Where do trees come from?

A slight digression on the "Schools of taxonomy"
Evolutionary school- Typified by textbook written in 1953 by Ernst Mayr, E. Gorton Linsley and Rober L. Usinger (last two of UCB Entomology).
Tree selection- "When there are several possible branching diagrams, the one that is most probable on the basis of all available data, not merely on the basis of parsimony, is preferred." (M&A).
Considered a synthetic method.

Phenetics- "extreme empiricists", Sokal and Sneath (1963) Principles of Numerical Taxonomy.
1. Complete exclusion of evolutionary considerations because phylogeny is unknowable.
2. Evolutionary school's methods were not sufficiently explicit and quantitative
3. Phylogenetic classifications are by their nature special purpose. Goal is to devise a general purpose classification based on overall similarity.
4. Mayr et al. had pushed the field toward population biology and phenetic revived and interest in higher-level groupings.

Proponents saw as beneficial qualities:
1. No previous knowledge of the taxa or related literature needed, only the ability to observe and measure. Work handled by technicians.
2. More characters leads to a more natural classification
3. Inductive, theory-free approach
4. Implementing computer algorithms allowed for new standards of rigor and repeatability

Problems (some listed by Mayr and Ashlock and others):
1. A large number of characters are required
2. A large number of characters are likely to be "noise"
3. Fails claim of objectivity- subjective character choice
4. Many methods exist with no means of selecting.

More on Cladistics in general and parsimony in particular to come…
>>
The Phenetic paradigm failed in general probably because it denied the ability to ask and answer interesting questions, namely evolutionary questions. However, much of the computational methodology remains or has been altered and is widely used for various distance and clustering methods.

Distance Methods: **Character data is transformed into a matrix of pairwise distances for all OTUs.** Generally, a distortion coefficient is calculated for the difference between pairwise *D(ij)* distance matrix and the path lengths *d(ij)* implied from trees.

Distance using optimality: (the initial distance matrix may be, and now usually is, corrected using one of various substitution models setting distance current distance methods apart from classic phenetic methods.)

Least Squares: Selects tree(s) with the smallest squared differences between *D(ij)*and *d(ij)* (more on heuristic searches and tree building next time)

Minimum Evolution (ME): Select the tree(s) with the smallest total branch length that are fit to the data using least squares.

Distance Clustering:

UPGMA: (Unweighted Pair Group Method with Arithmetic mean). Requires a rooted tree and the assumption that changes along branches are clock-like. The distance from the root to any tip is equal. This is an ultrametric tree. Violation of these assumptions results in the wrong topology.

-create a pairwise distance matrix
-select *i* and *j* with the smallest pairwise distance
-create group *(ij)* with distance to the node *p(ij)*/2 and count the number of members (OTUs)
-remove *i* and *j* from matrix and add in group *(ij)*
-calculate new distance matrix
-continue till only one item remains

Neighbor Joining: No clock assumption needed. No need to root. Acts as an approximation of ME. Computationally much faster than ME and so deemed better for larger matrices. However, found to be less accurate as number of taxa increases, so not as good as ME (!?).

-compute distance matrix
-join terminals into a star, this uses a special case of star decomposition (more on tree building and searching later)
-compute net divergence from each terminal to all others in the star tree (u_i)
-join the *i* and *j* that minimize D_{ij}-u_i-u_j
-compute the distance between the new node (*ij*) and remaining terminals
-remove *i* and *j* and replace them with (*ij*)
-if more than 2 nodes remain repeat

NJ Results in one fully resolved tree, always.

General Issues with Distance methods:
-Appropriate when the data are distances or pairwise measures between OUTs.
-For character data the distance matrix replaces the characters so our grasp of character evolution is lost.
-Completely different character matrices can result in the same distance matrix.
-IF the distance matrix is a reflection of the correct phylogeny and fundamental assumptions of a method are met, the result will be correct.
-The reported robustness to model and assumption violations and statistical consistency vary widely. Most are not good.
-Ties for distance scores are arbitrarily broken and this may alter results, i.e. input order matters!
-Systematically breaking ties can lead to false support for groups using bootstrapping, etc.

-Negative branch lengths. What do they mean biologically? Some variation set them to zero (pushing change into adjacent branches) or use absolute value. Still are not sensible in terms of any kind of character transformation process.
-The distances of *D(ij)*and *d(ij)* are treated as independent when they are not. In a tree (A(B(CD))) the path length from B—D shares an edge with B—C.
-Distance does not handle gaps in molecular data well. Usually it is ignored or discarded.
-Fractional changes don't make any sense. How can a node be 0.5 of "G" away from "A"?

"School"	Phylogeny	Rates	Characters	Grouping	Diagram	Taxa det.
Evolutionary	Yes	Yes	Usually discrete, special [selective and weighted]	special similarity	Evolutionary Tree/Phylogram	Amount difference
Phenetic	No or post-tree	Yes/method requirement	All	overall similarity	Phenogram	Amount difference
Cladistic	Yes	No/yes method dependent	Usually discrete, special [+- selective and weighted]	special similarity	Cladogram	Unique shared similarity
Pattern Cladists	post-tree interpretation	No	Usually discrete, special [+- selective and weighted]	special similarity	Cladogram	Unique shared similarity