Tracing the Evolution of Numerical Phylogenetics: History, Philosophy, and Significance

Adam W. Ferguson
Phylogenetic Systematics
26 January 2009

Inferring Phylogenies

- Historical endeavor
- Darwin - 1837
  - Evolutionary tree
- Haeckel - 1866
  - Coined "Phylogeny"
Algorithmic Methods

• Well-defined methodology
• Tree reconstruction
• “those that are well-defined enough to be carried out by a computer.”
  Felsenstein 2004, p. 123

Why Algorithms?

• How to count the number of changes of a character state?
• How to search among all possible trees?
• How to infer branch lengths?
• How to handle different types of characters?
  – DNA, morphological, protein sequences
What Is an Algorithm?

- “A step-by-step problem-solving procedure, especially an established, recursive computational procedure for solving a problem in a finite number of steps.”

- “A finite set of unambiguous instructions performed in a prescribed sequence to achieve a goal, especially a mathematical rule or procedure used to compute a desired result. Algorithms are the basis for most computer programming.”

Phylogenetic Tree

- "a decision-making process for computing cladograms, as in the Wagner algorithm" Schuh 2000
Algorithmic Phylogenetics

- Not that simple
- Controversial history (Felsenstein 2001)
- Extensive variety of methodologies
- Continually debated

Beginnings

- Late 1950s
- Early 1960s
- Computers
- Numerical taxonomy (Phenetics)
- Peter Sneath
- Robert Sokal
- Clustering algorithms
Michener and Sokal (1957)

- Journal of Evolution
- Morphological characters of bees (Proteriades)
- 1<sup>st</sup> Paper on numerical inference of phylogenies
- Clustering algorithms

### Numerical Taxonomy

- Sokal and Sneath (1963)
- Classification based on phenetic principles
- Measures of overall similarity
- No consideration for phylogenetic relationships
Phenetics vs. Evolutionary Systematics

- Systematists - outrageous!

- Ernst Mayr and G. G. Simpson

- Minority view

- Lay ground work for thinking algorithmically

Edwards & Cavalli-Sforza (1964)

- Students of R. A. Fisher
- Trees of human populations
- Gene frequencies of blood group alleles
- Two methods
  - Parsimony - least amount of “string”
  - Least-squares/distance - minimize difference between genetic and actual (geographic) distance
Edwards & Cavalli-Sforza (1964)

- First publication of a parsimony tree

Edwards & Cavalli-Sforza (1964)

Introduced three (four) major concepts

1) Parsimony

2) Maximum Likelihood

3) Statistical inference and phylogenetics

4) Distance matrix (could have)
Camin and Sokal (1965)

- Modern work on parsimony
- Organism with known phylogeny
- Caminalcules
- Best reconstructions
  - Minimized number of changes of state of a character

Camin and Sokal’s Algorithms

- Evaluate number of changes of a given tree
- Construct/rearrange a tree to search among topologies
- Assumptions
  - Multi-state characters with linear arrangement
  - Known ancestral state
  - Change irreversible
Eck & Dayhoff (1966)

- *Atlas of Protein Sequence and Structure*
- Parsimony analysis of protein sequences
  - Each amino acid change to any of the 19 others in a single step
  - Sequential addition strategy
  - First application of parsimony with unordered states

Fitch & Margoliash (1967)

- First published distance matrix phylogeny
- Cytochrome-c sequences
- Distance matrix based on least-squares
- Algorithm employed clustering methods (*ca.* Sokal and Sneath 1964)
Gomberg 1968 (unpubl.)

- Applied Bayesian approach
- Characters changed according to Brownian motion process
- Based on Bayes’ theorem (1790)
- Conditional probability

Kluge & Farris (1969)

- Unordered parsimony
- Algorithms
  - Reconstructing changes on a given tree
  - Searching among trees for most parsimonious
- Problems
  - Unknown ancestral character state
  - Reversible change
Farris (1971)

- Application of algorithms to DNA sequences
- Parsimony methods for unordered states
  - A, C, G, T
- Constructing phylogenies from nucleotide sequences by parsimony

Rapid Development

- Sneath and Sokal to Fitch, only 14 years

Computers + Molecular Biology
Advancements + Mathematics + Smart People
= Numerical Phylogenetic Inference
A Footnote- Hennig?

- Father of Parsimony?
- 1950, 1966
- Methods based on earlier work by Zimmerman (1931)
- Enforcement of monophyletic classification
- Neither introduced parsimony or any algorithmic approach

Where Are We Now?

Four Major Families of Phylogenetic Methods (all utilizing some form of algorithmic methodology)

1) Distance
2) Parsimony
3) Likelihood
4) Bayesian
Division of Families

Two Ways To Divide Methods

1) How they handle data?
   Distance vs. Discrete Data

2) Approach taken when building trees?
   Clustering vs. Search Methods

Distance Methods

• Introduced by Cavalli-Sforza & Edwards (1967) and Fitch & Margoliash (1967)
• Distance, clustering algorithms
• Least-squares, UPGMA, Neighbor-joining
• Fastest
• Loss of information, inability to correct using data from within tree
Parsimony Methods
- Edwards & Cavalli-Sforza (1963)
  - “the minimum net amount of evolution.”
- Discrete data, searching algorithms
- Inconsistency is strongest challenge
- Fairly well-behaved method (except in face of long-branch attraction)

Likelihood Methods
- Maximum likelihood (R.A. Fisher)
- Introduced by Edwards & Cavalli-Sforza (1964)
- Discrete data, searching algorithms
- Probability of obtaining observed data given specific hypothesis
- Requires model of evolution, tree, and observed data
- Computationally expensive
Bayesian Methods

- Dates to 1790s
- Introduced by Gomberg (1968)
- Based on posterior probabilities
- Discrete, searching algorithms
- Assumes prior knowledge about data
- Markov Chain Monte Carlo (Metropolis algorithm)
- All controversy involves prior knowledge

Summary

- Defining phylogenies algorithmically
- Mathematical/statistical approaches to phylogenetic inference
- Multiple ways to construct phylogenetic trees
- Invoke different assumptions
- Continuing debate
Questions?

Bayesian

Maximum Likelihood