6.047/6.878 Lecture 20: Molecular Evolution and Phylogenetics

Scribed by Andrew Cooper, Stephanie Chang, and Stephen Serene (2012)
Akashnil Dutta (2011)
Albert Wang and Mashaal Sohail (2010)
Guo-Liang Chew and Sara Baldwin (2009)

December 13, 2012

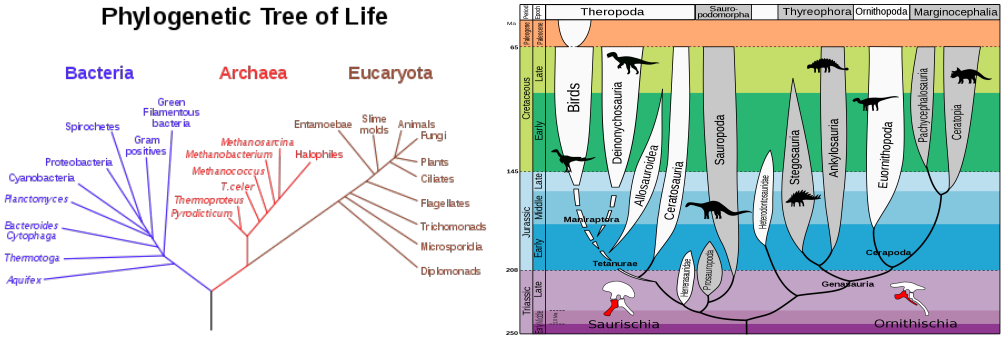
Contents

List of Figures

1 Introduction

Phylogenetics is the study of relationships among a set of objects having a common origin, based on the knowledge of the individual traits of the objects. Such objects may be species, genes, or languages, and their corresponding traits may be morphological characteristics, sequences, words etc. In all these examples the objects under study change gradually with time and diverge from common origins to present day objects.

In Biology, phylogenetics is particularly relevant because all biological species happen to be descendants of a single common ancestor which existed approximately 3.5 to 3.8 billion years ago. Throughout the passage of time, genetic variation, isolation and selection have created the great variety of species that we observe today. Not just speciation however, but extinction has also played a key role in shaping the biosphere as we see today. Studying the ancestry between different species is fundamentally important to biology because they shed much light in understanding different biological functions, genetic mechanisms as well as the process of evolution itself.



(a) Tree of Life (b) Divergence and Extinction of Dinosaurs

Figure 1: Evolutionary History of Life

2 Basics of Phylogeny

2.1 Trees

A tree is a mathematical representation of relationships between objects. A general tree is built from nodes and edges. Each node represents an object, and each edge represents a relationship between two nodes. In the case of phylogenetic trees, we represent evolution using trees. In this case, each node represents a divergence event between two ancestral lineages, the leaves denote the set of present objects and the root represents the common ancestor.

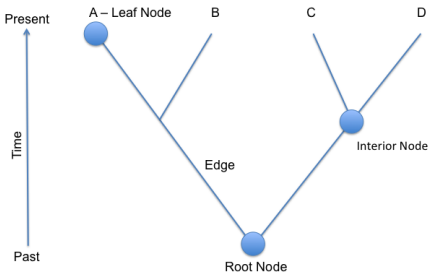


Figure 2: Defining tree terminology. A tree of branching nodes is depicted with leaves at the top and the root on the bottom. Time continues upward, toward the leaves.

However, sometimes more information is reflected in the branch lengths, such as time elapsed or the amount of dissimilarity. According to these differences, biological phylogenetic trees may be classified into three categories:

Cladogram: gives no meaning to branch lengths; only the sequence and topology of the branching matters.

Phylogram: Branch lengths are directly related to the amount of **genetic change**. The longer the branch of a tree, the greater the amount of phylogenetic change that has taken place. The leaves in this tree may not necessarily end on the same vertical line, due to different rates of mutation.

Chronogram (ultrametric tree): Branch lengths are directly related to **time**. The longer the branches of a tree, the greater the amount of time that has passed. The leaves in this tree necessarily end on the same vertical line (i.e. they are the same distance from the root), since they are all in the present unless extinct species were included in the tree. Although there is a correlation between branch lengths and genetic distance on a chronogram, they are not necessarily exactly proportional because evolution rates / mutation rates are not constant. Some species evolve and mutate faster than others, and some historical time periods foster faster rates of evolution than others.

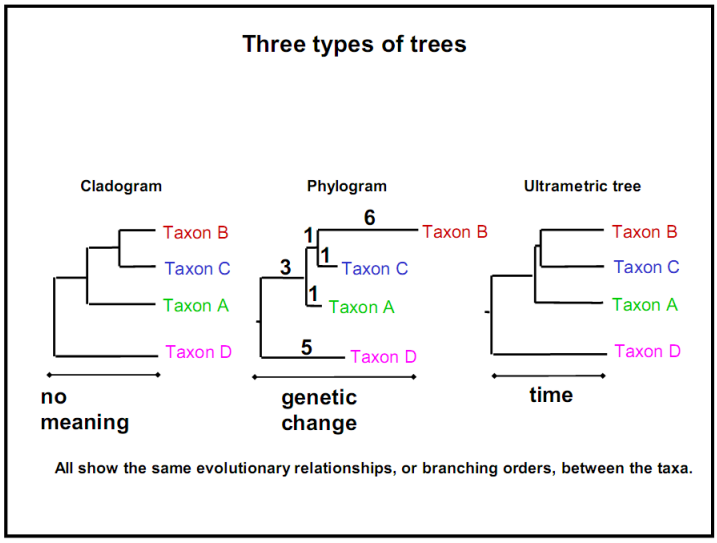


Figure 3: Three types of trees.

2.2 Traits

A trait is any characteristic that an object or species possesses. In humans, an example of a trait may be bipedalism (the ability to walk upright) or the opposable thumb. Another human trait may be a specific DNA sequence that humans possess. The first examples of physical traits are called **morphological traits**, while the latter DNA traits are called **sequence traits**. Each has its advantages and disadvantages to study. All methods for tree-reconstruction rely on studying the occurrence of different traits in the given objects. In traditional phylogenetics the morphological data of different species were used for this purpose. In modern methods, genetic sequence data is used instead. Each has its advantages and disadvantages.

Morphological Traits: Arise from empirical evaluation of physical traits. This can be advantageous because physical characteristics are very easy to quantify and understand for everyone, scientists and children alike. The disadvantages to this approach are that we can only evaluate a small set of traits, such as hair, nails, hoofs, teeth, etc. Further, these traits only allow us to build species. Finally, it is much easier to be "tricked" by convergent evolution. Species that diverged millions of years ago may converge again on the few traits that are observable to scientists, giving a false representation of how closely related the species are.

Sequence Traits: Are discovered by studying the genomes of different species. This approach can be advantageous because it creates much more data and allows scientists to create gene trees in addition to species trees. The primary difficulty with this approach is that DNA is only built from 4 bases, so back mutations are frequent. In this approach, scientists must reconcile the signals of a large number of ill-behaved traits as opposed to that of a small number of well-behaved traits in the traditional approach. The rest of the chapter will focus principally on tree building from gene sequences.

Since this approach deals with comparing between pairs of genes, it is useful to understand the concept of **homology**: A pair of genes are called **paralogues** if they diverged from a duplication event, and **orthologues** if they diverged from a speciation event.

FAQ

Q: Would it be possible to use extinct species' DNA sequences?

A: Current technologies only allow for usage of extant sequences. However, there have been a few successes in using extinct species' DNA. DNA from frozen mammoths have been collected and are being sequences but due to DNA breaking down over time and contamination from the environment, it is very hard to extract correct sequences.

2.3 Methods for Tree Reconstruction

Once we have found genetic data for a set of species, we are interested in learning how those species relate to one another. Since we can, for the most part, only obtain DNA from living creatures, we must infer the existence of ancestors of each species, and ultimately infer the existence of a common ancestor. This is a challenging problem, because very limited data is available. The following sections will explore the modern methods for inferring ancestry from sequence data. They can be classified into two approaches, distance based methods and character based methods.

Distance based approaches take two steps to solve the problem, i.e. to quantify the amount of mutation that separates each pair of sequences (which may or may not be proportional to the time since they have been separated) and to fit the most likely tree according to the pair-wise distance matrix. The second step is usually a direct algorithm, based on some assumtions, but may be more complex.

Charecter based approaches instead try to find the tree that best explains the observed sequences. As opposed to direct reconstruction, these methods rely on tree proposal and scoring techniques to perform a heuristic search over the space of trees.

Did You Know?

Occam's Razor, as discussed in previous chapters, does not always provide the most accurate hypothesis. In many cases during tree reconstruction, the simplest explanation is not the most probable. For example, a set of possible ancestries may be possible, given some observed data. In this case, the simplest ancestry may not be correct if a trait arose independently in two seperate lineages. This issue will be considered in a later section.

3 Distance Based Methods

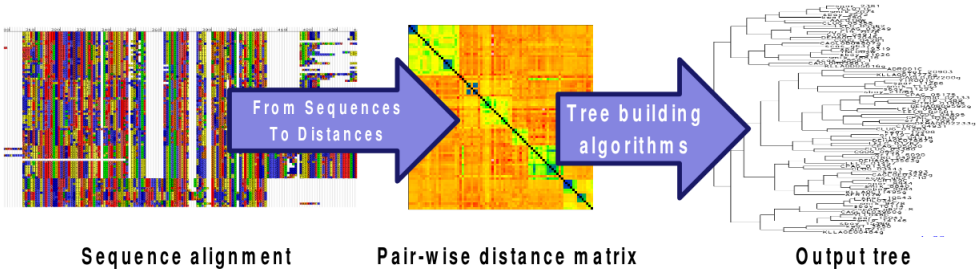


Figure 4: The two steps of distance based phylogenetic reconstruction.

The distance based models sequester the sequence data into pairwise distances. This step loses some information, but sets up the platform for direct tree reconstruction. The two steps of this method are hereby discussed in detail.

3.1 From alignment to distances

In order to understand how a distance-based model works, it is important to think about what distance means when comparing two sequences. There are three main interpretations.

Nucleotide Divergence is the idea of measuring distance between two sequences based on the number of places where nucleotides are not consistent. This assumes that evolution happens at a uniform rate across the genome, and that a given nucleotide is just as likely to evolve into any of the other three nucleotides. Although it has shortcomings, this is often a great way to think about it.

Transitions and Transversions This is similar to nucleotide divergence, but it recognizes that A-G and T-C substitutions are most frequent. Therefore, it keeps two parameters, the probability of a transition and the probability of a transversion.

Synonymous and non-synonymous substitutions This method keeps tracks of substitutions that affect the coded amino-acid by assuming that substitutions that do not change the coded protein will not be selected against, and will thus have a higher probability of occurring than those substitutions which do change the coded amino acid.

The naive way to interpret the separation between two sequences may be simply the number of mismatches, as described by nucleotide divergence above. While this does provide us a distance metric (i.e. $d(a,b) + d(b,c) \geq d(a,c)$) this does not quite satisfy our requirements, beecause we want **additive distances**, i.e. those that satisfy $d(a,b) + d(b,c) = d(a,c)$ for a path $a \to b \to c$ of evolving sequence, because the amount of mutations accumulated along a path in the tree should be the sum of that of its individual components. However, the naive mismatch fraction do not always have this property, because this quantity is bounded by 1, while the sum of individual components can easily exceed 1.

The key to resolving this paradox is **back-mutations**. When a large number of mutations accumulate on a sequence, not all the mutations introduce new mismatches, some of them may occur on already mutated base pair, resulting in the mismatch score remaining the same or even decreasing. For small mismatch-scores however, this effect is statistically insignificant, because there are vastly more identical pairs than mismatching pairs. However, for sequences separated by longer evolutionary distance, we must correct for this effect. The Jukes-Cantor model is one such simple markov model that takes this into account.

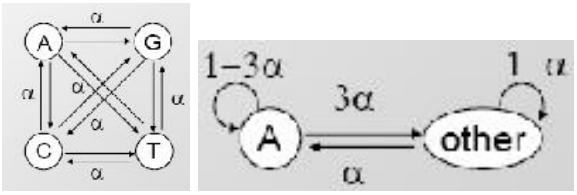
3.1.1 Jukes-Cantor distances

To illustrate this concept, consider a nucleotide in state 'A' at time zero. At each time step, it has a probability 0.7 of retaining its previous state and probability 0.1 of transitioning to each of the other three

states. The probability $P(B|t)$ of observing state (base) B at time t essentially follows the recursion

$$P(B|t+1) = 0.7P(B|t) + 0.1\sum_{b\neq B} P(b|t) = 0.1 + 0.6P(B|t)$$

.



(a) 4 States (b) 2 States

Figure 5: Markov chain accounting for back mutations

If we plot $P(B|t)$ versus t, we observe that the distribution starts off as concentrated at the state 'A' and gradually spreads over to the rest of the states, eventually going towards an equilibrium of equal probabilities. This progression makes sense, intuitively. Over millions of years, species can evolve so dramatically that they no longer resemble their ancestors. At that extreme, a given base location in the ancestor is just as likely to have evolved to any of the four possible bases in that location over time.

time:-	0	1	2	3	4
A	1	0.7	0.52	0.412	0.3472
C	0	0.1	0.16	0.196	0.2196
G	0	0.1	0.16	0.196	0.2196
T	0	0.1	0.16	0.196	0.2196

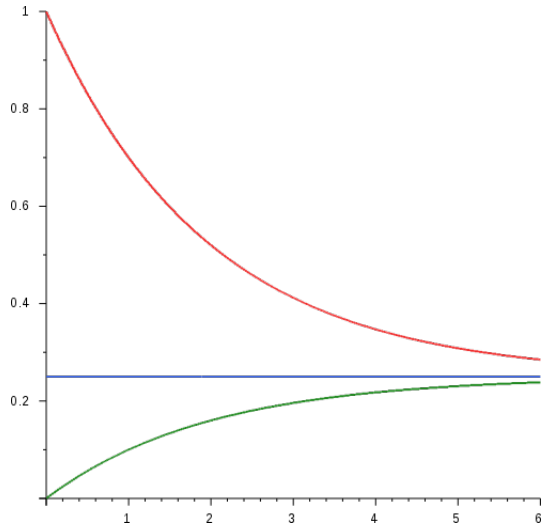


Figure 6: The y axis denotes probability of observing the bases - A(red), others(green). x axis denotes time.

The essence of the Jukes Cantor model is to backtrack t, the amount of time elapsed from the fraction of altered bases. Conceptually, this is just inverting the x and y axis of the green curve. To model this quantitatively, we consider the following matrix $S(t)$ which denotes the respective probabilities $P(x|y, \Delta t)$

of observing base x given a starting state of base y in time Δt.

$$S(\Delta t) = \begin{pmatrix} P(A|A,\Delta t) & P(A|G,\Delta t) & \cdots & P(A|T,\Delta t) \\ P(G|A,\Delta t) & \cdots & & \cdots \\ \cdots & & & \cdots \\ P(T|A,\Delta t) & \cdots & \cdots & P(T|T\Delta T) \end{pmatrix}$$

We can assume this is a stationary markov model, implying this matrix is multiplicative, i.e.

$$S(t_1 + t_2) = S(t_1)S(t_2)$$

For a very short time ϵ, we can assume that there is no second order effect, i.e. there isn't enough time for two mutations to occur at the same nucleotide. So the probabilities of cross transitions are all proportional to ϵ. Further, in Jukes Cantor model, we assume that all the transition rates are same from each nucleotide to another nucleotide. Hence, for a short time ϵ

$$S(\epsilon) = \begin{pmatrix} 1-3\alpha\epsilon & \alpha\epsilon & \alpha\epsilon & \alpha\epsilon \\ \alpha\epsilon & 1-3\alpha\epsilon & \alpha\epsilon & \alpha\epsilon \\ \alpha\epsilon & \alpha\epsilon & 1-3\alpha\epsilon & \alpha\epsilon \\ \alpha\epsilon & \alpha\epsilon & \alpha\epsilon & 1-3\alpha\epsilon \end{pmatrix}$$

At time t, the matrix is given by

$$S(t) = \begin{pmatrix} r(t) & s(t) & s(t) & s(t) \\ s(t) & r(t) & s(t) & s(t) \\ s(t) & s(t) & r(t) & s(t) \\ s(t) & s(t) & s(t) & r(t) \end{pmatrix}$$

From the equation $S(t+\epsilon) = S(t)S(\epsilon)$ we obtain

$$r(t+\epsilon) = r(t)(1-3\alpha\epsilon) + 3\alpha\epsilon s(t) \text{ and } s(t+\epsilon) = s(t)(1-\alpha\epsilon) + \alpha\epsilon r(t))$$

Which rearrange as the coupled system of differential equations

$$r'(t) = 3\alpha(-r(t) + s(t)) \text{ and } s'(t) = \alpha(r(t) - s(t))$$

With the initial conditions $r(0) = 1$ and $s(0) = 0$. The solutions can be obtained as

$$r(t) = \frac{1}{4}(1 + 3e^{-4\alpha t}) \text{ and } s(t) = \frac{1}{4}(1 - e^{-4\alpha t})$$

Now, in a given alignment, if we have the fraction f of the sites where the bases differ, we have:

$$f = 3s(t) = \frac{3}{4}(1 - e^{-4\alpha t})$$

implying

$$t \propto -\log\left(1 - \frac{4f}{3}\right)$$

To agree asymptotically with f, we set the evolutionary distance d to be

$$\boxed{d = -\frac{3}{4}\log\left(1 - \frac{4f}{3}\right)}$$

Note that distance is approximately proportional to f for small values of f and asymptotically approaches infinity when $f \to 0.75$. Intuitively this happens because after a very long period of time, we would expect the sequence to be completely random and that would imply about three-fourth of the bases mismatching

with original. But the uncertainty values of the Jukes-Cantor distance also becomes very large when f approaches 0.75.

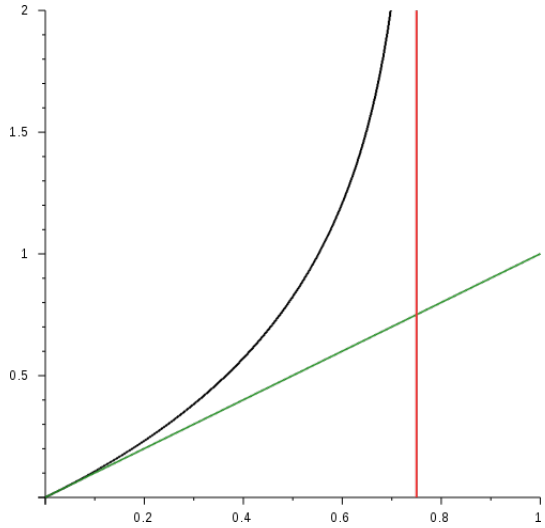


Figure 7: Fraction of altered bases (x axis) versus the Jukes Cantor distance(y axis).
Black line denotes the curve, green is the trend line for small values of f while the red line denotes the asymptotic boundary.

3.1.2 Other Models

The Jukes Cantor model is the simplest model that gives us theoretically consistent additive distance model. However, it is a one-parameter model that assumes that the mutations from each base to a different base has the same chance. But, changes between AG or between TC are more likely than changes across them. The first type of substitution is called transitions while the second type is called transversions. The Kimura model has two parameters which take this into account. There are also many other modifications of this distance model that takes into account the different rates of transitions and transversions etc. that are depicted below.

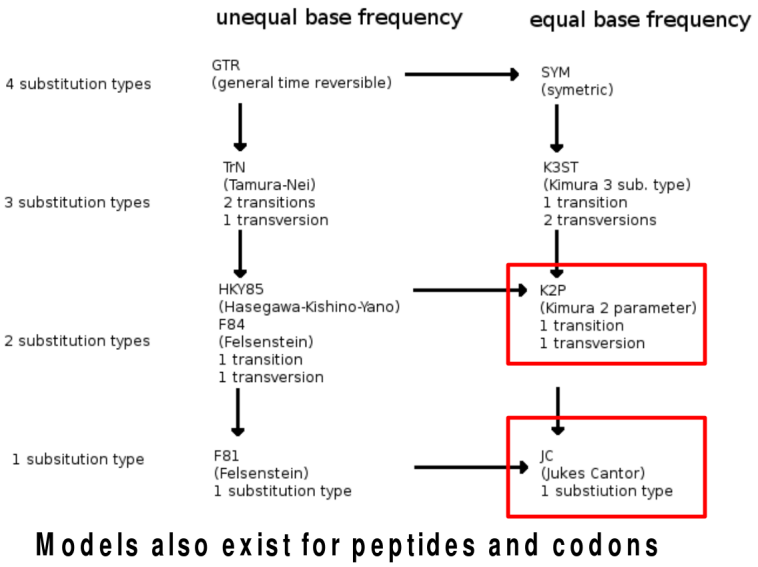


Figure 8: Distance models of varying levels of complexity(parameters).

FAQ

Q: Can we use different parameters for different parts of the tree? To account for different mutation rates?

A: Its possible, it is a current area of research.

3.2 Distances to Trees

If we have a weighted phylogenetic tree, we can find the total weight (length) of the shortest path between a pair of leaves by summing up the individual branch lengths in the path. Considering all such pairs of leaves, we have a distance matrix representing the data. In distance based methods, the problem is to reconstruct the tree given this distance matrix.

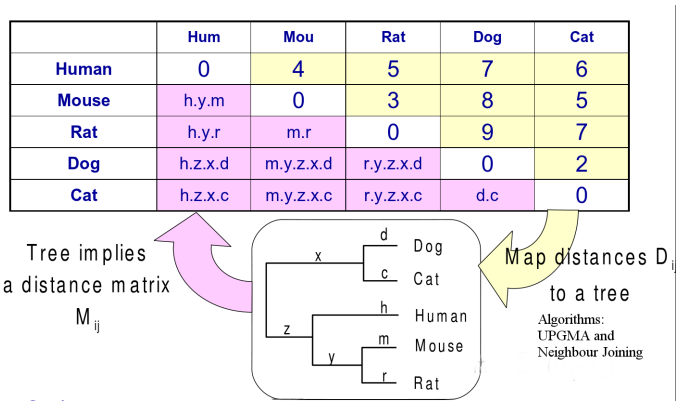


Figure 9: Mapping from a tree to a distance matrix and vice versa

FAQ

Q: In Figure 9 The m and r sequence divergence metrics can have some overlap so distance between mouse and rat is not simply m+r. Wouldn't that only be the case if there was no overlap?

A: If you model evolution correctly, then you would get evolutionary distance. It's an inequality rather than an equality and we agree that you can't exactly infer that the given distance is the precise distance. Therefore, the sequences' distance between mouse and rat is probably less than m + r because of overlap, convergent evolution, and transversions.

However, note that there is not a one-to-one correspondence between a distance matrix and a weighted tree. Each tree does correspond to one distance matrix, but the opposite is not always true. A distance matrix has to satisfy additional properties in order to correspond to some weighted tree. In fact, there are two models that assume special constraints on the distance matrix:

Ultrametric: For all triplets (a, b, c) of leaves, two pairs among them have equal distance, and the third distance is smaller; i.e. the triplet can be labelled i, j, k such that

$$d_{ij} \leq d_{ik} = d_{jk}$$

Conceptually this is because the two leaves that are more closely related (say i, j) have diverged from the thrid (k) at exactly the same time. and the time separation from the third should be equal, whereas the separation between themselves should be smaller.

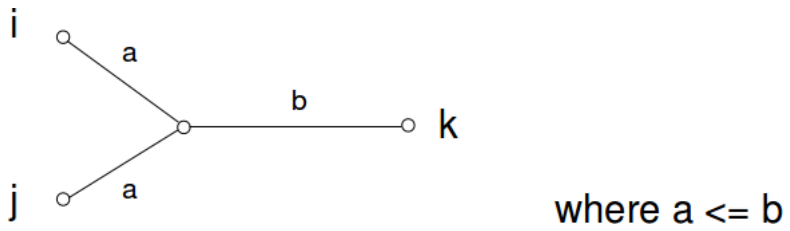


Figure 10: Ultrametric distances.

Additive: Additive distance matrices satisfy the property that all quartet of leaves can be labelled i, j, k, l such that

$$d_{ij} + d_{kl} \leq d_{ik} + d_{jl} = d_{il} + d_{jk}$$

This is in fact true for all positive-weight trees. For any 4 leaves in a tree, there can be exactly one topology, i.e.

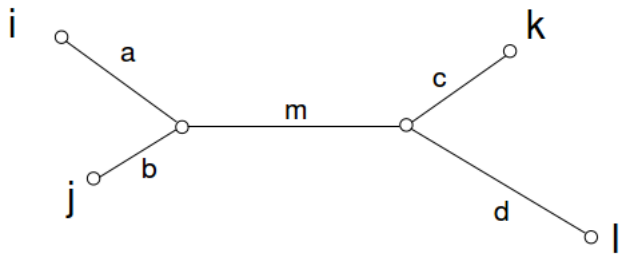


Figure 11: Additive distances.

Then the above condition is term by term equivalent to

$$(a + b) + (c + d) \leq (a + m + c) + (b + m + d) = (a + m + d) + (b + m + c)$$

. This equality corresponds to all pairwise distances that are possible from traversing this tree.

These types of redundant equalities must occur while mapping a tree to a distance matrix, because a tree of n nodes has $n - 1$ parameters, one for each branch length, while a distance matrix has n^2 parameters. Hence, a tree is essentially a lower dimensional projection of a higher dimensional space. A corollary of this observation is that not all distance matrices have a corresponding tree, but all trees map to unique distance matrices.

However, real datasets do not exactly satisfy either ultrameric or additive constraints. This can be due to noise (when our parameters for our evolutionary models are not precise), stochasticity and randomness (due to small samples), fluctuations, different rates of mutations, gene conversions and horizontal transfer. Because of this, we need tree-building algorithms that are able to handle noisy distance matrices.

Next, two algorithms that directly rely on these assumptions for tree reconstruction will be discussed.

3.2.1 UPGMA - Unweighted Pair Group Method with Arithmetic Mean

This is exactly same as the method of **Hierarchical clustering** discussed in Lecture 13, Gene Expression Clustering. It forms clusters step by step, from closely related nodes to ones that are further separated. A branching node is formed for each successive level. The algorithm can be described properly by the following steps:

Initialization:

1. Define one leaf i per sequence x_i.
2. Place each leaf i at height 0.
3. Define Clusters C_i each having one leaf i.

Iteration:

1. Find the pairwise distances d_{ij} between each pairs of clusters C_i, C_j by taking the arithmetic mean of the distances between their member sequences.
2. Find two clusters C_i, C_j such that d_{ij} is minimized.
3. Let $C_k = C_i \cup C_j$.
4. Define node k as parent of nodes i, j and place it at height $d_{ij}/2$ above i, j.
5. Delete C_i, C_j.

Termination: When two clusters C_i, C_j remain, place the root at height $d_{ij}/2$ as parent of the nodes i, j

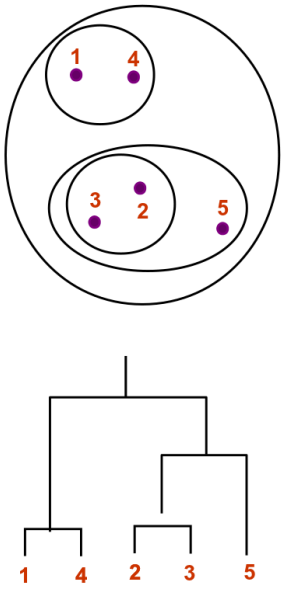


Figure 12: UPGMA / Hierarchical Clustering

Weaknesses of UPGMA

Although this method is guaranteed to find the correct tree if the distance matrix obeys the ultrameric property, it turns out to be a inaccurate algorithm in practice. Apart from lack of robustness, it suffers from the molecular clock assumption that the mutation rate over time is constant for all species. However, this is not true as certain species such as rat and mice evolve much faster than others. The following figure illustrates an example where UPGMA fails:

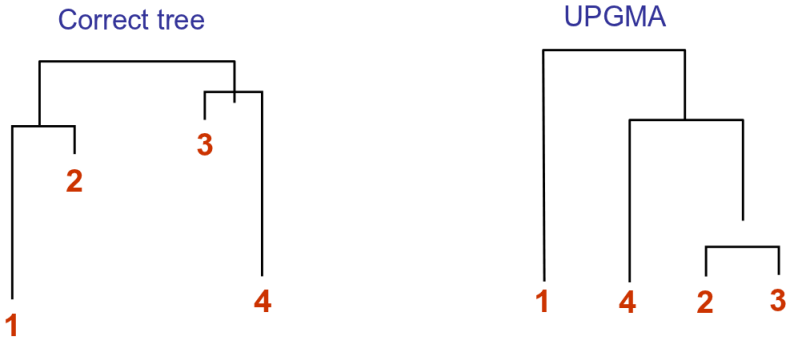


Figure 13: UPGMA fails to find the correct tree in this case

3.2.2 Neighbor Joining

The neighbor joining method is guaranteed to produce the correct tree if the distance matrix satisfies the additive property. It may also produce a good tree when there is some noise in the data. The algorithm is described below:

Finding the neighboring leaves: Let

$$D_{ij} = d_{ij} - (r_i + r_j) \text{ where } r_a = \frac{1}{n-2}\sum_k d_{ak},\ a \in \{i, j\}$$

Here n is the number of nodes in the tree; hence, r_i is the average distance of a node to the other nodes. It can be proved that the above modification ensures that D_{ij} is minimal only if i, j are neighbors. (A proof can be found in page 189 of Durbin's book).

Initialization: Define T to be the set of leaf nodes, one per sequence. Let $L = T$

Iteration:

1. Pick i, j such that D_{ij} is minimized.
2. Define a new node k, and set $d_{km} = \frac{1}{2}(d_{im} + d_{jm}d_{ij}) \forall m \in L$
3. Add k to T, with edges of lengths $d_{ik} = \frac{1}{2}(d_{ij} + r_i r_j)$
4. Remove i, j from L
5. Add k to L

Termination: When L consists of two nodes i, j, and the edge between them of length d_{ij}, add the root node as parent of i and j.

4 Character-Based Methods

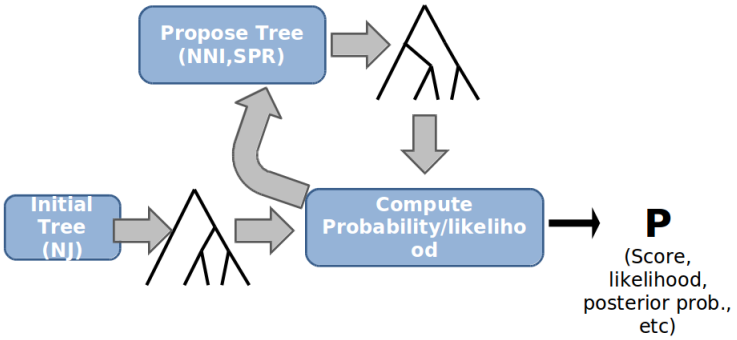


Figure 14: An overview of the character based methods

In character-based methods, the goal is to first create a valid algorithm for scoring the probability that a given tree would produce the observed sequences at its leaves, then to search through the space of possible trees for a tree that maximizes that probability. Good algorithms for tree scoring, and while searching the space of trees is theoretically NP-Hard (Due to the large number of possible trees), tractable heuristic search methods can in many cases find good trees. We'll first discuss tree scoring algorithms, then search techniques.

4.1 Scoring

There are two main algorithms for tree scoring. The first approach, which we will call parsimony reconstruction, is based on Occam's razor, and scores a topology based on the minimum number of mutations it implies, given the (known) sequences at the leaves. This method is simple, intuitive, and fast. The second approach is a maximum likelihood method which scores trees by explicitly modeling the probability of observing the sequences at the leaves given a tree topology.

4.1.1 Parsimony

Conceptually, this method is simple. It simply assigns a value of for each base pair at each ancestral node such that the number of substitutions is minimized. The score is then just the sum over all base pairs of that minimal number of mutations at each base pair. (Recall that the eventual goal is to find a tree that minimizes that score.)

To reconstruct the ancestral sequences at internal nodes on the tree, the algorithm first scans up from the (known) leaf sequences, assigning a set of bases at each internal node based on its children. Next, it iterates down the tree, picking bases out of the allowed sets at each node, this time based on the node's parents. The following illustrates this algorithm in detail (note that there are $2N-1$ total nodes, indexed from the root, such that the known leaf nodes have indices $N-1$ through $2N-1$):

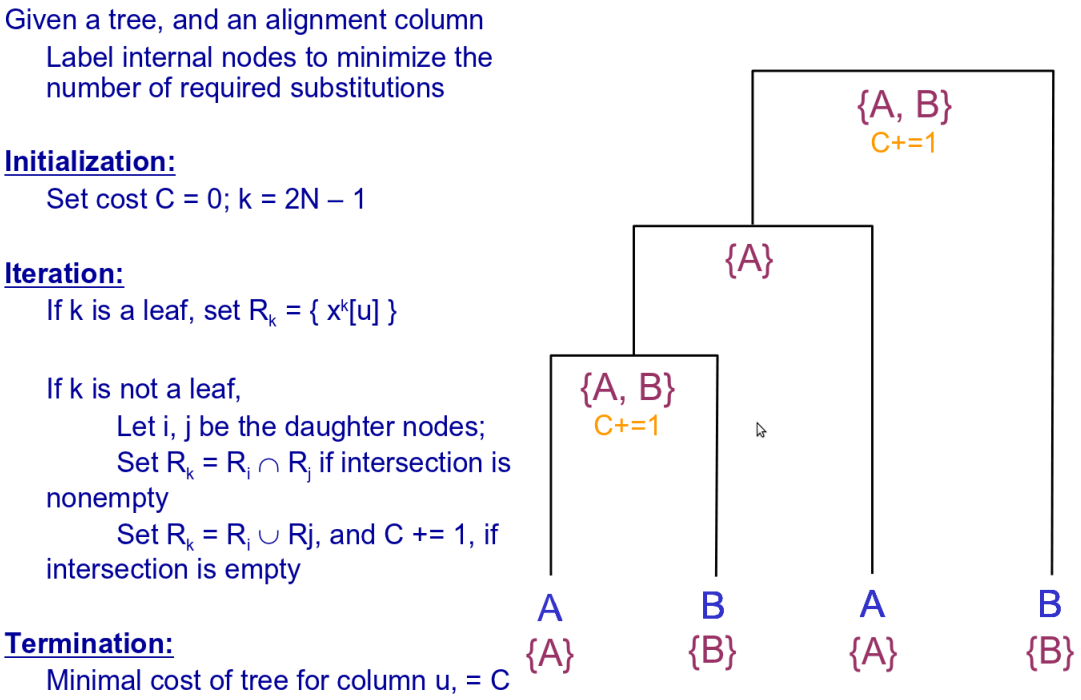


Figure 15: Parsimony scoring: union and intersection

Traceback:

1. Choose an arbitrary nucleotide from R_{2N-1} for the root

2. Having chosen nucleotide r for parent k,
   If $r \in R_i$ choose r for daughter i
   Else, choose arbitrary nucleotide from R_i

Easy to see that this traceback produces some assignment of cost C

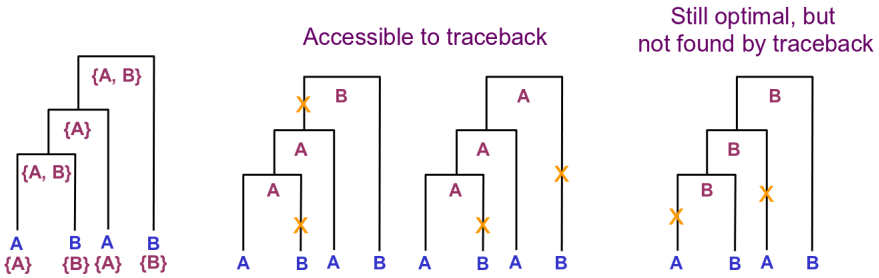


Figure 16: Parsimony traceback to find ancestral neucleotides

	M	R	B1	H	B2	D	B3
A	0	1	1	0	1	1	2
C	1	1	2	1	3	1	4
G	1	0	1	1	2	0	2
T	1	1	2	1	3	1	4

- Each cell (N,C) represents the min cost of the subtree rooted at N, if the label at N is C.
- Update table by walking up the tree from the leaves to the root, remembering max choices.
- Traceback from root to leaves

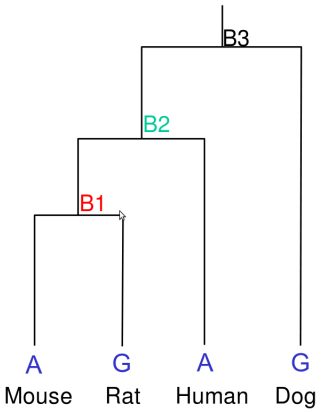


Figure 17: Parsimony scoring by dynamic programming

As we mentioned before, this method is simple and fast. However, this simplicity can distort the scores it assigns. For one thing, the algorithm presented here assumes that a given base pair undergoes a substitution along at most one branch from a given node, which may lead it to ignore highly probably internal sequences that violate this assumption. Furthermore, this method does not explicitly model the time represented along each edge, and thus cannot account for the increased chance of a substitution along edges that represent a long temporal duration, or the possibility of different mutation rates across the tree. Maximum likelihood methods largely resolve these shortcomings, and are thus more commonly used for tree scoring.

4.1.2 Maximum Likelihood - Peeling Algorithm

As with the general Maximum likelihood methods, this algorithm scores a tree according to the (log) joint probability of observing the data and the given tree, i.e. $P(D, T)$. The peeling algorithm again considers

individual base pairs and assumes that all sites evolve independently. As in the parsimony method, this algorithm considers all base pairs independently: it calculates the probability of observing the given characters at each base pair in the leaf nodes, given the tree, a set of branch lengths, and the maximum likelihood assignment of the internal sequence, then simply multiplies this probabilities over all base pairs to get the total probability of observing the tree. Note that the explicit modeling of branch lengths is a difference from the previous approach.

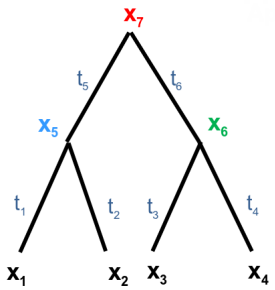


Figure 18: A tree to be scored using the peeling algorithm. n=4

Here each node has a character x_i and t_i is the corresponding branch length from its parent. Note that we already know the values $x_1, x_2 \cdots x_n$, so they are constants, but $x_{n+1}, \cdots x_{2n-1}$ are unknown characters at ancestral nodes which are variables to which we will assign maximum likelihood values. (Also note that we have adopted a leaves-to-root indexing scheme for the nodes, the opposite of the scheme we used before.) We want to compute $P(x_1 x_2 \cdots x_n|T)$. For this we sum over all possible combinations of values at the ancestral nodes. this is called marginalization. In this particular example

$$P(x_1 x_2 x_3 x_4|T) = \sum_{x_5} \sum_{x_6} \sum_{x_7} P(x_1 x_2 \cdots x_7|T)$$

There are 4^{n-1} terms in here, but we can use the following factorization trick:

$$= \sum_{x_5} \sum_{x_6} \sum_{x_7} P(x_1|x_5,t_1)P(x_2|x_5,t_2)P(x_3|x_6,t_3)P(x_4|x_6,t_4)P(x_5|x_7,t_5)P(x_6|x_7,t_6)P(x_7)$$

Here we assume that each branch evolves independently. And the probability $P(b|c,t)$ denotes the probability of base c mutating to base b given time t, which is essentially obtained from the Jukes Cantor model or some more advanced model discussed earlier. Next we can move the factors that are independent of the summation variable outside the summation. That gives:

$$= \sum_{x_7} \left[P(x_7) \left(\sum_{x_5} P(x_5|x_7,t_5)P(x_1|x_5,t_1)P(x_2|x_5,t_2) \right) \left(\sum_{x_6} P(x_6|x_7,t_6)P(x_3|x_6,t_3)P(x_4|x_6,t_4) \right) \right]$$

Let T_i be the subtree below i. In this case, our $2n-1 \times 4$ dynamic programming array computes $L[i,b]$, the probability $P(T_i|x_i = b)$ of observing T_i, if node i contains base b. Then we want to compute the probability of observing $T = T_{2n-1}$, which is

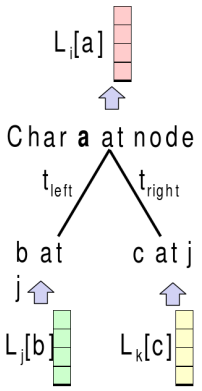
$$\sum_b P(x_{2n-1} = b)L[2n-1, b]$$

Note that for each ancestral node i and its childer j, k, we have

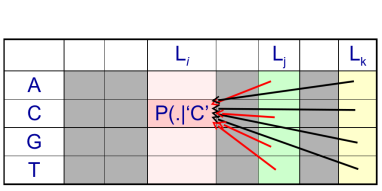
$$L[i,b] = \left(\sum_c P(c|b,t_j)L[j,c] \right) \left(\sum_c P(c|b,t_k)L[k,c] \right)$$

Subject to the initial conditions for the leaf nodes, i.e. for $i \leq n$:

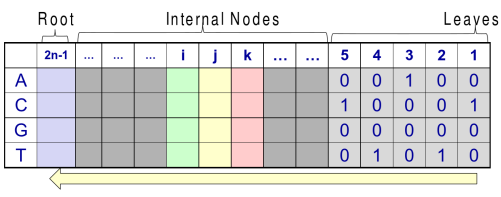
$$L[i,b] = 1 \text{ if } x_i = b \text{ and } 0 \text{ otherwise}$$



The recurrence



(a) Filling the matrix



(b) Inititalization and direction

Note that we still do not have the values $P(x_{2n-1} = b)$. It is usually assigned equally or from some prior distribution, but it does not affect the results greatly. The final step is of course to multiply all the probabilities for individual sites to obtain the probability of observing the set of entire sequences. In addition, once we have assigned the maximum likelihood values for each internal node given the tree structure and the set of branch lengths, we can multiply the resulting score by some prior probabilities of the tree structure and the set of branch lengths, which are often generated using explicit modeling of evolutionary processes, such as the Yule process or birth-death models like the Moran process. The result of this final multiplication is called the a posteriori probability, using the language of Bayesian inference. The overall complexity of this algorithm is $O(nmk^2)$ where n is the number of leaves (taxa), m is the sequence length, and k is the number of characters.

There are addvantages and disadvantages of this algorithm. Such as

Advantages:

1. Inherently statistical and evolutionary model-based.
2. Usually the most consistent of the methods available.
3. Used for both character and rate analyses
4. Can be used to infer the sequences of the extinct ancestors.
5. Account for branch-length effects in unbalanced trees.
6. Nucleotide or amino acid sequences, other types of data.

Disadvantages:

1. Not as simple and intuitive as many other methods.
2. Computationally intense Limited by, number of taxa and sequence length).
3. Like parsimony, can be fooled by high levels of homoplasy.
4. Violations of model assumptions can lead to incorrect trees.

4.2 Search

A comprehensive search over the space of all trees would be extremely costly. The number of full rooted trees with $n + 1$ leaves is the n-th catalan number

$$C_n = \frac{1}{n+1}\binom{2n}{n} \approx \frac{4^n}{n^{3/2}\sqrt{\pi}}$$

Moreover, we must compute the maximum likelihood set of branch lengths for each of these trees. Thus, it is an NP-Hard problem to maximize the score absolutely for all trees. Fortunately, heuristic search algorithms can generally identify good solutions in the tree space. The general framework for such search algorithms is as follows:

Inititalization: Take some tree as the base of iteration (randomly or according to some other prior, or from the distance based direct algorithms).

Proposal: Propose a new tree by randomly modifying the current tree slightly.

Score: Score the new proposal according to the methods described above.

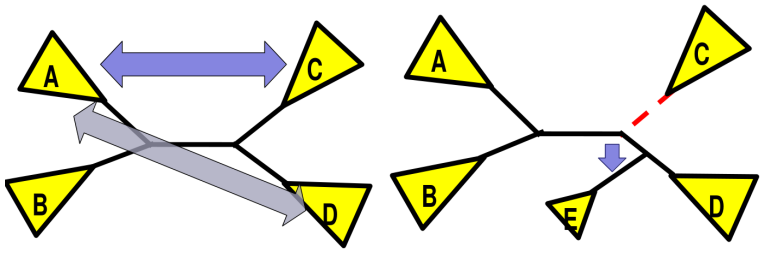
Select: Randomly select the new tree or the old tree (corresponding probabilities according to the score(likelihood) ratio.

Iterate: Repeat to proposal step unless some termination criteria is met (some threshold score or number of steps reached.

the basic idea here is the heuristic assumption that the scores of closely related trees are similar, so that good solutions may be obtained by successive local optimization, which is expected to converge towards a overall good solution.

4.2.1 Tree Proposal

One method for modifying trees is the Nearest Neighbor Exchange (NNI), illustrated below.



(c) Before (d) After

Figure 20: An unit step using Nearest Neighbor Interchange scheme

Another common method, not described here, is Tree Bisection and Join (TBJ). The important criteria for such proposal rules is that:

(a) The tree space should be connected, i.e. any pair of trees should be obtainable from each other by successive proposals.

(b) An individual new proposal should be sufficiently close to the original. So that it is more likely to be a good solution by virtue of the proximity to an already discovered good solution. If individual steps are too big, the algorithm may move away from an already discovered solution (also depends on the selection step). In particular, note that the measure of similarity by which the measure these step sizes is precisely the difference in the likelihood scores assigned to the two trees.

4.2.2 Selection

Choosing whether or not to adopt a given proposal, like the process of generating the proposal itself, is inherently heuristic and varies. A general rules of thumb is:

1. If the new one has a better score, always accept it.
2. If it has a worse score, there should be some probability of selecting it, otherwise the algorithm will soon fixate in a local minima, ignoring better alternatives a little far away.
3. There should not be too much probability of selecting an worse new proposal, otherwise, it risks rejecting a known good solution.

It is the trade-off between the steps 2 and 3 that determines a good selection rule. Metropolis Hastings is a Markov Chain Monte Carlo Method (MCMC) that defines specific rules for exploring the state space in a way that makes it a sample from the posterior distribution. These algorithms work somewhat well in practice, but there is no guarantee for finding the appropriate tree. So a method known as bootstrapping is used, which is basically running the algorithm over and over using subsets of the base pairs in the leaf sequences,. then favoring global trees that match the topologies generated by using only these subsequences.

5 Possible Theoretical and Practical Issues with Discussed Approach

A special point must be made about distances. Since distances are typically calculated between aligned gene sequences, most current tree reconstruction methods rely on heavily conserved genes, as non-conserved genes would not give information on species without those genes. This causes the ignoring of otherwise useful data. Therefore, there are some algorithms that try to take into account less conserved genes in reconstructing trees but these algorithms tend to take a long time due to the NP-Hard nature of reconstructing trees.

Additionally, aligned sequences are still not explicit in regards to the events that created them. That is, combinations of speciation, duplication, loss, and horizontal gene transfer (hgt) events are easy to mix up because only current DNA sequences are available. (see [9] for a commentary on such theoretical issues) A duplication followed by a loss would be very hard to detect. Additionally, a duplication followed by a speciation could look like an HGT event. Even the probabilities of events happening is still contested, especially horizontal gene transfer events.

Another issue is that often multiple marker sequences are concatenated and the concatenated sequence is used to calculate distance and create trees. However, this approach assumes that all the concatenated genes had the same history and there is debate over if this is a valid approach given that events such as hgt and duplications as described above could have occurred differently for different genes. [8] is an article showing how different phylogenetic relationships were found depending on if the tree was created using multiple genes concatenated together or if it was created using each of the individual genes. Conversely, additional [4] claims that while hgt is prevalent, orthologs used for phylogenetic reconstruction are consistent with a single tree of life. These two issues indicate that there is clearly debate in the field on a non arbitrary way to define species and to infer phylogenetic relationships to recreate the tree of life.

6 Towards final project

6.1 Project Ideas

1. Creating better distance models such as taking into account duplicate genes or loss of genes. It may also be possible to analyze sequences for peptide coding regions and calculate distances based on peptide chains too.

2. Creating a faster/more accurate search algorithm for turning distances into trees.

3. Analyze sequences to calculate probabilities of speciation, duplication, loss, and horizontal gene transfer events.

4. Extending an algorithm that looks for HGTs to look for extinct species. A possible use for HGTs is that if a program were to infer HGTs between different times, it could mean that there was a speciation where one branch is now extinct (or not yet discovered) and that branch had caused an HGT to the other extant branch.

6.2 Project Datasets

1. 1000 Genomes Project http://www.1000genomes.org/

2. Microbes Online http://microbesonline.org/

7 What Have We Learned?

In this chapter, we have learnt different methods and approaches for reconstructing Phylogenetic trees from sequence data. In the next chapter, its application in gene trees and species trees and the relationship between those two will be discussed, as well as modelling phylogenies among populations within a species and between closely related species.

References

[1] 1000 genomes project.

[2] et al Ciccarelli, Francesca. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311, 2006.

[3] Tal Dagan and William Martin. The tree of one percent. *Genome Biology*, Nov 2006.

[4] Ochman Howard Daubin Vincent, Moran Nancy A. Phylogenetics and the cohesion of bacterial genomes. *Science*, 301, 2003.

[5] A.J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein familes. *Nucleic Acids Research*, 30(7):1575–1584, Apr 2002.

[6] Stephanie Guindon and Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systems Biology*, 52(5):696–704, 2003.

[7] Sanderson MJ. r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2):301–302, Jan 2003.

[8] R. Thane Papke, Olga Zhaxybayeva, Edward J Fiel, Katrin Sommerfeld, Denise Muise, and W. Ford Doolittle. Searching for species in haloarchaea. *PNAS*, 104(35):14092–14097, 2007.

[9] Douglas L Theobald. A formal test of the theory of universal common ancestry. *Nature*, 465:219–222, 2010.