

## Checklist for Bioinformatics 1

# Weeks 7,14 :

### lecture 13 -- Mother, Caring for 7 billion

There will be no exam questions on this movie. But here's food for thought:

✓ How does the empowerment of women in poor countries lead to slower population growth?

### lecture 14, 15 -- phylogenetics trees

Phylogenetic trees are models for evolution, generated from sequence distances, using the neighbor-joining method. There are two ways to root a tree.

We learned how to calculate tree lineage distances using the Fitch-Margoliash method and by the least-squares method.

Character-based maximum parsimony can be used to select between a set of proposed trees the one that has the fewest mutations.

A set of trees can be generated by branch-swapping algorithms (NNI, SPR, TBR). Maximum likelihood selects from possible trees by summing over all possible ancestor sequences. (We did not learn how to do M-L.)

Trees can be constructed in random order from subsets of 4 taxa, or "quartets". There are  $N$ -choose-4 quartet trees, and each has one of three possible topologies.

The confidence of each lineage (or branchpoint) in a tree can be found by "bootstrapping", in which many trees are generated by sampling and the percentage presence of lineages (or branchpoints) is determined. Bootstrapping branchpoints requires a root, bootstrapping lineages does not.

✓ What are the properties that pairwise sequence distances must obey? Given distances and a tree, can you identify a violation of the additivity property? A violation of triangle inequality? A non-neighbor-joining tree?

✓ What is a "split"? How do you calculate the symmetric distance between two trees that have different topologies?

✓ How do we calculate the similarity between a set of pairwise sequence distances and the corresponding patristic (tree) distances?

✓ Given several trees and a MSA, can you select the most parsimonious tree?

✓ How do we solve for the branch lengths given distances?

✓ What does branch swapping accomplish? How do we select the best tree topology?

✓ What is quartet puzzling? How do we assemble quartets to make a tree?

✓ Given a set of bootstrap trees, can you calculate the bootstrap values for each lineage?

✓ Given a MSA can you calculate a bootstrap tree in UGENE? Can you infer bootstrap values for a sub-tree?

### lecture 16 -- tree of life

Taxonomic naming roughly follows the course of evolution.

Homologs can be orthologs (originating from a speciation event) or paralogs (originating from a duplication event). Duplicate genes (paralogs) are free to evolve other functions and often do.

Clades that conserve classifications (names or phenotypes) are "monophyletic" (originating once). A clade that contains multiple classifications is "paraphyletic".

Trees based on sequence analysis can be compared to classification schemes, such as the tree of life and its associated taxonomic names. When they disagree, it could be because the naming scheme is paraphyletic (i.e. classifying birds as separate from reptiles when they are actually descended from reptiles). They can also disagree due to paralogy between taxa, causing the sequence distance to be much larger than what you would expect from the species similarity. Or, the genes could be related by horizontal transfer, making their distances much lower than you would expect from the species similarity.

Pruning a MSA to remove distant homologs, nearly-identical sequences, terminal extensions and insertions can reduce a large set of sequences to a manageable number of well-aligned regions. Trees based on pruned/trimmed MSAs have sequence distances based on mutations within ungapped blocks.

✓ Can you recognize monophyly/paraphyly given a tree and taxon names?

✓ Can you infer paralogy given a tree with species classifications and sequence distances? Horizontal transfer?

✓ Can you apply the principles of parsimony to select the best species tree based on anatomical/physiological phenotypes?

✓ What is it about indels and the chain termini that add errors into the sequence distance calculations (and therefore tree calculations)?

### lecture 17 -- Markov models and HMMs

Markov models are directed graphs where edges have transition probabilities and node (or states) "emit" discrete values. In Markov models, the discrete values emitted by the states map 1-to-1 with the states, but in Hidden Markov models (HMM) they do not. HMM can emit the same character from multiple states, and states can emit letters based on a probability distribution.

There are three algorithms associated with HMMs. The Viterbi algorithm finds the most probable pathway through the states given the sequence, calculated using dynamic programming. In other words, it saves the maximum probability pathway at each stage. The forward algorithm, instead of taking only the maximum, sums over all pathways through states given sequence, giving the total probability of all pathway up to state  $q$  and sequence position  $t$ . The Backward algorithm does the same task as the Forward, but starting from the end and working backward. The Forward/Backward algorithm is the product of all forward values and backward values, giving the "a posteriori" value ( $\gamma$ ) for each state  $q$  at each sequence position  $t$ , the sum of all paths through a point  $(q,t)$ . The expectation maximization (EM, or Baum-Welsh) algorithm uses the  $\gamma$  values to estimate the maximum likelihood parameter settings of the model, such as the emission probabilities and the transition probabilities. EM is used to optimize or train a HMM based on a training dataset.

Profile HMMs are HMMs made up of three types of states, Match, Delete and Insert, with conventional transitions. The topology (state-state connectivity) or a profile HMM is defined using a MSA, choosing columns as match states if they contain high information content. (Remember, information is the opposite of random chance.) All sequences in the MSA must have a path through the profile HMM with non-zero probability.

- ✓ Can you write the Viterbi algorithm in pseudocode?
- ✓ Can you write the Forward algorithm?
- ✓ What is the essential difference between the Viterbi and Forward algorithms?
- ✓ Can you define the topology of a profile HMM given a MSA?
- ✓ Can you estimate new model parameters given old model parameters and the output of the Forward/Backward algorithm?

#### lecture 18 -- motif finding

Motifs are recurrent short sequence patterns in protein and DNA. MEME and Gibbs Sampling solve the simultaneous problem of What is it? and What am I looking for? K-means clustering can separate sequences into recurrent classes.

Different methods for predicting something may be compared using the ROC score, which is sensitive to the order of trues and falses in a list sorted by score.

Repeat sequences may be long or short (and can sometimes have short repeats within long repeats). Short repeats are low complexity, like tongue-twisters. Many such repeats are the results of retrotransposon activity and make up a large fraction of the genome for corn and human, among other species. Repeat sequences create problems with database searches, high random scores and erroneous significance values. HMMs may be used to model random scores of repeat alignments, providing a basis for

modeling the extreme value distribution for low-complexity sequences (an unsolved problem!).

- ✓ Can you construct an HMM for a tongue-twister?
- ✓ Can you describe the difference(s) between MEME and the Gibbs Sampler?
- ✓ How would you align low complexity sequences?
- ✓ How would you generate random low complexity sequences?
- ✓ How would obtain a e-value for a repeat-sequence alignment?
- ✓ What was the role of retrotransposons in creating repeats?

#### lecture 21-- RNA folding

Guest lecture by Michael Zuker. RNA forms arrangements of double-helices and loops called secondary structure. All sorts of base pairings are found, not just Watson-Crick. SS can be represented as a dot plot, or using a circle plot. Energies of SS (foldings) are calculated considering the base-pairs and also the base stacking energies. The best folding can be found using energy calculations. Dynamic programming can be used to implicitly try all foldings and select the best.

Pseudoknots (base pairs non-local to an internal loop) break the rules for dynamic programming, so they are ignored in predictions (but they really exist in nature).

Multiple RNA sequence alignments provide conservation and covariation data that is the most powerful evidence for base pairing, therefore for secondary structure prediction.

- ✓ What is a Hoogsteen base pair?
- ✓ Can you draw a pseudoknot?
- ✓ In a circle plot, what does it mean when the lines cross?
- ✓ Given a RNA MSA can you plot covariance on a dotplot? How is it calculated?
- ✓ Do you know all the IUPAC characters for DNA/RNA?

#### lectures 22, 24, 25-- Splicing, gene finding, gene ontology, metagenomics

Gene finding in eukaryotes is the problem of identifying introns and exons. Programs that do this use HMMs that use various motifs, constraints and signals to assign predictions such as intergenic, initial exon, short intron, long intron, internal exon,

terminal exon, with 3 possible intron frames, on both forward and reverse strands. Known splicing constraints include a donor site, acceptor site and branch adenosine, each with corresponding sequence motifs. Splicing enhancers and silencers exist in both exonic and intronic DNA, found by methods such as MEME and Gibbs (and simple counting methods), leading to regulatory mechanisms that produce alternately-spliced isoforms.

Gene ontology, like all ontologies, is a controlled language, defined by experts to classify genes. Classifications (GO terms) are related to each other by "part-of" and "is-a" relationships, proceeding in a hierarchy that goes from vague to specific. GO terms are assigned to genes via evidence, and the evidence is assigned a code, which implies a confidence. For example, GO provides a structure for knowledge. GO enables and facilitates analysis of high-throughput datasets, for example microarray data containing expressed DNA sequences from cell cultures or tissue samples.

Metagenomics is the study of uncultured, unamplified environmental samples by high throughput (sometimes called "next gen") DNA sequencing. This is the discovery field of the future, opening up possibilities ranging from personalized medicine, to monitoring the health of the soil and water, to microbial community bioengineering. There are outstanding technical and computational problems, including how to assemble short reads with high error rates into reliable contigs. Without the sequence scaffolds used in shotgun sequencing, errors will abound. Also, the number of species present is not known, and close homolog species may confuse the assembly.

- ✓ How are genes spliced by the spliceosome?
- ✓ What signals enhance or silence splicing? What can happen to the protein sequence when splicing is silenced?
- ✓ Do you know how to find introns/exons in NCBI Gene Viewer?
- ✓ Can you navigate AmiGO? Can you construct a paragraph of text given a graph of GO terms?
- ✓ Do the part-of and is-a relationships make sense to you? Given a pair of related terms, can you tell if the relation is is-a or part-of?
- ✓ How would you use GO to determine the presence or absence of a metabolic function in a metagenomic dataset?
- ✓ How must short reads be assembled? How is this different from whole genome shotgun sequencing using BAC and YAC scaffolds?