Phylogenetics:
Building Phylogenetic Trees

COMP 571

Luay Nakhleh, Rice University

Four Questions Need to be Answered

* What data should we use?
* Which method should we use?
* Which evolutionary model should we use?
* Which test should we use to assess the robustness of the prediction of particular tree features?

Desired Properties of the Data Used in a (Species) Phylogeny Reconstruction

* As we discussed before, reconstructing a tree that reflects the evolutionary history of a set of species is a hard task, and great care must be taken in the choice of the data used
* An ideal choice is a genomic region that appears exactly once in every species and whose evolutionary history is "identical" to that of the species
* The region should have little, if any, traces of HGT
* The rate of change in the region should be fast enough to distinguish between closely related species, but not too fast that regions from very distantly related species cannot be reliably aligned

Small Ribosomal Subunit rRNA

* The DNA sequence specifying the small ribosomal subunit rRNA (called 16S RNA in prokaryotes) has been found to be one of the best genomic segments for this type of analysis, despite occurring in several copies in some genomes

* We'll later illustrate the reconstruction of the evolutionary history of 38 bacterial species from the Proteobacteria phylum using the 16S RNA sequence

Choice of a Method

* Many computational methods exist for phylogeny reconstruction
* These can be divided into two categories: distance-based methods and sequence-based methods
* Distance-based methods first compute pairwise distances from the sequences, and then use these distances to obtain the tree
* Sequence-based methods use the sequence alignment directly, and usually search the tree space using an optimality criterion that is defined on the columns of the alignment

Choice of a Method

* Examples of distance-based methods include UPGMA (unweighted pair-group method using arithmetic averages), NJ (neighbor joining), and Fitch-Margoliash

* Examples of sequence-based approaches include maximum parsimony (MP), maximum likelihood (ML), and Bayesian inference

Properties of Methods

Method	distances?	tree type	single tree?	tree score?	tree test?
UPGMA	yes	ultrametric	yes	no	no
NJ	yes	additive	yes	no	no
Fitch-Margoliash	yes	additive	yes	no	no
Minimum evolution	yes	additive	no	yes	yes
MP	no	additive	no	yes	yes
ML	no	additive	no	yes	yes
Bayesian	no	additive	no	yes	yes

Choice of a Model of Evolution

* As we discussed before, the p-distance is usually an underestimate of the true evolutionary distance

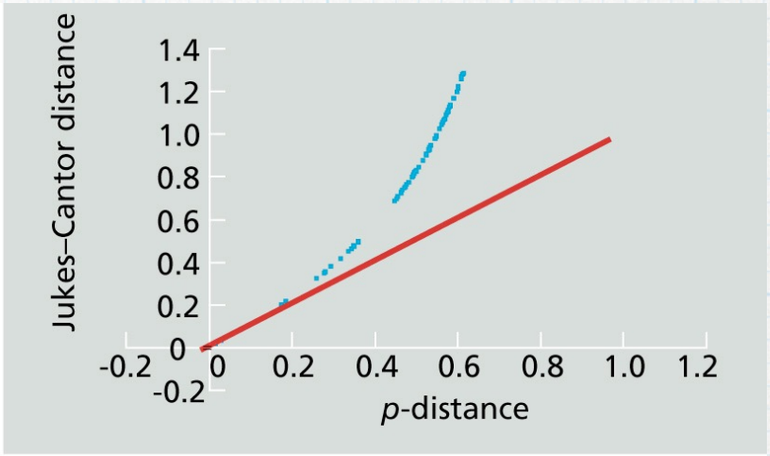
* Therefore, a correction of the p-distance is necessary for phylogenetic methods

* Models of evolution can be used to derive such distance corrections (we'll see some of the formulas later)

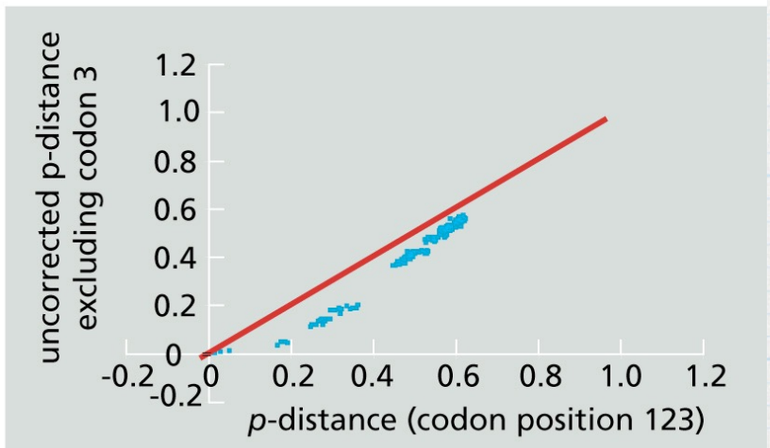
Choice of a Model of Evolution

Correlation between corrected and uncorrected distances

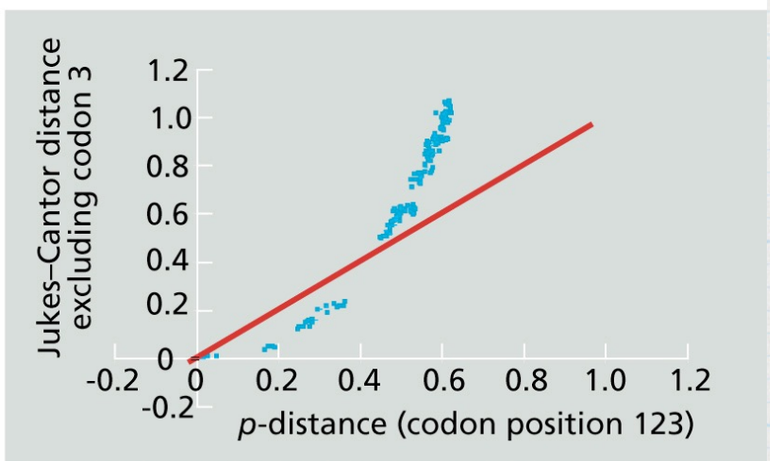
(A)



(B)



(C)



Choice of a Model of Evolution

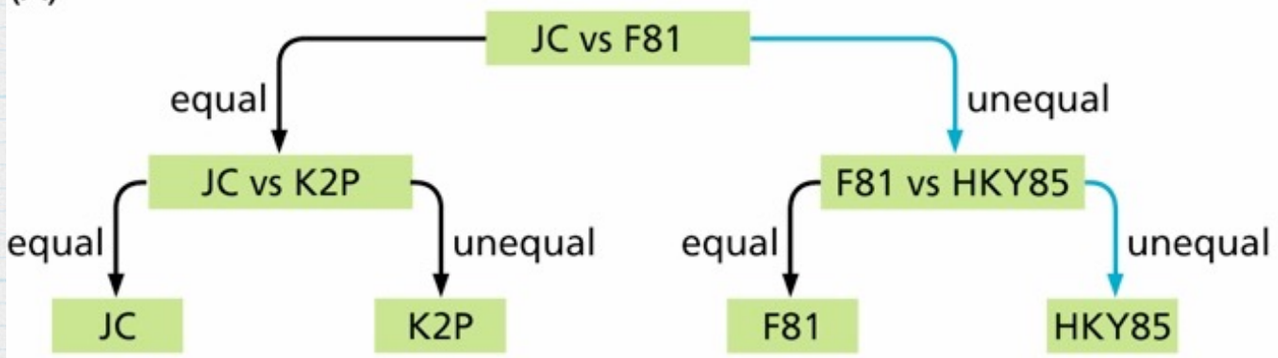
Model	Base composition	R=1?	Identical transition rates?	Identical transversion rates?	Reference
JC	1:1:1:1	no	yes	yes	Jukes and Cantor (1969)
F81	variable	no	yes	yes	Felsenstein (1981)
K2P	1:1:1:1	yes	yes	yes	Kimura (1980)
HKY85	variable	yes	no	no	Hasegawa et al. (1985)
TN	variable	yes	no	yes	Tamura and Nei (1993)
K3P	variable	yes	no	yes	Kimura (1981)
SYM	1:1:1:1	yes	no	no	Zharkikh (1994)
GTR	variable	yes	no	no	Rodriguez et al. (1990)

Choice of a Model of Evolution

* Choosing the evolutionary model is not a simple task
* One approach to choose a model is to try several of them and select the tree that best fits the data (doesn't work for methods such as NJ that don't provide a way to test goodness of fit)
* A way of comparing two evolutionary models has been proposed, called the **hierarchical likelihood ratio test** (hLRT), which can be used when one of the models is contained in the other
* In this method, both models must be assessed with the same tree topology
* Using the ML method for each evolutionary model in turn, the tree branch lengths are found that give the highest likelihood of the trees
* These two likelihood values can then be compared using hLRT, which is a chi-squared test based on the difference in likelihoods and the differences in numbers of parameters in the two models

Choice of a Model of Evolution

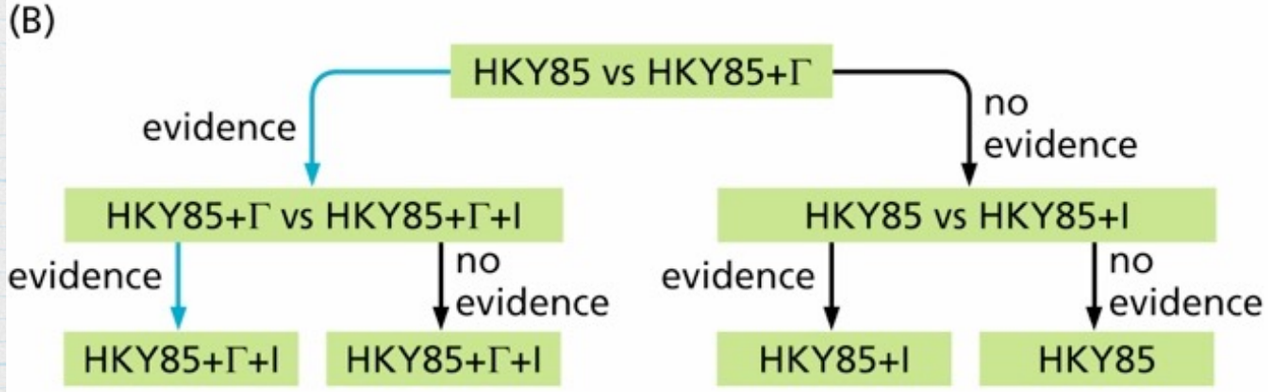
(A)



check for equal base frequencies

check for equal transition and transversion rates

(B)



check for evidence of rate variation at different sites

check for evidence of invariant class

(C)

Checking for equal base frequencies:

H_0: JC	Log-likelihood = -19864.051	17 parameters
H_1 : F81	Log-likelihood = -19859.027	20 parameters

Likelihood ratio statistic:	10.048
Degrees of freedom:	3
Probability:	0.018156 (< **0.05**)

Null hypothesis H_0 rejected: F81 chosen

Choice of a Model of Evolution

* Two problems with the hLRT test are (1) the requirement that models are nested, and (2) the decision on the order of testing and significance levels to use when deciding among many models

* Two other tests that have been proposed are:

* The **Akaike information criterion** (AIC), which is defined as $-2 \ln L_i + 2p_i$, where L_i is the maximum likelihood value of the optimal tree obtained using model i in a calculation which has p_i parameters (including the branch lengths as parameters)

* The **Bayesian information criterion** (BIC), which is defined as $-2 \ln L_i + p_i \ln(N)$, where N is the size of the data set

Choice of a Test of Tree Features

* One of the most commonly used tests for the reliability of the tree topology is the bootstrap
* It has been shown, however, that bootstrap values may be conservative
* The bootstrap measures the degree of support within the data for the particular branch, given the evolutionary model and tree reconstruction method
* The bootstrap values give no indication of the robustness of these features to changing the model or method

Two Examples

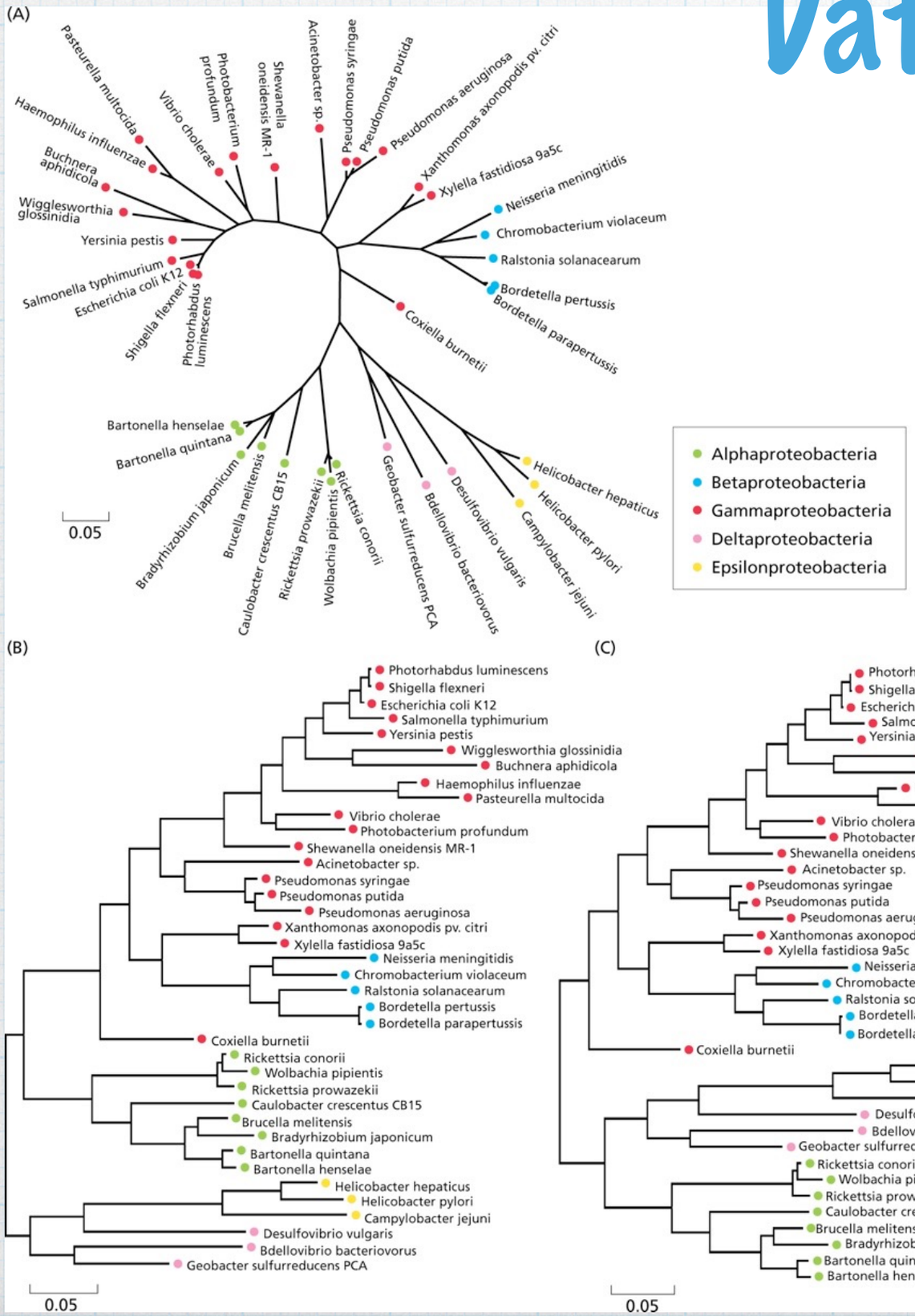
* Phylogenetic analysis of a data set of 38 species of bacteria from the phylum Proteobacteria

* A gene tree for a superfamily of enzymes

Analysis of the Proteobacteria Data Set

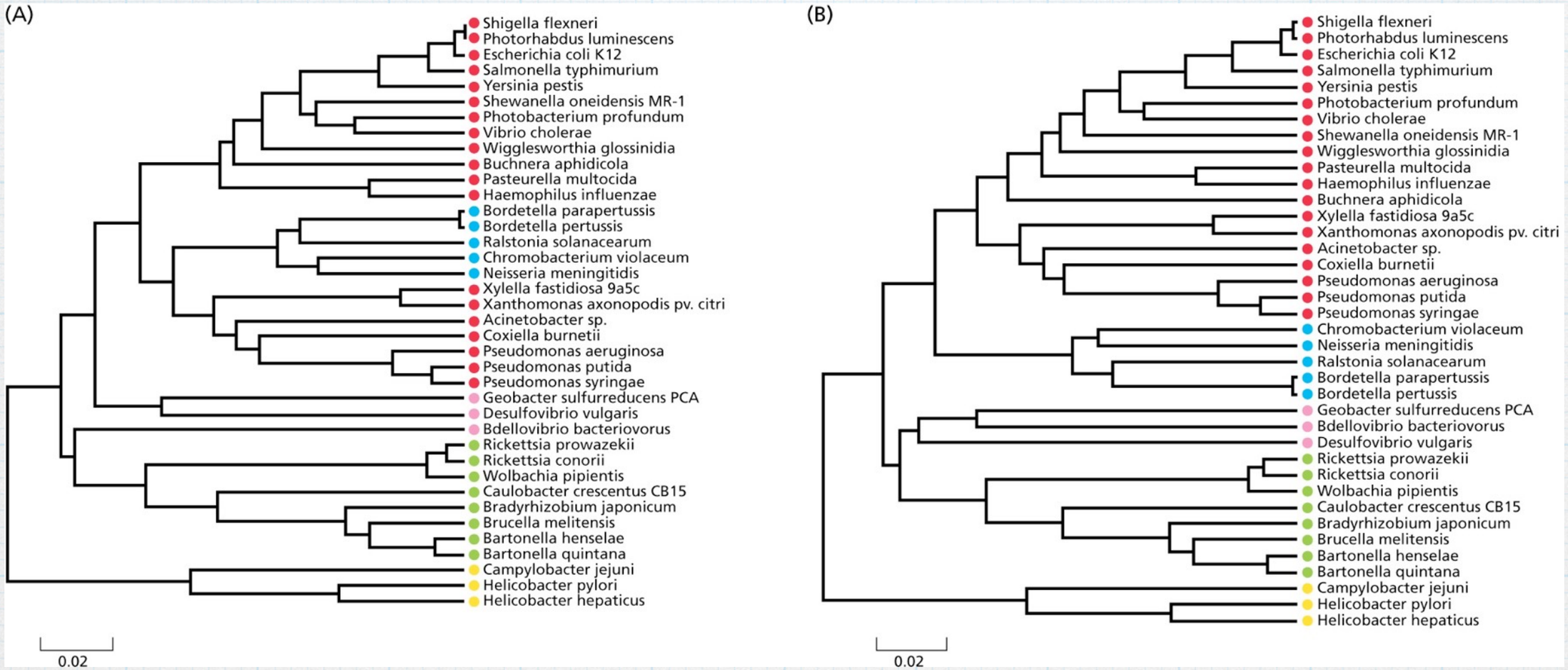
* Data: 16S RNA sequences from all 38 species
* Using the AIC method, as implemented in MrAIC, the GTR+Gamma was found to be the most appropriate model
* Using GTR+Gamma, an unrooted ML tree was generated using the tool PHYML

Analysis of the Proteobacteria Data Set



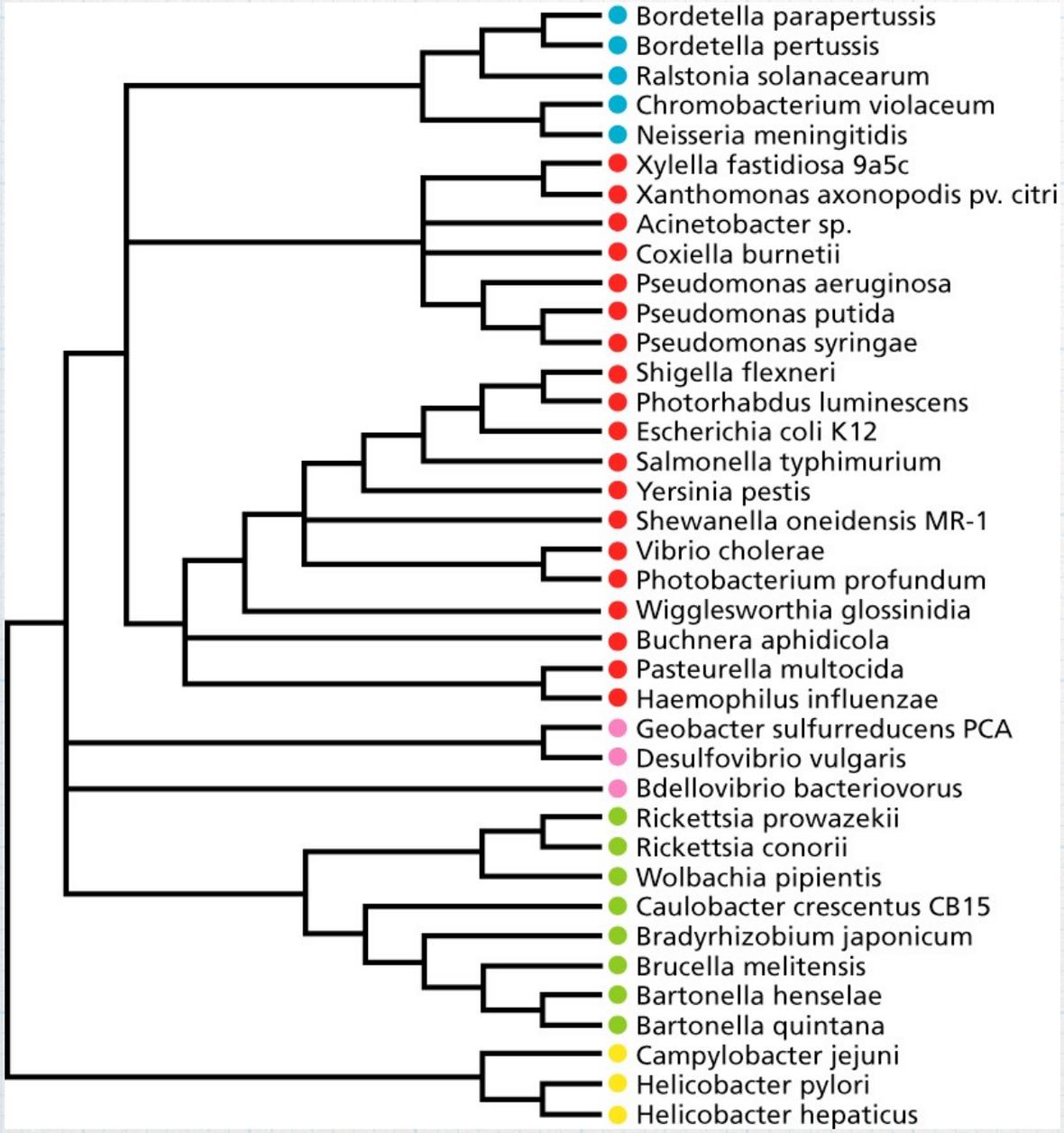
(A) the unrooted tree
(B) the same tree rooted so that the Deltaproteobacteria and Epsilonproteobacteria diverge from the other classes first
(C) the same tree rooted using the midpoint method

Analysis of the Proteobacteria Data Set



UPGMA trees reconstructed from the same 16S RNA data set with JC correction.
(A) Alignment columns that have at least one gap are ignored
(B) Alignment columns are only ignored when one of the two sequences being compared has a gap

Analysis of the Proteobacteria Data Set



The condensed tree obtained by a bootstrap analysis of the tree in (A) on the previous slide.
Branches with bootstrap values below 75% have been contracted

Analysis of the TDP-dependent Enzyme Superfamily

* An example of a phylogenetic analysis of an enzyme superfamily

* This analysis would address two important issues:

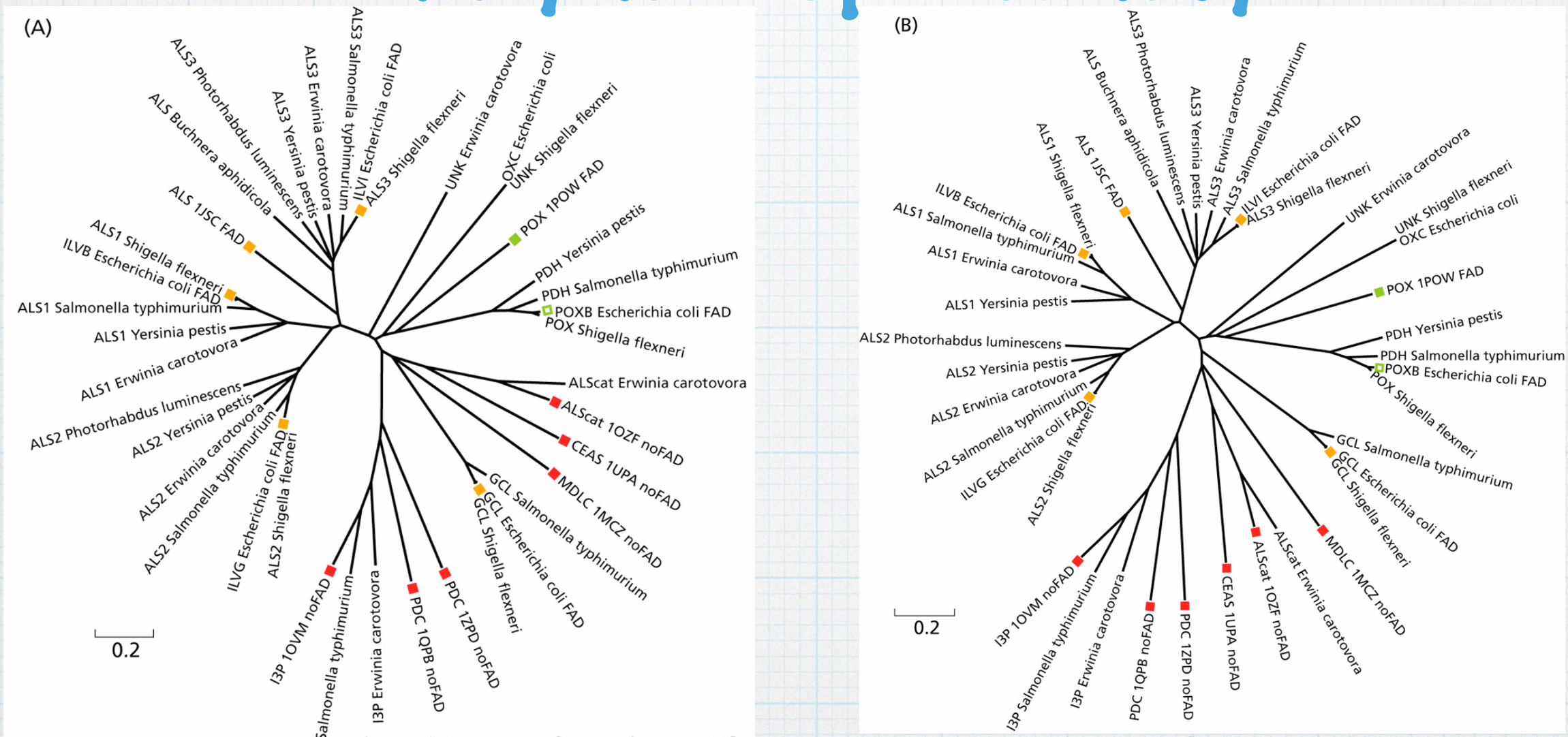
  * Predicting the function of an enzyme that belongs to the superfamily, by knowing only its sequence

  * Elucidating the evolution of enzyme function within the superfamily

Analysis of the TDP-dependent Enzyme Superfamily

* The cofactor thiamine diphosphate (TDP) superfamily
* Some of the constituent families also use the cofactor flavin adenine dinucleotide (FAD)
* Unusually, in some cases, FAD does not participate in the reaction but it must still be bound for the enzyme to be active
* The selected data set contains enzymes from various families
* The evolution of the superfamily chosen in this analysis must have involved either the development or loss of an FAD-binding site (or both), possibly on more than one occasion, and a correct reconstruction of the evolutionary history would reveal this

Analysis of the TDP-dependent Enzyme Superfamily

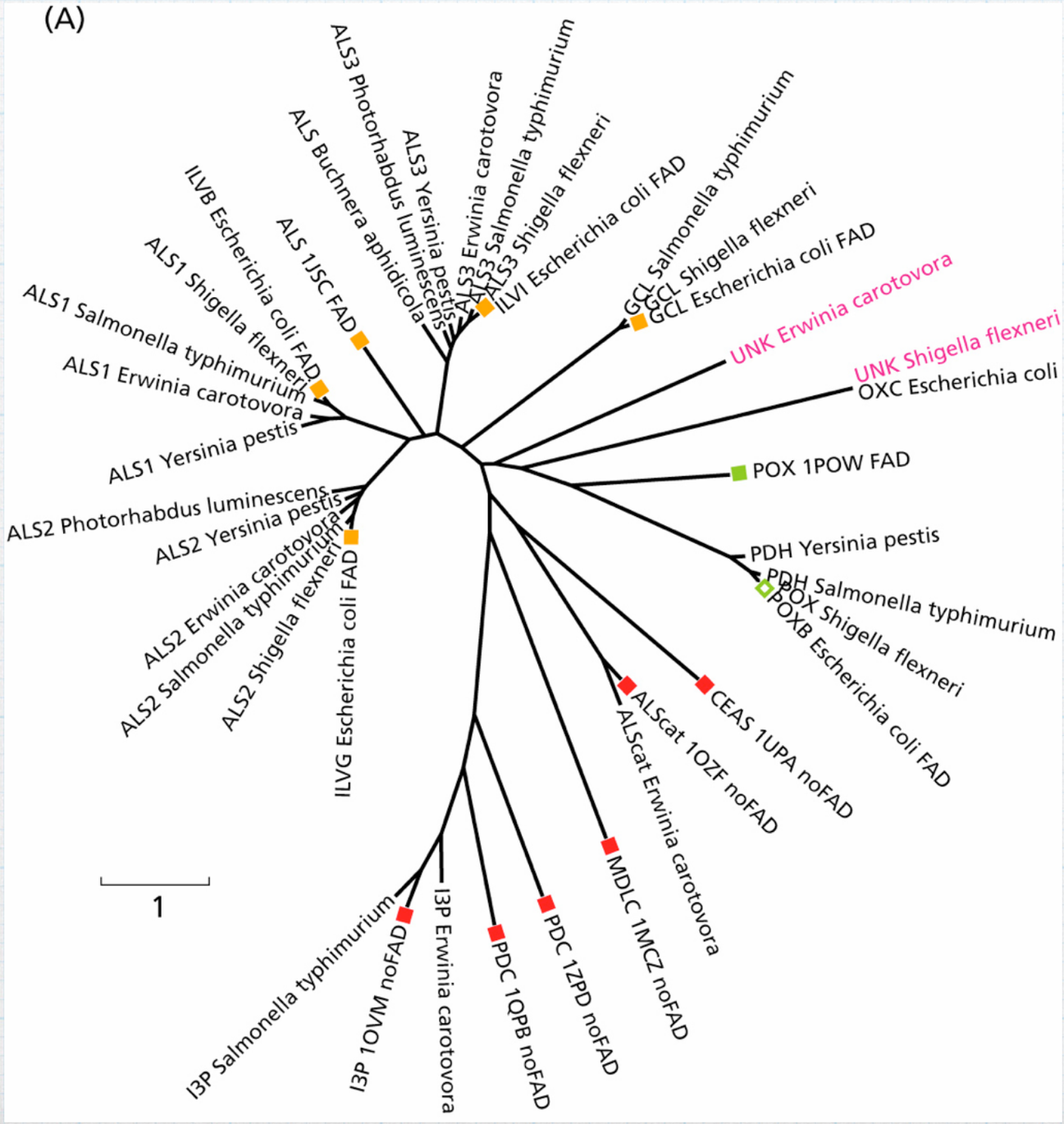


The NJ tree for the TDP-dependent enzyme superfamily
(A) JC correction and (B) K2P correction

red squares: proteins whose experimental structure is known and does not contain a bound FAD cofactor
green squares: proteins known to bind FAD and use it as an active cofactor
orange squares: proteins known to bind FAD but do not use in catalysis

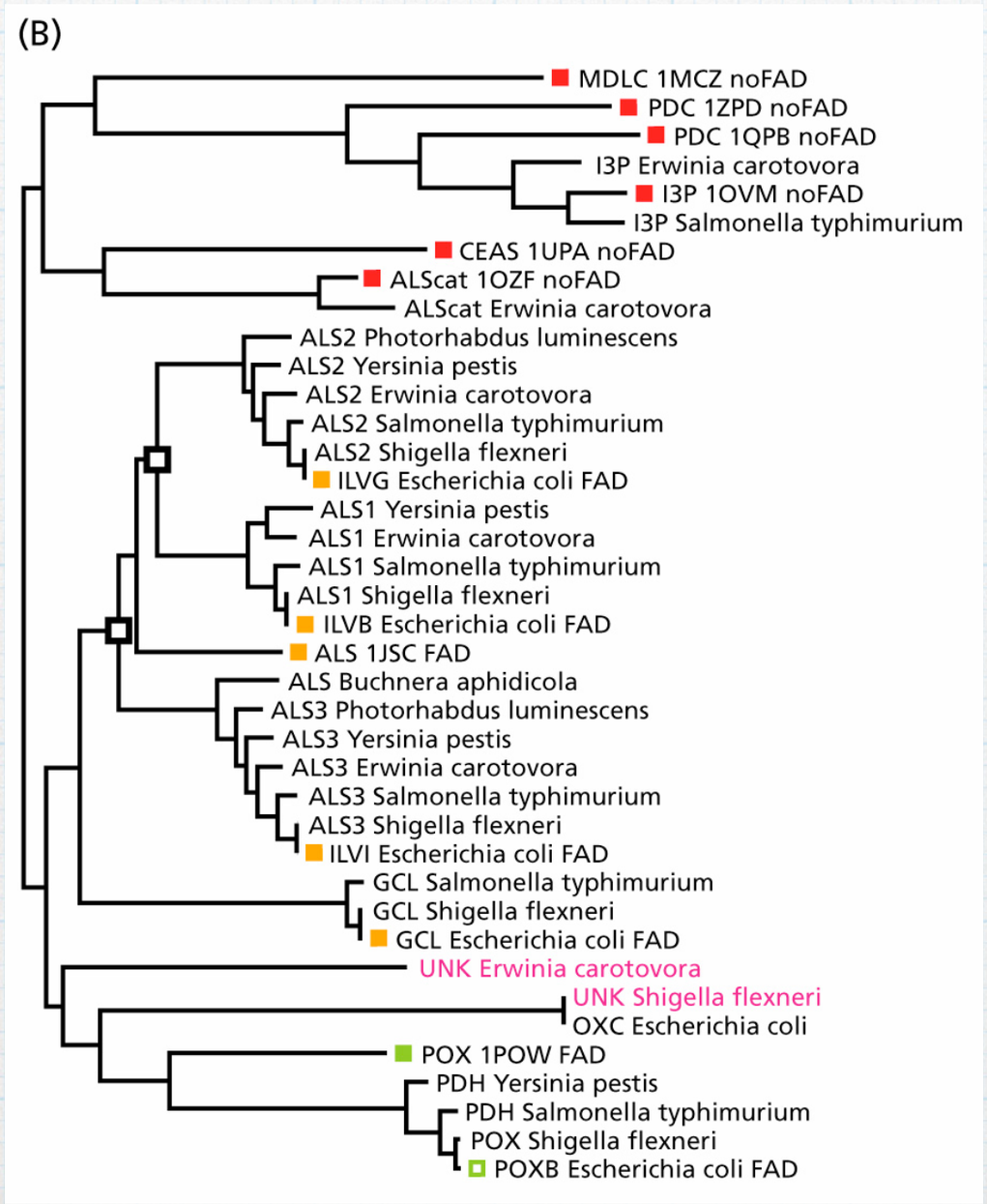
Analysis of the TDP-dependent Enzyme Superfamily

* Analysis of the same data set using the GTR+Gamma+I model and ML



Analysis of the TDP-dependent Enzyme Superfamily

* Rooting the ML tree along the branch that separates the FAD-binding and non-FAD-binding proteins



Questions?