Phylogenetic Reconstruction Methods

- Distance-based Methods
- Character-based Methods
  - non-statistical
    - a. parsimony
  - statistical
    - a. maximum likelihood
    - b. Bayesian inference

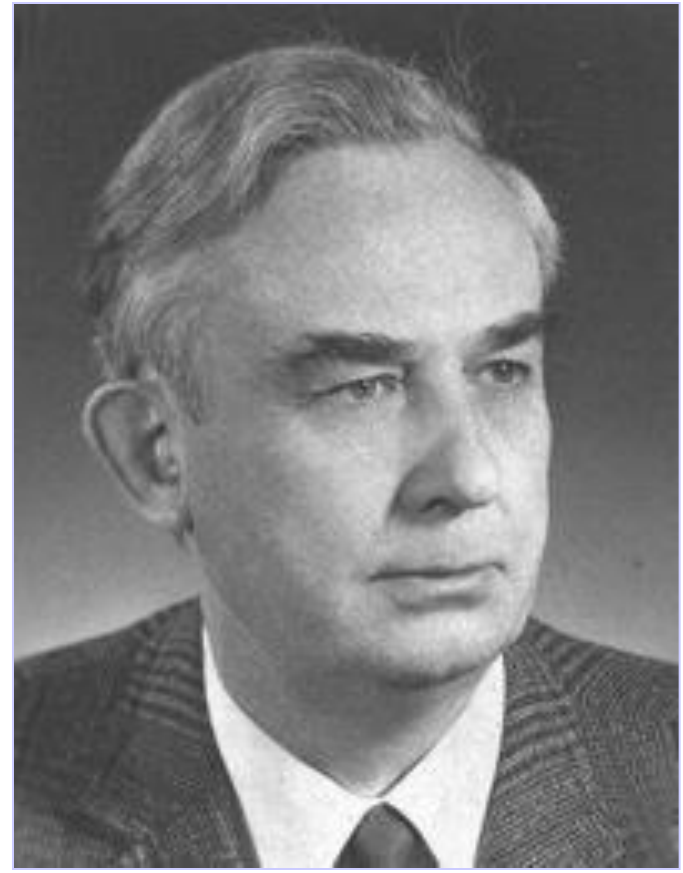
Parsimony has its roots in Hennig's phylogenetic systematics (cladistics)

1950: *Grundzage einer Theorie der Phylogenetischen Systematik*

1966: Davis and Zangerl's published English translation: *Phylogenetic Systematics*

1966: Brundin published monograph "*Chironomids of the Transantarctic Continents*"

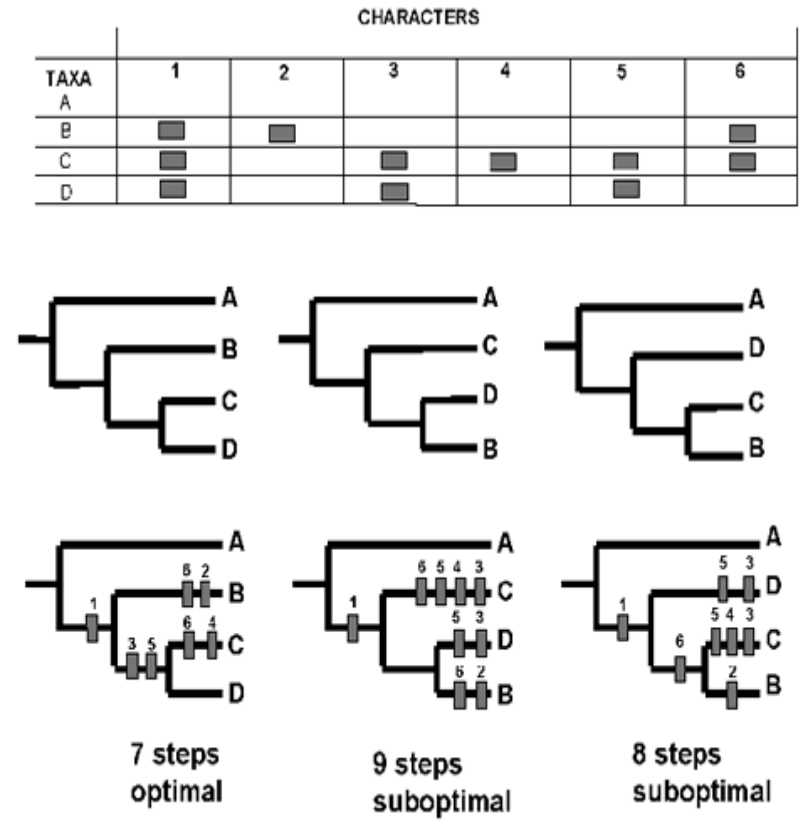
- branches in midge cladogram mirrored hypothesized break up of Gondwanaland
- identical pattern obtained across three subfamilies



Willi Hennig

Phylogenetic Systematics

* phylogenetic systematics refers to Hennig's classification philosophy
* we collectively refer to cladistic methodology as parsimony or maximum parsimony methods (for phylogenetic inference
* build a taxon x character matrix
* seek tree(s) with the fewest number of evolutionary changes (= parsimony)
* characters usually plotted on branches



Cladistics for Palaeontologists:
http://images.palass.org/newsletters/clastics/Forey.fig5.gif

Parsimony

* primary phylogenetic inference method of cladists
* *optimality criterion* is shortness: tree with fewest number of evolutionary steps (= most parsimonious tree, MPT)
* not saying that it is true
* not claiming that nature/evolution is parsimonious
  - only that parsimonious hypotheses can be defended without resorting to special knowledge, authoritarianism, **a priorisms**
* non-statistical approach

Parsimony

* tree length depends on
  1) the data (number of taxa, number of characters, amount of homoplasy)
  2) costs (steps) associated with character transformations

Number of possible unrooted trees

$$\frac{(2n-3)!}{2^{n-2}(n-2)!}$$

Number of possible rooted trees

$$\frac{(2n-5)!}{2^{n-3}(n-3)!}$$

Number of sequences	Number of unrooted trees	Number of rooted trees
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425

For 50 taxa there are 3 x 10^{74} rooted trees

...more possibilities than there are than atoms in the universe

A. Exhaustive algorithms

- search all possible trees
- below 12 or so taxa: search all trees
- PAUP yields a frequency histogram of tree by length

B. Branch and bound

- 18 to 25 taxa
- counts steps as it builds tree
- tree length can never decrease as a result of adding taxa
- suboptimal solutions that exceed the shortest trees in memory are excised
- all reasonable trees searched
- shortest trees assured
- upper bound (25 taxa) for clean data; fewer if lots of homoplasy

C. Heuristic algorithms

- 25+ taxa, you must employ **heuristic algorithms**
- different programs have different heuristic algorithms
- initially find a short tree then try trail and error (heuristic) methods to seek shortest tree

1) randomize input order of taxa
2) branch rearrangements
    * local and global rearrangements
      e.g., in parsimony algorithms use sub-tree pruning and tree bisection and reconnection

- heuristic methods can get trapped on local optima (called *tree islands*)

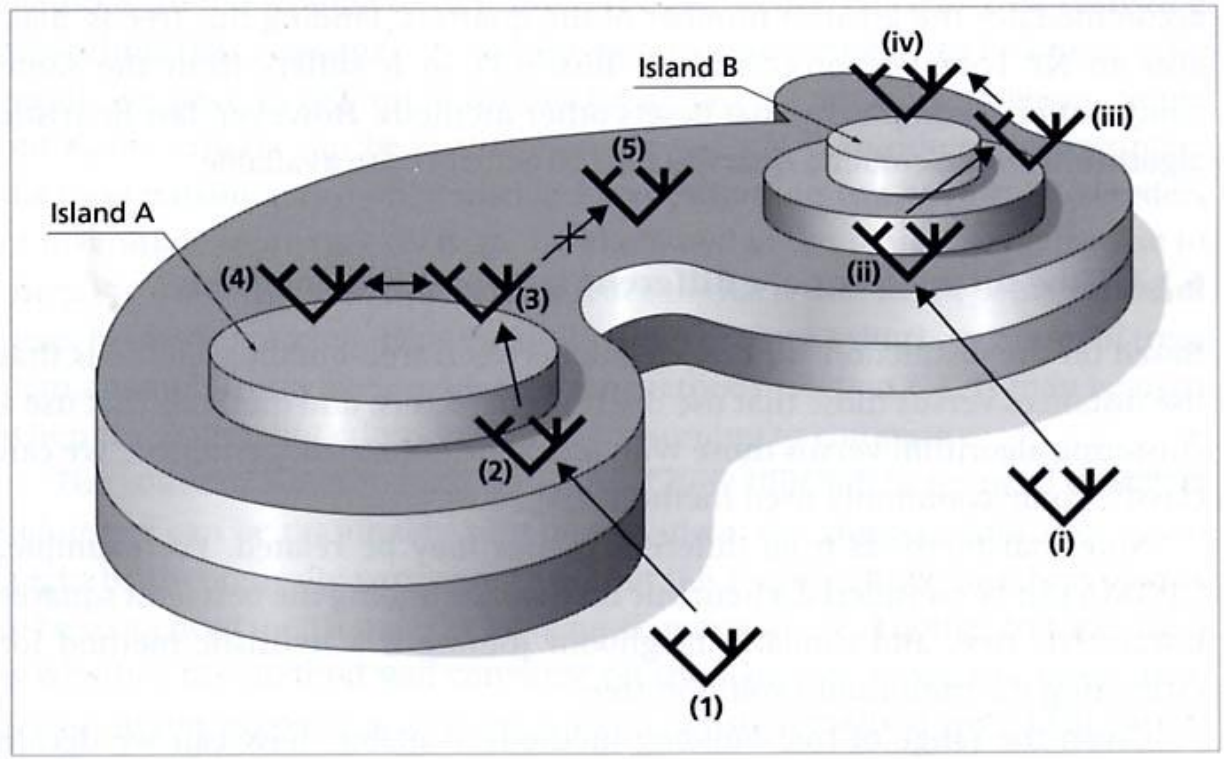


Fig. 6.4 Landscapes and the problem of islands of trees (locally optimal sets of trees). A hill climbing algorithm that started from tree 1 would succeed in finding trees 3 and 4 (which comprise island a), but would fail to discover that tree iv (island b) was even better, because to get to that tree it would have to cross a plateau of trees that were worse than trees 3 and 4. However, a search starting from tree i would succeed in finding the best tree. If the set of possible trees contains more than one island then heuristic methods may land on a suboptimal island, and the optimal island will not be discovered. After Maddison (1991).

Parsimony

* Widely used method in phylogenetic reconstruction for morphological data

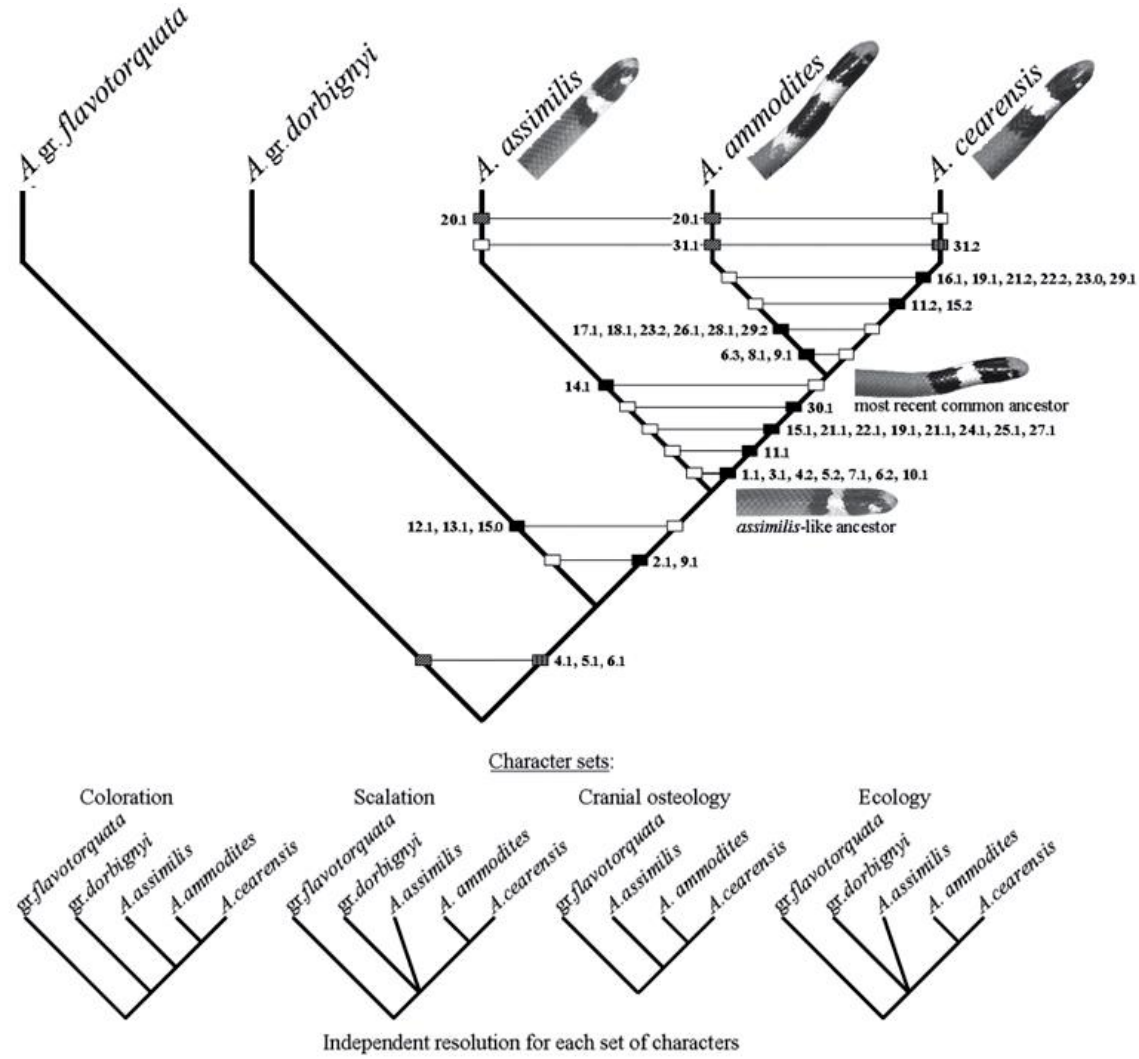
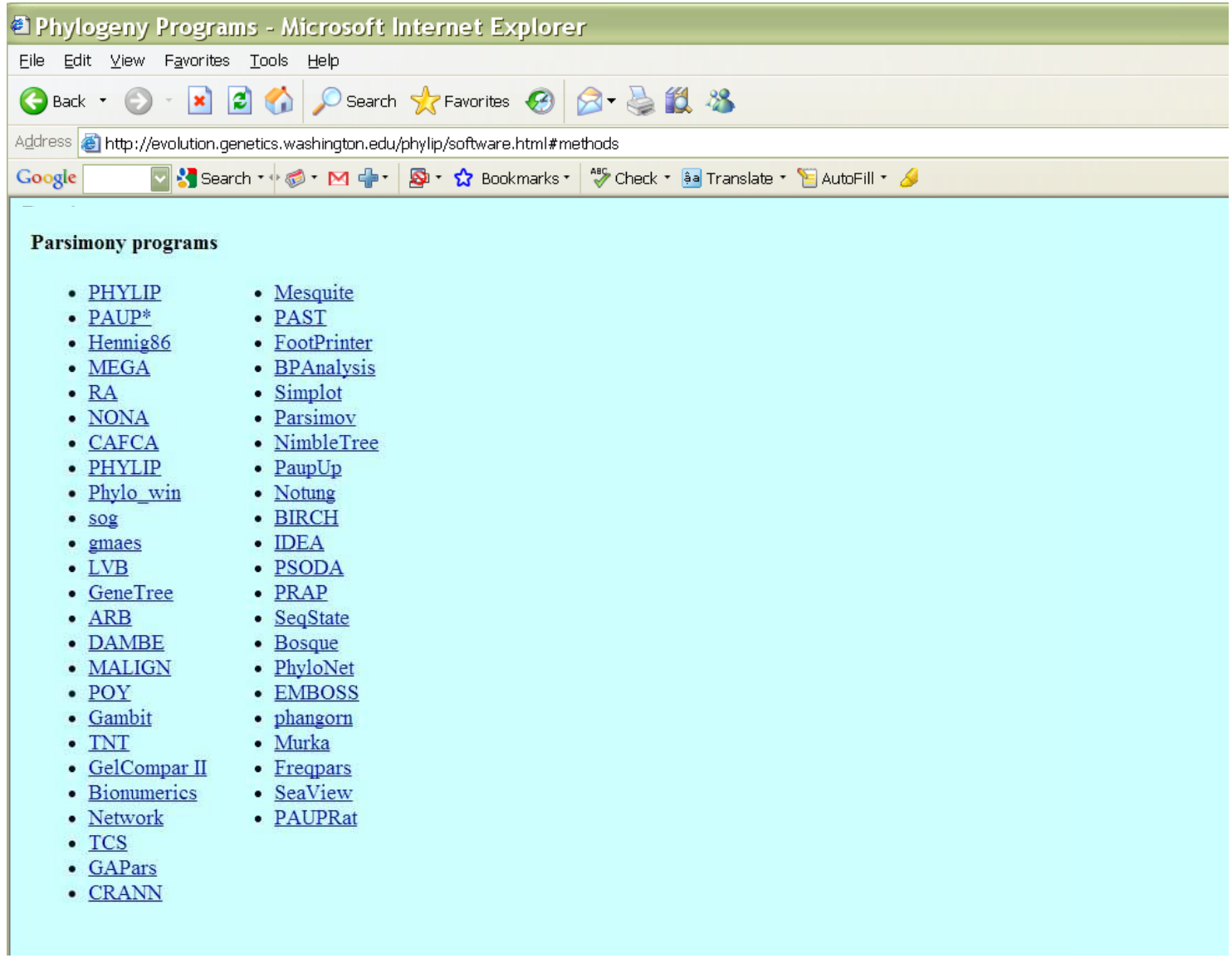


FIGURE 11. Cladogram depicting the phylogenetic relationships among the components of the *assimilis* species group of *Apostolepis*. This is the single most parsimonious tree derived from the cladistic analysis of the data matrix of Table 3 (length = 43; CI = 0.97, RI = 0.97 under both Wagner and Fitch parsimony options applied to ordered multistate characters). The head and neck external morphology was reconstructed for the hypothetical common ancestors, as a corollary of our theory about cladistic relationships and character state polarities. Differently colored numbers represents four sets of characters. The four small cladograms below represents the tree topology derived from independent analyses of each character set.


RP Listen Links - Radi
Daily Weather Foreca
Untitled Message - O
THETA - Dave Work -
Phylogeny Programs
evolution.genetics.washington.edu/phylip/software.html
Owing to other pressures on my time, I cannot devote much time to searching for new programs, so their authors are begged to (please!) use the submission form instead.
Methods
By computer
Cross-referenced
Data types
Web servers
New programs
Submitting
Phylogeny Programs
http://evolution.genetics.washington.edu/phylip/software.html
Changes
Waiting list
Other lists
Old programs
Not listed
News

Parsimony Programs



http://evolution.genetics.washington.edu/phylip/software.html

PAUP*
Sinauer Associates, Inc. Publishers
Sunderland, Massachusetts

By
David Swofford

PAUP* version 4.0 is a major upgrade and new release of the software package for inference of evolutionary trees, for use in Macintosh, Windows, UNIX/VMS, or DOS-based formats. The influence of high-speed computer analysis of molecular, morphological and/or behavioral data to infer phylogenetic relationships has expanded well beyond its central role in evolutionary biology, now encompassing applications in areas as diverse as conservation biology, ecology, and forensic studies. The success of previous versions of PAUP: Phylogenetic Analysis Using Parsimony has made it the most widely used software package for the inference of evolutionary trees. In addition, the PAUP manual has proven to be an essential guide, serving as a comprehensive introduction to phylogenetic analysis for beginning researchers, as well as an important reference for experts in the field. With the inclusion of maximum likelihood and distance methods in PAUP* 4.0, the new version represents a great improvement over its predecessors. In addition, the speed of the branch-and-bound algorithm has been enhanced and a number of new features have been added, from agreement subtrees to tests for combinability of data and permutation tests for nonrandomness of data structure. These, along with many other improvements, will make PAUP* 4.0 an even more indispensable tool in comparative biological analysis than were previous editions of the program and manual. PAUP* 4.0 and MacClade 3 use a common data file format (NEXUS), allowing easy interchange of data between the two programs.

http://paup.csit.fsu.edu/about.html

TNT (Tree Analysis Using New Technology)

Pablo Goloboff, of INSUE - Fundación e Instituto Miguel Lillo 205, 4000 S. M. de Tucumán, Argentina, (pablogolo@csnat.unt.edu.ar) together with J. S. Farris of the, Laboratory of Molecular Systematics of the Naturhistoriska Riksmuseet, Stockholm, Sweden and Kevin Nixon of the L. H. Bailey Hortorium, Cornell University, Ithaca, New York, have produced **TNT** (Tree analysis using New Technology), version of August 2008. This is a parsimony program intended for use on very large data sets. It makes use of the methods for speeding up parsimony searches introduced by Goloboff in the paper: Goloboff, P.A. 1999. Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics* **15:** 415-428, and the highly effective "parsimony ratchet" search strategy introduced by Nixon in the paper: Nixon, K.C. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* **15:** 407-414. It can handle characters with discrete states as well as continuous characters. The program is distributed as Windows, Linux, and both PowerMac and Intel Mac OS X executables. The program and some support files including documentation is available from its web page at http://www.zmuc.dk/public/phylogeny/TNT

It is free, provided you agree to a license with some reasonable limitations.

From: http://evolution.genetics.washington.edu/phylip/software.pars.html#TNT

TNT (Goloboff et al. 2008) [Optional]

* Tree fusing: subgroups exchanged between different groups of trees
* Sectorial searching: analyzes separately different sectors of the tree
* Tree drifting: an extension of branch swapping, that places limits on suboptimal solutions which accelerates speed of algorithm
* TNT combines the above to yield a fast and efficient algorithm (significantly superior to parsimony algorithm in PAUP)

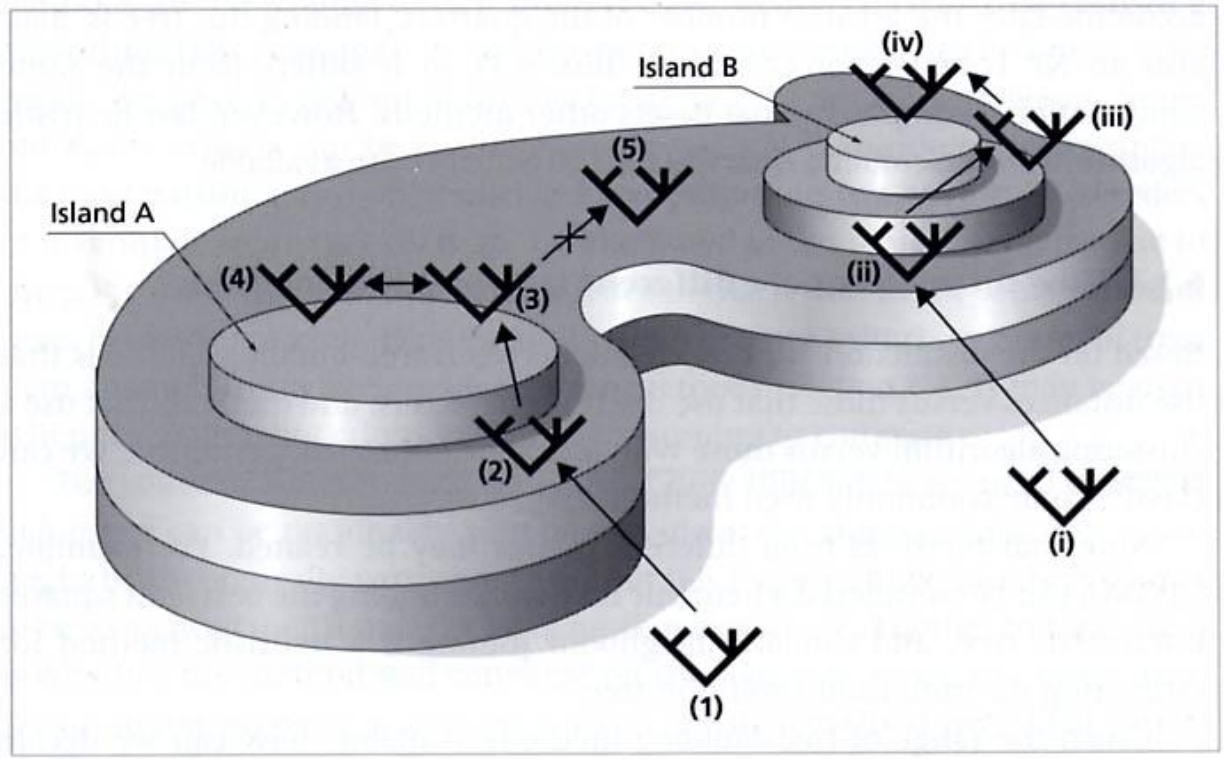


Fig. 6.4 Landscapes and the problem of islands of trees (locally optimal sets of trees). A hill climbing algorithm that started from tree 1 would succeed in finding trees 3 and 4 (which comprise island a), but would fail to discover that tree iv (island b) was even better, because to get to that tree it would have to cross a plateau of trees that were worse than trees 3 and 4. However, a search starting from tree i would succeed in finding the best tree. If the set of possible trees contains more than one island then heuristic methods may land on a suboptimal island, and the optimal island will not be discovered. After Maddison (1991).

The Parsimony Ratchet

(1) Generate a starting tree
(2) Randomly select a subset of characters, each of which is given additional weight (e.g., add 1 to the weight of each selected character).
(3) Perform branch swapping (e.g., "branch-breaking" or TBR[1]) on the current tree using the reweighted matrix, keeping only one (or few) trees.
(4) Set all weights for the characters to the "original" weights (typically, equal weights).
(5) Perform branch swapping (e.g., branch-breaking or TBR[1]) on the current tree (from step 3) holding one (or few) trees.
(6) Return to step 2. Steps 2–6 are considered to be one iteration, and typically, 50–200 or more iterations are performed.

1. TBR = tree bissection and reconnection: cut tree into two subtrees and then reconnecting the subtrees by creating a new branch that joins them

Weighted Parsimony

A priori

* assign weights prior to analysis based on data inside or outside of the data set
  - e.g., transversions get greater weight than transitions
  - e.g., nucleotide bias
* weights may be general, estimated from related taxa, or estimated from the data itself
* six parameter parsimony: nucleotide bias and transition vs transversion

A posteriori

* weights calculated after the analysis is run.
* successive approximations used to select among MPTs
  - downweights homoplastic characters in iterative runs

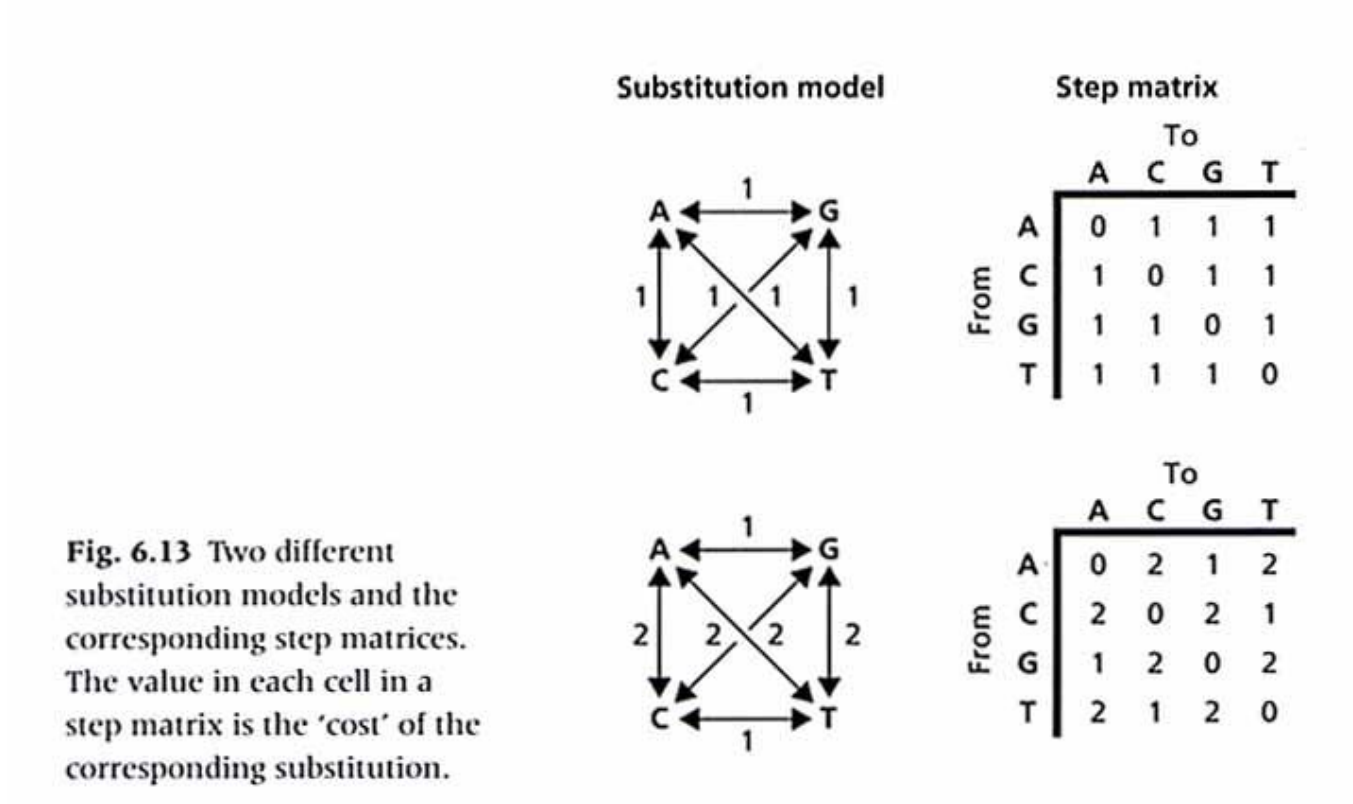


Fig. 6.13 Two different substitution models and the corresponding step matrices. The value in each cell in a step matrix is the 'cost' of the corresponding substitution.

from Page and Holmes. 1998. Molecular Evolution: A Phylogenetic Approach.

Parsimony informative sites

Data reporting from parsimony analyses

* number of characters/nucleotides examined
* number of variable sites/characters (among all taxa)
* number of ***parsimony informative*** sites = numbers of characters shared by two or more (but not all taxa in an analysis)

Why the distinction?

Invariants sites are not useful in a parsimony reconstruction

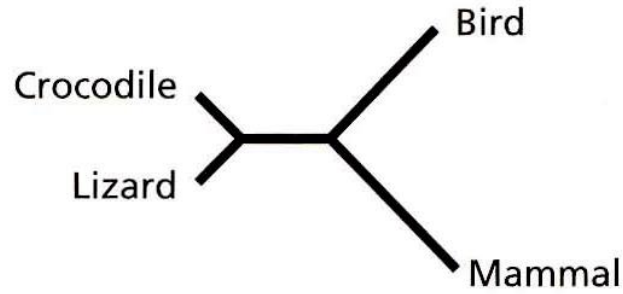
Autapomorphies are not useful in a parsimony reconstruction

- they convey no information about relationships among taxa

- note: they do come into play in likelihood and distance methods as these are part of branch lengths

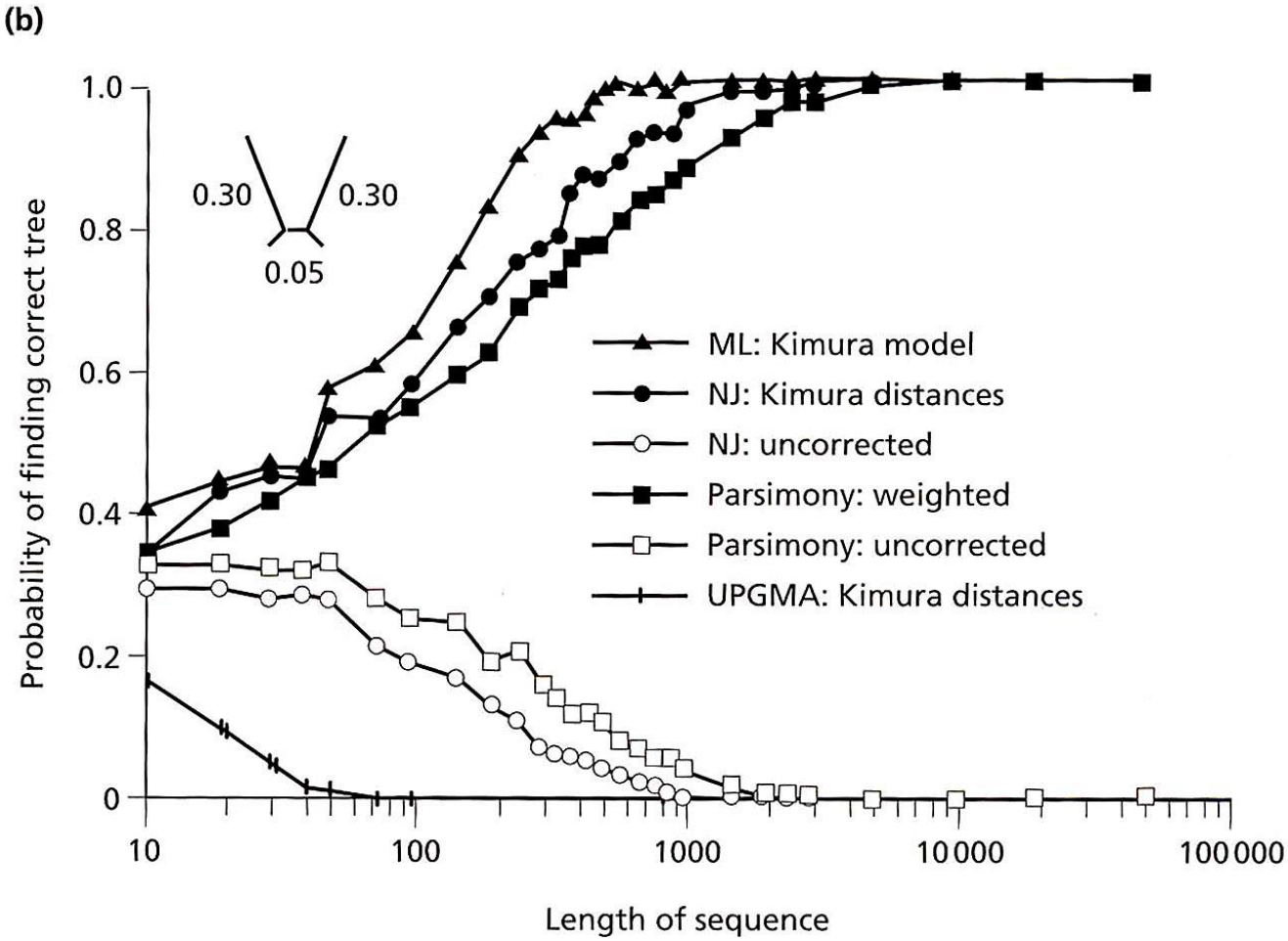
Two of the biggest problems in phylogenetic reconstruction are

1. long branches
2. short internal branches



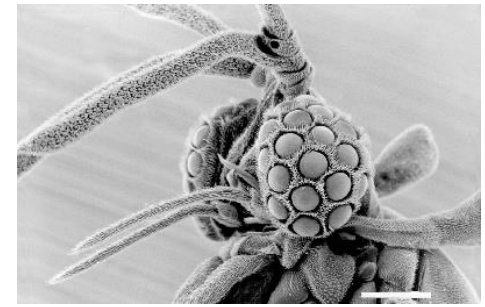
18s RNA sequences

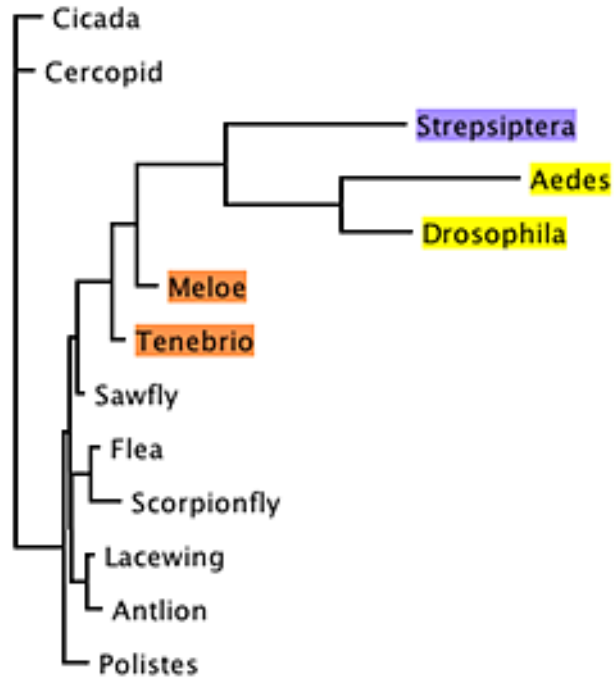
Long branch attraction is a plague for all methods but especially parsimony



from Page and Holmes. 1998. Molecular Evolution: A Phylogenetic Approach.

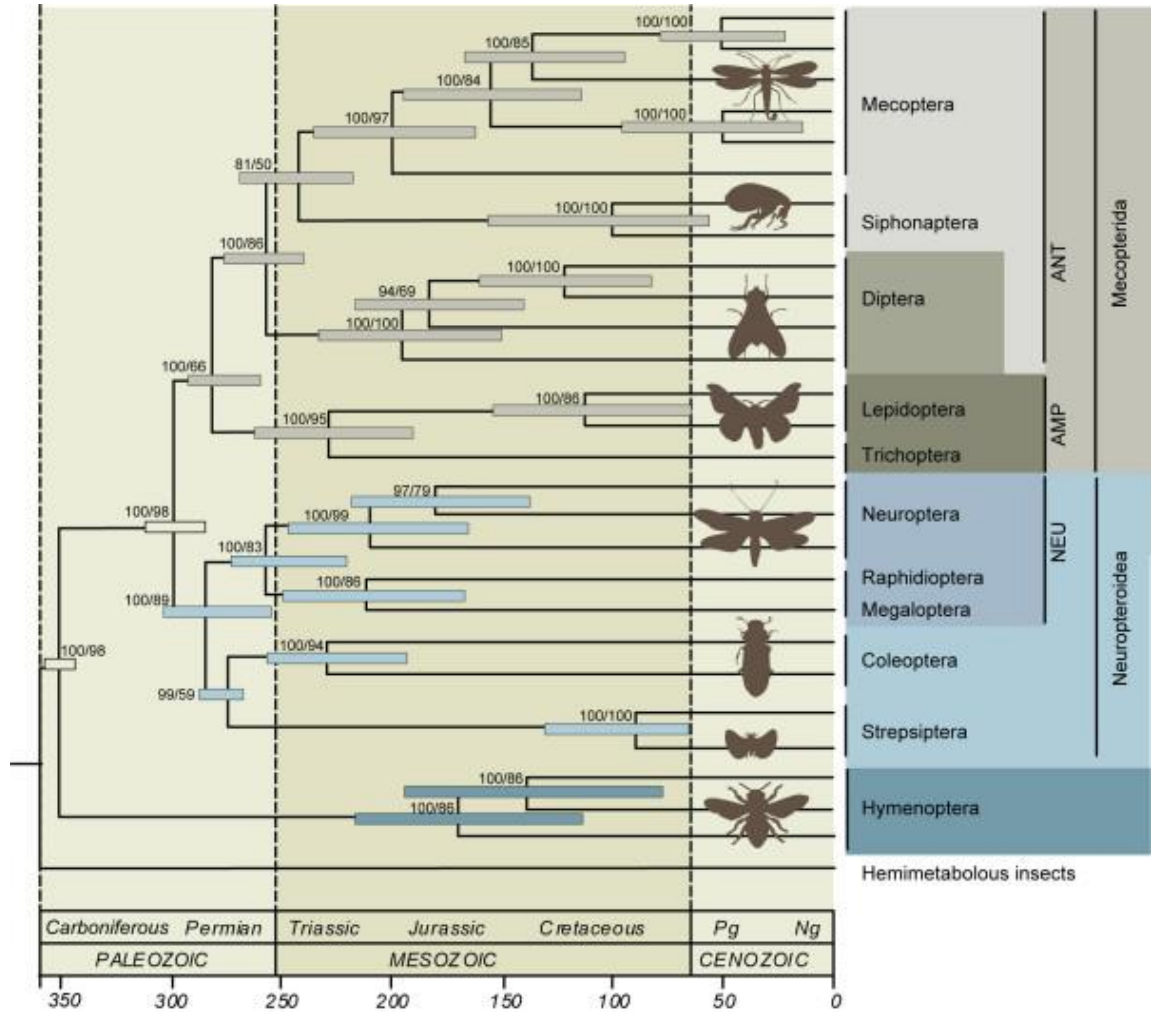
Long Branch Attraction: Strepsiptera





Maddison, D.R. 2004. Are strepsipterans related to flies? Exploring long branch attraction. Study 2 *in* Mesquite: a modular system for evolutionary analysis, version 2.54, http://mesquiteproject.org.

Long Branch Attraction: Strepsiptera



Wiegmann et al. 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC Biol. 7: 34

Dealing with long branch attraction

Things you can do:

1. compare results across various likelihood/Bayesian runs
2. (wisely) add taxa to break long branches
3. delete rapidly evolving characters to shorten branch lengths, if you have reason to believe signal maybe randomized