Systematics and molecular evolution: some history of numerical methods

Joe Felsenstein

Department of Genome Sciences and Department of Biology
University of Washington, Seattle
Irrelevant?

In the 1960’s molecular biologists commonly considered all population biology as an irrelevancy at best: mere “stamp collecting"

In science there is only physics, all else is stamp collecting

*Ernest Rutherford*

By contrast, while molecular biologists ignored evolutionary biology until the 1990s, population biologists had been using molecular methods since the late 1960’s, but to a degree limited by their limited funding.
The problem

Population biologists have been working on numerical methods for inferring phylogenies (evolutionary trees) since the 1960’s. Their importance has only recently been recognized by molecular biologists.

The inference of, and use of phylogenies is a central problem in computational molecular biology, though in bioinformatics textbooks they usually are covered in only a small part of the book (usually just a few pages).
Why evolution?

Nothing in biology makes sense except in the light of evolution.

_Theodosius Dobzhansky, 1973_

Perhaps that is an overstatement (what about physics and chemistry?), but evolution is how the data came to be, and taking evolution into account is the only efficient way of analyzing it.
Ernst Mayr and George Gaylord Simpson


- Major figures in the completion of the “modern synthesis” or “Neodarwinian synthesis” in the 1940s.
- Leaders of the “evolutionary systematics” approach to taxonomic classification, dominant until the 1970s.
A pattern of grades involving very unequal rates of overall evolution is implicit in the use of paraphyletic groups in Mayr and Simpson’s practice.
A horse tree drawn by Simpson

The tree seems to be fading away – it is getting obese and nonspecific. Also notice that genera arise from other genera.
Willi Hennig (1913-1976)

The major advocate and developer of “phylogenetic systematics” which advocates that all groups be monophyletic. Outlined a simple method for inferring phylogenies that can be used when characters do not conflict.
Positions on classification as of about 1960

- **Evolutionary systematics.** George Gaylord Simpson and Ernst Mayr led a movement that allowed non-monophyletic (paraphyletic) groups such as reptiles, on the assumption that groups could be separated by real differences of rates of evolution (sometimes “grades” rather than “clades”). But did molecular data show similar differences of rates of evolution as morphology?

- **Phylogenetic systematics.** Willi Hennig advocated purely monophyletic classification.

- **Phenetics.** Sokal and Sneath advocated making a classification without reference to evolution, using numerical clustering methods.
Technological change post World War II

- Former physicists found molecular biology (first protein sequence, 1951)
- Former codebreakers and atomic bomb builders build the early computers (first stored-program digital computer, 1949)
- Most U.S. universities got their first computer about 1957.
- First sequences of same gene in multiple species in late 1950s.
Molecular evolution gets off the ground

Zuckerkandl and Pauling in 1962 discussed using trees to infer ancestral sequences, and named this “chemical paleogenetics”. They were about 30 years ahead of their time.

(But then, you expect Linus Pauling to be 30 years ahead of his time).
Developers in the late 1950s and early 1960s of numerical clustering methods and chief originators and advocates of the “phenetic” approach to classification, which clusters organisms by similarity without reference to evolutionary history.
The first numerical phylogeny: Sokal and Michener 1957

A tree of bees, which Michener intended as an inference of the phylogeny.
Cavalli-Sforza and Edwards, 1963; Edwards, 1970

Introduced (in 1963-1964) the parsimony and likelihood methods for inferring phylogenies for models of gene frequency evolution. They were also co-inventors (with Fitch and Margoliash) of the distance matrix methods. These are the three major methods for reconstructing phylogenies.
The first phylogeny by parsimony

Gene frequencies of human populations, the tree of minimum length in gene frequency space, inferred by Edwards and Cavalli-Sforza.
Camin noticed (in 1965) that students who did the best job recovering the true “phylogeny” of the Caminalcules made the reconstruction which required the fewest changes of state.
J. S. Farris and Arnold Kluge in the 1980s

Margaret Dayhoff (1925-1983)

Margaret Dayhoff in about 1966. *Courtesy of Ed Dayhoff.*

- A major pioneer of molecular databases (starting in 1965)
- (With Richard Eck) made the first numerical phylogenies using molecular data
- Presented trees organized by gene families in the *Atlas of Protein Sequences* (later the PIR database) in 1966.
- Compiled the first empirical substitution rate matrices for amino acids, intended to form the basis of a probabilistic model of protein evolution.
Walter Fitch (1929-2011):

- The first major distance matrix method (1967)
- Developed algorithm (1971) that counts changes of state in DNA parsimony.
- Introduced the terms and concepts of orthology and paralogy.
- Co-founded the journal MBE and the SMBE.
Fitch and Margoliash’s 1967 distance tree

Fig. 3 (right above). A gene phylogeny as reconstructed from observable mutations in several heme-containing globins. See Fig. 2 for details. The percent “standard deviation” (7) for this tree is 1.33.
In 1969 Jukes and Cantor introduced the first stochastic process model of DNA change, in one paragraph buried in the midst of a giant review of protein sequence evolution.

Cantor later made important technical discoveries in genomics. Jukes was a nutritional biochemist who was the primary person responsible for insisting that pregnant women get folic acid in their diet.
Jerzy Neyman

A major figure in mathematical statistics (confidence intervals, Neyman-Pearson testing theorems), Neyman was enticed in 1971 into doing the first likelihood analysis of molecular sequence data for protein sequences with a 3-species tree and a Jukes-Cantor-like symmetrical model of change among 20 amino acids.
Further development of statistical methods

- “Pruning” algorithm for efficient calculation of likelihoods on trees (me, 1973, but not yet applied to molecules)
Further development of statistical methods

- “Pruning” algorithm for efficient calculation of likelihoods on trees (me, 1973, but not yet applied to molecules)
- Proof of inconsistency of parsimony (me, 1978; James Cavender too, 1978)
Further development of statistical methods

- “Pruning” algorithm for efficient calculation of likelihoods on trees (me, 1973, but not yet applied to molecules)
- Proof of inconsistency of parsimony (me, 1978; James Cavender too, 1978)
- Programs distributed (WAGNER78, JS Farris 1978; PHYLIP, me, 1980; PAUP, Swoford, 1983)
Further development of statistical methods

- “Pruning” algorithm for efficient calculation of likelihoods on trees (me, 1973, but not yet applied to molecules)
- Proof of inconsistency of parsimony (me, 1978; James Cavender too, 1978)
- Programs distributed (WAGNER78, JS Farris 1978; PHYLIP, me, 1980; PAUP, Swofford, 1983)
- Pruning algorithm used to make ML work for DNA (me, 1981)
Further development of statistical methods

- “Pruning” algorithm for efficient calculation of likelihoods on trees (me, 1973, but not yet applied to molecules)
- Proof of inconsistency of parsimony (me, 1978; James Cavender too, 1978)
- Programs distributed (WAGNER78, JS Farris 1978; PHYLIP, me, 1980; PAUP, Swofford, 1983)
- Pruning algorithm used to make ML work for DNA (me, 1981)
- Bootstrap and other sampling methods assess confidence of clades (me, 1985; Penny and Hendy, 1985)
Further development of statistical methods

- “Pruning” algorithm for efficient calculation of likelihoods on trees (me, 1973, but not yet applied to molecules)
- Proof of inconsistency of parsimony (me, 1978; James Cavender too, 1978)
- Programs distributed (WAGNER78, JS Farris 1978; PHYLIP, me, 1980; PAUP, Swofford, 1983)
- Pruning algorithm used to make ML work for DNA (me, 1981)
- Bootstrap and other sampling methods assess confidence of clades (me, 1985; Penny and Hendy, 1985)
- Neighbor-Joining method (Nei and Saitou, 1987)
Further development of statistical methods

- “Pruning” algorithm for efficient calculation of likelihoods on trees (me, 1973, but not yet applied to molecules)
- Proof of inconsistency of parsimony (me, 1978; James Cavender too, 1978)
- Programs distributed (WAGNER78, JS Farris 1978; PHYLIP, me, 1980; PAUP, Swofford, 1983)
- Pruning algorithm used to make ML work for DNA (me, 1981)
- Bootstrap and other sampling methods assess confidence of clades (me, 1985; Penny and Hendy, 1985)
- Neighbor-Joining method (Nei and Saitou, 1987)
- Codon models (Goldman and Yang, 1994; Muse and Gaut, 1994)
Further development of statistical methods

- “Pruning” algorithm for efficient calculation of likelihoods on trees (me, 1973, but not yet applied to molecules)
- Proof of inconsistency of parsimony (me, 1978; James Cavender too, 1978)
- Programs distributed (WAGNER78, JS Farris 1978; PHYLIP, me, 1980; PAUP, Swofford, 1983)
- Pruning algorithm used to make ML work for DNA (me, 1981)
- Bootstrap and other sampling methods assess confidence of clades (me, 1985; Penny and Hendy, 1985)
- Neighbor-Joining method (Nei and Saitou, 1987)
- Codon models (Goldman and Yang, 1994; Muse and Gaut, 1994)
- Use of AIC model comparison (Adachi and Hasegawa, 1996; Posada and Crandall, 1998)
A noticeable tendency is for much of the influence to come from outside systematics, and for biochemists (as opposed to “molecular biologists”) to be an important influence on work in molecular evolution.
Conflict over phylogeny methods, 1970s to 1980s

Conflict over phylogeny methods, 1970s to 1980s


- Advocacy of his views is taken up by a number of systematists (Brundin, Nelson, Platnick) in the 1970s.
Conflict over phylogeny methods, 1970s to 1980s


- Advocacy of his views is taken up by a number of systematists (Brundin, Nelson, Platnick) in the 1970s.

- Some numerical systematists, chiefly J. S. Farris, ally with them.
Conflict over phylogeny methods, 1970s to 1980s

- Advocacy of his views is taken up by a number of systematists (Brundin, Nelson, Platnick) in the 1970s.
- Some numerical systematists, chiefly J. S. Farris, ally with them.
- They adopt an evangelical and messianic style and attack entrenched evolutionary systematists, calling themselves “phylogenetic systematists”.
Conflict over phylogeny methods, 1970s to 1980s


- Advocacy of his views is taken up by a number of systematists (Brundin, Nelson, Platnick) in the 1970s.

- Some numerical systematists, chiefly J. S. Farris, ally with them.

- They adopt an evangelical and messianic style and attack entrenched evolutionary systematists, calling themselves “phylogenetic systematists”.

- Tensions between phylogenetic systematists and others within the Numerical Taxonomy Conference lead to major conflicts at its 1979 meeting at Harvard University. Not a fun meeting.
Conflict over phylogeny methods, 1970s to 1980s

- Advocacy of his views is taken up by a number of systematists (Brundin, Nelson, Platnick) in the 1970s.
- Some numerical systematists, chiefly J. S. Farris, ally with them.
- They adopt an evangelical and messianic style and attack entrenched evolutionary systematists, calling themselves “phylogenetic systematists”.
- Tensions between phylogenetic systematists and others within the Numerical Taxonomy Conference lead to major conflicts at its 1979 meeting at Harvard University. Not a fun meeting.
- They found a separate organization, the Willi Hennig Society, in 1980 (today their house journal is *Cladistics*)
More conflict over phylogeny methods, 1980s to today

- Many other numerically-oriented people found themselves outside this group.
More conflict over phylogeny methods, 1980s to today

- Many other numerically-oriented people found themselves outside this group.
- Molecular evolutionists mostly avoided these conflicts and have remained pragmatic and eclectic.
More conflict over phylogeny methods, 1980s to today

- Many other numerically-oriented people found themselves outside this group.
- Molecular evolutionists mostly avoided these conflicts and have remained pragmatic and eclectic.
- After about 1990, the SSB and the SMBE became the gathering points for all the folks who were outside of the Will Hennig Society. Many were phylogenetic systematists, but they wanted normal scientific communication.
More conflict over phylogeny methods, 1980s to today

- Many other numerically-oriented people found themselves outside this group.

- Molecular evolutionists mostly avoided these conflicts and have remained pragmatic and eclectic.

- After about 1990, the SSB and the SMBE became the gathering points for all the folks who were outside of the Will Hennig Society. Many were phylogenetic systematists, but they wanted normal scientific communication.

- In the 1990s use of likelihood and Bayesian methods (and distance methods interpreted statistically) gradually became recognized as “statistical phylogenetics”, an alternative to the Hennig Society’s parsimony-only approach to reconstructing phylogenies.
More conflict over phylogeny methods, 1980s to today

- Many other numerically-oriented people found themselves outside this group.
- Molecular evolutionists mostly avoided these conflicts and have remained pragmatic and eclectic.
- After about 1990, the SSB and the SMBE became the gathering points for all the folks who were outside of the Will Hennig Society. Many were phylogenetic systematists, but they wanted normal scientific communication.
- In the 1990s use of likelihood and Bayesian methods (and distance methods interpreted statistically) gradually became recognized as “statistical phylogenetics”, an alternative to the Hennig Society’s parsimony-only approach to reconstructing phylogenies.
- The Hennig Society spurns compromise and has centers of strength among morphological systematists (e.g. parts of AMNH in New York; The Hortorium (herbarium) at Cornell University).
Positions on classification nowadays

- **Phylogenetic systematics.** Willi Hennig advocated purely monophyletic classification. Now the (strongly) dominant approach.

- **Evolutionary systematics.** Has almost faded away. Its adherents were complacent and did not bother to make their methods algorithmic.

- **Phenetics.** Although Sokal and Sneath strongly influenced the field of numerical clustering, their approach to biological classification has few adherents.

- **IDMVM** One person (me) takes the view that It Doesn’t Matter Very Much, as we use the phylogeny and, given that, we never use the classification system. *Warning:* this is widely regarded as a marginal crackpot view [“A bizarre thumb in the eye to systematists” – Michael Sanderson].
Classification versus phylogenies

It is critically important to realize that the task of making a classification system and the task of making inferences about phylogeny are logically separable. You can infer the phylogeny without yet deciding how it will be used (or not used) in determining the classification.

Many biologists do not understand this. Systematists *insist* on not understanding it.

Most textbooks muddle it thoroughly.

Historians of science and philosophers of science do the same.
Rise of interest in things phylogenetic

Web of Science citations that have "phylogen*"

Systematics and molecular evolution: – p.29/34
Rise of phylogenetic considerations in genomics

fraction of Web of Science citations with "genom*" that have "phylogen*"

Systematics and molecular evolution: – p.30/34
Rise of statistical-model-based methods

Fraction of those Web of Science citations that have "phylog*" or "cladist*" that also have "likelihood" or "bayesian"
References


References


References


