

GLOBEX Bioinformatics (Summer 2015)

Review Session

Basics of Molecular biology

Topics covered after midterm are highlighted in blue

- Central dogma
 - Transcription
 - Translation
 - Genetic code
- DNA
 - Double helix
 - Watson-Crick binding
- RNA
 - Secondary structures
- Genes
 - Compositional Structure
 - Reading frames
- Proteins
 - Secondary structure: alpha helices, beta sheets, and coils
 - 3-d structure
- Gene Regulations (Operons)
- DNA Microarray
- Cloning
- PCR
- Gel electrophoresis
- 2D gel + MS
- Yeast 2 hybrid system
- Gene orders

Computational methods

Dynamic programming

- Problems can be dissect to sub-problems
- Recurrence equations
- DP Table
- Traceback

Maximum Likelihood (ML), and EM

Pairwise alignments

Multiple alignments

Hidden Markov models

- viterbi, forward, and viterbi training

Phylogenetic trees

- distance-based, character-based, and probability-based
- Bootstrap

Structural predictions

- secondary (RNA: Nussinov, Zuker; Proteins: ANN)
- lattice model for proteins.
- genetic algorithm

Clustering Profiles and inferring genetic networks

- K-means
- Boolean networks
- Bayesian (conditional independence, joint probability, factorization)
- Pearson correlation

Computational methods

Dynamic programming

- Problems can be dissect to sub-problems
- Recurrence equations
- DP Table
- Traceback

Maximum Likelihood (ML), and Expectation Maximization (EM)

Greedy Algorithms

Genetic Algorithms

Sequence alignments

- Motivations: evolution and homology
- Concepts and assumptions
 - Substitution scores: log-odds, PAM, BLOSUM, transition, transversion
 - gap penalties
 - Position independence
- Dynamic programming
 - Problems can be dissected to sub-problems
 - Recurrence equations
 - DP Table
 - Traceback
- Global v.s. local alignments
- Pairwise v.s. multiple alignments
 - Progressive approaches: guide tree, seq-profile, profile-profile
- Score statistics: E-value, p-value.

Hidden Markov models

- Motivations
- Concepts and assumptions
- Three tasks
 - Decoding (Viterbi algorithm)
 - Recurrence
 - Dynamic programming
 - Model likelihood (Forward algorithm, Backward algorithm)
 - Recurrence
 - Dynamic programming
 - Model building
 - Model architecture: profile HMM for sequence family
 - Model parameters
 - Labeled data
 - Unlabeled data (Viterbi training, Baum-Wunsch algorithm)

Phylogenetic Trees

- Motivations
- Concepts and assumptions
 - Gene trees versus species trees
 - Rooted v.s. unrooted
- Methods
 - Character-based
 - Maximum Parsimony: conventional and weighted
 - Traceback
 - Distance-based
 - UWPGA
 - Neighbor-joining
 - Metric distance, ultrametric distance, molecular clock, additive distance
 - Probability-based

RNA secondary structures

- Concepts
 - Single strand
 - Intra-molecular pairings
 - Structural categories
 - Hairpins, bulge loops, interior loops, and multi loops
 - Pseudo knots
 - Similarity in secondary structure plays a more important role than primary sequence in determining homology for RNAs
- Algorithms for predicting structures
 - Nussinov's
 - Maximum base pairing
 - Zuker's
 - Minimum energy

Protein structures

- Concepts
 - Anfinsen hypothesis
 - Primary, secondary, tertiary, and quaternary
- Algorithms for predicting structures
 - Secondary structure
 - encoding
 - Artificial neural networks
 - 3-dimensional
 - Lattice models
 - HP models: Genetic algorithm

Systems Biology

- Concepts
 - Gene expression
 - Regulation, protein-protein interactions, ...
 - Operons, co-regulation
- Clustering and Modeling
 - Profiling, distance, hierarchical cluster, dendrograms.
 - K-mean
 - Genetic networks modeling
 - Boolean networks
 - Bayesian

About the Final Exam

- Time and Place:
 - July 25, 8:00AM-11:00AM
 - 3Rd Teaching Building, Rm 106
- closed-book (One sheet of A4 size is allowed)
- Exam is cumulative.
- Four parts
 - Primary sequence analysis [~25 points]
 - Phylogenetic trees [~30 points]
 - Structural prediction [~25 points]
 - Systems biology [~20 points]: gene expression analysis, and regulation, genetic networks.