

# Molecular Evolution and Phylogeny

Daniel F. Simola

13 May 2005

## 1 Molecular Evolution

Molecular evolution is composed of three aspects:

1. Patterns of change (generative processes) in genetic material and its products. Origins of study in molecular biology.
2. Evolutionary history of macromolecules and organisms (molecular phylogenetics). Origin of study of microevolution in population genetics.
3. Origin of life

**Macroevolution** is the study of the evolutionary history of species (above the species level). Thus a phylogeneticist operates on species level homology. Some see this as an extension of microevolution, and others as a completely different process. Today macroevolution most commonly means neo-Darwinian theory plus the neutral theory, although others (Gould) contend that the theory of punctuated equilibria is more reasonable than smoothly graded evolution (Dawkins).

The **neo-Darwinian theory of evolution** is the unification of Darwin's theory of evolution and Mendel's laws of inheritance; the primary concern of evolution is the distribution of a trait in a population (NB, evolution occurs at the population level). Mutation (insertion, deletion, substitution, transposition) is seen as one major source of genetic variation (it changes the trait values directly), but natural selection (positive/directed or negative/purifying) is the "creative" force in shaping the genetic makeup of populations (it changes the trait distribution by interacting with the trait values). Note selection acts directly on the *phenotype*, and only indirectly on the genotype.

This is extended by the **neutral theory of molecular evolution** (King, Jukes, Kimura), which states that the majority of evolutionary changes and much of the variability within species is the result of random drift of selectively neutral alleles (by changing the trait distribution independent of its value). This implies not so much the exact selective neutrality of alleles, rather that stochastic effects dominate selection in the determination of allele frequency. For this to be true, the selective advantage of an allele must be less than  $O(1/N)$ ,  $N$  is the population size. Factors such as mating patterns and migration/geography also indirectly affect trait distributions.

The neutral theory assumes most of the harmful mutations were eliminated initially by purifying selection. This would leave selectively neutral bases unaffected, and thus capable of being mutated, in addition to even fewer bases which could be selected for positively. Thus the rate of evolution of a molecule is proportional to the amount of functional constraint applied to it (ie how tolerant is a protein to change?).

Selectionists favor the idea that natural selection is the overall dominant force in evolution, whereas neutralists contend that neutral mutations and random effects dominate. Both agree that random drift cannot be neglected, and genetic polymorphism and molecular evolution are generated by the same process. The key is determining their relative contributions.

**Microevolution** is the study of the evolutionary history of macromolecules (at or below the species level (but always of a population)). Typically this involves changes in gene frequencies in a population over time (which is specified in terms of generations). Several processes are responsible for such change, including mutation, gene flow, genetic drift, and natural selection. Microevolution is directly observable (we do it in lab all the time).

**Gene flow:** the movement of genes from one population to another. Often due to dynamics of population migration and geographic barriers.

**Genetic drift (random):** aka DIFFUSION. A stochastic effect, acting on populations, arising from the role of random sampling in the production of offspring. Random effects on gene frequency are most easily seen when population size is small, since the number of possible combinations of genes generated is limited (sampling). Drift and diffusion are always coupled/dependent on each other.

**Natural selection:** aka DRIFT. A process by which the frequency of particular traits increase (and decrease) over time, due to natural events such as the food chain, geographic and climatic changes, etc. Drift is an active process which drives trait frequency in populations.

## 1.1 Terminology

**Genetics:** information decoding

**Development:** information decoding process

NB, an organism is realized by a generative decoding process.

**Homology:** Some characteristic of a group of organisms which originated once in history in a unique ancestor of the group (monophyletic lineage).

**Homoplasy:** False homology due to convergent similarity. eg, fins of fish and whales, or wings of bats and birds.

NB, homology is the basis of phylogenetic classification, as the only unique tool for determining true causal relationships (read: descent with modification). Thus phylogenetic estimation operates on monophyletic lineages.

**Sequence homology:** Two sequences are homologous if they exist as a result of some bifurcating and mutational process, beginning with a single sequence.

**Fitness:** A measure of the selective advantage of a genotype. This is typically used as a relative value, where one compares the relative fitnesses of alleles in a genome or across genomes (organisms). Fitness also has associated direction:

**Positive selection:** Favoring of a trait which has greater relative fitness than another version of the trait.

**Negative selection:** aka purifying selection. Elimination of a trait which has lower relative fitness than another version of the trait.

**Neutrality:** Modification of a trait with no change in relative fitness. The maintenance of this trait over time is subject to chance (stochastic) effects.

**Overdominance:** aka heterozygote advantage. A situation occurring when the fitness of a heterozygote of a trait is greater than the fitness of either homozygote.

## 1.2 Generation and alteration of genes

A genome's population of genes can change in content and increase or decrease in number through the following events:

- Intragenic mutation: Point mutations (transition/transversion), insertions, and deletions within the coding region of a gene
- Gene duplication
- Segment shuffling: Two or more genes can be broken and hybridized, or transposons can insert into an existing gene to add, substitute, or delete sequence. In the case of multiply exonic genes, exons can be shuffled and exchanged with other exons through the above mechanisms.

- Horizontal transfer: The introduction of genetic material from one organism to another (within or between species). Cf vertical transfer, the inheritance of genetic material from parent to child cell/organism.

### 1.3 Characteristics of molecular evolution

**Genome size is extremely variable** The vast majority of DNA of most organisms is noncoding; in fact only about 5% is typically coding. Since mutations strike randomly (or at least anywhere) along a genome, molecular evolution at the level of the genome must be largely non-adaptive.

**Molecular evolution is sometimes decoupled from morphological evolution** When comparing the morphological and genetic similarity between two taxa, there are generally four possible situations:

	Morphological	Genetic	Comment
1)	low	low	This outcome is expected.
2)	high	high	This outcome is expected.
3)	high	low	This crops up occasionally for morphologically similar groups, such as frogs and salamanders, suggesting genetic selection becomes highly directed/positive, whereas morphological evolution halts.
4)	low	high	This outcome is seen especially between chimps and humans, where our genomes are over 99% identical, but are phenotypically strikingly different. This indicates rapid morphological changes can occur over short time periods, and with only a trivial amount of genetic mutation. This suggests perhaps heterochronic changes at the gene expression level can be altered to produce dramatic phenotypic changes.

**Many genes appear to evolve at a roughly uniform rate over evolutionary time** ie evolution of a particular gene follows the same “molecular clock” for all species. To see this, you can grab a protein sequence from a bunch of species, estimate sequence divergence, and plot divergence vs time. You typically see a linear relationship. The “clock” ticks at different rates for different proteins. Many think this is the result of neutral mutations.

**Silent mutation** The ratio of silent to replacement DNA mutation rates is typically 5 to 10 > 1.

The above evidence suggests the role of natural selection may be less important than that of genetic drift and hence, the neutral theory.

## 2 Phylogeny

While Molecular evolution describes the bifurcating process of population-based change over time, phylogenetics seeks to reconstruct this evolutionary history of terrestrial life, since we in no way can observe this historical process. Thus phylogenetics is really a tool used to classify the set of organisms on earth, assuming that organismal diversity is the product of evolution, beginning with a single life form.

### 2.1 The need for phylogeny

1. To distinguish between hypotheses regarding evolutionary history. ie to identify the tree of life. This is critical, since homology is the basis of classification.
2. To distinguish between homology and homoplasy (see above)
3. To determine the direction of the evolution. Every ontogeny is a product of selection and chance (see above), so to understand mechanism, one must identify how it came to be.

Table 1: Differences between Evolutionary Biologists and Developmental Biologists in Their Views of Major Biological Qualities. Borrowed from “The Shape of Life”, Raff.

Quality	Evolutionary biologists	Developmental biologists
Causality	Selection	Proximate mechanisms
Genes	Source of variation	Directors of function
Variation	Central role of diversity and change	Importance of universality and constancy
History	Phylogeny	Cell lineage
Time scale	$10^1 - 10^9$ years	$10^{-1} - 10^{-9}$ years

## 2.2 Terminology

**Taxonomy:** (taxis + nomos = name/law of order) the science of biological classification, with the basic unit the species. It consists of three parts: identification, nomenclature, classification.

- Identification is the assignment of an organism to a taxon
- Classification is the arrangement of organisms into taxa

A ‘taxon’ is a phylogenetic grouping of organisms, of which there are three kinds:

1. **Monophyletic** taxon is a group of organisms which arose from a single ancestor and includes **all** descendents of that ancestor
2. **Polyphyletic** taxon is one whose members all derive from a common ancestor, where at least one common ancestor is not in the same group. This is typically due to homoplasy.
3. **Paraphyletic** taxon is one in which all members share a common ancestor but at least one known descendent is excluded. ie this taxon is monophyletic after including the restricted subset. eg the monophyletic group (reptiles, birds, mammals) has a paraphyletic taxon (reptiles, birds).

**Systematics:** the study of organismal diversity and their evolutionary relationships. NB systematics does not say anything about evolutionary *processes*. Often interchangeable with taxonomy, but technically slightly broader in scope, in that systematics encompasses any study on the nature of organisms to be used in taxonomy. This thus includes most modern (and some extinct) biological fields: morphology, ecology, epidemiology, biochemistry, molecular biology, physiology, etc. To the point, there are two ways to classify: by phenotype (phenetic) and by phylogeny (phyletic). Phenetic classification considers gestalt-type features of a group, and can thus lead to an incorrect grouping as a result of homoplasy.

Thus there are two schools of systematics:

**Cladistics:** dominant school of systematics, which traces evolutionary history by constructing a network of clades. A clade is a monophyletic lineage (basically an evolutionary module). Clades indicate true descent with modification. The assignment of a clade comes from identification of homologies among organisms.

**Evolutionary systematics:** grouping is based on all shared phenetic features, and this approach is typically less rigorous and more intuitive.

**Phylogeny:** (phulon + genesis = birth of a race) is the evolutionary development and diversification of a species or group of organisms, or of a particular feature of an organism. Cf ontogenesis. ‘phyla’ means ‘races’. NB a phylogeny then describes a causal relationship among taxa using shared processes (ie homologies). Clade is synonymous with phylogeny.

**Ontogeny:** (ont + genesis = birth of a being) is the development of an individual organism or anatomical or behavioral feature from the earliest stage to maturity.

**Homology:** Some feature of a clade, necessarily based on phylogenetic (phyletic) evidence. NB homologies do not indicate the direction of evolution. There are three kinds of homology:

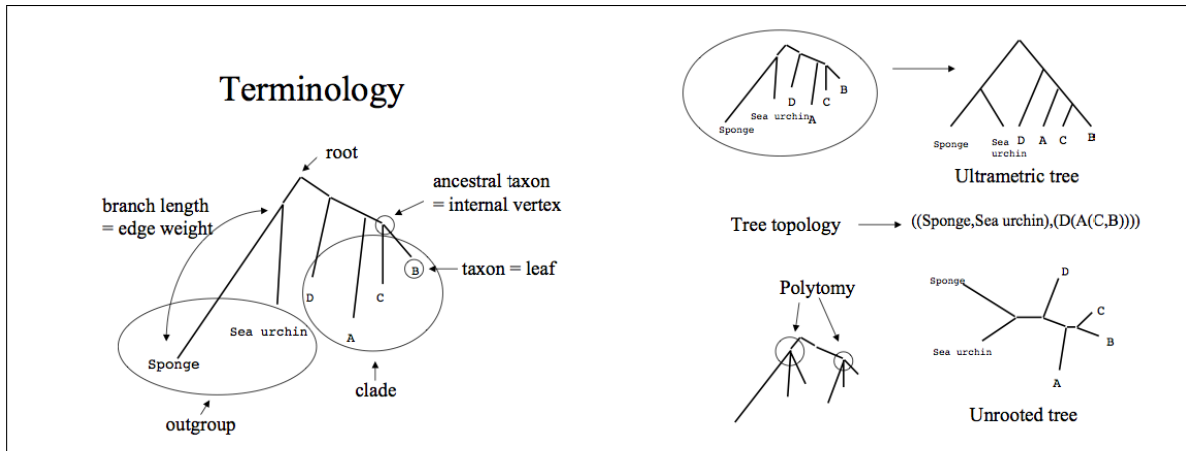


Figure 1: Examples of phylogenetic trees

1. Primitive features shared by all members (symplesiomorphies): such features are not informative about the relationships among members within a clade.
2. Features shared by a single lineage within a clade (automorphies): used to distinguish between lineages, but can't infer branching patterns.
3. Derived features shared by some lineages within a clade (synapomorphies): the informative type of homology, used to infer the hierarchy of relationships among lineages.

We often want to distinguish between homology resulting from a speciation event (orthology) and homology from gene duplication (paralogy).

### 2.2.1 Taxonomic hierarchy

Domain → Kingdom → Phylum/Division → Class → Order → Family → Genus → Species

NB Phylum is a major subgroup of the animal kingdom, whereas Division is a major subgroup of the plant kingdom.

Favorite mnemonic: **Do Keep Privates Clean Or Forget Getting Sex**

### 2.3 Phylogenetic tools

Data structures are needed to present the characters and historical relationships within a taxon and among taxa. One structure includes no prior assumptions about the data, and the other may introduce such assumptions.

**Cladogram:** aims to describe the branching pattern between a group of organisms, but not evolutionary direction or ancestor/descendent relationships. The inferred knowledge is the splitting pattern of character states. A cladogram may or may not incorporate ancestral/extinct organisms. Cladograms are constructed solely from inferences. Technically cladograms are undirected and unrooted trees (networks).

**Phylogenetic tree:** a graph structure representing the evolutionary history of a set of taxa. Such trees are directed (indicating direction of evolution) and often rooted (ancestral order) and weighted (indicating length of time between speciation events). These trees are often constructed using prior inferences, hypotheses, or assumptions, since there are many possible phylogenetic trees for a given cladogram.

**Leaf (taxon):** A degree one vertex representing present day object (group, organism, gene, etc)

**Internal vertex (ancestral taxon):** A vertex with degree > 1 representing some ancestral object

**Binary tree:** A tree graph where all non-leaf vertices have degree 3, corresponding to the idea that genetic lineages bifurcate. One special vertex called the root, may have a degree 2 vertex.

**Leaf-labeled tree graph:** A tree graph with unique identifiers associated with each leaf vertex (degree one vertex).

**Tree topology:** The branching structure of a leaf-labeled tree graph ignoring other information such as edge weights

**Rooted tree:** A tree graph with a special degree two vertex from which directed edges are assigned representing time flow.

**Unrooted tree:** A tree graph with no temporal directionality.

**Edge weight (branch length):** A numerical value associated with the edges of the tree, often denoting elapsed time or expected number of mutations under some stochastic model.

**Path length:** For two vertices in a tree graph the sum of edge weights along the unique edge-path that connects the vertices.

**Ultrametric tree:** A tree graph such that the path length from all of the leaves of the tree to the root vertex are equal

**Polytomy ('many cuts'):** Internal vertices with greater than degree 3, representing simultaneous split of more than two lineages (or inability to resolve the split).

**Most common ancestor:** Given a collection of leaves, L, the internal vertex that is the closest to the leaves and is ancestral to all L.

**Clade:** For each ancestral vertex, the collection of all leaves subtending from the vertex. Each leaf will be members of a hierarchically nested series of clades.

**Outgroup:** A set of taxa (leaves) that are known to be "outside" the common ancestor of the set of taxa of interest (in-group). Often used to estimate the position of the root of the tree. You want to choose an outgroup that is outside the clade, but most similar to the common ancestor in that homologies exist in a primitive form.

**Leaf Coloring (character):** The assignment of some measured state to the leaves of the tree. For example, an aligned position of a sequence assigns nucleotide identity to all the leaves of the tree, constituting a "coloring" of the leaves or a "character" in biological terminology

## 2.4 Tree estimation

Inferring a phylogeny is an estimation procedure, where the topology (edges, weights) and the ancestral states of a phylogenetic tree are estimated, given a set of taxa and associated characters. One can either establish an algorithm for phylogeny generation, or define an optimality criterion in the form of an objective function used to compare alternative phylogenies and select the best candidate. Algorithmic methods implicitly define optimality criteria and are typically faster (poly time vs NP time) as they skirt the need to evaluate numerous competing tree topologies.

A basic procedure for phylogenetic inference is the following:

1. Collect character data for each taxa
2. Align data using multiple sequence alignment
3. Estimate a tree either using a model of evolution or by transforming the data into a distance matrix
4. Assess significance using bootstrap or MCMC
5. Return single best or consensus tree
6. Subsequent hypothesis testing

### 2.4.1 Enumerating trees

	num leaves	num binary trees
	3	1
	4	$1 * 3 = 3$
A tree with $t$ leaves has $2t-3$ edges, yielding the series:	5	$1 * 3 * 5 = 15$
	6	$1 * 3 * 5 * 7 = 105$
	$\vdots$	$\vdots$
	$t$	$1 * 3 * 5 * 7 * \dots * (2t - 5) = 2^{O(t \log(t))}$

More exact, there are  $(2t - 3)!!$  rooted binary trees with  $t$  leaves. Tree enumeration is NP-hard, thus any algorithm which requires tree estimation is NP-hard.

### 2.4.2 Data

Two categories of data: discrete characters and similarity or distance measurements. Character data are often transformed into the latter. A character is a variable which can take on a finite number of mutually exclusive states. It is typically assumed characters are independent and homologous. Character states may have known relationships, and so one may restrict possible character transformations (eg  $A \rightarrow \{T\}$ , but not  $A \rightarrow \{G, C\}$ ).

**Sequence:** nucleotide, amino acid, binary. Typically requires positional homology (from an alignment)

**Restriction endonuclease:** binary states, indicating presence/absence of endonuclease site

**Gene order:** Order along chromosomes, etc. Useful for very distance phylogenies, since gene rearrangements occur less frequently than nucleotide changes, making homoplasy less likely. A serious difficulty is the independence assumption, since it is very unlikely that characters evolve independently in this fashion.

### 2.4.3 Optimality and algorithms

Formally an optimality criterion is written as an objective function which measures the fit of a data set to a binary tree (you provide the topology). The configuration space is the space of possible binary tree topologies for your data set. The combination of objective function and configuration space defines a combinatorial optimization problem.

There are three major estimation approaches:

1. Evolutionary distance: neighbor-joining/additive, clustering/ultrametric (algorithmic optimality)
2. Hierarchical pattern of character evolution: maximum parsimony, maximum compatibility (algorithmic optimality)
3. Stochastic evolutionary models: maximum likelihood, bayesian estimator (criterion based optimality)

## 2.5 Distance-based methods

The main idea is to transform available data into pairwise distances. Unless distances are measured directly, this results in a loss of information. Although distance methods typically perform worse than maximum likelihood methods (despite consistency), they are especially useful for large data sets, as they run in polynomial time since trees do not need to be enumerated. There are two conditions for distance methods: additive distances and ultrametric distances.

### 2.5.1 Transformations

Distance measurements may come from direct measurements, eg DNA hybridization strengths or protein affinity, but are usually determined from sequence alignment methods. We want distance to be additive over

evolutionary time. Unfortunately this is difficult as additive measurements underestimate evolutionary time because of unseen multiple mutant sites.

Hamming distance (uncorrected distance) is good if there is a constant and identical mutation rate over all positions in a sequence; these assumptions are not typically met, warranting a correction method using a nucleotide substitution model (divergence matrix  $F$ ). One solution is to take LogDet:  $-\log(\det(F))$ , a model-free correction. The transition matrices used include the common models: JC (1 parameter), Kimura (2 param), HKY, etc. Rate heterogeneity can be accounted for using a gamma distribution for each position in the sequence.

### 2.5.2 Additive distances and neighbor-joining

True evolutionary distances are additive. The distance between a pair of taxa is equal to the sum of the lengths of all pairs along this path. An additive distance matrix is defined as one which satisfies Buneman's four-point metric criterion for any four taxa A, B, C, D:

$$d_{AB} + d_{CD} \leq \max(d_{AC} + d_{BD}, d_{AD} + d_{BC})$$

Having an additive distance matrix  $\iff$  an edge-weighted tree. However not all distance matrices are additive. To create an additive matrix then, we simply find the additive matrix which is in some sense closest to the original data matrix. We can then use the Neighbor-joining algorithm to generate an edge-weighted, unrooted tree in  $O(L^2)$  time:

1. Initialize with blank edge-weighted tree  $T$
2. Until two taxa remain in the list  $L$ :
  - (a) Choose the nearest pair of nodes:  $\min\{D_{ij}\}$
  - (b) Define a new node  $k$  whose distance to all other nodes  $m$  is defined by additivity:  $d_{km} = 1/2(d_{im} + d_{jm} - d_{ij})$
  - (c) Add  $(i, j)$  to tree, and connect  $k$  with  $d_{ik}, d_{jk}$  defined by additivity similar to the above step
  - (d) Add  $k$  to  $L$ , and remove  $i$  and  $j$

**Distance metrics** The distance between two  $L \times L$  matrices takes the following form:

$$d(A, D) = \sum_{j=1}^{L-1} \sum_{i=j+1}^L w_{ij} (a_{ij} - d_{ij})^\alpha,$$

$w_{ij} = 1, \alpha = 2$ : ordinary least-squares,  $L^2$  metric

$w_{ij} = 1/d_{ij}, \alpha = 2$ : weighted least-squares

$w_{ij} = 1/d_{ij}^2, \alpha = 2$ : Fitch-Margoliash (1967)

$w_{ij} = 1/\text{var}(d_{ij}), \alpha = 2$ : error weighted least-squares

$\alpha = \infty$ : min-max distance, also called  $L^\infty$  metric

NB can solve using least-squares linear equations, but typically want to bound edge-weights to be non-negative, so quadratic-programming or iterative numerical methods are used.

### 2.5.3 Ultrametric distances and UPGMA clustering

UPGMA means unweighted pair group method using arithmetic averages. The basic idea is to iterate over pairs of sequences, joining the pair at each step and replacing it with a single node. The distance metric is

$$d_{ij} = \frac{1}{|C_i| + |C_j|} \sum_{p \text{ in } C_i, q \text{ in } C_j} d_{pq}$$

Algorithm:

1. Initialize by making each sequence a leaf, and its own cluster  $C_i$
2. Iterate until two cluster remain:
  - (a) Take the  $(C_i, C_j)$  with minimal distance and define a new cluster  $C_k = C_i \cup C_j$ , and define  $d_k$  by 2.5.2
  - (b) Define new node  $k$  in tree, at height  $d_{ij}/2$
  - (c) remove  $C_i$  and  $C_j$ , replacing with  $C_k$

The distances created by UPGMA exhibit the ultrametric distance property:

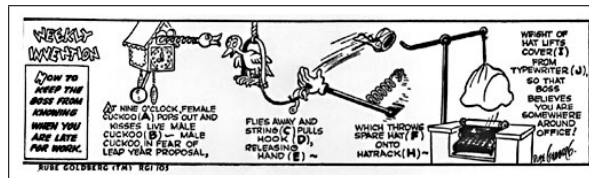
$$d_{AC} \leq \max(d_{AD}, d_{BC})$$

, ie the distances follow a “molecular clock” and are produced as if evolution were occurring at a constant rate. In other words, the sum of the lengths along any path from root to leaf is the same. In other words, the expected number of changes for a site is proportional to time. Having an ultrametric distance matrix  $\iff$  correct edge-weighted tree construction  $\iff$  molecular clock. The ultrametric property is a tighter constraint on distance matrices than the additive property, in that ultrametricity  $\implies$  additivity.

## 2.6 Hierarchical optimality criteria

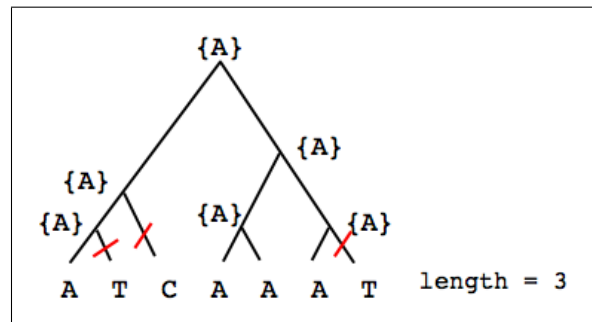
Hierarchical character states follow the basic principle of parsimony: the simplest model is the most reasonable. There are two underlying assumptions:

1. Evolution operates parsimoniously: based on our observations of the natural world (the fact that we look for and can find homology) indicates this is a reasonable principle
2. Systematically, it is convenient and intuitive to work with the simplest possible model (to avoid those silly Rube Goldberg machines).



### 2.6.1 Parsimony

A parsimony assignment of ancestral character states is a state assignment of the internal vertices such that the sum of the edge weights is minimized. This sum is called the parsimony length of the character or the ML length. Parsimony assumes independent characters (sites). Parsimony is disadvantageous because tree estimation is based on minimizing a single score, and many possible tree reconstructions that can yield an equivalent score. In this way parsimony is not always consistent. eg, parsimony does not take into account possible convergence along long branches, hence trees with two branches longer than the rest will be treated as most similar accounting for length alone. The parsimony criterion, which minimizes the sum of state changes, is similar to the minimum evolution criterion, which aims to minimize the sum of branch lengths on a tree.



MP length for this tree is 3.

Fitch derived the basic unweighted parsimony algorithm, which permits free state transformations (state transformation is unordered). Thus the parsimony length is the sum of the state transformations over all characters.

Weighted parsimony places a cost on state changes, and so the parsimony score is the sum of costs over characters. Wagner parsimony adds constraints on character-state transformations using a symmetric and constant cost matrix (the state transformation is ordered):

$d(x, y)$	A	C	G	T
A	0	a	a	a
C	a	0	a	a
G	a	a	0	a
T	a	a	a	0

Weighted parsimony reduces to basic Fitch parsimony when  $d(x, x) = 0$  and  $d(x, y) = 1, \forall x \neq y$  (ie when  $a = 1$ ).

Maximum parsimony returns the set of trees whose sum of MP lengths over all states is minimum:

$$\{T\}_{MP} = \underset{T_j}{\operatorname{argmin}} \left\{ \sum_{c_i \in C} mp(c_i, T_j) \right\}$$

The Fitch-Hartigan algorithm for weighted parsimony can determine the minimum length under the parsimony criteria with a single post-order traversal of the tree (left subtree, right subtree, root, etc), ie bottom-up fashion:

- For all vertices  $v$ :

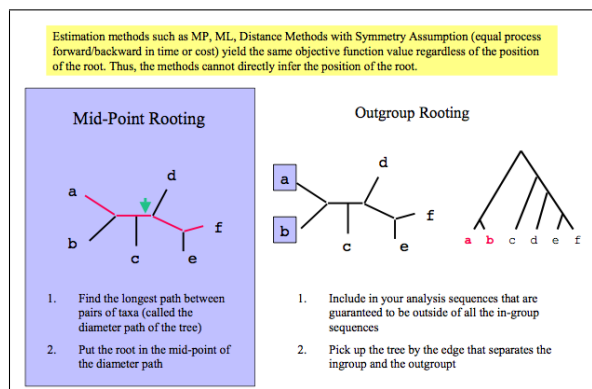
$$S(v) = \begin{cases} L(v) \cap R(v) & \text{if } L(v) \cap R(v) \neq \emptyset \\ L(v) \cup R(v) & \text{if } L(v) \cap R(v) = \emptyset \end{cases}$$

This doesn't actually return the most parsimonious tree reconstruction (MPR), since the ancestral states may not be unique. To do this, a pre-order (top-down) traversal is necessary (root, left subtree, right subtree, etc) to enumerate all ancestral state assignments. The many possible MPRs can be reduced by assigning further restrictions to the choice of substitution states.

Start at the root:

1. if the root state set has more than one state, choose one arbitrarily
2. if the daughter vertex state set shares the parent state, assign this shared state
3. else choose any state among the possible states
4. repeat for all daughter vertices, iterating in prefix order

You may choose a root node arbitrarily, because for symmetric cost matrices, every possible rooting returns the same MP length (and thus the root cannot be determined):



Also, certain characters have the same MP length for all possible trees:

- invariant characters
- characters with only a single state having more than one leaf assignment

There is a proof of correctness for the parsimony algorithm, but I'm assuming we don't need to know it.

## 2.6.2 Other parsimony methods

**Dollo parsimony:** Wagner and Fitch parsimony methods assume a symmetric character state transition matrix. Dollo parsimony relaxes this assumption, and requires binary, or linearly ordered states (eg  $A \rightarrow T \rightarrow C \rightarrow G$  are the possible transitions). The basic assumption is that character states must be uniquely derived in the tree (a state can only appear once in the tree), and so the only possible type of homoplasy (false derivation) is a reversal of state. Dollo parsimony only requires knowing the polarity of characters over each branch, and does not require use of rooted trees. Instead one may use outgroup taxa instead of knowing ancestral sequence. This type of parsimony is particularly suited for estimating the history of restriction-site characters.

**Amino acid parsimony:** You can also run parsimony on amino acid sequences, but this is trickier since the sequences are degenerate and parsimony tries to minimize these changes, thus tending to ignore the genetic code. The idea is to correct for these "silent" substitutions.

## 2.6.3 Maximum compatibility

For a given character  $C_i$  in a tree  $T$ , let  $n_i$  be the number of distinct values at the leaves for  $C_i$ . Regardless of the labeling of internal nodes at site  $i$ ,  $T$  must contain at least  $n_i - 1$  edges over which state changes from leaf to root. If  $C_i$  changes  $n_i - 1$  times on this path, this character is compatible for the tree. Let  $nc(T)$  be the number of compatible states for  $T$ . A tree  $T$  is the maximum compatibility tree if  $T = \underset{T_j}{\operatorname{argmax}}\{nc(T_j)\}$ .

## 2.7 Stochastic optimality criteria: Maximum likelihood

Notably the optimality criterion for stochastic models of evolution is maximum likelihood:

$$P(M | D) = \frac{P(M, D)}{P(D)} = \frac{P(D | M)P(M)}{P(D)} \sim P(D | M)$$

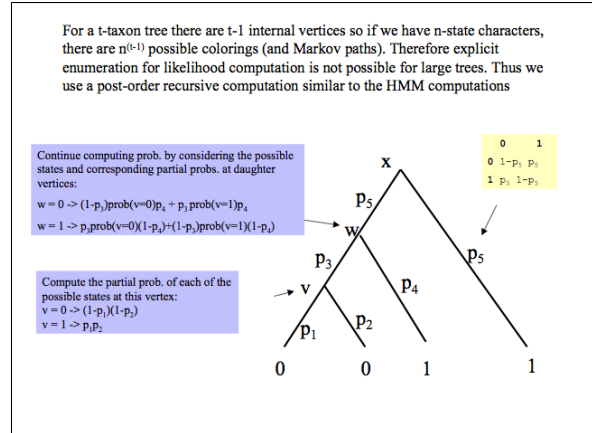
$P(D | M)$  is the pdf/pmf of a sample of data  $D$ . But since the data is observed, we tend to write this as  $P(M | D)$  (technically the likelihood function), and optimize this function of  $M$ . However  $M$  is a set of parameters, not random variables, so we define  $L(M | D) = f(D | M)$  as the likelihood function. The likelihood formula is proportional to the joint distribution, so maximizing the joint is equivalent to finding the highest point on the parameter landscape. A Bayesian approach would be to evaluate the posterior, which is proportional to the likelihood times the prior:  $P(D | M)P(M)$ . Bayesian optimality is equivalent to maximizing the probability volume under the marginal distribution curve of the parameter landscape.

The model is specified by a

1. tree topology
2. initial/stationary state distribution
3. transition matrix for each branch

The transition matrix used is a substitution matrix of nucleotide evolution, defined as a Markov chain. This includes the Jukes-Cantor, Kimura, etc. models (see section on stochastic processes). Just as in parsimony, proper estimation of branch length is critical. There are two kinds of substitution matrices, instantaneous rate matrices ( $Q = R\Pi$ ) and time dependent substitution matrices ( $S$ ). See below.

Typically you need to estimate both the branch length and the time-scale over which these lengths make sense. However one has the option of assuming a molecular clock hypothesis for ML. This is done by setting all branch lengths to 1, and estimating only the time passed from root to leaf.



Basic ML algorithm:

1. Determine optimal topology (search through tree space efficiently):
  - (a) Determine ML value for a given tree:
    - i. Estimate transition matrix parameters using numerical/stochastic optimization
    - ii. Compute likelihood using these estimates
    - iii. compare current tree to previous, using  $\chi^2$  or likelihood ratio statistics
    - iv. Repeat
2. compare current tree to previous, using  $\chi^2$  or likelihood ratio statistics
3. Repeat

The likelihood of a root node at site/index  $i$  is the following:

$$L_i = \sum_{\text{states } s} \pi_s \prod_{\text{children } c} \{(1-p)L_i(c=s) + (p)L_i(c \neq s)\},$$

$L_i(c=s)$  is the likelihood that a child of the root has the same state ( $c=s$ ) or a different state ( $c \neq s$ ). This assumes binary characters (1/0) instead of (A,T,C,G), in which case there are four terms in the sum there, and the probabilities are taken from the nucleotide substitution matrix. Also note the recursive call for nodes that are not the root do not use  $\pi_s$  the initial probability of state  $s$ .

Likelihood is appealing because all possible mutational pathways are considered, implicit in the optimization procedure of ML estimation.

### 2.7.1 Probabilistic models of evolution

Markov models have a general structure, but often we want to construct a model that describes stochastically how a nucleotide  $x$  changes over the course of evolutionary time. In order to calculate the probability of observing  $x$ , given its previous state  $y$ , having changed over a time span  $t$ , ( $P(x | y, t)$ ), we need a model of evolution. The interpretation of  $t$  can change. It may mean literal time, but often is used as a value proportional to the mutation rate x evolutionary time. The models described here work on ungapped character state transformations (the alphabet is A, T, C, G, no gap character).

You can describe such a substitution matrix  $S$  in terms of a rate matrix  $R$  and some amount of time  $t$ :

$$S(t) \simeq (I + Rt) \tag{1}$$

The rows of a rate matrix must sum to 0. Another way of writing this is  $S(t) = e^{tR}$ .

**Jukes-Cantor:** In this way the JC rate matrix  $R$  is defined to be:

$$\begin{matrix} & A & C & G & T \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \end{matrix}$$

and we determine the substitution matrix  $S$  from this using equation 1:

$$\begin{matrix} & A & C & G & T \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} 1-3\alpha t & \alpha t & \alpha t & \alpha t \\ \alpha t & 1-3\alpha t & \alpha t & \alpha t \\ \alpha t & \alpha t & 1-3\alpha t & \alpha t \\ \alpha t & \alpha t & \alpha t & 1-3\alpha t \end{pmatrix} \end{matrix}$$

If you set  $t = 1$ , then you assume the molecular clock is ticking.

**Kimura:** One limitation of the JC model is that it assumes transitions and transversions occur with equal frequency (fully symmetric matrix). The Kimura model uses an additional parameter to handle this situation:

$$\begin{matrix} & A & C & G & T \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} -2\beta - \alpha & \beta & \alpha & \beta \\ \beta & -2\beta - \alpha & \beta & \alpha \\ \alpha & \beta & -2\beta - \alpha & \beta \\ \beta & \alpha & \beta & -2\beta - \alpha \end{pmatrix} \end{matrix}$$

The major assumption with both JC and Kimura is that the equilibrium frequencies are identical (when  $t = \infty$ ):  $P(A) = P(C) = P(G) = P(T) = 1/4$ . A necessary and sufficient condition of stationarity is the equation:

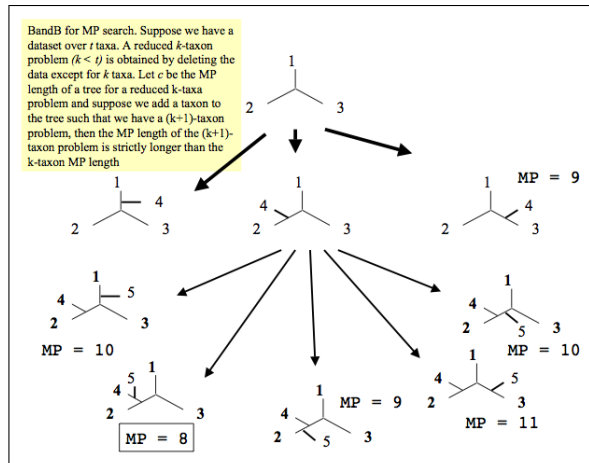
$$\pi_i p_{ij} = \pi_j p_{ji}$$

. In addition to JC and Kimura, continuant matrices have stationary distributions.

**More parameter models:** There are other models which require fewer assumptions, and which notably allow arbitrary  $\pi_i$  (HKY model). The most general model is the general time reversible model (GTR) which has 12 parameters (the most allowed for a 4x4 matrix).

## 2.7.2 Tree constructing/searching

- Exact methods
  - Exhaustive search:  $(2t - 3)!!$  trees,  $t =$  number of leaves
  - Branch-and-bound search: works when the optimality criterion strictly decreases as you traverse a search tree using DFS. Let  $B$  be the best value of the objective function found thus far. While searching, if any solution's OF value is worse than  $B$ , you don't need to search the daughter branches of that part of the search tree, since you are guaranteed that all subsequent trees using that subtree will have a worse OF value. Thus the key to branch-and-bound is to construct a tree so that you maximize the probability of cutting off search branches as you search.



- Heuristic tree searching
  - Branch swapping (heuristic search): choose an internal node, and either rotate neighbors or swap them (repeated swapping can enumerate all trees).
  - Hill-climbing: Starting with a single tree, branch swap and compare. If a better tree is found repeat on this new tree. This can either be strict (if new tree better, always accept) or stochastic (accept better tree with certain probability).
- Heuristic tree construction
  - Sequential tree construction: choose an initial 3-taxon tree. Add taxa st MP length is minimal each iteration. Complexity:  $O(t^3)$
  - Star decomposition: Connect all taxa to a known or ancestral “hub” sequence. Neighbor-joining is a star decomposition method using an approximation to the minimum evolution optimality criterion.

## 2.8 ML vs parsimony

The following are true of parsimony but not ML:

- Cost of change are not a function of branch lengths
- Returns the single best solution
- See above section on optimality

## 2.9 Significance

If a particular tree is returned as optimal for a given data set, one still wants to assess the significance or statistical support for this answer. Typically something of a randomization procedure is performed to get an estimate of what a “random” or null probability distribution for parameter space looks like. This involves either permuting a distance matrix randomly, or swapping nucleotide positions or tree branches numerous times, then going along with the same estimation procedure as for the test data set. The basic idea is that something is significant if it is difficult to generate randomly, or using a null model.

## 2.10 Theoretical issues: assumptions, consistency and convergence rate

Each one of these methods has particular assumptions about how evolution works. In distance methods you either assume additivity or the molecular clock. In parsimony you may assume molecular clock or simply a generalized notion of parsimony, which implicitly assumes that the occurrence of convergent homoplasy is low. In ML, the assumptions are built into the model of sequence evolution. These assumptions can be organized into a few general specifications:

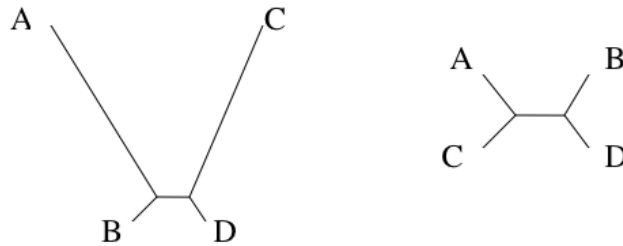
- independent sites within a taxon
- whether rate of change can vary across sites within a taxon
- whether rate of change can vary across taxa (molecular clock)
- other specifications built into Markov type evolutionary models

Generally speaking if you choose to accept an optimal tree under certain conditions, you must believe that these conditions are unlikely to be violated.

**Consistency** is the idea that given an infinite sample of data, a statistic will yield the correct estimate for the unknown quantity being estimated (ie a parameter). Thus consistency is convergence in probability of an estimate to a true value, in other words the weak law of large numbers. In this sense maximum likelihood estimation and distance-based methods yield a consistent tree estimate (the true tree) when solved exactly using the general time reversible Markov-type model of evolution or its submodels (JC, Kimura, etc). This is very difficult to achieve in practice, although numerical estimation with ML is generally more reliable than with distance methods.

Parsimony, on the other hand, is theoretically inconsistent, even for small samples. This is seen for a four-taxon tree which contains two long peripheral branches (the evolutionarily closest taxa either have different rates of change or different lengths of time between them). This results because one of the implicit assumptions in parsimony is that convergent evolution is unlikely: the chance that the same character will arise from mutation more than once in a tree is very low. General parsimony assumes the molecular clock, and if your data set features two evolutionarily related branches each of which has a very different mutation rate, the molecular clock hypothesis is invalid. In the 4-taxon tree example, this situation arises if two non-similar branches generate the same character state independently, but their true neighboring branches do not. Parsimony will claim that these long branches are most closely related, despite the truth that they are not.

- Jukes-Cantor evolution
- The Felsenstein zone



A variant of parsimony does not assume the molecular clock, and in this situation there is no specific model of sequence evolution, in contrast to any of the Markovtype models. This actually corresponds to a situation where every branch in a tree has its own parameter(s). This situation also arises in a fully generalized maximum likelihood model, and thus although ML is consistent, consistency holds only when ML is used in combination with a general time reversible markov model. The general ML method uses the no common mechanism model, and so is equivalent to maximum parsimony in this situation.

In another sense become inconsistent if the assumptions used are not properly met in the data. For example,

- if you assume the molecular clock but your distance data does not pass the ultrametric criterion, UPGMA may return an incorrect tree.
- For ML, if a given sample of data is sparse, the ML estimate can be considerably poor, because ML is not an *unbiased* statistic given a small sample. This bias decreases as data are added, and disappears given infinite data.

- Parsimony is generally unbiased if the branch lengths are short (indicating few changes or short time spans).

**Convergence rate** is the amount of data that a method needs to return the true tree with high probability, as a function of the model tree. We probably don't need to go into too much detail about this.

## 2.11 Other things

Things like assessment of systematic error, significant branch assignments, all alternative methods of parsimony, model-based distance correction methods, etc., were not really discussed. see references below.

## 2.12 References

1. Swofford, D, Olsen, G., Waddell, P., Hillis, D. "Phylogenetic Inference". Molecular Systematics.
2. Raff, R. The Shape of Life.
3. Durbin, R. et al. "Biological Sequence Analysis".