Phylogenetic Tree Computation Tutorial

Frank Olken

Lawrence Berkeley National Lab

Presentation to PGA Course

May 3, 2002

Berkeley, California

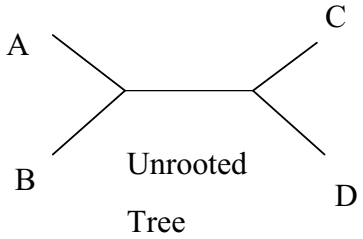
Overview

- Introduction: What? Why? …
- Multiple Sequence Alignment
- Computing Phylogenetic Trees
- Merging Trees
- Resources
- U.S. Funding Sources

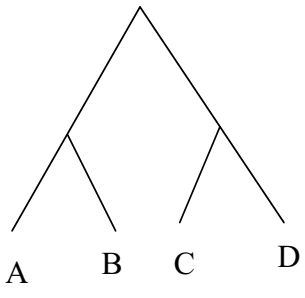
Introduction

Example

- Seq. A = A A C C G G T T
- Seq. B = A A C C G G T G
- Seq. C = A C C C G G T C
- Seq. D = A C C C G G T A



Unrooted Tree



Rooted Tree

Similarity vs. Homology

- Similar
  - sequences resemble one another
- Homolog
  - sequences derived from common ancestor
- Ortholog
  - homologous sequences within a species
- Paralog
  - homologous sequences between species

Ortholog vs. Paralog

- Ortholog – genomic variation occurs after speciation – hence can be used for phylogeny of organism
- Paralog – genetic duplication occurs before speciation – hence not suitable for phylogeny of organism

Homoplasy

- Sequence similarity NOT due to common ancestry
- May arise due to parallelism or convergent evolution
- Parallelism
- Convergent evolution

Phylogenetic Tree

- Binary tree (i.e, fan-out from nodes = 2)
- Variously rooted or unrooted
- Tree topology
- Branch lengths (evolutionary time)

Phylogeny of what?

- **Organisms**
  - **Whole genome phylogeny**
  - **Ribosomal RNA (surrogate for whole genome)**
- **Strains (closely related microbes)**
- **Individual genes (or gene families)**
- **Repetitive DNA sequences**
- **Metabolic pathways**
- **Secondary Structures**
- **Any discrete character(s)**
- **Human languages**
- **Microbial communities**

Why compute phylogenetic trees?

- **Understand evolutionary history**
- **Map pathogen strain diversity for vaccines**
- **Assist in epidemiology**
  - **Of infectious diseases**
  - **Of genetic defects**
- **Aid in prediction of function of novel genes**
- **Biodiversity studies**
- **Understanding microbial ecologies**

Rooted vs. Unrooted Tree

- Root - ancestor of all taxa considered
- Unrooted tree - typical result, unknown common ancestor
- Rooted tree - known common ancestor
- Specify root by means of outgroup
- Outgroup is distant from all other taxa
  - example: mammals and a salamander
  - ancestor of outgroup is presumed root

Which Sequences ?

- DNA
  - Very sensitive, non-uniform mutation rates
- cDNA/RNA
  - Useful for more remote homologies
- Protein Sequences
  - Useful for most remote homologies, deep phylogenies, more uniform mutation rates, more character states

Ribosomal RNA 16S Sequences

- These sequences exist in all organisms
- They are highly conserved
- Hence suitably for broad, very deep phylogeny studies
- Compiled for tens of thousands of organisms, mostly microbial
- Unsuited to fine grained phylogeny

What is to be computed?

- Tree topology
  - branching order
  - root
- Branch lengths (evolutionary time)
- Ancestral sequences
- Tree figure of merit (e.g., likelihood)
- Tree reliability

Computational Process

- Get DNA/RNA/Protein Sequences
- Construct multiple sequence alignment
- Compute pairwise distances
  - (for distance methods)
- Build tree: topology + branch lengths
- Estimate reliability
- Visualize

Multiple Sequence Alignment

Multiple Sequence Alignment

- Align sequences so that corresponding positions are in same columns
- Pad missing nucleotides as nulls where needed
- Each “column” then becomes a single character in the phylogeny computation

Multiple Sequence Alignment

- NP-hard
  - Worst case computational complexity is exponential in number of problem size
- In practice, often use greedy algorithms
  - choose best incremental change in solution
  - no backtracking
- Figure of merit:
  - Sum of pair-wise edit distances

Multiple Sequence Alignment Algorithms

- Dynamic Programming
- Hidden Markov Models
- Stochastic Context Free Grammars

Dynamic Programming MSA - Sequential Addition

- Compute pairwise edit distances between sequences via dynamic programming
- Merge closest pair
- Generate consensus sequence
- Merge sequence closest to consensus sequence alignment

Dynamic Programming MSA - Bottom Up Merging

- Compute pairwise edit distances between sequences via dynamic programming
- Merge closest pair
- Replace merged pair with consensus sequence
- Recompute pairwise edit distances, and merge closest pair ….

Hidden Markov Model MSA

- Estimate HMM for set of sequences
- Compute most probable alignment of each sequence to HMM
- Use this as basis for MSA
- HMM allows probabilistic representation of "consensus sequence"

SCFG MSA Algorithm

- Estimate Stochastic Context Free Grammar for set of ribosomal RNA sequences (e.g., 16S sequences)
- Compute most probable sequence alignment to SCFG for each sequence
- Use this as basis of multiple sequence alignment
- Preserves RNA secondary structure in MSA

Algorithms for computing phylogenetic trees

Evaluation Criteria for Tree Computational Methods

- Accuracy
- Explicit statistical model of evolution ?
- Efficient use of data
- Computation Time
- Branch lengths ?
- Quality measure ?
- Reliability measure ?

Computational Approaches to Phylogenetic Tree Computation

- Distance-based methods
  - UPGMA, Neighbor joining
- Maximum Parsimony Method
- Maximum Likelihood Methods
- Tree merginng
  - Consensus trees, supertrees

Distance vs. Character State Methods

- Distance Methods
  - UPGMA, Neighbor Joining, Min. Evol., ….
  - Requires distance measures between sequences
  - Suitable for continuous characters
- Character State Methods
  - Max. parsimony, Max. Likelihood, …
  - Requires discrete characters

Similarity Measures vs. Distances

- Similarity Measure
  - bigger value = more similar
- Distances
  - bigger value = less similar
  - triangle inequality
    - $|x,y| + |y,z| \leq |x,z|$
  - often assumed additive for distance-based phylogenetic tree construction

Can distance-based methods use similarity measures?

- Maybe …
- Depends on whether distance methods uses:
  - triangle inequality
  - additive distance measure

Simple distances between sequences

- Number of differing positions
- Weighted differences
- Edit distances (weighted sum of inserts, deletes, substitutions)
- Weighted Substitution Cost Matrices
  - PAM, BLOSUM
- Poisson Corrections (next slide)

Distance Metric Between Sequences

- p = n_d / n
- = number of characters which differ / total number of characters
- p is not proportional to evolutionary time
- Reason: sites can mutate more than once
- Poisson correction:
- d = - ln (1-p)

Similarity Measure for Protein Structures

- **Construct contact map (graph) for each protein structure**
  - **vertex = residue (amino acid)**
  - **edge = distance between AA's is less than 5 Angstroms**
- **Compute pairwise alignment between structures**
  - **nonoverlapping matching of residues**
- **Similarity Measure**
  - **= number of shared edges from contact graphs**

UPGMA Distance Method

- Unweighted Pair Group Method Using Arithmetic Mean
- by Sokal and Michener, 1958
- Merge closest pair of taxa (by distance)
- Recompute distances to merged node via mean of pairwise distances to leaves
- Repeat

UPGMA Method

- Fast to compute
- Implicitly assume molecular clock
  - i.e., uniform mutation rates across sites and branches

Neighbor Joining Distance Method

- Compute pairwise distances, d(i,j), set L = all leaves T
- Compute D(i,j) = d(i,j) - (r(i)+r(j))
- r(i) = average distance to other leaves
- Merge closest pair of sequences i and j
  - for new k, set d(k,m) = 1/2 (d(i,m)+d(j,m)-d(i,j)) for m in L
  - Add k to T with
    - set d(i,k) = 1/2 (d(i,j)+r(i)-r(j))
    - set d(j,k) = d(i,j)-d(i,k)
  - replace i and j with k in L
- Repeat

Neighbor Joining

- Generates unrooted trees
- Assumes additive distances in tree

Merits of Distance Methods

- Fastest method
- Not very accurate or efficient use of data
- Can use continuous data (not just sequences)
- No statistical model of evolution
- No figure of merit
- No branch lengths

Maximum Parsimony

- Minimize total number of state changes along all paths in tree
- Intermediate computational cost
- Figure of merit = number of state changes
- No statistical model
- No branch lengths

Maximum Likelihood

- L = p (data | tree, branch lengths, model)
- Search trees, branch lengths to find max L
- This is MLE tree
- Algorithms:
  - generate tree topology
  - optimize branch lengths
  - retain if best seen
  - loop

Maximum Likelihood Methods

- Explicit statistical model
- Figure of merit = log likelihood
- Computes branch lengths
- Very expensive to compute, use heuristics
- Reliability estimation by resampling
- Efficient use of sequence data
- Examples: Phylip/dnaML, fastdnaML

Maximum Likelihood Estimaton - Assumptions

- Characters (nucleotide positions) evolve independently
- Mutation Rate variation:
  - Molecular clock ==> uniform rates across positions and branches
  - We can allow rate to vary by position (usually assume Gamma distribution)
  - Requires that estimate more parameters

MLE Phylogenetic Tree Construction Algorithms

- Must compute both tree topology, branch lengths, rates of evolution (for each position)
- Too many tree topologies to search exhaustively
- Hence heuristic tree search
  - Order taxa (randomly?)
  - Build tree incrementally
  - For each taxa, consider all possible locations in existing tree to insert
  - Compute likelihood for each possible insertion point.
  - Choose best insertion point
  - Get next taxa

MLE Phylogenetic Tree Algorithms

- Branch swapping
  - Given tree
  - Consider all possible pairwise branch swaps, where branches within distance “k” in tree
  - For each possible swap
    - reoptimize branch lengths
    - compute likelihood of resulting tree

MLE Phylogenetic Tree Algorithms

- Overview of algorithm
  - Pick next taxa
  - Generate possible topologies
  - Perform branch optimization for each topology
  - Compute likelihood of resulting tree
  - Retain best “k” trees
  - Get next taxa

Popular MLE Codes

- dnaML - Joe Felsenstein (U. Washington)
- fastdnaML - Gary Olsen (UIUC)
- PAUP - Dave Swofford (Florida State U.)
- PAML

Merging Phylogenetic Trees

Merging Phylogenetic Trees

- Consensus Trees
  - All trees computed over same taxa
- Supertrees
  - Trees computed over overlapping sets of taxa

Consensus Trees

- All trees computed over same set of taxa
- Used for cases when there exist many tree with similar figures of merit (parsimony scores, likelihood estimates)
- Goal: identify robust result

Consensus Trees

- Strict Consensus
  - Only report branch orders which occur in all trees
  - Use multi-way branching when branch order is not consistent across all trees
  - Very strong consensus criteria

Supertrees

- Supertrees are constructed from candidate trees computed over overlapping sets of taxa
- Supertree construction is a way to compute very large trees.
- Usually need significant pairwise overlaps among at least some pairs of trees taxa

Majority Consensus Trees

- Use voting algorithms on branch orders
- Report branch orders which supported by a majority (or supermajority, e.g., 2/3) of candidate trees
- More relaxed that strict consensus trees
- Widely used ….

Supertrees

- Merge overlapping phylogenetic trees
- DCM Algorithm
  - Tandy Warnow at UT Austin

Tree Reliability

- Bootstrapping (Felsenstein, et al.)
  - Resample characters (nucleotide positions)
  - Recompute phylogenetic tree
  - Repeat
  - Compare trees, compute consensus tree
  - Expensive
  - Sample size = 100 (?)

Whole Genome Phylogeny

- Done at gene level (not sequence)
- Look at larger genome rearrangements
- Good for deep phylogeny
- Consider inversions, translocations
- Currently primarily suitable for microbes
  - most whole genomes sequences have been for microbes
- See papers by Sankhoff, et al.

Choosing an algorithm

- **Continuous characters, lots of data, few computing resources ==> Neighbor joining**
- **Discrete characters, few mutations / homoplasy ==> Maximum Parsimony**
- **Discrete characters, limited sequence lengths, some homoplasy ==> Maximum Likelihood Estimation**
- **Discrete characters, many taxa ==> Supertree**
- **Complete genomes ==> Whole genome phylogeny**

Choosing an algorithm

- For low numbers of mutations - MP and MLE generate similar results

Practical problems

- Often difficult to resolve deep branch orders
  - (ancient branching of evolutionary tree)
- Tradeoff between:
  - Sensitivity (rapid genome evolution)
  - Detection of ancient branching (slow genome evolution)

Data Management for Phylogeny

- Sequence Data Management
- Multiple Sequence Alignments
- Tree Data Management
- Tree Alignments
  - consensus subtrees

Phylogeny for Vaccine Studies

- Phylogeny of disease strains
- Need to assure that vaccine confers resistance to all strains
- Phylogeny is used to select diverse strains for constructing vaccine
- Example: influenza

Phylogeny and Epidemiology

- Pathogen phylogeny used to assist epidemiological studies
- Example: HIV
  - rapid evolution of virus
  - use phylogeny to verify source of infection of particular individual
- Co-evolution of pathogens and hosts
- See Crandall, *Evolution of HIV*

Resources on Phylogeny

Web Resources

- **Felsenstein's Phylogenetic Program Directory**
  - **http://evolution.genetics.washington.edu/phylip.html**
- **UT Austin Phylogenetics Lab**
  - **http://kristin.csres.utexas.edu/**
- **Woese Lab**
  - **http://www.life.uiuc.edu/micro/woese.html**
- **Tree-of-life web site**
  - **http://tolweb.org/tree/phylogeny.html**

Bibliographies and Search Engines

- **Reading lists to IB 200A + IB 200B**
  - **by Mishler at UC Berkeley**
  - **http://ib.berkeley.edu/courses/ib200a/**
- **Medline (a.k.a. PubMed) at NLM**
- **BIOSIS (Biological Abstracts)**
- **INSPEC (by IEE)**
- **NEC Citeseer (CS papers)**
- **MathSciNet (Mathematical Reviews)**

Books on Phylogeny

- Graur, Li. *Fundamentals of Molecular Evolution,* Sinauer
- Hall, *Phylogenetics Made Easy*, Sinauer
- Hillis, Moritz, Mable. *Molecular Systematics*, 2nd edition, Sinauer, 1996
- Kitching, Forey, Humphries. Cladistics: *The Theory and Practice of Parsimony Analysis*, 1998
- Kimura, M. *The Neutral Theory of Evolution*, Cambridge, 1983
- Li. *Molecular Evolution*, Sinaeur
- Nei, M. & S. Kumar. *Molecular Evolution and Phylogenetics*, Oxford, 2000
- Page & Holmes. *Molecular Evolution: A Phylogenetic Approach,* 1998
- Smith, J.M., *Evolutionary Genetics*, 1998
- Wheeler & Meier. *Species Concepts and Phylogenetic Theory*, 2000
- Wilkins. *Evolution of Developmental Pathways*, Sinaeur, 2001

More Books on Phylogeny

- Harvey, Leigh Brown, Smith, Nee. *New Uses for New Phylogenies*, Oxford, 1966
- Crandall, K. (editor) *The Evolution of HIV,* Johns Hopkins Univ. Press, 1999
- Mount, D.W. *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, 2000, Chapter 6 - Phylogenetic Prediction
- Doolittle, R.F. *Computer Methods for Macromolecular Sequence Analysis*, Methods in Enzymology, vol. 266, 1996, Academic Press

Journals on Phylogeny

- **Cladistics**
- **Molecular Biology and Evolution**
- **Molecular Phylogenetics and Evolution**
- **Systematic Biology**
- **Systematic Zoology**
- **Evolutionary Biology**
- **Taxon**
- **Bioinformatics**
- **J. of Computational Biology**
- **J. of Theoretical Biology**
- **Ecology and Evolutionary Biology**

Conferences on Phylogeny

- **RECOMB**
- **ISMB (Intelligent Systems for Molecular Biology)**
- **Evolution 2002**
- **Classification Society of N. America Annual Mtg.**
- **Conf. Of the Int'l. Federation of Classification Societies (IFCS)**
- **ICSEB (Intl. Conf. On Systematics & Evolutionary Biology)**

Phylogenetics Organizations

- **http://www.ucmp.berkeley.edu/subway/phylo/phyloorg.html**
- **Classification Society of North America**
- **Society for Molecular Biology and Evolution**
- **Will Hennig Society**
- **Classification Society**
- **Green Plant Phylogeny Group**
- **European Society for Evolutionary Biology**
- **Society for Systematic Biology**
- **Systematics Society (Britain)**

Funding for Phylogeny Research

U.S. Funding Sources for Computational Phylogeny

- NSF
- NIH
- CDC
- DOE
- DOD

Funding - NSF

- **NSF Biology Directorate**
  - Primary funder for phylogeny in U.S.
  - Both specific programs in phylogeny
  - and **ITR (Info. Tech. Research) Program** (large interdisciplinary projects)
  - Contact Sylvia Spengler
  - **Tree of Life Program** - Diana Lipscomb

Funding - NIH

- **National Institutes of Health**
- **Primarily interested in phylogeny and human health**
- **Phylogeny of pathogens, disease genes, …**
- **NIAID - National Institute of Allergies and Infectious Diseases**
  - **phylogeny of pathogens of infectious diseases**
- **NIGMS - National Institute of General Medical Sciences**
  - **phylogeny of gene families**

Funding - CDC

- Center for Disease Control (Atlanta)
- Concerned with infectious diseases and food toxins
- Phylogeny of pathogens
- Phylogeny and epidemiology

Funding - DOE

- U.S. Department of Energy
- Microbial Research Program
  - Prokaryotes primarily
  - Bioremediation
  - Carbon sequestraton
  - Biofuels
- Microbial Phylogeny & Ecology
- Bioterrorism / Biowarfare Program (NN20)

Funding - DOD

- Biowarfare / Bioterrorism concerns
- Phylogeny of pathogens
- Understanding Pathogenicity
- DARPA Biowarfare Defense
  - **DARPA DSO Pathogen Genomic Sequencing (Dr. Eric Eisenstadt)**

Acknowledgements

- This research supported by U.S. Department of Energy, Office of Biological and Environmental Research

Contact Information

- **Frank Olken**
  - **Lawrence Berkeley National Laboratory**
  - **1 Cyclotron Road, Mailstop 50B-3238**
  - **Berkeley, CA 94720**
  - **Tel: 510-486-5891**
  - **Fax: 510-586-4004**
  - **Email: olken@lbl.gov**
  - **URL: http://www.lbl.gov/~olken**