

Bioinformatics Prereqs and Topics Crowdsourced Commenting Assignment 0 for CBB752

Go to :

<https://goo.gl/7MDP1I>

(while you are logged into Google with your Yale.edu email account)

Then edit & comment as described in

<http://cbb752b16.gersteinlab.org/assignments/homework0>

In particular, an edit consists of addition of new topic, a deletion of a topic (preferably with a brief comment) or a just an extended comment of at least 1 sentence (>4 words). (See example edits below.)

Categories of Knowledge for Bioinformatics Education

U = Undergrad. level

G = Grad. level

C = CS

S = Stats/Math

B = Bio/Chem/Phys

I = Intro. Bioinformatics Topic (i.e., in a class like Yale CBB752)

A = Advanced Bioinformatics Topic (i.e., maybe beyond Yale CBB752)

Combining Abbreviations - viz:

UC, GC, US, GS, GI, GA, GB, UB

Prerequisites for Bioinformatics: Stats & CS

These go beyond Basic Math (calculus), Biology, Chemistry & Physics taught in pre-medical education and in undergraduate majors such as Yale MB&B

Programming Topics [UC]

- Interpretative & compiled languages
- Recursion
- Data structures - lists, arrays, hashes, stacks
- Computational complexity - related to operations as sorting
- Basics of computer architecture (caches, disks, bottlenecks)

- Practical Programming: modularization (OOP), version control, debugging, APIs & libraries, software carpentry, reproducible research/coding
- Basics of web programming - stateless access, web protocols
- Databases
 - Basics of SQL, with concept of indices & joining, schema
 - non-relational architectures - NoSQL
- Regular expression & string processing
- Numeric computing issues: FP arithmetic & random numbers
- Encryption & compression

Programming Topics [GC]

- VMs & cloud computing
- Computational optimization & integration of functions
- Relational database concepts
- DB interoperation
- Privacy & security
- Ontologies
- Distributed and high-performance computing (parallel computing)

Statistical Topics [US]

- Knowledge of distributions, hypothesis testing & inference (includes mult. testing, t-test)
- Permutation Testing (bootstrapping, cross-validation)
- Regression
- Power analysis (Type 1 & 2 errors)
- Non-parametric vs. Parametric methods
- Bayes Rule

Statistical Topics [GS]

- Regularization
- Unsupervised Methods (PCA, clustering)
- Supervised Methods (SVM, Kernels)
- Bayesian Analysis
- Graphical Models
- Causal inference
- Missing data, imputation & EM algorithm
- Hierarchical Modelling
- Information theory - mut. information, complexity & entropy
- HMMs (Viterbi, Forward and Backward Algorithms)
- MCMC
- Feature Selection
- Assessing Predictions

- Training and testing data
- Cross validation
- ROC curves

Specific Bioinformatics Topics

Classical Sequence Analysis

- String Matching
 - Pairwise Alignment via Dynamic Programming [GI]
 - Local vs. Global Alignment & Suboptimal Alignment [GI]
 - Hashing & Indexing to increase speed (BLAST, FASTA) [GI]
 - Suffix arrays & BWT
 - Substitution scoring matrices (e.g., for amino acids) [GI]
 - Scoring schemes & matching statistics
 - Score Distributions (e.g., EVD)
- Multiple Alignment and Consensus Patterns
 - Identifying genomic regions such as genes & promoters with various statistical methods (e.g., HMMs) [GI]
 - HMMs applied to biology, Profiles, Position dependent subst. matrices [GI]
 - Motifs [GI] --GB: Searching genomes for protein motifs (to find evolutionarily active proteins unidentifiable by homology screening alone?)
 - EM & Gibbs Sampling [GI]
- Whole-Genome analysis
 - Genome Assembly [GA]
 - De Bruijn graphs
 - Characterizing Repeats in Genomic DNA [GA]
 - Identification Duplications in the Genome [GA]
 - Whole-Genome Comparisons & large scale genomic alignments [GA]
 - Synteny
 - Orthologs & Function Classification
 - Genome Annotation
 - Gene Prediction
 - Regulatory site and network prediction
 - miRNA prediction and targeting site prediction
 - Pseudogene prediction and functional prediction
- Cup cake baking [GA]
 - ~~Cream whipping~~
 -

Next-Gen Sequencing Data Processing

- Variant Calling
 - Germline
 - Somatic
 - Structural variation & rearrangements
- RNA-seq
 - Recognizing and correcting batch effects
 - Transcript assembly & splicing
 - Quantification [GI]
 - eQTLs & allelic transcription
 - Normalization
 - Expression Analysis
 - Time Course clustering/longitudinal clustering
 - Differential expression
- DNA methylation & epigenetic gene regulation
 - ChIP-seq
 - Peak calling
- Metagenomics (microbiome) [GA]

Statistical Genetics [GA]

- Population Genetics & Allele Freq.
- Genotype-Phenotype Associations
 - Case-control & GWAS
 - Correlation vs. causality
 - QTLs
- Survival Analysis
- Evolutionary Issues
 - Rates of mutation and change
 - Clustering & Trees [GI]
 - Distance vs. maximum likelihood tree methods

Processing Other Big Data Sets

- Flow Cytometry & CyTOF data analysis
 - Causal network
- Proteomics (Mass Spec)
- Metabolomics [GA]
- Literature & Text Mining
 - Topic analysis
 - Ontologies for terms
- Structural Genomics
- EM image analysis

- B/T cell repertoire sequencing

Data Integration & Mining

- Information integration and fusion
 - Dealing with heterogeneous data
- Ensemble Learning
- Dimensionality Reduction (PCA etc.) in a biological setting
- Network Analysis
 - Pathway analysis
 - Topology Analysis (Hubs & Bottlenecks)
 - Prediction of linkages
 - Global structure vs. local network motifs
- Meta-analysis
 - Fisher's Method for combining p-values

Sequence to Structure

- Secondary Structure Prediction
 - via Propensities
 - TM-helix finding
- Tertiary Structure Prediction
 - Homology Modeling & Protein Threading (Fold Recognition)
 - Ab initio
- Direct Function Prediction
 - Active site identification

3D Structure Analysis

- Molecular Geometry
 - Distances, Angles, Axes, Rotations
 - Calculating a helix axis
 - Molecular Graphics
 - Calculation of Volumes & Surfaces
 - Hinge prediction
 - Packing Measurement
- Structural Comparison & Alignment
 - Basic Protein Geometry and Least-Squares Fitting
 - Aligning sequences on the basis of 3D structure
- Docking and Drug Design [GA]

Simulation & Modelling

- Molecular Mechanics
 - Basic interactions, potential energy functions

- Geometry => Energy => Forces
 - Covalent Bonds
 - Bonds & Angles (as springs)
 - Dihedral
 - Noncovalent interactions
 - Electrostatics
 - VDW Forces
 - Energy Minimization
 - Steepest Descent & Conjugate Gradient
 - Molecular Dynamics & MC
- Simplifications
 - Poisson-Boltzmann Equation
 - Lattice Models
- Signaling & Pathway Modeling
 - Population Dynamics w/ ODEs
 - Regulatory network modeling via Boolean networks, ODEs
 - Flux-balance calculations
 - Agent based modeling
 - Stochastic modeling: Extrinsic and intrinsic noise

Prominent Stat/CS Topics NOT included

- Interrupts
- Advanced Parallel programming
- Machine language
- Compiler design
- Computer Graphics
- Advanced Crypto

Place for Test Comments

[Donghoon Lee \(test comment\)](#)