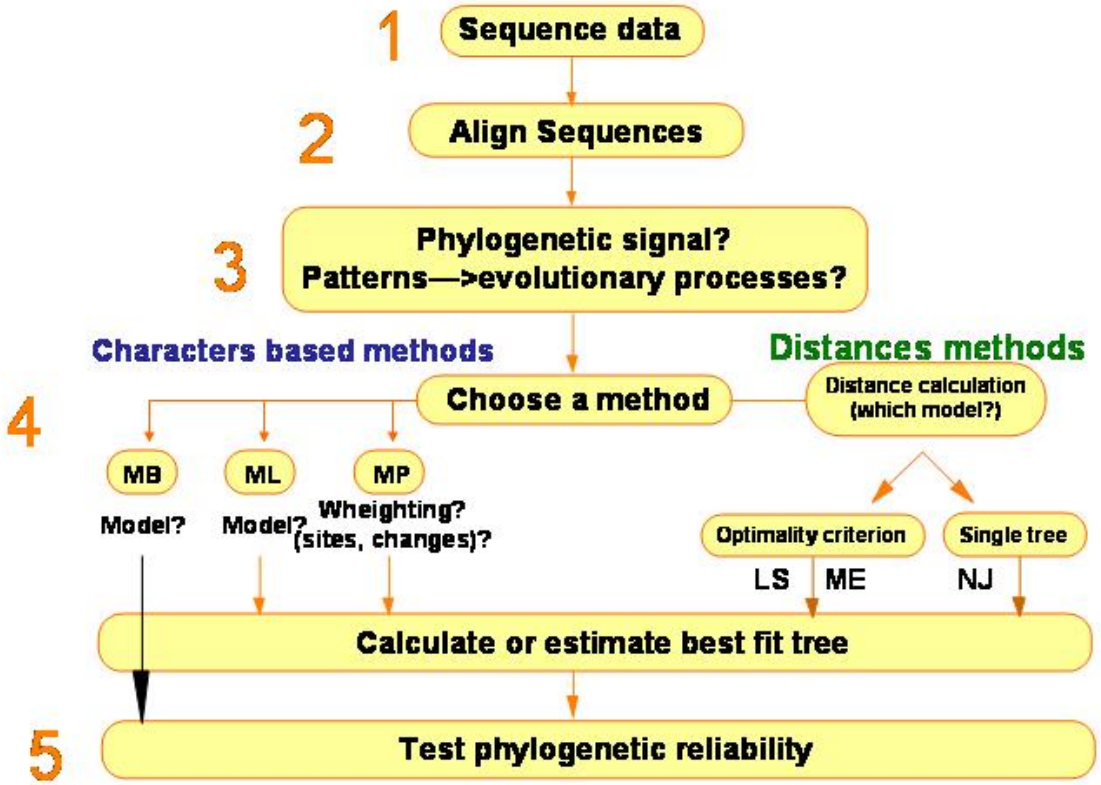
Stat 536

Molecular Evolution and Phylogenetics

Misha Rajaram

December 3, 2008

From DNA/protein sequences to trees



Modified from Hillis et al., (1993). Methods in Enzymology 224, 456-487

Recall

- Recall that in inferring phylogenies, we start with representation of data in the form of a multiple sequence alignment data matrix D.

Alpha	**A**	C	G	T
Beta	**G**	C	G	C
Gamma	**C**	C	A	A
Delta	**C**	C	G	T

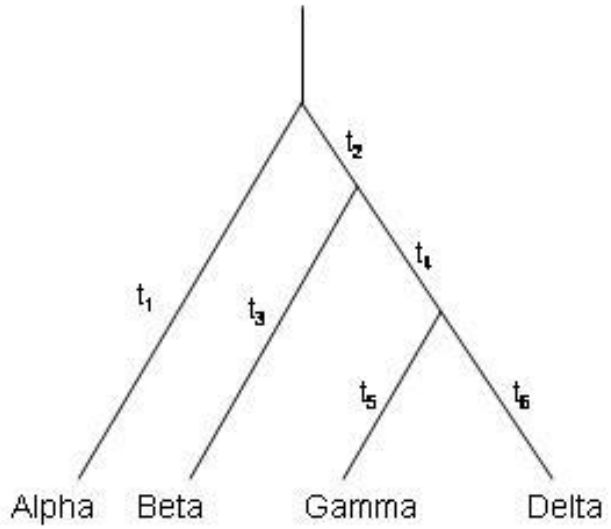
- In this example $D_{1,2} = A$ and so on.

Stat 536

Misha Rajaram

Recall

- Likelihood methods compute the likelihood of the relationship between these sequences (as presented by the data) to be summarized by a specific structure of the tree (topology), more importantly by specific branch lengths.



Recall

- We also use specific models of base substitution that have additional parameters.
- Suppose we use the JC69 model, we have a Q matrix

$$Q = \begin{pmatrix} - & \mu & \mu & \mu \\ \mu & - & \mu & \mu \\ \mu & \mu & - & \mu \\ \mu & \mu & \mu & - \end{pmatrix}$$

- The likelihood of a particular phylogeny (represented by a specific topology) then becomes

$$L(D|\tau, \mu)$$

Computing the likelihood of a tree

- Recall that given a tree with topology τ representing the relationship between the sequences of interest, the likelihood can be computed by assuming that
  - Sites are iid.
  - Lineages evolve independently conditional on the ancestral nucleotide.
- These assumptions help us compute the likelihood of the tree ($L(D|\tau, \mu)$), as the product of site-wise likelihood $L(D^{(i)}|\tau, \mu)$ which is in turn, a product of likelihood of observing particular nucleotides at each leaf node of the tree given the topology of the tree, summed over all possible states(nucleotides/amino acids) that the internal nodes can be in.

Using Gamma Rates

- Recall that rates of evolution can vary among sites in a nucleotide or protein sequence, so now there is a separate rate $\mu_1, \mu_2 \ldots \mu_m$ for a sequences with m sites.
- In computing distances we discussed the use of gamma distributed rates across sites. In principle, the same thing can be done for likelihood on trees. (Yang 1993)
- We independently generate rate μ_i from a gamma distribution function $f(\mu; \alpha)$. The overall likelihood would then be

$$L(D|\tau, \alpha) = \prod_{i \leq m} \int_0^\infty f(\mu_i; \alpha) L(D^{(i)}|\mu_i, \tau) d\mu_i$$

where $L(D^{(i)}|\tau, \mu_i)$ is the likelihood of this tree for site i given the rate of evolution at site i is μ_i.

- The main problem with this approach is that the Site-wise likelihood $L(D^{(i)}|\tau, \mu_i)$ needs to be computed for all possible values of μ_i , increasing the computational load many fold.
- Many alternatives have been suggested to get around this problem.
- The easiest came from Yang (1994). It was proposed to compute the site-wise likelihoods for a set of k rates $r_1, r_2 \ldots r_k$.
- The likelihood at the site i can then be approximated by a weighted sum

$$L(D^{(i)}|\tau, \mu_i) = \int_0^\infty f(\mu_i; \alpha) L(D^{(i)}|\tau, w_1, w_2 \ldots, w_k, r_1, r_2 \ldots r_k) d\mu_i \approx \sum_{l=1}^{k} w_l L(D^{(i)}|\tau, r_l)$$

- The weights w_l and points r_l need to be chosen to best approximation.
- The Yang 1994 paper uses rates corresponding to quantile points of the gamma distribution for a fixed k categories as r_k. For example, is $k = 10$ is chosen, then r values corresponding to every 10th quantile were used. All received equal weights of $w_l = 1/k$ so as to maximize the α parameter associated with the gamma distribution.

Stat 536

Misha Rajaram

Comparability of likelihoods

- Topology is a discrete parameter and there is controversy over what it means to compare likelihoods inferred on two different trees.
- Nei, Saitou and Li argue that different topologies are different hypotheses and therefore likelihoods computed on different trees are equivalent to likelihood being computed under different hypotheses.
- Counter arguments by Felsenstein and others state that the likelihood is the probability of the same event (data in this case).

- They are, therefore, on the same scale and one number being larger than another is indicative of probability of the data being higher under this topology.
- Assigning equal prior probabilities to both trees, these likelihoods become directly comparable.
- The conventional likelihood ratio test is however, not possible for this comparison since they are discrete and definitely not nested.
- Techniques like bootstrap are used to assess confidence in topology or branches in a topology.
- Note that LRTs are valid for comparison of nested base substitution models . For example JC69 is a subcase of K80 with a difference of 1 degree of freedom.

Bayesian inference of phylogenies

- Since as early as 1970, partly Bayesian inference of phylogenies have been proposed by various researchers.
- Most were prevented by computational difficulties from extending their methods to a fully Bayesian inference.
- In 1996, Rannala and Yang proposed the first fully Bayesian inference method.
- Recall that Bayesian inference of trees will involve sampling from the posterior distribution of the trees given by

$$P(\tau|D) \approx P(D|\tau)P(\tau)$$

- Bayesian inference will therefore need specification of a prior distribution $P(\tau)$ and computation of the likelihood $P(D|\tau)$.

Rannala and Yang's method

- They assumed a molecular clock with its inherent restriction on branch lengths.
- They used a birth-and-death prior on the trees which involves birth rate λ of new lineages and death rate μ of old branches.
- For a known, fixed interval of time t_1 since the start of the process, they inferred birth, death (of nodes) and substitution rates as well as the transition-transversion ratio for the HKY model of base substitution.
- The posterior probability of a single tree is the amount it contributed to the overall posterior probability.
- They used numerical integration to compute the posterior probabilities

$$P(\tau|D) = \int_Q \int_b P(\tau, b, Q|D) dQ db$$

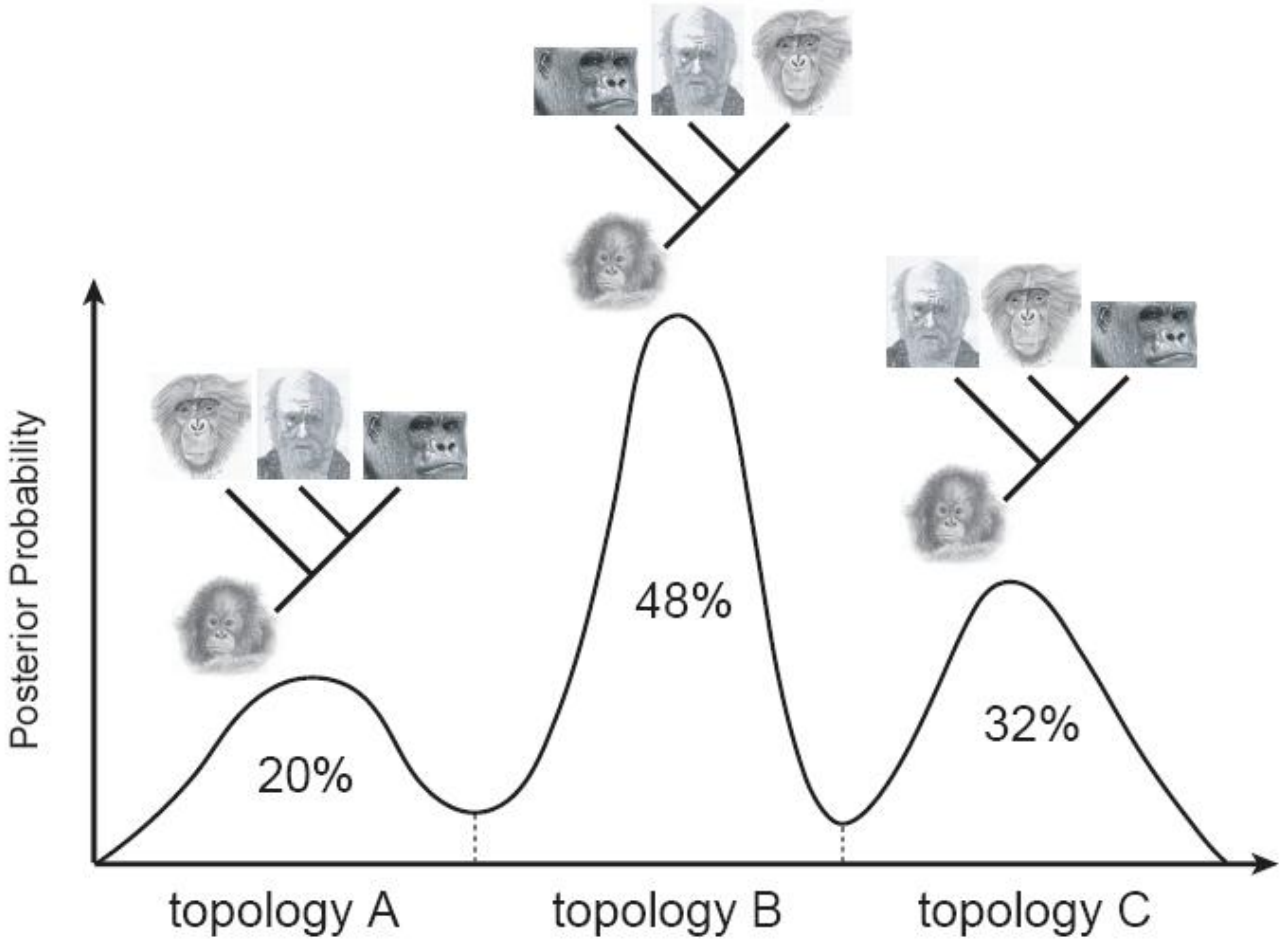
- With a large number of possible topologies and the need to integrate over many dimensions, this method becomes impractical for more than a few sequences at one time.

MCMC methods

- Markov Chain Monte Carlo methods are a class of algorithms that help sample from an arbitrary probability distribution by constructing a Markov chain whose stationary distribution is the desired distribution.
- In the context of tree topologies, it is helpful to think of all possible tree topologies residing in a tree space (may be as a connected graph).
- There is a distribution $f(\tau)$ associated with the topology vector τ that we want to be able to ultimately approximate.
- The idea of the MCMC is to randomly "wander" in the tree space till it is visiting each tree in proportion to $f(\tau)$, its probability.

Stat 536

Misha Rajaram



Bayesian MCMC

- Adapting the Metropolis algorithm to Bayesian inference involves, essentially, sampling from the posterior distribution

$$P(\tau|D) \propto P(D|\tau)P(\tau)$$

- To use, we know how to compute likelihoods $P(D|\tau)$. We need to specify priors $P(\tau)$.
- The Metropolis Algorithm involves designating the current tree τ_i and picking a new tree τ_j
- We then compute a ratio of probabilities ,R as :

$$R = \frac{P(\tau_j)P(D|\tau_j)}{P(\tau_i)P(D|\tau_i)}$$

- If $R \geq 1$, accept the τ_j as the current tree else generate a uniform random number $u \in (0, 1)$ and accept the proposal if $R \geq u$.
- Essentially in the inference of phylogenies we have three main concerns:
  - Prior distributions on trees
  - Proposal Distribution used for changes in phylogeny: Used for picking anew tree at each iteration.
  - Summarizing the posterior distribution

Priors

- Yang and Rannala used birth-and-death process like the one described earlier.
- For a fully Bayesian analysis they would have to also estimate birth and death rates for this process.
- They use empirical Bayes methods instead where they choose birth and death rates that maximize the sum of all posterior probabilities.
- Mau and Newton assumed that all possible tree topologies are equiprobable thus assigning a uniform or "flat" prior on the trees.
- Additionally they assume that for each topology, the placement of interior nodes also came from a uniform distribution.
- A third method places an exponential prior on the branch lengths.

Proposal Distributions

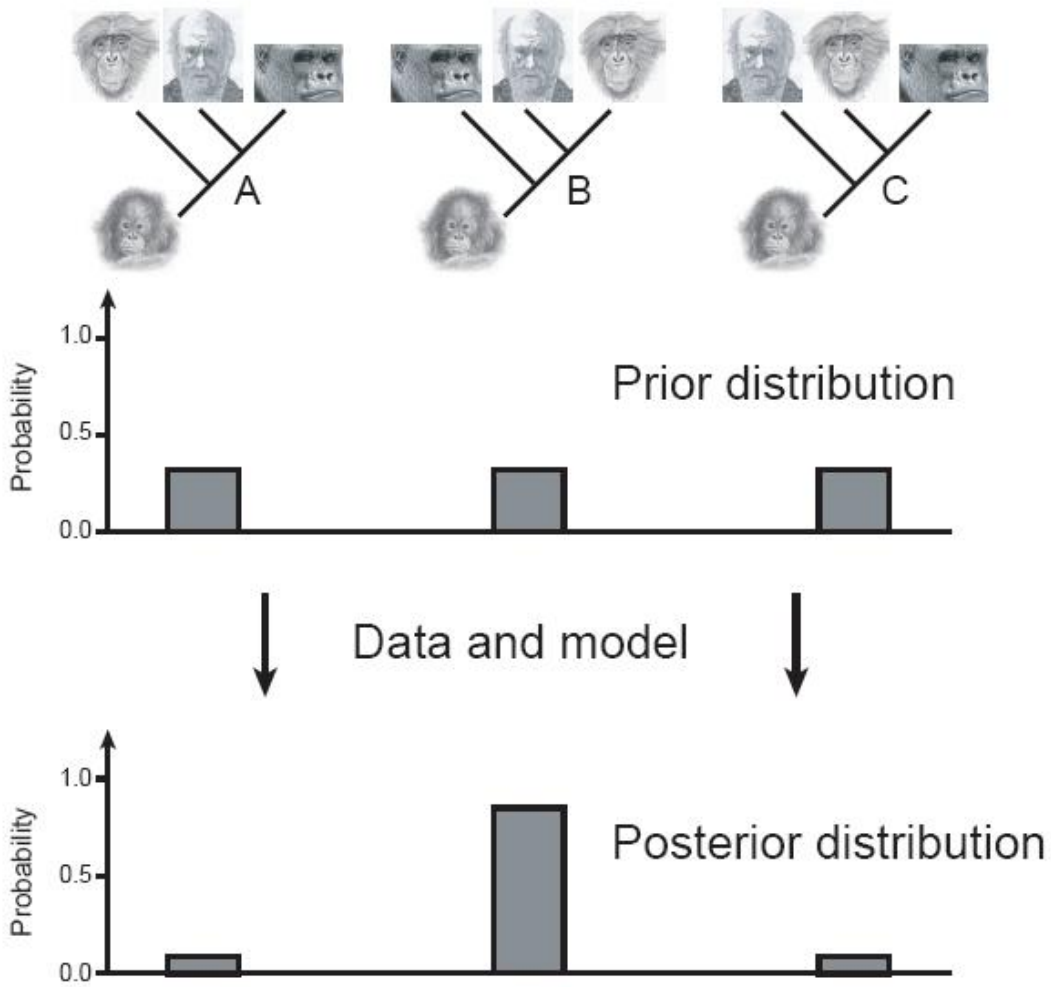
- These are distributions that give rise to the new proposal tree τ_j at each step i.e The transition probability of the MC $P(\tau_j|\tau_i)$.
- In principle any distribution can be used as long as all states communicate.
- In practice, it is not such an easy problem. A proposal distribution that proposes trees that are radically different from the current tree will end up with most trees rejected.
- A proposal distribution that moves very slowly may not adequately explore the tree space.
- Choice of Prior and proposal distributions are currently more a matter of practicality than much else with debates still on , on what the most appropriate distribution for either is.

Computing Likelihoods

- Recall that likelihoods are products of likelihoods at individual nodes of a tree for each site.
- When proposal trees are very close in structure to the current tree i.e. most of the tree is the same,computation time of likelihood can be reduced by exploiting the similarity in topology.
- We reuse conditional likelihoods for the part of the tree similar to the old tree , adding only computation from the changed parts.

Summarizing the Posterior

- Once we have a good sample of the trees from the posterior distribution, how do we best summarize the information they hold?
- One solution is to use clade probabilities.
- Recall that taxa that get placed in groups with a shared most recent common ancestor (MRCA) are called monophyletic groups or clades.
- For a particular clade of interest, we can sum the posterior probabilities of all the trees that contain that clade.


A
B
C
1.0
0.5
0.0
Probability
Prior distribution
Data and model
1.0
0.5
0.0
Probability
Posterior distribution

Testing Phylogenetic Hypotheses

- Hypotheses on a phylogeny can arise as a matter of wanting to test the reliability of a phylogeny produced from a method to then draw biologically relevant relatedness relationships with some certainty.
- Three most popular tools will be reviewed
  - Likelihood Ratio Tests
  - AIC and BIC
  - Bootstrap methods

Stat 536

Misha Rajaram

Likelihood and Tests

- Recall that nested models can be tested using Likelihood Ratios that have nice asymptotic properties that allow us to use them as test statistics.
- Suppose we have a maximum likelihood estimate from our data, of p parameters such as branch lengths in a phylogeny represented in vector form as $\hat{\theta} = (\theta_1, \theta_2 \ldots . \theta_p)$.
- Suppose also,that the true value of the parameters is θ_{null}
- Then the ratio of likelihoods under the two parameter vectors

$$2[lnL(\hat{\theta}) - lnL(\theta_{null})] \sim \chi^2_q$$

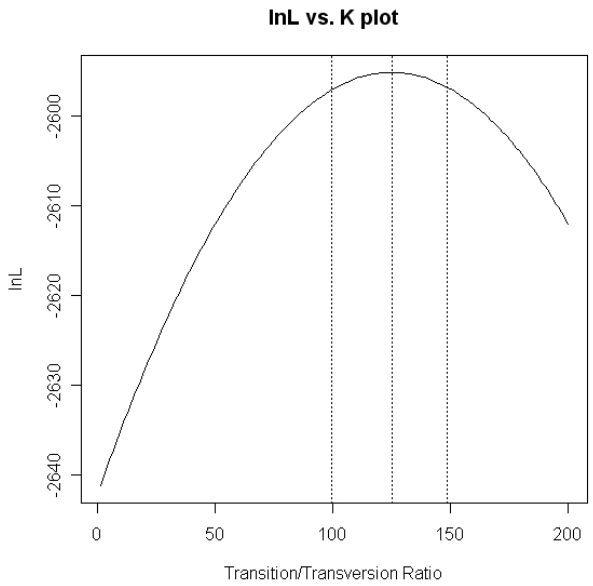
where q is the number of free parameters.

Stat 536

Misha Rajaram

Using LRTs for Interval Estimates

- LRTs can also be used to construct interval that may contain the true value.
- This is achieved by first finding the maximum likelihood estimate and then finding all values that cannot be rejected compared to it.
- Suppose we have a set of DNA sequences and a model of base substitution such as HKY85 model.
- For this model we have three parameters for base frequencies and a transition/transversion ratio κ.
- Suppose that we compute the Ln L values for different values of κ and find a maximum at $\kappa = 125$ with corresponding Ln L -2593.205



- Suppose that we compute the Ln L values for different values of κ and find a maximum at $\kappa = 125$ with corresponding Ln L -2593.205
- We are only fixing the value of one parameter thus the LRT has an asymptotic χ^2_1 distribution.
- We therefore reject all points that have a Ln less than $-2593.205 - \chi^2_1(0.95)/2$.
- This yields $(99.5, 148.5)$ as the 95% interval for κ.

Choosing among non-nested hypothesis

- As models get more and more complex, LRTs may not be the best solution to compare them.
- It will always be the case that a more general model will have higher likelihood than a restricted subcase.
- Choosing the model with highest likelihood may lead to an unnecessarily complex model.
- We can use standard measures like AIC and BIC to compare models.

AIC and BIC

- AIC is the Akaike Information Criterion and is computed as

$$AIC = -2lnL + 2p$$

- BIC is Bayesian Information Criterion and is computed as

$$BIC = -2lmL + pln(n)$$

where p is the number of parameters and n is the sample size.

AIC and BIC

- Notice that both penalize for added parameters. In BIC this penalty is dependent on the sample size.
- The model with lowest AIC or BIC is the best.
- AIC is more frequently used in literature.
- Completely resolved tree topologies all have the same number of parameters (branch lengths) and hence choosing the topology with the best AIC value is the same as choosing the one with the highest likelihood.
- The ModelTest software (Posada et al) is particularly useful for such comparisons.

Stat 536

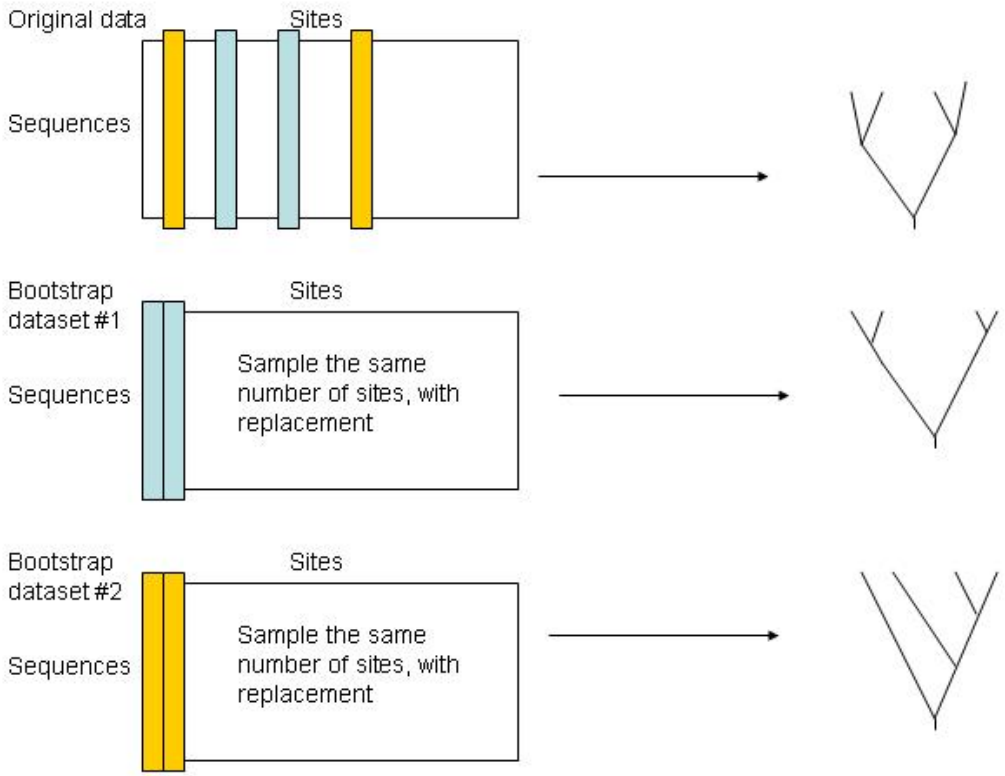
Misha Rajaram

Bootstrap Methods

- Bootstrap involves resampling from one's sample with replacement to make a pseudo sample of the same size.
- To use the bootstrap to assess the uncertainty of our estimate of the phylogeny, the data should be a series of independently sampled points.
- We typically have a matrix of species and characters.
- Species can not be considered independent sampled.Characters, however, can.
- We sample whole characters from the set of n characters, with replacement and we do so n times. The result is a new data matrix with the same size as the original.

Stat 536

Misha Rajaram



- Once we have resampled our dataset, we have a cloud of trees, estimated from the bootstrap samples.
- For a real-valued parameter, we can simply make a histogram of values and compute an interval.
- How do we do this for phylogenies?
- We could do something analogous for branch lengths. If a branch is present in all the bootstrap samples, we could construct the histogram for its length and a 95% interval for its true value. If this interval does not include zero, we can be sure that the branch is real.
- It is rare, though, for a single branch to appear in every bootstrap tree.
- Alternate methods include doing the same thing for the fraction of bootstrap trees that have the branch and considering the ones that have the branch missing , to have a zero branch length for that branch.

- These methods require one, to scan through all the trees and this can become tedious if there is a large number of branches.
- Consensus tree methods can come to the rescue.
- Usually, a majority-rule consensus tree is found by tabulating all groups that occur in all trees and retaining those that occur in a majority of trees.
- The single resultant tree has groups that appear in a majority of trees.
- Additionally, we could note, next to each branch, in what fraction of bootstrap replicates it appeared.

Stat 536

Misha Rajaram

Criticisms of Bootstrap

- The most telling criticism is the assumption of independence of characters may not always be true.
- Imagine the case where pairs of characters are identical i.e. we have, by chance collected two characters that provide the exact same phylogenetic information.
- Suppose that we have done this so often that each character now has an identical partner in our data.
- To counter this, block bootstrap methods have been suggested.
- Block bootstrap method suggests drawing blocks of data whose start site is random.
- Instead of drawing n characters , we now draw n/B blocks at B sites.

Dealing with invariant characters

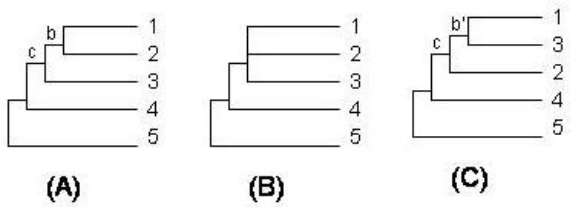
- Bootstrap is argued to be particularly sensitive to invariant characters included in the data set.
- Using methods such as parsimony will produce significantly different bootstrap values by omitting the invariant characters.
- Consider a single variant site. For a total of N sites, this site has a probability $1/N$ of being chosen in a bootstrap replicate.
- Conversely it has a probability of $(1 - 1/N)^N$ of being entirely left out.
- It can be shown that adding M invariants does result in a decrease of inclusion probability but Harshman (1994) showed that this decrease is not very much.
- In fact it stabilizes to almost a constant at $e^{-1} - 0.36788$ and is insensitive to the number of invariants.

Testing Internal branches

- Sometimes we are less interested in whole phylogenies and more interested in a part.
- We can compute, using various methods, the variance on the estimate of length of a branch in the interior of a tree.
- Approximations help expedite computation by not considering the total topology of the tree.
- The argument is that it the length of the branch is significantly different from zero, then the branch must be regarded as real.
- We can , however, do some standard tests for internal branches.

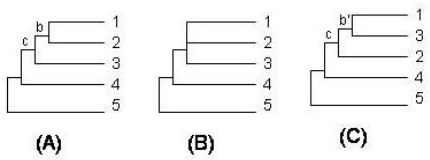
Normal Deviate (Z) Test

- This method was developed for trees constructed by distance methods.
- Consider the following example



- For 5 taxa, 15 unrooted topologies are possible, each with 5 leaf nodes.

- Suppose that topology A is correct. It can then be shown that the estimate of branch lengths in topology A are all positive or zero.
- Additionally, at least one branch in each of the other 14 topologies will have a negative estimated branch length.
- This branch gives rise to a incorrect partition of the sequences.
- This is true for any number of sequences as long as distance estimates are unbiased.
- Thus, if a tree have at least one branch whose length estimate is significantly negative, then the topology of the tree is likely wrong.

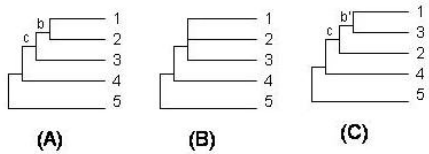


- Suppose that topology A was obtained by distance matrix methods and can be considered reliable if internal branches b and c are shown to be real.
- The null hypothesis then is that $b \leq 0$ and $c \leq 0$.
- We can test this by computing the standard error of the estimate $\hat{b}$,$s[(\hat{b})]$.
- For sufficiently large number of substitutions, $\hat{b}$ follows a normal distribution.
- Using this we can now perform a one-tailed test for the normal deviate $Z = \frac{\hat{b}}{s[(\hat{b})]}$.

Stat 536

Misha Rajaram

- In phylogenetic analysis, it is often very important to test whether or not a group of taxa is monophyletic.
- This is equivalent to testing if the branch leading to the group is real.
- We may also want to test that two very similar topologies are, in fact, same.
- Consider trees A and C.



- Testing that they are same is equivalent to testing that branches b and b' are both equal to 0.
- If this is true then it leads to the topology exhibited by tree B, which is therefore called the *null tree*.
- In the case of Parsimony or Likelihood methods, all branch lengths are necessarily positive so it is difficult to develop a suitable test for the null hypothesis.
- Alternately, a **confidence probability** (P_C) can be computed that $b > 0$ using the Z test.
- If $P_C > 95\%$ or 99%, branch b is considered real.

Likelihood Ratio Test

- Specific scenarios can be tested using a Likelihood Ratio Test.
- For example the hypotheses $b = 0$ can be considered to be nested within the hypothesis $b > 0$.
- In the example before, tree B is a special case of tree A.
- However, using unresolved trees in the hypothesis may lead to violation of asymptotic assumptions of the test since changing branch lengths leads to changing discrete parameter τ and 0 branch lengths are on the boundary of the parameter space.

Internal Branch Bootstrap tests

- As in the case of the bootstrap test, the dataset is resampled and bootstrap replicates are obtained.
- The lengths of all branches are estimated using a given method, for the same topology as the original one but using the bootstrap replicate dataset.
- This is repeated several times for the same topology.
- The length estimate for a branch ($\hat{b}$) will then vary from replicate to replicate, even become negative sometimes.
- We can compute the mean and standard error for $\hat{b}$ and conduct a Z test.

Models with molecular clock

- There are two forms of the molecular clock hypothesis.
- One asserts that all lineages have the same rate of evolution which does not change over time (strict clock).
- The other, less restrictive version asserts that all lineages have the same rate of evolution but allows them to change with time, as long as the changes apply simultaneously to all lineages.
- Unless we have sampled individuals at different times, it is difficult to distinguish between the two forms of the molecular clock.

Stat 536

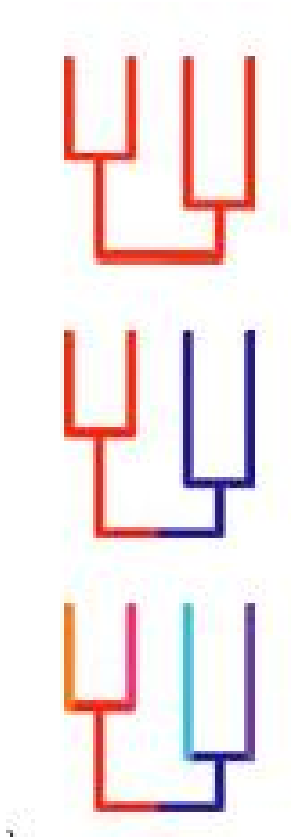
Misha Rajaram

Models with clocks

- So far we have allowed construction of trees with any non-negative branch length.
- This implies that the tree may or may not have clock-like behavior.
- In order to find a maximum likelihood tree under a molecular clock, we must place a set of constraints on the growing tree such that each of the tips/leaves is equidistant from the root.

Types of molecular clocks

- Global/Strict clocks have a single rate of evolution enforced across all branches of the tree.
- Local Clocks : Local regions within a tree follow different rates of evolution.
- Relaxed Clocks: Rate allowed to vary among branches
  - Autocorrelated relaxed clocks: Adjacent branches have related rates
  - Uncorrelated relaxed clocks: Rates are iid across branches.



Relaxing Molecular Clocks

- The notion of molecular clocks has been controversial right from the beginning.
- For the species level analyses it may be reasonable to assume a molecular clock should we find a gene that seems to evolve more or less neutrally.
- The concern is how much beyond the species levels can be successfully carry this idea on before it starts to break down seriously?
- A number of researchers have successfully developed methods using relaxed molecular clocks, ones whose rate of evolution is slow enough to still enable them to be useful.

Models with Autocorrelated relaxed clocks

- Autocorrelated relaxed clocks treat the rate also as a heritable trait that "evolves" along the tree .
- These models assume that rate is tied to Life history traits, DNA proof reading mechanism, cellular environment and so on.
- Specifying a lognormal prior on the rates of evolution produce autocorrelated relaxed clock models. An autocorrelation parameter ν constrains the standard deviation of the distribution i.e how much rate r_j can depart from initial rate r_0.

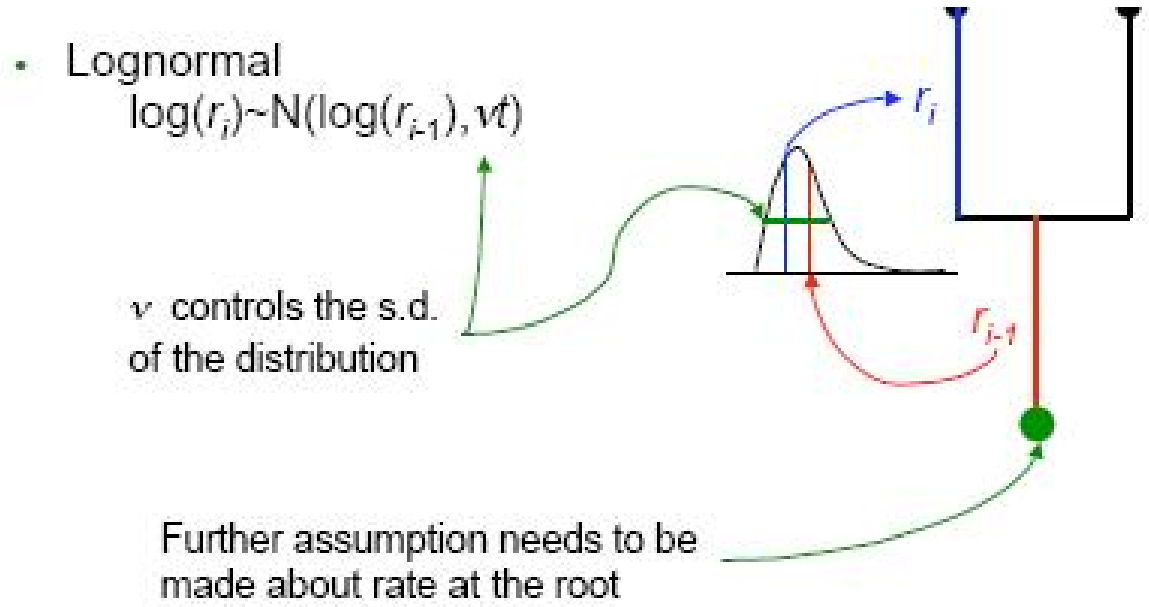


Figure from Simon

Ho's notes

Stat 536

Misha Rajaram

Models with Uncorrelated relaxed clocks

- Lognormal or Exponential distributions can be used.
- Rates tend to cluster around the mean of the distribution.

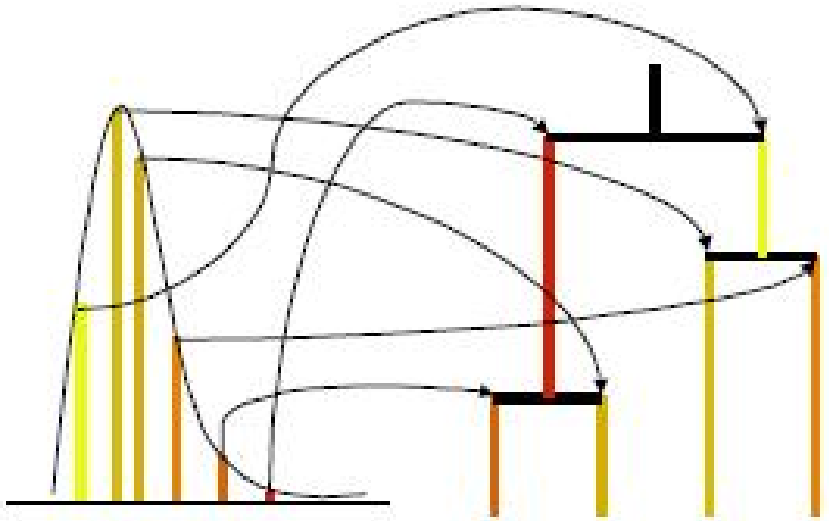


Figure from Simon Ho's notes

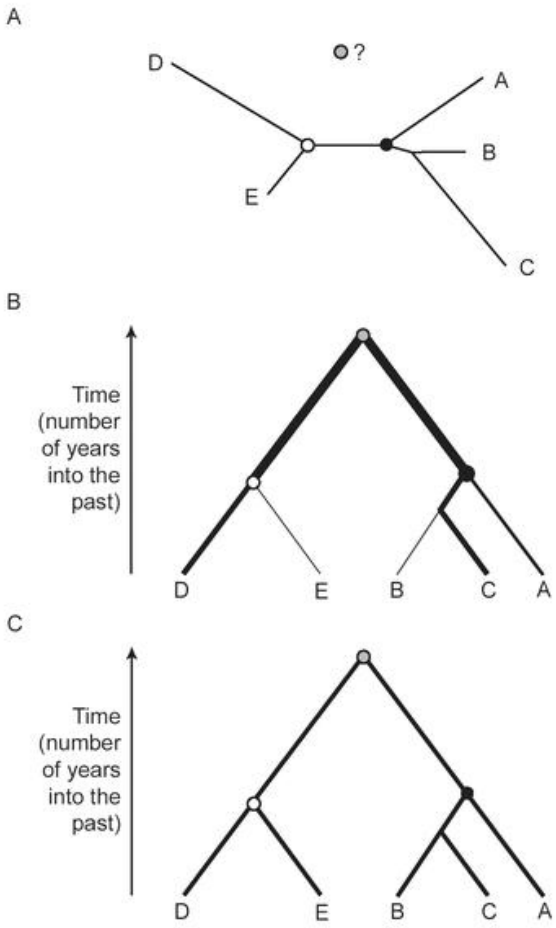


Figure from Pybus OG (2006) Model Selection and the Molecular Clock. PLoS Biol 4(5)

Testing the Molecular clock - Parsimony based methods

- The first attempt was made by Langley and Fitch (1973-74).
- Using amino acid sequences, they assigned replacements to branches in a known tree i.e. assigned ancestors.
- They then used these as if they were known data and estimated branch length by maximum likelihood.
- They then constructed a chi-square test of whether the number of substitutions on each branch were proportional to these clock like branch lengths.
- This test relies on the accuracy of parsimony reconstruction , however and is therefore limited by it.

Testing the Molecular clock- Distance Based Methods

- One of the ways of fitting a distance tree is by a least squares method.
- Say you have a distance matrix D.
- Any particular tree leads to a set of predicted distances $\hat{D}_{ij}$.
- The least squares method seeks to solve for branch lengths so as to minimize

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(D_{ij} - \hat{D}_{ij})^2$$

Testing the Molecular clock- Distance Based Methods

- One way to test the molecular clock hypothesis is to quantify the increase in least squares when a clock is assumed over the least squares when no clock is present.
- This can be done when the two tree topologies are the same.
- The test is an F ratio with $n-1$ and $(n-2)(n-3)/2$ degrees of freedom when a triangular distance matrix is used.
- The tests assumes independence of the distances, which may not always be true.

Testing the Molecular clock- Likelihood based Methods

- The likelihood ratio test can be used to test the molecular clock.
- Suppose that we estimate a phylogeny under the molecular clock and also without it.
- Suppose further that these turn out to be the same unrooted tree topology.
- For n sequences, the clock-like tree is specified by knowing the branch lengths from the leaf nodes to each of the $n-1$ internal nodes.
- The Full tree without the clock is specified by $2n-3$ branch lengths.
- The LRT thus has $(2n-3)-(n-1)=n-2$ degrees of freedom.

Stat 536

Misha Rajaram

Relative Rates Test

- Introduced by Sarich and Wilson and later developed as a statistical test by Wu and Li.
- For a tree of 3 species A,B and C, the number of sites in which A and B differ and B and C differ is counted.
- The variance of the difference between differences can be approximated.
- The statistical test of the difference of differences (D) being significantly different from zero can now be conducted by approximating D to a normal distribution.
- If the test rejects the null hypothesis of $D = 0$, it suggests different rates of evolution in the two lineages.

Two cluster Tests

- This is an extension of the relative rates test and can be used to test the molecular clock hypothesis for a tree topology obtained without making the assumption of rate constancy.
- The principle is to test whether or not the average substitution rates for two clusters of sequences created by a node in the given tree are the same.
- Let us consider a tree with 3 clusters A, B and C and let b_A and b_B be estimates of the average number of substitutions per site (distance) from a node N to the leaves of clusters A and B respectively.
- Under the assumption of rate constancy, the expectation of difference (D) between b_A and b_B is zero.

- Let L_{AB} be the average distance between clusters A and B. Therefore,
- Similarly we define L_{AC} and L_{BC} as distances between clusters A and C and B and C respectively.
- Using these b_A and b_B can now be estimated as
  $b_A = (L_{AB} + L_{AC} - L_{BC})/2$ and
  $b_B = (L_{AB} + L_{BC} - L_{AC})/2$
- Difference D can now be computed as
  $D = b_A - b_B = L_{AC} - L_{BC}$
- Variance can also be computed and a Z test can be done.

Stat 536

Misha Rajaram

Building Consensus Trees

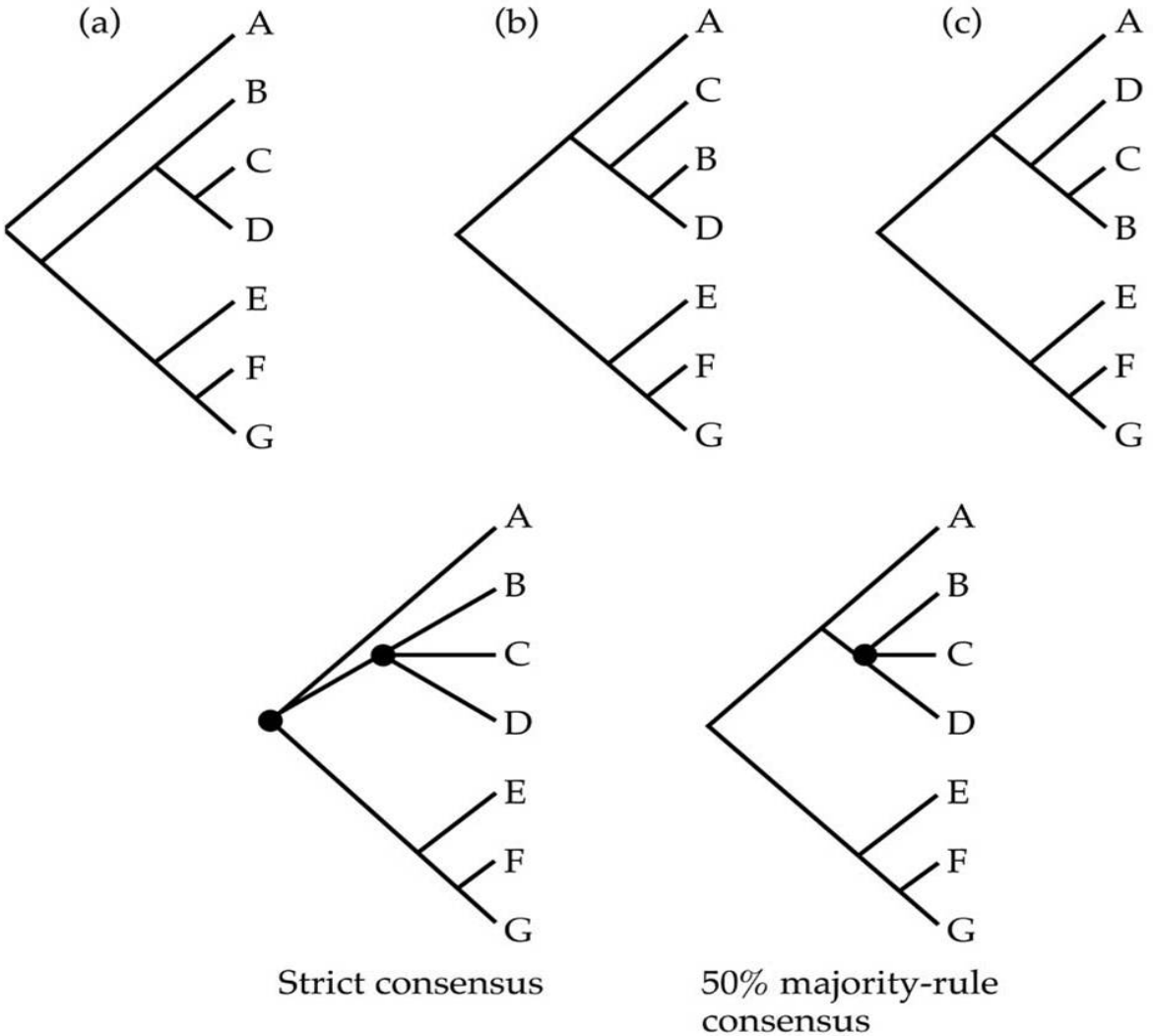
- Trees that are inferred from the data are called fundamental trees.
- Often, methods yield more than one best fundamental trees.
- We can use Consensus trees to summarize the phylogenetic information contained in the fundamental trees.
- There are three main types of consensus trees based on how much agreement within the fundamental trees they need, to include the information as a consensus information.

Types of consensus trees

- Strict consensus- A frequency based method. Only monophyletic groups found in all source trees are found in the resultant tree. The tree excludes a subset of all possible trees and conversely includes a subset of possible trees, whether or not they are part of the source set. In some sense the most conservative consensus.
- Semi-strict- A frequency based method- Only monophyletic groups found in at least one of the source trees and compatible (not in conflict) with all other source trees are found in the resultant tree, i.e. if a clade is never contradicted, but not always supported, then it is still included in this compromise tree.
- Majority-rule- Again, a frequency based method. Shows groups that appear on pre-specified percentage of source trees, usually $> 50\%$. Used for summary of searches where plurality is important. Can result in a tree that contains two groups that are simultaneously found in only one of the source trees

Stat 536

Misha Rajaram



From Dr.Graur's lectures on consensus trees.

Other methods of consensus

- Greedy consensus. Frequency based method. Groups ordered by frequency like in Majority-rule, then added in to the consensus tree as long as they are compatible.
- Adams . An intersection method. Inconsistently placed taxa are moved to the first node that summarizes the possible topologies. Groups can appear in Adams consensus that are not found in any source tree. Adams trees have no biological or phylogenetic interpretation. They do point to .wildcard. taxa. Those taxa may be experimentally removed from the matrix and the resulting analysis compared to when they are included.
- Matrix representation with parsimony (MRP). A recoding consensus method that can be used for trees with different sets of taxa. Both topology and frequency are important.