

Bioinformatics and Molecular Evolution

BIO 454 & APSC 454/654 — Spring 09

Gregory D. Smith (Applied Science) greg@as.wm.edu

Bioinformatics and Molecular Evolution is an introduction to the computational analysis of molecular sequence variation and mathematical modeling of sequence evolution.

Topics covered include nucleotide and amino acid sequence alignment, the causes of sequence variation within and between species, probabilistic models of nucleotide substitutions and molecular sequence evolution, methods of phylogenetic tree construction and inference, and selected topics from genome evolution (e.g., mobile DNA elements, organelle genomes, and the effect of gene duplication, loss, and rearrangement on genome structure.)



Fundamentals of Molecular Evolution. Dan Graur and Wen-Hsiung Li. Sinauer Associates. 2000. [Required]

Bioinformatics and Molecular Evolution. Paul G. Higgs and Teresa K. Attwood. Blackwell, 2005. [Recommended]



Bioinformatics and Molecular Evolution is a required course in the Applied Science *Computational Biology* minor.

Bioinformatics and Molecular Evolution

BIO 454 & APSC 454/654 — Spring 09

Gregory D. Smith (Applied Science) greg@as.wm.edu

Bioinformatics and Molecular Evolution is an introduction to the computational analysis of molecular sequence variation and mathematical modeling of sequence evolution.

Topics covered include nucleotide and amino acid sequence alignment, the causes of sequence variation within and between species, probabilistic models of nucleotide substitutions and molecular sequence evolution, methods of phylogenetic tree construction and inference, and selected topics from genome evolution (e.g., mobile DNA elements, organelle genomes, and the effect of gene duplication, loss, and rearrangement on genome structure.)

- **Fundamentals of Molecular Evolution.** Dan Graur and Wen-Hsiung Li. Sinauer Associates. 2000. [Required]
- **Bioinformatics and Molecular Evolution.** Paul G. Higgs and Teresa K. Attwood. Blackwell, 2005. [Recommended]
- **The Origins of Genome Architecture.** Michael Lynch. Sinauer Associates, 2007. [Recommended]
- **The Causes of Molecular Evolution (Oxford Series in Ecology and Evolution).** John H. Gillespie. Oxford University Press, 1994. [Recommended]

Syllabus

- **Introduction**

(0) What is bioinformatics? What is computational biology? What is molecular evolution? Two areas of study in molecular evolution: evolution of macromolecules and molecular phylogenetics. Population genetics plus molecular biology equals molecular evolution. Population genetics provides the theoretical foundation. Molecular biology provides the empirical data.

- **Genes, Genetic Codes, and Mutation**

(1) The central dogma of molecular biology. Nucleotide sequences as strings over alphabets with four characters. Prefixes, suffices, and substrings. Genomes and DNA replication. Genes and gene structure. Genic and nongenic DNA. Protein-coding genes, RNA-specifying genes, and untranscribed genes. Pseudogenes. Three domains, many kingdoms.

(2) Categorization of amino acids. Proteins as strings over an alphabet of twenty characters. Primary, secondary, tertiary, quaternary structure. Transfer RNAs. Genetic codes and translation. Role of the ribosome. Set theory, union, intersection. Probability, outcomes, events. Frequency vs. axiomatic probability. Conditional probability. Random nucleotide sequences.

(3) Mutation. Substitution mutations, recombinations, deletions, insertions, inversions. Transitions vs. transversions. Synonymous vs. nonsynonymous mutations. Silent mutations. Homologous recombination. Replication slippage. Chromosomal inversions. Hot spots of mutation.

- **Dynamics of Genes in Populations**

(4) Allele frequencies in haploid and diploid populations. Hardy-Weinberg equilibrium and relative genotype frequencies. Natural selection, fitness of genotypes, and allele frequency. Deterministic discrete time maps, finding equilibria, classifying equilibria, and cobwebbing. Dominant, co-dominant, and overdominant selection. Balancing or stabilizing selection vs. directional selection.

(5) Random genetic drift. Gamete sampling in a diploid population. Binomial probability distribution. Stochastic model of allele frequencies with no selection. Cumulative behavior, fixation, and loss. Example Matlab scripts. Expectation and variance of a random variable. Effective population size. Dynamics of gene substitution. Fixation probability and fixation time for a new allele. Conditional fixation time. Rate of gene substitution for neutral mutants vs. advantageous mutants. More Matlab examples.

- **Evolutionary Change in Nucleotide Sequences**

(6) Mathematical modeling of nucleotide substitution in a DNA sequence. Derivation of the Jukes-Cantor model. Matrix-vector notation for continuous-time Markov chains.

(7) Kimura's two-parameter model. Number of nucleotide substitutions between two DNA sequences that share a common origin is greater than the Hamming distance. Modeling nucleotide divergence between two sequences that share a common origin using Jukes-Cantor and/or Kimura's two-parameter model.

(8) Edit distance between two sequences. Similarity scores and distance between two sequences. Are methods of scoring alignments arbitrary? Alignment of nucleotide and amino acid sequences. A dynamic programming algorithm for longest common subsequence problem.

(9) A dynamic programming algorithms for global, semi-global, and local sequence alignments.

(10) Multiple alignments. Sum-of-pairs scoring. Induced pair-wise alignments. Star alignments.

(11) Scoring schemes used when comparing amino acid sequences. Point accepted mutation (PAM) matrices. Constructing a 1-PAM matrix from list of accepted mutations and probability of occurrence of each amino acid. PAM matrices define discrete-time Markov chains and evolutionary models. PAM matrices can be used to construct scoring matrices that reflect important chemical and physical properties of amino acids.

- **Rates and Patterns of Nucleotide Substitution**

(12) The driving forces in evolution. Mutationism, neutralism, and selectionism. The synthetic theory of evolution (selectionism). Adaptation. The Panglossian paradigm. The

neutral theory of molecular evolution (neutralism). Neutrality tests. The distribution of fitness values of mutant alleles.

(13) Pattern vs. rate of substitution. Causes of variation in substitution rates. Case study of lysozyme and forgut fermenters. Patterns of substitution in pseudogenes. Methylation, deamination, and loss of CG in pseudogenes. Detection of strand inequalities in mutation rates. Patterns of amino acid replacement. Conservative vs. radical replacements. Degree of conservation of bulkiness, hydrophobicity, polarity, optical rotation, and charge. Stability index for amino acids. Amino acid frequency is determined by nucleotide composition and the number of codons for the amino acid! Codon-usage bias.

(14) The molecular clock hypothesis. Relative-rate tests. Causes of variation in substitution rates among evolutionary lineages (generation times, metabolic rate). Local clocks. Rates of substitution in organelle DNA RNA viruses. The tempo of evolution (phyletic gradualism vs. punctuated equilibrium). No clear relationship between rates of molecular and morphological evolution!

- **Molecular Phylogenetics**

(15) Objectives of phylogenetics. Species concepts (intuitive, morphological, phenetic, biological, evolutionary/phylogenetic). The Linnaean-Simpsonian hierarchy. Reproductive barriers and species creation. The impact and advantages of molecular data on phylogenetic studies.

(16) Terminology of phylogenetic trees (root, internal nodes, internal branches, external branches, terminal nodes). Bifurcation vs. multifurcation. Rooted and unrooted trees. Cladograms vs. phylograms. Newick format. Cladogenesis vs. anagenesis. True and inferred trees. Monophyletic groups and natural clades. Sister taxa. Paraphyletic taxa. Convenience taxa (e.g., reptiles).

(17) Data types and phylogenetic trees. Characters vs. distances. Unordered, ordered, and polar characters. Sympleiomorphy, synapomorphy, autapomorphy, and homoplasy. Methods of tree reconstruction. Distance matrix methods (UPGMA, neighbor-relations, neighbor joining). Maximum parsimony methods. Variant, invariant, informative, uninformative traits. Maximum likelihood methods.

(18) Bayesian inference of character evolution.

- **Readings**

Readings and critique of primary literature on student-selected topics from molecular evolution drawn from <http://workshop.molecularevolution.org/resources/references/> or by permission of instructor. For example, in Spring 2007 we reviewed:

Ronquist F. **Bayesian inference of character evolution.** *Trends Ecol Evol.* 19(9):475–81, 2004.

Linz B et al. **An African origin for the intimate association between humans and *Helicobacter pylori*.** *Nature.* 445(7130):915–8, 2007.

Chao L et al. **The advantage of sex in the RNA virus phi6.** *Genetics.* 147(3):953–9, 1997.

Peterson SN, Fraser CM. **The complexity of simplicity.** *Genome Biol.* 2(2):comment2002, 2001.

Kuang D et al. **Ancestral reconstruction of the ligand-binding pocket of Family C G protein-coupled receptors.** *Proc Natl Acad Sci U S A.* 103(38):14050–5, 2006.

Endo T, Ikeo K, Gojobori T. **Large-scale search for genes on which positive selection may operate.** *Mol Biol Evol.* 13(5):685–90, 1996.

Yokoyama S, Radlwimmer FB. **The “five-sites” rule and the evolution of red and green color vision in mammals.** *Mol Biol Evol.* 15(5):560–7, 1998.

Salzberg SL, White O, Peterson J, Eisen JA. **Microbial genes in the human genome: lateral transfer or gene loss?** *Science.* 292(5523):1903-6, 2001.

Catalog Description

An introduction to computational molecular biology and molecular evolution including nucleotide and amino acid sequence comparison, DNA fragment assembly, phylogenetic tree construction and inference, RNA and protein secondary structure prediction, and substitution models of sequence evolution. (Cross-listed with BIOL 454.)

Prerequisites: MATH 111 and MATH112/113 (Calculus I &II); BIO203 (Principles of Biology: Molecules, Cells, and Development); or consent of instructor.

Bioinformatics and Molecular Evolution is a required course in the Applied Science *Computational Biology* minor.