm; Viterbi algorithm; posterior decoding; ML estimation of parameters; Expectation Minimization (EM) algorithm Baum-Welch algorithm; Baldi-Chauvin approach; gene structure prediction: VEIL; GENESCAN; profile HMM for multiple sequence alignment
4. PROTEIN STRUCTURE PREDICTION (17/03, 24/03)  
GOR approaches to protein secondary structure (PSS) prediction; artificial neurons; discrete and continuous perceptron; gradient-descent learning; multi-layer neural networks; back-propagation algorithm; PHD method for PSS prediction; protein tertiary structure prediction; accuracy of prediction: precision-recall trade-off; ROC curves; cross validation; bias-variance trade-off

5. RECOGNITION OF PROTEIN FEATURES (31/03)

Support vector machines (SVM), Recognition of transmembrane helices; turn propensity scale for TM helices; hydrophobicity approach; TMHMM – a HMM approach; ENSEMBLE; topography prediction; subcellular localization; secretory pathway; compartments and sorting; PSORT-B, TargetP, SCL-BLAST, adaptation of protein surfaces; amino acid composition: SubLoc

6. PREDICTION OF GENE FEATURES (07/04)

Signal selection; recognition of translation initiation sites; recognition of transcription start sites; promoter finders; neural network based systems for recognizing gene features: Promoter 2.0, NNPP, Promoter Inspector, Grail's Promoter, DIANA-TIS, LVQ for TATA Recognition, Netstart

7. MOTIF DETECTION (14/04, 21/04)

Data model for motif finding problem; approaches: profile analysis, word counting method, Gibbs sampling, MLME and HMM; Subtle Weak motif detection problem, graphical approaches: WINNOWER; SP-STAR; random projection, dynamic programming, motifs with gaps

8. Revision (TBD)

Final Grading: 75% exam and 25% project

Reading:

1. *Bioinformatics: machine learning approach*, P. Baldi and S. Brunak, The MIT Press, Cambridge, Second Edition, 2001
2. *Computational molecular biology: an introduction*, P. Clote and R. Backofen, John Wiley and Sons, Ltd., Chichester, 2000
3. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, Cambridge, 2001
4. *Manuscripts and Notes provided at the class*