PATTERNS OF AMINO ACID SUBSTITUTION AND BACTERIAL EVOLUTION

By

HYE WON LEE

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2013

To my family

ACKNOWLEDGMENTS

My deepest appreciation goes to Dr. Luciano Brocchieri who guided me with advisement and support as a mentor. His financial support and intellectual guidance allowed me to explore the world of molecular evolution.

I am deeply grateful to my committee members. Dr. Edward Braun and Dr. Michael Miyamoto always provided me with critical questions and suggestions on my work so that my work could be more enriched with solid concepts and backgrounds on molecular evolution. Dr. Marta Wayne gave me insightful comments and warm encouragement regarding my models and theories. I also appreciate her assistance at the beginning of my Ph.D. program as a coordinator of Genetics and Genomics Graduate Program. Constructive comments and suggestions given by Dr. Lei Zhou have been great help in application of my models and methodologies. I appreciate Dr. Byung-Ho Kang who was willing to join my committee as an external member.

I especially thank Hope Parmeter who has arranged and prepared everything related to administration for me. I'd like to thank the UF HPC center. Without the resources and permissions of the center, I could not have my work done in time.

Most importantly, I would like to thank my family. I owe a very important debt to my husband, Seungjun. He always showed his faithful love and great support, and gave thoughtful advices and the warmest caring to me. Finally, I want to thank my precious babies, Shimok and Shihyun, who were born and have given me great joy during my Ph.D. program. I could not finish this dissertation without their patience.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACD	Amino acid Conservation Degree
aLRT	approximate Likelihood Ratio Test
CAS	Conserved Amino acid Sequence
CNM	the Constrained Neutral Model
COG	Cluster of Orthologs
FCB	Fibrobacteres, Chlorobi, and Bacteroidetes
GTR	General Time-Reversible model
HGT	Horizontal Gene Transfer
LBA	Long Branch Attraction
ML	Maximum Likelihood
NJ	Neighbor Joining
OTU	Operational Taxonomic Unit
PVC	Planctomycetes, Verrucomicrobia, and Chlamydiae
RI	Retention Index

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

PATTERNS OF AMINO ACID SUBSTITUTION AND BACTERIAL EVOLUTION

By

Hye Won Lee

August 2013

Chair: Luciano Brocchieri
Major: Genetics and Genomics

Even though the bacterial phylogeny based on 16s RNA has been used as a standard for bacterial classification, the evolutionary relationship between phyla has been difficult to resolve due to low resolution of deep-branch topologies.

In this study I investigated bacterial phylogeny based on a large collection of 169 protein families conserved across most of 1814 species. To minimize the problem of uncertainty in the alignment of distantly related sequences, I developed CONSALIGN, by which alignments obtained from two or more alignment methods are compared and only consistently aligned positions and characters are retained. From the alignment, I only considered characters (amino acid types) conserved within taxonomic groups. Conserved characters were identified based on definition and calculation of “cluster size” (Expected Genetic Diversity), a new measure of diversity equivalent to the expected number of leaf-states not identical by descent within a cluster. Maximum likelihood trees obtained based on conserved characters, resulted in consistent clustering of the different taxonomic groups reproduced with high statistical support over trees obtained with different stringency in the definition of conserved characters,

confirming the bush-like shape of the evolution of bacterial clades, and consistent with the clustering obtained in other studies.

Furthermore, this study explored what extent rates and patterns of protein evolution can be described by a constrained neutral model of protein evolution. With our application of a constrained neutral model (CNM) position-specific amino acid usage profiles, substitution matrices, and expected evolutionary rates were derived based on nucleotide and codon substitution models, and on alignment data. It was found that pair-wise patristic distances calculated from a bacterial phylogenetic tree obtained with a generalized (non-position specific) time-reversible matrix are more consistent with expectations based on the CNM than with expectations from the generalized model used to derive the tree itself. However, it was also found that the CNM is not sufficient to fit the observed relation between sequence similarities and inferred evolutionary distances, to explain the distribution of amino acid types at most alignment position, suggesting that more complex, lineage-specific evolutionary events must be considered to describe short- and long-range evolution of bacterial proteins.

CHAPTER 1
INTRODUCTION

Protein Evolution and Phylogenetic Analysis

Advances in molecular biology and sequencing technology accelerate the development of molecular phylogenetics where evolutionary relatedness of any organism can be evaluated using nucleotides or amino acids (Whelan et al. 2001). Molecular phylogenetics has some advantages over classical morphology-based approaches. As molecular characters can be scored unambiguously and multiple alignments of sequences of many genes and proteins are available, molecular characters can be used to compare any group of organisms even distantly related.

Traditional phylogenetic analysis based on nucleotide or amino acid sequences required identification of homologous characters and reconstruction of evolutionary trees. The accuracy of the inference is heavily dependent on the quality of models for the evolution of characters (Delsuc et al. 2005; Sleator 2011).

Mathematical models of protein evolution have been greatly improved by the incorporation of biological, biochemical and evolutionary knowledge so that statistical testing has become an indispensable tool for phylogeneticists. These improvements allow for robust evaluation of complex evolutionary hypotheses (Whelan and Goldman 2001). To choose an appropriate model for a given dataset or to improve the previously developed model in order to extract more evolutionary information, it is essential to understand protein evolution and dynamics.

Protein Evolution and Protein Dynamics

It has been observed that different patterns of protein evolution are caused by many factors that influence the evolution in different ways.

Incorporation of amino acids on protein sequences is determined by translating mRNA sequences with a rule called the genetic code. The degeneracy of the genetic code plays a role in providing for more tolerance of point mutations on nucleotide sequences. If an organism has biased usage of nucleotides, compositional bias, a certain set of codons is favored and usage of amino acids is also biased. Point mutations in the nucleotide sequence altering the amino acid sequence, non-synonymous mutation, can affect the functioning of the protein and diverging proteins between species when the mutation is beneficial. Synonymous mutation is not expected to affect the fitness of the protein but may affect efficiency of transcription or translation (Komar et al. 1999; Sauna and Kimchi-Sarfaty 2011).

It is easily observed that evolution rates are varied across proteins that are encoded within the same genome (Pál et al. 2006). The functional structure of a protein can affect mutation pattern and speed of amino acid substitution. A slight mutation of amino acids of a protein may affect protein folding and stability. Motifs and domains, combinations of secondary structures are essential structures for a protein's function. Therefore, the functional constraints strongly limit the substitution of amino acids on structures responsible for the protein function. Pal et al. (2006) showed the interaction sites of a DNA-repair enzyme are among the least mutable and most conserved sites. Expression level and pattern of genes expression are factors affecting protein evolution. Kuma et al. (1995) and Duret and Mouchiroud (2000) proposed that proteins expressed in a tissue-specific manner tend to evolve more rapidly than broadly expressed ones (Kuma et al. 1995; Duret and Mouchiroud 2000). These findings can be explained by considering that the proteins expressed in many tissues interact with more molecules

and may also have to function under various cellular conditions. Pál and his colleagues concluded that highly expressed genes have a low evolution rate in yeast by showing the strong correlation between expression level and evolution rate in yeast (Pál et al. 2001).

Phylogenetic Analysis of Protein Sequences

Several alternative approaches for the reconstruction of phylogenetic trees have been used. Commonly used traditional approaches to reconstructing phylogenies are distance methods, exemplified by the commonly used neighbor-joining algorithm. In distance methods, evolutionary distances between pairs of sequences are inferred based on the similarity observed between those sequences, and are represented in a pairwise-distance matrix, which is used to infer the evolutionary tree. Among distance methods, the neighbor-joining (NJ) method is very fast in estimating phylogenetic trees. Distance methods, however, rely on the overall similarity between sequence pairs not considering the detailed nature and dynamics of the differentiation process. Thus, it is expected that distance methods will be out-performed by methods that consider more detailed information on the processes of sequence differentiation. Parsimony methods are another traditional approach. Parsimony methods seek the tree compatible with the minimum number of mutations that can generate the data. These methods give reliable results when dense sampling of sequences can be done, i.e., when the true tree does not contain long branches. Otherwise, by principle, parsimony is prone to underestimating the extent of multiple mutation events and to generating long-branch attraction (LBA) artifacts. Maximum-likelihood (ML) methods identify the tree that has the highest probability of producing the observed data. ML methods are generally considered superior to distance and parsimony methods because their tree

reconstructions are based on a detailed probabilistic description of the evolutionary process. However, the success of ML tree reconstructions depends on the correctness of the probabilistic model. When applied to models that are too rich in parameters, ML can lead to over-fitting artifacts; furthermore, assessments of the robustness of the tree through bootstrap analysis can be computationally costly. Bayesian approaches provide faster measures of support than ML bootstrapping and allow implementing complex and parameter-rich models that can be limited in ML approach (Holder and Lewis 2003). ML and Bayesian methods base evolutionary distance estimates on a matrix of substitution rates from which change probabilities along the tree branches can be computed (Le and Gascuel 2008). Therefore, the choice of a good model of sequence evolution is one of the critical issues to get reliable reconstruction of phylogenetic trees (Thorne 2000; Wróbel 2008). Understanding the processes driving protein evolution is essential to the development of realistic probabilistic models of amino acid substitution and correct phylogenetic tree reconstructions (Thorne 2000).

General Approaches for Modeling Protein Evolution

The General Time-Reversible Substitution Model based on a Markov Process in Probabilistic Approaches

In probabilistic phylogenetic methods, amino acid substitutions are generally described as a time-reversible continuous Markov process. The Markov process is a mathematical model of the transition between (discrete) states over time, where transition to the next state depends only on the current state, and not on the history of how that state was reached. In molecular phylogenetics, the states of the Markov process are the possible amino acid types (20 states) present at a given time on a sequence position and state transitions correspond to amino acid substitutions at

sequence positions. The evolutionary process is assumed to be stationary, i.e., the overall frequency of amino acid types in the sequence is supposed not to change over time. The stationary (or equilibrium) distribution of amino acids is represented as a vector of frequencies denoted as π.

With assumptions that protein sites evolve independently and that the substitution occurs in a time-continuous manner, protein evolution can be described by an amino acid substitution matrix Q = $\{q_{ij}\}$, where q_{ij} ($i \neq j$) indicates the instantaneous rate of transition of amino acid type i into amino acid type j in a time-continuous process (Liò and Goldman 1998). Q is derived by combining a symmetric substitutability matrix R, and a vector of amino acid equilibrium frequencies π, to obtain transition rates $q_{ij} = r_{ij}\pi_j$ ($i \neq j$). The diagonal terms are calculated as $q_{ii} = -\sum_{i=j} q_{ij}$. To scale Q so that the time corresponding to one unit of evolutionary distance (d = 1) corresponds to 1.0 expected substitution per site, Q is divided by the normalizing factor C, defined by $C = -\sum q_{ii}\pi_i$. Based on Q, the transition probability matrix P representing amino acid changes over a period of time d, P(d) = $\{p_{ij}(d)\}$, is derived as $P = e^{Qd}$. The relation between P and Q matrices can be used to derive a relation between evolutionary distance d and expected sequence similarity S:

$$S(d) = \sum_{i=1}^{20} \pi_i p_{ii}(d) \qquad (1\text{-}1)$$

from which the evolutionary distance between two sequences can be inferred based on their sequence similarity. Most importantly, in ML and Bayesian methods, the relation between Q and P matrices can be used to calculate the likelihood of a phylogenetic tree or the ratio of the posterior probabilities of two phylogenetic trees in order to find the optimal tree.

Empirical and Mechanistic Approaches for Amino Acid Substitution Models

There are two different approaches to estimate substitution rates and other parameters for developing amino acid substitution models: empirical and mechanistic approaches. Empirical models estimate relative substitution rates from a certain set of alignment data and reuse them on every other dataset. Therefore, empirical models cannot explicitly consider the nature of the factors that influence the evolutionary process (Yang 2006). They are often constructed by analyzing a large collection of sequence-alignment data to get more generalized substitution rates. That is, the rate of substitution from one amino acid to another represented in substitution matrices is obtained by averaging over a large dataset of homologous positions and sequence families and implicitly averages a mixture of processes of neutral evolution, functional differentiation, mutational biases and position- and sequence-specific events.

In mechanistic models, amino acid substitutions are represented as a function of parameters which are estimated in a given dataset so that the biological process involved in amino acid substitutions such as mutational biases in the DNA and selective constraints resulting in allowance of a specific subset of amino acids in a certain position of the amino acid sequences can be described. Assumptions of nucleotides and codon structures can increase computational cost and make it not practical in some cases.

Traditionally, empirical models have been employed for capturing the substitution process of amino acid and based on substitution matrices derived from alignment datasets (Dayhoff et al. 1978; Jones et al. 1992; Whelan and Goldman 2001; Le and Gascuel 2008). Most of the nucleotide substitution models favored mechanistic models. Codon substitution models have been developed by both a mechanistic approach (Yang

1994; Yang et al. 1998; Miyazawa 2011) and empirical approach (Schneider et al. 2005; Kosiol et al. 2007).

Site Specificity of The Evolutionary Process

Models of Site-specific Substitution Rates

An obvious limitation of the amino acid substitution matrix based on a GTR model is that it does not take into consideration site-dependencies of the evolutionary process. On the biological basis, site-specific substitution rates can easily explain the effects of site-dependency on rates due to various biological factors. Mutation rates on sites can be changed depending on relation to DNA/protein structures that affect solvent accessibility and base reactivity. Sites related to protein functions are more likely to be resistant to mutations due to positive selection. A protein family may and generally will contain a fraction of sites that are invariant and that can affect estimates of evolutionary distances. More generally, different protein sites are observed to evolve at different rates, a feature that is not described by a unique substitution matrix. Therefore, the assumption that different sites have different evolution rates, called site- heterogeneity, can be considered as more biologically plausible.

A common method to capture the site-heterogeneity of the evolutionary process, is based on the assumption that rates are distributed among sites according to a Gamma distribution $\Gamma(\alpha, \beta)$ with expected rate 1.0 ($\beta = \alpha$) (Nei and Gojobori 1986; Tamura and Nei 1993; Yang 1993; Liò and Goldman 1998). The shape of the Gamma distribution varies depending on the single parameter *α*, so that the choice of *α* allows great flexibility to fit the theoretical distribution to phylogenetic data (Figure 1-1). For computational efficiency, Yang suggested a “discrete-gamma model” where the continuous Gamma distribution is approximated by a discrete distribution with a limited

number of rate categories (usually 4-12 categories) (Yang 1994). The Gamma distribution model is often also combined with the assumption that only for a fraction of sites rates follow the Gamma distribution, whereas the remaining fraction of sites is invariant (e.g., Churchill et al. 1992; Reeves 1992; Ota and Nei 1994; Hillis et al. 1996)

With incorporation into a GTR model, site-heterogeneity of the evolutionary rates can be implemented across n positions assigning to each position k a site-specific coefficient v_k so that $Q(k) = Qv_k$, with $(\sum_k v_k) / n = 1.0$, where the coefficient v_k follows a discretized gamma distribution with a shape parameter α estimated from the data (Uzzell and Corbin 1971; Nei et al. 1976; Jin and Nei 1990; Yang 1993; Foster 2004). In a model including both invariable and gamma distributed sites ($I + \Gamma$), a rate-coefficient $v = 0$ is assigned to a fraction I of sites, and gamma-distributed positive rates are assigned to the remaining fraction ($1 - I$) of sites. Evolutionary rates that substantially vary across sites have a significant effect on the relation between evolutionary distance and sequence similarity (Figure 1-2), as substitutions that would otherwise uniformly spread across all sites, tend instead to cumulate at fewer and fast evolving sites (Lee and Brocchieri 2013).

Models of Site-specific Compositional Properties (Mixture Models and Background Distribution)

Site heterogeneity of the evolutionary process may affect not only the rate of substitution at each site but also the pattern of substitution. That is, substitution patterns and amino acid background distributions at each site are determined by site-dependent structural, functional and evolutionary constraints. While site-specific rates affect the speed of evolution, they do not affect the evolutionary pattern of each position. In previous formulations, position-specific compositional distributions have been inferred

empirically as the position-specific profile using Dirichlet distributions (Sjölander et al. 1996; Lartillot and Philippe 2004; Crooks and Brenner 2005).

In another way, different evolutionary patterns can be fitted to individual sites by deriving site-specific Q matrices. Remembering how Q is constructed, this can be accomplished by allowing site-specificity to R, to π, or to both. The first choice, implemented in the QMM model (Wang et al. 2008), is computationally challenging, requiring the optimization of 189 parameters per site-class. A relatively simpler approach is to allow for site-specific stationary frequencies $\pi^{(k)}$. This approach also appears to be consistent with the observation from multiple sequence alignments that different subsets of amino acid types are typically seen at different sites. Site-specificity of equilibrium frequencies has many interesting repercussions on the features of the evolutionary process, on phylogenetic tree reconstruction, and on the relation between sequence conservation and mutational saturation.

Position Specific Profiles of Amino Acid Usage

Possibly, the most successful implementation of the idea of site-specificity of amino acid stationary distributions is the CAT mixture model of Lartillot and collaborators (Lartillot and Philippe 2004; Blanquart and Lartillot 2006; Lartillot et al. 2007; Lartillot and Philippe 2008; Lartillot et al. 2009). In the CAT (category) model, amino acid equilibrium frequency $\pi^{(k)}$ was empirically identified using a Bayesian approach (Lartillot and Philippe 2004). To speed up computation, sets of preassembled profiles of amino acid frequencies are provided in ML and Bayesian phylogenetic reconstruction implementations (Quang le, Lartillot, et al. 2008; Quang le, Gascuel, et al. 2008). Profiles $\pi^{(k)}$ specific to each site k are used in combination with a general substitutability matrix R, to construct site-specific normalized $Q^{(k)}$ with

$$q_s^{(k)} = C^{(k)} r_{ij} \pi_j^{(k)},\ q_{ii}^{(k)} = -\sum_{j \neq i} q_{ij}^{(k)}, \quad (1\text{-}2)$$

and $C^{(k)}$ such that

$$-\pi_i^{(k)} \sum_i q_{ii}^{(k)} = 1.0. \quad (1\text{-}3)$$

In comparison to the global vector of stationary frequencies, as implemented, for example in the LG model (Le and Gascuel 2008), profiles of the CAT model tend to favor different subsets of amino acid types with similar physico-chemical properties (Figure 1-3). As a consequence, while under a generalized Q matrix amino acid substitutions tend to wander over time across all 20 types, divergence is highly constrained within a few amino acid types within a profile, no matter how much evolution occurs, increasing the probability of homoplasy. Furthermore, the reduced effective size of the amino acid alphabet at each site produces higher expected similarity between sequences even at a high evolutionary distance. For example, the generalized LG model (Le and Gascuel 2008) predicts that over time, sequence similarity divergences to the asymptotic value of 5.996%. Profiles in the C20 set implemented in the Phylobayes (Quang le, Gascuel, et al. 2008) and PhyML (Quang le, Lartillot, et al. 2008) methods predict instead, on average, sequence divergence to 18.37% similarity, with a range for individual profiles from 7.54% to 33.56% similarity. Thus, the CAT model estimates that generalized models under-estimate the evolutionary distances of sequences of low similarity (Figure 1-2), providing an explanation for the phenomenon of long branch attraction (Lartillot et al. 2007).

Profiles, Fuzzy proteins, and Neutral Constrained Amino Acid Replacements

Position-specific equilibrium frequency profiles are justified by the idea that functionality and structural stability of a protein require certain residue types at certain

positions, with different degrees of stringency, depending on functional constraints. For example, a position corresponding to an active site may correspond to a profile with one amino acid type, whereas different hydrophilic amino acid types may be allowed to substitute in loops exposed at the protein surface. This suggests an interpretation of profiles based on a model of neutral constrained evolution (Brocchieri 2001; Lee and Brocchieri 2013). According to this interpretation, the profile associated with a particular position defines a subset of amino acid types that can be substituted at that position, without affecting the fitness of the protein (i.e., its functionality). This model asserts that a protein can be described as a functional unit and possibly a large set of alternative sequences, each functionally equivalent to the other. Thus, from a functional perspective, a protein would be described, rather than by a sequence of amino acids, by a sequence of amino acid subsets, whose size describes different degrees of "fuzziness" of different positions. A "fuzzy protein" can evolve within the limits imposed by the sequence of amino acid subsets that describe it with no effect on functionality. With this interpretation, position-specific profiles can explain not only the evolutionary pattern, but also the speed of evolution. The reasoning is that random substitutions will be retained only if they result in substitutions allowed by the profile. Thus, if the profile is stringent, most substitutions will be rejected slowing the evolutionary process; if the profile is permissive, most substitutions will be accepted, resulting in fast evolution. To model the effect of purifying selection on evolutionary rates, I first consider a general, normalized substitution-rate matrix, whose coefficients are derived from nucleotide and codon substitution matrices. At each position k, the substitution matrix is filtered by a position-specific "occupancy vector" that defines the subset of amino acid types allowed

at that position, so that equilibrium frequencies and transformation rates towards amino acid types not represented in the occupancy vector are set to zero. The result is a reduced $Q^{(k)}$ matrix, with a generally slower average transition rate $t^{(k)} = -\pi_i^{(k)} \sum_i q_{ii}^{(k)} \leq 1.0$. To preserve the scaling of the evolutionary process, where one unit is equal to one substation per site, all matrices are finally renormalized without loss of generality to an average one substitution per site multiplying them by the factor $n/\sum_{k=1}^{n} t^{(k)}$. With this model, selection against not-allowed transformations generates site-specific profiles of amino acid equilibrium frequencies, and a distribution across sites of different site-specific evolutionary rates, approximately proportional to the size of the profiles (Figure 1-4). The profiles that define a fuzzy protein are not likely to correspond to those identified by the CAT model, which, in terms of the constrained neutral evolution model, can be interpreted as combining substitutions within a profile with substitutions between profiles, taking into account events of profile evolution. An interesting question is how each process contributes to protein evolution.

Causes of Artifacts that Affect Reconstruction of Phylogenetic Trees

Technically, phylogenetic analysis based on a homologous sequence alignment estimates evolution rates based on the number of amino acid substitution on sites. Some artifacts are observed that affect the estimation of evolutionary distance between sequences. The most well known factors are nucleotide compositional biases on genomes and mutational saturation on protein sites.

Nucleotide Compositional Biases

One of the causes of artifacts affecting phylogenetic reconstruction is compositional bias of nucleotides. In reconstructing phylogenetic trees, compositional

biases are responsible for a systematic error in grouping unrelated species with similar nucleotide composition (Lockhart et al. 1992; Galtier and Gouy 1995; Tarrío et al. 2001; Blanquart and Lartillot 2006). As I have seen, global nucleotide compositional biases of genomes also reflect on amino acid usages: because of the bias, a limited set of codons are favored, e.g., according to the GC content, and the corresponding amino acids are favored in the protein sequence. Hence, nucleotide biases also affect tree reconstructions based on protein sequence data (Foster and Hickey 1999; Jermiin et al. 2004).

Different homologous sites of a protein family have different conservation patterns as well as different evolutionary rates. When some positions are observed to have non-random sequence variation, it often corresponds to the functional differentiation whereas other positions may be shown to have random substitutions within a certain set of amino acids as a neutral process (Brocchieri 2001). Therefore, sequence divergence in the tree can be different depending on which positions are considered for analysis. It is expected that this distinction between neutral and selective processes in modeling protein evolution will result in a better understanding of the relation between sequence similarity and evolutionary distance.

Mutational Saturation by Multiple Substitution and Reversion

Evolutionary distance, the number of substitutions per site, is related to the number of substitutions observed among different homologous sites (differentiation) between sequences. However, the relation between evolutionary distance and sequence differentiation is not expected to be linear, due to the increasing occurrence of unobservable multiple substitutions and substitution reversions as evolutionary distance increases. As mutational events accumulate, they eventually become so

widespread that further mutations accumulate over already mutated positions and cannot be measured by appreciable further sequence divergence. At this stage, the sequence is said to be mutationally saturated. This phenomenon may lead to under-estimation of the evolutionary distance of distantly related sequences (Rodríguez-Ezpeleta et al. 2007). This can result in erroneous grouping of diverged sequences in evolutionary tree reconstructions, a well-known effect called long-branch attraction (LBA). High conservation between sequences does not always indicate that sequences are not mutationally saturated, nor does it indicate that the sequences are evolutionarily closely related (Brocchieri 2001).

The Reconstruction of Bacterial Phylogeny

Bacterial Phylogeny Reconstruction Using Multiple Protein Families

Efforts to resolve the deep phylogeny of bacteria have been sustained for a long time after the introduction of 16s RNA-based phylogeny (Woese and Fox 1977; Woese et al. 1985; Woese 1987; Ludwig and Klenk 2001) (Figure 1-5), but the deep phylogeny of bacteria still remains uncertain despite of the large quantity of bacterial genomes that have been accumulated. Horizontal gene transfer (HGT) is one of the most important factors, complicating the reconstruction(Bapteste et al. 2004; Gogarten and Townsend 2005; Choi and Kim 2007). HGT introduces a different phylogenetic history into the genome genes so that the resulting tree based on molecular sequences is not congruent with the species tree (Gogarten and Townsend 2005). HGT is considered to be so widespread that it might have blurred the phylogenetic signal for a prokaryotic phylogeny (Doolittle 1999; Delsuc et al. 2005; Gogarten and Townsend 2005).

The widely employed methods based on multiple protein families are the "supertree" methods and "supermatrix" methods (Delsuc et al. 2005; McInerney et al. 2008). In the "supertree" approach, A consensus phylogeny is constructed from multiple gene trees (Brochier et al. 2002; Daubin et al. 2002; Puigbò et al. 2009). Each tree from each protein family can be combined into a big tree based on common branches. In the "supermatrix" approach, sequence alignments for individual gene families are concatenated into a single alignment that is used to construct a tree. In this approach, only common genes or proteins should be used (Teichmann and Mitchison 1999; Brown et al. 2001; Ciccarelli et al. 2006; Boussau et al. 2008; Wu and Eisen 2008; Oshima et al. 2012). These approaches are used to combine multiple datasets into a larger dataset, but there are some drawbacks because it assumes all the data have followed the same evolutionary history (Swithers et al. 2009). Moreover, one should be cautious when interpreting the results because of increased noise in the datasets such as "homoplasy," shared characters found in a different phylogenetic tree not directly inherited from a common ancestor (Rokas and Carroll 2008). Even when thousands of sequences without noises are in hand, conventional methods for generating a phylogenetic tree such as Maximum likelihood and Bayesian approach are not feasible to handle such a large amount of sequences.

Conserved Amino Acids within a Taxonomic Group as a Phylogenetic Signal

Rather than combining data, signature-based analysis (Gupta 2004; Gao et al. 2009) uses conserved insertion/deletions (indels) shared within a group, which is considered as a class of rare genomic changes (RGC) with a low level of homoplasy (Rokas and Holland 2000). Conserved indels contain distinct and considerable information showing the relation among groups, but group-specific indels are not

available for all groups and taxon sampling error may limit the ability to find group-specific indels. Thus, although conservation of insertions and deletions in proteins for a specific group is a valuable source of information, it does not appear to be sufficient to fully resolve the phylogenetic relations of bacterial phyla.

An alternative source of rare character changes is protein sites with a slow rate of evolution. Among these I am interested in sites that are conserved within phyla but whose state “switches” among subsets of phyla. What distinguishes these characters from other slowly evolving protein sites is association of their state with specific phyla and their conservation over a great evolutionary time covered by the ramification of individual lineages within the phylum. It can be hypothesized that this pattern of evolution identifies functionally important sites and functional diversification in different phyla. Therefore, lineage-related changes in character state have been used to identify functional divergence associated with a particular site (Lichtarge et al. 1996; Gu 1999). In this perspective, conserved amino acids within the taxonomic group along long evolutionary time would be identical by descent to their common ancestral state and subject to negative selection, whereas positive selection would be responsible for the evolution of the site between phyla. Gu (1999) defined this pattern of evolution, with sites that are conserved within subfamilies but diverged between them may be relevant to functional divergence as type-II functional divergence (Gu 1999; Gu 2006). Rogozin et al. (2008) suggested RGC_CAM (Conserved Amino acids-Multiple substitutions) as a new class of rare genomic changes to use for phylogenetic tree construction (Rogozin et al. 2007; Rogozin et al. 2008). Both analyses assume that the amino acid residues

conserved within the group but different between groups would be useful to obtain reliable phylogenetic information.

Research Plan

Phylogenetic inference based on sequence data reconstructs the evolutionary history of sequences from the comparison of characters (amino acids or nucleotides). As available sequence data have been accumulated rapidly, it is questioned how to reduce the noise from the large data set and identify meaningful characters containing evolutionary history to derive a reliable phylogenetic tree. Even though probabilistic methods such as ML or Bayesian approaches have been considered to be very successful, the resulting trees depend on a choice of amino acid substitution model.

This study was intended to reconstruct bacterial phylogenetic trees with refined tree construction methods such as more reliable positions and characters, and a newly developed protein evolution model. The dataset was constructed using NCBI bacterial genome sequences and COG database and contains 1892 bacterial sequences in 169 protein families (Figure 1-6). At first, the complete bacterial phylogenetic tree was constructed similarly as the conventional ways: Maximum Likelihood methods with LG model, an empirical model of amino acid substitution. The concatenated alignment of reliable positions filtered by the CONSALIGN method was used for the complete tree.

Calculation of support values on internal branches showing relations between phylum-level groups is computationally demanding in a ML approach. In this study, a conserved amino acid sequence (CAS) for a group was defined as concatenated positions with amino acid types that are conserved within a group. When multiple groups are involved in the same alignment, some positions may have conserved amino acid types within a group but be different or “switched” between groups. These positions

were called “switch positions” (Figure 1-7). These events of “switch” between different groups on the same alignment position can occur very rarely so it can be considered as a class of rare genomic changes. In identifying switch positions, it is necessary to consider sets of evolutionarily related sequences (evolutionary groups) sufficiently “numerous” to allow detection of conservation. A possible measure of the “size” of the set could be, for example, the number of leaves corresponding to a given phylogenetic group. With this choice, a rule could set by which, for example, a position is considered conserved within a group including N leaves if the same amino acid type is observed at all leaves, and $N \geq N_{min}$, where N_{min} is the minimum number of leaves. However, a difficulty of this choice is that leaves of different groups are related by different branch patterns and lengths. Thus, whereas two leaves connected by long branches may show the same amino acid type because of evolutionary constraints, if the same leaves were connected by very short branches, they may exhibit the same amino acid type simply because there has not been enough time to accumulate substitution during the evolutionary history separating them. In the case of bacterial evolution, this problem is particularly relevant considering the very unbalanced sampling of sequences across evolutionary groups in public databases. For example, a large number of protein homologs are available in the *Gammaproteobacteria* group, but many of these sequences belong to closely related strains of the same species. In an effort to develop a measure of cluster size that took into consideration of the evolutionary relations among sequences, I propose as a measure of cluster size the Expected Genetic Diversity (E_{gd}), which estimates the expected number of leaf-states not identical by descent. This measure does not directly depend on the amino acid types actually

observed within a cluster at a given alignment position, but only on the phylogenetic tree corresponding to the cluster. The expected genetic diversity (E_{gd}) depends on the details of the evolutionary model and on the data used to infer the tree. However, this detailed information is not usually available. Thus, E_{gd} is defined as an Index, which is calculated based on the assumption of a phylogenetic tree (defined by its topology and branch lengths) and on the approximation that evolution along branches is a homogeneous Poisson process with a rate parameter λ dictated by the length of each branch. Associated with this measure of cluster size, a weight w_f is defined and applied to each leaf, which measures the degree of "uniqueness" of the expected state at each leaf. These weights sum up to E_{gd} and are used to define the fraction of a cluster of size E_{gd} that is occupied at a given position by the same amino acid type (Amino acid Conservation Degree, ACD). With application of two thresholds, for E_{gd} and for ACD, alignments of sequences containing group-specific conserved amino acids could be obtained and used for generating phylogenetic trees of bacterial phylum. To find the differences between phylogenetic trees depending on the protein functions, the complete alignment was divided into two sub-alignments based on the functional category of the family each position belongs to: (1) Information Storage and Processing, and (2) Non-informational such as Metabolism, Cellular Process etc. Therefore, a total of three different alignments were generated: All-family alignment, Informational-protein alignment, and Non-informational alignment (Figure 1-8).

It is assumed that the amino acid substitutions in the group-specific conserved amino acid sequence (CAS) alignment would behave as in the ancestral sequences of groups. So a GTR matrix should be derived rather than using a conventional empirical

amino acid substitution matrix. RAxML was used to obtain a draft phylogenetic tree with estimation of a GTR matrix from the CAS alignment and then PhyML was used to optimize the tree and obtain the approximate Likelihood Ratio Test (aLRT) support values. To measure homoplasy on the CAS alignment, Retention Index (RI) was calculated on each alignment with a corresponding tree and then compared to RI of the complete alignment with the ML supertree.

Another bacterial phylogenetic tree was reconstructed with development of an alternative protein evolution model based on the concept of position-specific amino acid "profiles" of background frequency as in Lartillot and Philippe (2004) CAT model and Crooks and Brenner (2004) amino acid replacement model. An alternative model developed with a mechanistic approach in this study could explicitly interpret position-specific profiles and their properties in terms of evolutionary processes. The nucleotide and codon substitution models were developed to incorporate nucleotide base content information and codon structure into the new model. In this model, the concept of amino acid usage profiles at each position of a protein family was substituted by a combination of "occupancy" profiles defining position-specific processes of purifying selection at the protein level, and concomitant "distribution" profiles reflecting mutational biases and selective processes at the nucleic-acid level.

Position-specific profiles and Q matrices are obtained from the bacterial sequence alignment based on nucleotide and codon substitution models. Phylogenetic-tree reconstruction strategies based on profiles and Q matrices are developed where the relation of sequence similarity and evolutionary distance is derived on each position and then averaged over the alignment positions. Pairwise evolutionary distance of

sequences are estimated and then a Neighbor-Joining method was used for generating a phylogenetic tree. This tree was compared to the ML tree based on the same set of sequences, in terms of tree topology, running time of computation and model fitting.

Validity of reduction of a codon substitution matrix to an amino acid substitution matrix was discussed using the lumpability test. Profiles are analyzed using conditional likelihood analysis to check if amino acids are randomly distributed over the tree and to estimate how likely the amino acid subset observed in a given position could be changed.

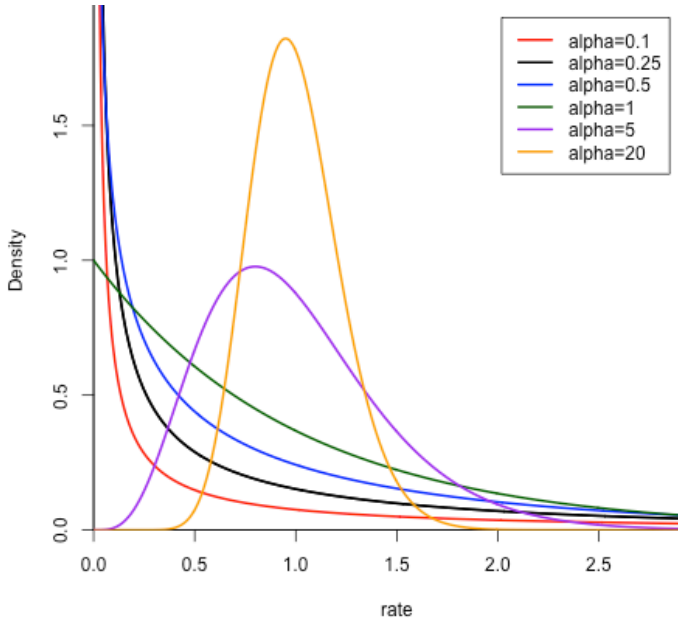


Figure 1-1. Gamma distribution with different shape parameter (alpha).

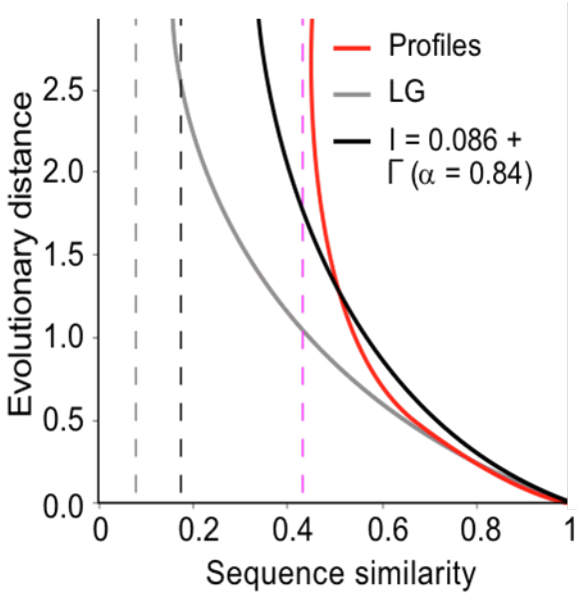


Figure 1-2. Expected relation between sequence similarity and evolutionary distance, as predicted by the LG model (Le and Gasquel, 2008), by the LG model with estimated proportion of invariable sites and gamma distributed rates (I+Γ), and by profiles of amino acid occupancies from an alignment of bacterial DnaK sequences.

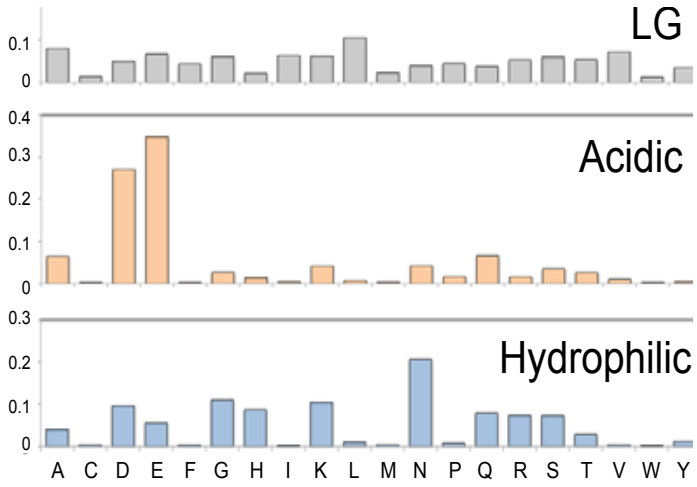


Figure 1-3. Amino acid equilibrium frequencies from the LG model, compared to those described by two profiles from the C20 set (Quang le, Gascuel, et al. 2008).

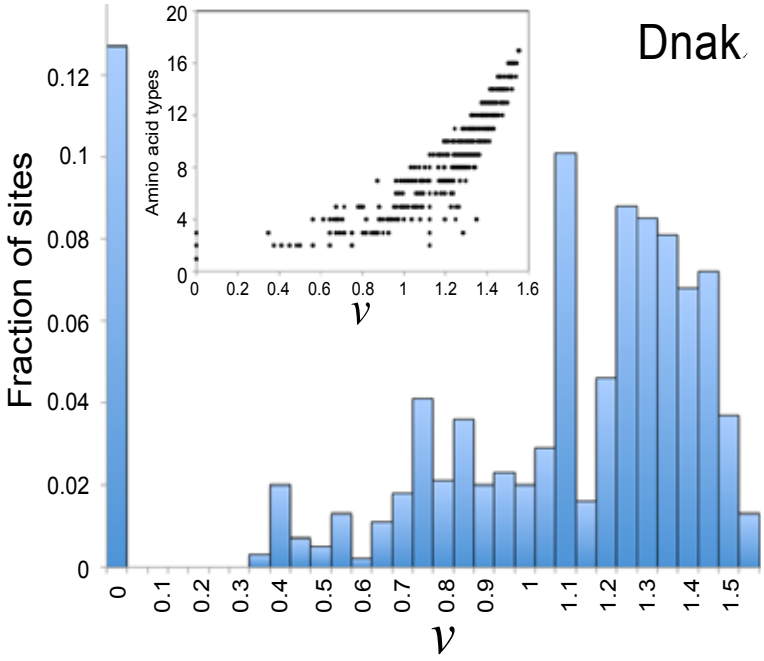


Figure 1-4. Distribution of site-specific relative rates of evolution obtained from an alignment of about 500 DnaK proteins. The insert shows the relation between these rates and the number of different amino acid types observed at corresponding positions.

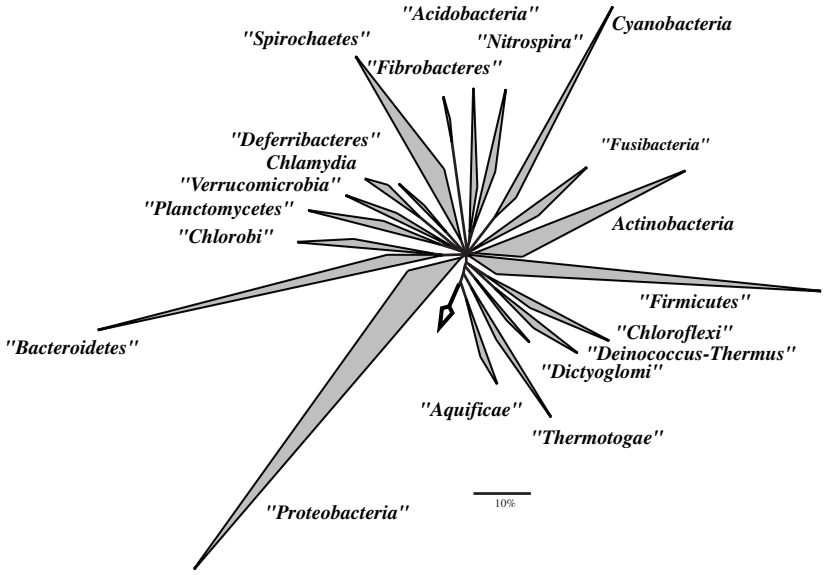


Figure 1-5. Bacterial phylogeny based on16S rRNA where major bacterial phyla were defined well (Ludwig and Klenk 2001).

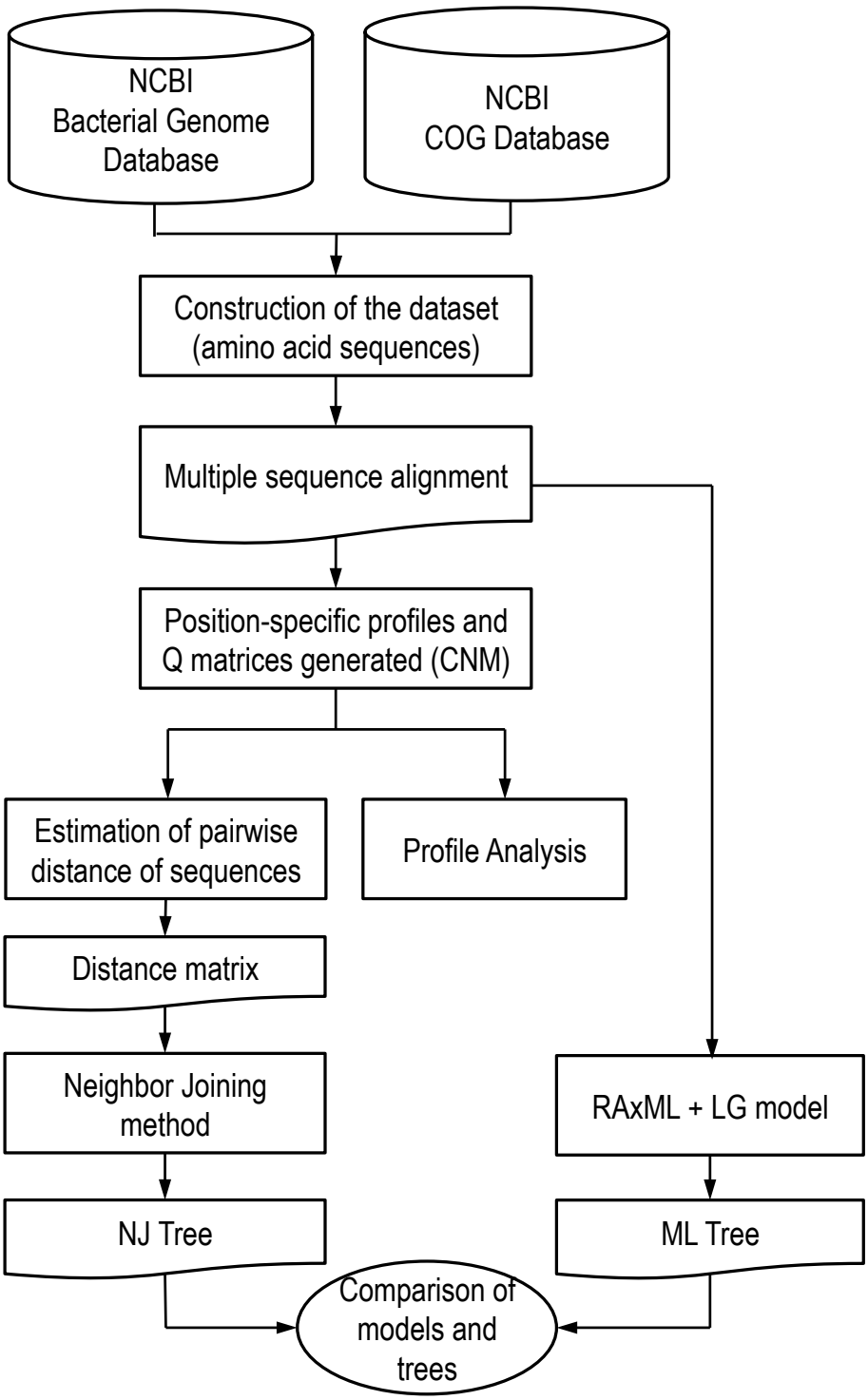


Figure 1-6. A workflow to construct maximum likelihood tree and position-specific profile based tree from a bacterial sequence alignment.

seq A1	F	D	S	A	S	S	I	K	A	H	G
seq A2	F	E	S	P	T	D	V	K	A	H	G
seq A3	F	E	S	P	S	D	V	K	A	H	G
seq B1	F	D	K	A	S	L	V	R	S	H	G
seq B2	F	D	R	A	S	Q	V	K	S	H	G
seq B3	F	D	K	A	S	S	V	R	S	H	G
seq B4	F	D	K	A	S	S	I	R	S	H	G
seq C1	F	P	H	-	-	S	I	K	S	H	G
seq C2	F	P	H	-	-	S	V	K	S	H	G
seq D1	F	P	H	G	-	-	-	R	A	H	G
seq D2	F	S	H	A	-	-	-	K	A	H	G
seq D3	F	P	H	A	-	-	-	K	A	H	G
seq D4	F	P	H	-	-	-	-	K	A	H	G
seq D5	F	P	H	G	-	-	-	K	A	H	G

Figure 1-7. An example of switch positions. The same colored sequences belong to the same group and have positions with conserved amino acids. The 1st, 10th, and 11th positions are totally conserved and the 2nd, 3rd, and 9th are “switch” positions where different type of conserved amino acids are represented.

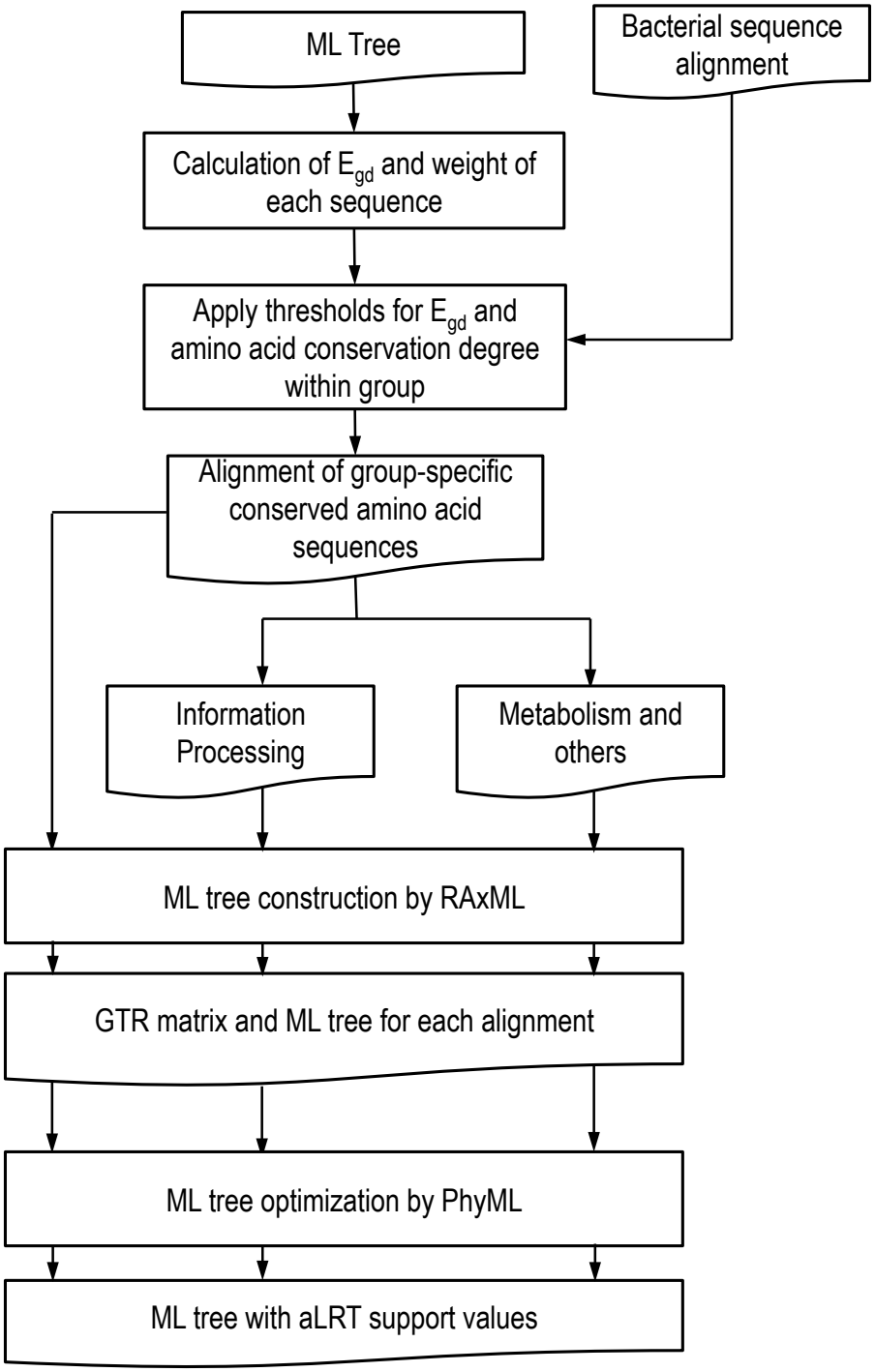


Figure 1-8. A workflow to construct bacterial phylogenetic trees based on group-specific conserved amino acid sequences. The COG functional categories are used to divide the complete alignment into two partial alignments: Informational alignment (Information Storage and Processing) and Not-informational alignment (Metabolism and others)

CHAPTER 2
RECONSTRUCTION OF BACTERIAL PHYLOGENETIC TREE

Data Construction

Collection of Sequences

The Clusters of Orthologs (COG, http://www.ncbi.nlm.nih.gov/COG/) is a manually curated database of 4873 protein families from 66 unicellular genomes covering 14 phyla and 28 classes (Tatusov 1997; Tatusov et al. 2003). The COG protein families contain orthologous sequences generated by comparative genomic study on microbial genomes. The COG database provides taxonomic information and protein sequences of species for each family. All COG families are assigned with COG IDs and classified into 25 classes based on the function of proteins. These classes belong to 4 divisions: Information storage and processing, Cellular processes, Metabolism, and Poorly characterized. For this study, a total of 169 COG families were chosen which are common to almost all prokaryotes.

At the time of this study, 2,056 prokaryotic genome sequences were available at the NCBI microbial genome database, ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria. This collection contains coding sequences and amino acid sequences covering 129 archaeal and 1,927 bacterial organisms in 5 phyla for Achaea and 38 for Bacteria.

The COG database is composed of a relatively small number of species compared to all available genome sequences in NCBI. Therefore, homologous sequences could be added to each family based on sequence similarity in order to fully utilize the resources of NCBI (Figure 2-1).

The phylum was chosen for taxonomy group in this study except for the *Firmicutes* and *Proteobacteria* that have more than 300 sequences (Table 2-1). The

class was instead chosen as taxonomic groups for sequences belonging to the *Firmicutes* and *Proteobacteria* so that the distribution of number of sequences among groups are more even (Table 2-1).

Searching Protein Sequences Homologous to Nearly Universal COG Families

Taxonomic information of every organism in each COG family is compared to those of a given bacterial organism. After choosing the closest organism in the COG family, its protein sequence is queried against all protein sequences in the bacterial organism. At most one homologous sequence per genome was found using sequence similarity. The sequence similarity was calculated using SSPA (Karlin and Brocchieri 1998). The protein with the highest sequence similarity and with length between ½ and 2 times the query sequence length was selected and added to this COG family.

Excluding Outliers and Horizontally Transferred Genes

Automatically generated homologous sequences might have some outliers. In this study outliers were identified using sequence similarities, size of proteins and group information (Figure 2-2). In order to measure how similar the query sequence is to other sequences in the same group, the “similarity score” was calculated for the query sequence. For a given sequence as query, all the sequences were divided into two groups: *In-Group*, sequences belonging to the same group, and *Out-Group*, sequences whose groups are different from the query. A sequence from the *In-Group*, M_i, is paired with a sequence from the *Out-Group* and the sequence similarity is calculated by

$$\text{Pair-wise sequence similarity} = c / (c + v) \qquad (2\text{-}1)$$

where c = number of conserved sites and v = number of non-conserved sites. Once pair-wise sequence similarity of every pairs containing the sequence M_i was calculated,

its average over all pairs of the sequence M_i against sequences in the Out-group is similarity score of M_i.

To identify outliers from automatically generated homologous sequences in each group, a plot of similarity score and length of protein was used. Proteins that are located too far from the crowd of the majority of sequences in the same group, or proteins that have too low of a similarity score are regarded as outliers of this group. To consistently identify outliers in terms of two axes, a library *mvoutlier* (Filzmoser et al. 2005) in the R package was used. A function, *aqplot*, in *mvoutlier* was applied to sequences belonging to the same group in each family. Identified outliers were removed from the data set.

Horizontal gene transfer (HGT) means conveyance of genetic material not from ancestors to descendants. Horizontal gene transfer contrasts with vertical gene transfer, which is the traditional method of passing genetic material from one generation to the next. HGT is frequently observed and complicates the process of inferring correct phylogenetic relationships (Bapteste et al. 2005). Even though protein families in the data set are thought to be hardly horizontally transferred because they are universally distributed in most of prokaryotes and most of them are involved in informational processing and storage (Jain et al. 1999; Wu and Eisen 2008), some proteins were susceptible based on three studies where some species or organisms were suggested to horizontally transferred in regard of COG protein families (Wolf et al. 1999; Ciccarelli et al. 2006; Kanhere and Vingron 2009). These proteins were removed from the data set. The list of taxonomic groups or species removed from the dataset is shown in Table 2-2.

Protein sequences in each family were re-aligned after excluding outliers and sequences horizontally transferred. The data set contains distantly related sequences, which are difficult to align correctly. Therefore, it is necessary to develop or find a way to minimize the problem of uncertainty in the alignment of distantly related sequences by filtering out unreliable positions or characters.

Bacterial Phylogenetic Tree Construction

Filtering Positions and Identify Reliable Characters to Reduce Noises

Multiple sequence (protein or nucleic acid) alignments strive to faithfully recognize homologous positions from sets of evolutionary related sequences, based on sequence similarity and other information. Because of the vast range of applicability of alignments and of their usage as basic data in a variety of applications, correctness of the alignment is a crucial prerequisite to the analysis and evaluation of the processes underlying sequence evolution. For example, based on patterns of conservation in the alignment it is possible to determine which positions are evolutionary constrained and crucial for functionality and for functional differentiations. Patterns of differentiation across alignment positions are analyzed in the process of phylogenetic tree reconstruction (Karlin and Brocchieri 1998).

Different approaches have been devised to ameliorate the quality of multiple sequence alignments (Brocchieri 2001). One “global alignment” approach is to consider different alignment methods combining their information to recover a “consensus” alignment that is expected to ameliorate the result of each individual alignment method (T-coffee, Notredame et al. 2000). A second, “local alignment” approach is to consider in an alignment only positions that are more robustly aligned. This can be accomplished either as part of the algorithm leading to the alignment, as in local alignment methods

(Needleman and Wunsch 1970; BLOCKS, Henikoff and Henikoff 1991; Iteralign, Brocchieri and Karlin 1998), or by post- processing of the alignment, selecting only alignment positions that pass some criterion of conservation (Gblocks, Castresana 2000). Here, a simple procedure, CONSALIGN, is developed which is devised to obtain alignment positions by comparison of alignments generated with different methods to generate a “local alignment” of sequence positions consistently aligned across a set of any number of alternative alignments.

Collection of Reliable Positions (CONSALIGN)

CONSALIGN takes as input a set of multiple alignments in FASTA format, considers each vector of aligned positions and verifies that the same positions are aligned by all alignment methods. If the alignment is “agreed” upon by all methods it is retained, otherwise it is discarded. The definition of agreement depends on a single parameter, chosen by the user, that determines its stringency. The parameter (α) is the fraction of positions that are commonly aligned by all methods, and can be chosen in the interval [0.5, 1.0]. Values of alpha less than 0.5 are excluded to avoid inconsistencies. Consider an alignment of *N* sequences and length *M* as a *N* x *M* matrix composed of column vectors of size *N*, whose elements are the positions aligned from each of the sequences or indicate a gap (e.g., -1). One alignment is equivalently chosen as reference and each of its columns are in turn used as queries to search the other alignments. From each of the other alignment matrices a column-vector that contains more than a fraction alpha of the sequence-specific positions contained in the query alignment-column is identified, if it exists. To determine this fraction, the number of common positive (non-gap) sequence-specific positions is divided by the total number of elements that are not negative in both vectors, where are no gaps in both alignment

positions. If such column-vectors are not identified in all alignments, the vector is discarded. Otherwise, from all collected vectors the size of the intersection of conserved alignment positions (N_c) is calculated to determine how many positions are commonly observed in all alignment columns, as well as the number N_g of vector elements which are occupied by gaps in all sequences. If $N_c/(N - N_g) \leq \alpha$, the position is discarded. Otherwise, all elements that are not conserved across all vectors are set to gaps, and the modified vector is retained. An example of obtaining reliable position from two alignment methods by COSNALIGN is shown in Figure 2-5.

The resulting matrix of modified vectors represents a collection of alignment positions robustly aligned across all alignments. With this approach, only a subset of sequence positions is represented in the final alignment. A gap in the alignment should be interpreted as “missing data”, as commonly assumed in probabilistic methods (ML and Bayesian) for phylogenetic tree reconstruction.

Basic Analysis of Data

The dataset contains 1,814 bacterial sequences belonging to 25 groups. These sequences are represented over the 169 nearly universal protein families in the dataset (Table 2-3). The concatenated alignment, referred to the complete alignment, contains 72,700 positions that were retained in both alignments from two different alignment methods to avoid alignment artifacts and obtain more reliable positions.

Among 72,700 positions, about 4,549 positions are all-gapped and there are 12,383 single-sequence positions or 17% of all positions (Figure 2-3). The number of positions containing 3 to 17 amino acid types are close to each other but the number of positions having 18 to 20 amino acid types rapidly decrease as the number of amino

acid types increase (Figure 2-3). The distribution of number of sequences represented on positions is similar to the normal distribution except for single-sequence positions (Figure 2-4). On average, 7.39 amino acid types were seen at positions. 15,226 sequences have only one amino acid type but 12,383 (81%) of them have only one sequence. The distribution of subset size and sequences on positions showed the complete alignment contains enough positions and a fair distribution of amino acid subsets.

For the COG functional categories, 89 families of 118 Informational protein families belong to *J* (Translation, ribosomal structure and biogenesis) and 8 and 21 families belong to *K* (Transcription) and *L* (Replication, recombination and repair), respectively (Table 2-4). Non-informational protein families are 51 families that belong to Cellular process and signaling, Metabolism, and others (Table 2-5). Some families are assigned multiple functional categories.

A ML Tree Based on the Alignment of Reliable Positions

From 169 protein families, sequences in each family were aligned using MUSCLE (Edgar 2004) and OPAL (Wheeler and Kececioglu 2007). Using CONSALIGN with $\alpha = 0.5$, reliable positions from each family were obtained and concatenated over protein families.

Based on the alignment of reliable positions containing 2,000 sequences with 74,857 positions, a Maximum Likelihood tree is generated using RAxML (Stamatakis et al. 2005) based on the LG model (Le and Gascuel 2008) with estimation of the average amino acid frequencies and four classes of rates determined from gamma distribution (LG+Γ+I). Generating a tree based on such a big alignment requires intensive computing power, so RAxML v.7.3.2 (Stamatakis 2006) was run on 16 nodes of a

cluster with 150G RAM memory at the UF High Performance Computing Center (command= raxmlHPC-PTHREADS-SSE3 f –a –m PROTLGIF –c 4 –N 1 –T 16). The support values based bootstrapping or aLRT estimation could not be obtained due to extremely high computational cost and running time.

The resulting tree is rooted using a group of Archaeal sequences. Most of sequences are clustered consistent with their taxonomic groups. Some extremely long branches were observed and some sequences were clustered with sequences from a different group. Such sequences and Archaeal sequences were removed from the tree and alignment. A total of 1,814 bacterial sequences were selected (Figure 2-6) and the size of the alignment was 72,700 positions after removing positions having all gaps. The ML tree was used as a supertree and the resulting alignment as the complete alignment for further analysis.

Handling such a big alignment to generate a tree is computationally demanding in ML approach, and in particular multiple repetitions of the procedure on resampled data to calculate support values (bootstrap) is prohibitively time consuming. This means that any advantage that might be provided by the availability of a large quantity of data, may be upset by limitations in computational power. Moreover, although the analysis can be based on a very large number of homologous sequences, and on a relatively large number of gene families, the distribution of sequences across phylogenetic groups (phyla) is not uniform, with some groups (e.g. the *Proteobacteria*, *Firmicutes*, and *Actinobacteria*) represented by a large amount of sequences, and others by only one or a few. Finally, although a particular phylum may be represented by a large number of sequences, many of them are often very closely related and only add a limited amount

of diversity to the dataset. Thus, methods for efficiently generating optimal sub-samples of sequences are likely to become more and more relevant as more sequences become available. In this dissertation a new procedure is presented to evaluate sample density and to assign weights to individual sequences, which can be used to select “optimal” subsamples of sequences.

Deep-Branching Bacterial Phylogeny

Conserved Amino Acid Types within a Group and Diversity of a Tree

I was interested in identifying protein sites subject to constraining evolution within phylogenetic groups, which could be detected by observing their conservation over a sample of sequences belonging to the group. For example, an amino acid type may be considered as “conserved” when its fraction over all sequences in the data is greater than a given threshold. If multiple groups with different levels of representation are present, this strategy may lead to a systematic error in defining conserved sites. For example, to satisfy a threshold of more than 80% of sequences conserved, in a 100-sequences group at least 81 sequences have to share an amino acid as while a 6-sequence group needs to have only 5 conserved sequences. At the time of the study, protein homologs from some bacterial phyla could be collected from a large number of genomes (e.g. the *Proteobacteria* and *Firmicutes*, with up to 400 genomes) while most other phyla had less than 10 genomes available per group (e.g. the *Verrucomicrobia* with 4 genomes, and *Nitrospirae* with 3 genomes). Representation of groups by similar number of sequences does not necessarily make group sampling balanced. Another factor that should be considered in constructing a balanced sample of sequences is the degree of diversity represented by those sequences. Counts of number of sequences

within a group, therefore, should be replaced by some measure of group "size" that reflects the genetic diversity of the sequences. Given the same number of sequences, the size of a group can be quite different depending on the shape of the underlying phylogenetic tree. Thus, the tree having leaves connected by very short internal branches to the last common ancestor may correspond to a smaller "size" (lower diversity) than the tree with fewer leaves connected by much longer branches (Figure 2-7).

The degree of difference or similarity among samples, called as "biodiversity" in species, can be quantified by assessment of the species' evolutionary relationship in the form of taxonomy or a phylogeny (Vellend et al. 2010). To quantify the evolutionary distinctness of individual sample in a group, indices of conservation have been developed in two types. The first metrics are calculated at the level of species/individuals (taxa) as in Taxonomic Diversity (Vane-Wright et al. 1991), Taxonomic Distinctness (Warwick and Clarke 1995), and Evolutionary Distinctness (Isaac et al. 2007). These metrics measure species' contribution to the diversity as distinctness score and then calculating some function of these scores (typically the sum) for particular focal subsets of species to yield a metric of phylogenetic diversity (Vellend et al. 2010).

$$ED(T, i) = \sum_{b \in s(T,i,r)} (\lambda_b \cdot \frac{1}{S_b}) \quad (2\text{-}2)$$

where *b* is a branch of length λ in the set connecting species *i* to the root, *r* and S_b is the number of species that descend from branch *b* (Isaac et al. 2007; Cadotte et al. 2010). The second type of metrics for phylogenetic diversity is to measure how much evolutionary history is contained in a given area at the level of community/ecosystem.

Phylogenetic Diversity (Faith 1992) calculates sum of all branch lengths in the portion of a phylogenetic tree connecting samples in a group and Mean Phylogenetic Distance (Webb 2000; Webb et al. 2002) calculates mean phylogenetic distance between each pair of species in a group.

As sequences of a group represented changed over different positions in the alignment, it is need to measure each sequence's weight (contribution) to the diversity of the complete tree to measure diversity of each group of sequences per position such as *ED*. As alternative of evolutionary distinctness, Expected genetic diversity, referred to E_{gd}, was developed in this study. This measure calculates the expected number of leaf-states not identical by descent based on the probability of no substitution along the branch. Therefore, it depends on substitution process as well as tree topology.

Expected Number of States Not identical by Descent (*Expected genetic diversity, E_{gd}*)

In this section a measure of genetic diversity (as opposed to phenotypic diversity) was developed, named by "Expected genetic diversity" (E_{gd}). Expected genetic diversity is an estimate of how many states not identical by descent are expected to be represented at the leaves of a phylogenetic tree (or of a cluster), where in our implementation states correspond to amino acid types.

Derivation of E_{gd} of a given tree

For each node *i* of a rooted tree the two quantities of the descendent cluster are evaluated (Figure 2-8):

1. The expected number $E_i \geq 1.0$ of leaf-states not identical by descent;
2. The probability $r_i (0.0 < r_i \leq 1.0)$ that one of the leaf-states is identical by descent to the state at node *i*. N_a and r_a were recursively calculated for the cluster associated with parental node *a*, given the two descendant clusters (nodes) *i* and

j, with corresponding values E_i, r_i, N_j and r_j , and corresponding branches d_i and d_j connecting the clusters to *a*.

The tree is initialized as $E_i = r_i = 1$ at each external node (leaf) and then values for the parental node of pairs of clusters are recursively calculated using previously calculated values. If $p(d)$ is the probability of no substitutions along a branch of length d, the values for the cluster associated with parental node *a* are calculated as:

$$E_a = \left(E_i + E_j\right)\left[1 - p\left(d_i + d_j\right)\right] + \left[E_i + E_j - r_i r_j\right]p\left(d_i + d_j\right)$$

$$= E_i + E_j - p(d_i)p\left(d_j\right) \quad (2\text{-}3)$$

$$r_a = \left[1 - (1 - r_i)\left(1 - r_j\right)\right]p\left(d_i + d_j\right) + r_j[1 - p(d_i)]p(d_i) + r_i\left[1 - p\left(d_j\right)\right]p\left(d_j\right)$$

$$= r_i p(d_i) + r_j p\left(d_j\right) - r_i r_j p(d_i)p(d_j) \quad (2\text{-}4)$$

The new values are used to traverse the tree from the leaves to the root and the final value N_{root} corresponds to the expected number of states not identical by descent for the whole tree (E_{gd}). Assuming that substitutions follow a Poisson distribution, then $p = e^{-d}$ and the previous relations become:

$$E_a = E_i + E_j - r_i r_j e^{-(d_i + d_j)} \quad (2\text{-}5)$$

$$r_a = r_i e^{-d_i} + r_j e^{-d_j} - r_i r_j e^{-(d_i + d_j)} \quad (2\text{-}6)$$

Based on the definition of E_{gd}, to each OTU can be assigned a weight, calculated as follows.

To each leaf *f* belonging to a (sub)tree *T* rooted by *root* can be assigned a weight w_f, such that $\sum_f w_f = E_{root}$, the expected genetic diversity of the (sub)tree. For each node *a* of the tree, the *k*-th element $\rho_k^{(a)}$ of vector $\rho^{(a)}$ indicates the probability that *k* leaves within the cluster identified by descent from node *a*, are identical by descent to

the state at node *a*. The vector $\rho^{(i)}$ associated to each leaf *f* is initialized to $\rho_1^{(f)} = 1.0$ and $\rho_k^{(f)} = 0.0 \; \forall k \neq 1$.

Given nodes *i* and *j* directly descending from node *a* through branches d_i and d_j, with associated $\rho^{(i)}$ and $\rho^{(j)}$ vectors, respectively, vector $\rho^{(a)}$ can be recursively calculated as:

$$\rho_0^{(a)} = \; [1 - p(d_i)]\big[\, 1 - p\big(d_j\big)\big] + [1 - p(d_i)]p\big(d_j\big)\rho_0^{(j)} + p(d_i)\big[1 - p\big(d_j\big)\big]\rho_0^{(i)}$$

$$+p\big(d_i + d_j\big)\rho_0^{(i)}\rho_0^{(j)} \;, \qquad (2\text{-}7)$$

and:

$$\rho_k^{(a)} = \; [1 - p(d_i)]p\big(d_j\big)\rho_0^{(j)} + \; p(d_i)\big[1 - p\big(d_j\big)\big]\rho_0^{(i)}$$

$$+ \, p\big(d_i + d_j\big) \textstyle\sum_{h=0}^{k} \rho_h^{(i)}\rho_{k-h}^{(j)} \;, \forall k \geq 1. \qquad (2\text{-}8)$$

Rooting the tree in coincidence to leaf *f*, *root* ≡ *f*, with *j* = *f* it is $d_f = 0.0$ and $p\big(d_f\big) = 1.0$. The vector $\rho^{(root)}$ is then calculated using the above relations, simplified as:

$$\rho_0^{(root)} = \; 0.0, \qquad (2\text{-}9)$$

$$\rho_1^{(root)} = [1 - p(d_i)] + p(d_i)\rho_0^{(i)}, \qquad (2\text{-}10)$$

and:

$$\rho_k^{(root)} = p(d_i)\rho_{k-1}^{(i)}, \forall k > 1 \qquad (2\text{-}11)$$

The weight w_f of leaf *f* is finally calculated as:

$$w_f = \textstyle\sum_{k=1} \frac{\rho_k^{(root)}}{k} \qquad (2\text{-}12)$$

The procedure is repeated for each leaf, obtaining weights w_f for all leaves. It can be shown that

$$\sum_f w_f = E_{gd}. \quad (2\text{-}13)$$

In this way, E_{gd} of a cluster can be calculated by estimating the weight of each leaf.

Measure of group density and relatedness derived from E_{gd}

The definition of E_{gd} suggests definition of the following related quantities:

$$sample\ density\ (\rho) = \frac{N}{E_{gd}} \quad (2\text{-}14)$$

$$group\ relatedness\ (\gamma) = \frac{N - E_{gd}}{N - 1} \quad (2\text{-}15)$$

where *N* is the number of leaves. Because $1 \leq E_{gd} < N$ ($E_{gd} = N$ only if $N = 1$), sampling density is always at least 1.0 and never greater than N. Group relatedness is 1.0 if all states share the same state identical by descent. Group relatedness is 0 if all members of the group are unrelated (connected by branches of infinite length).

When applied to homologous sequences from a large collection of complete bacterial genomes, E_{gd} and ρ provide information accounting for phylogenetic relations on which groups are more diversely represented and on the density with which different bacterial groups have been sampled.

Sizing the Bacterial Phylogenetic Tree and Subtrees of Groups

The weight of each branch on the complete tree was calculated. The E_{gd} of the complete tree is the sum of all weights, was $E_{gd} = 218.31$. The "size" of the complete tree is smaller than the total number of sequences (1,814 sequences) and greater than the number of groups (26 groups). To calculate E_{gd} for each group, the weights of sequences belonging to a group were summed (Table 2-6). This measure of group size shows a positive but not linear correlation with the number of sequences within a group (Figure 2-9). E_{gd} of the biggest group, the *Gammaproteobacteria* with 415 sequences, is 28.292 and the smallest groups, the *Nitrospirae* with 3 sequences and the

Deferribacteres with 4 sequences, had E_{gd} values around 1.5. The *Chlamydiae* clade is composed of 39 sequences but its expected genetic diversity was E_{gd} = 2.922. The E_{gd} of the *Lactobacillales* and of *Bacillales* are smaller than those of the *Clostridia*, *Bacteroidetes*, and *Betaproteobacteria* even though the number of sequences in those groups is larger, indicating that sequences in *Lactobacillales* and of *Bacillales* are more closely related (connected by shorter branches) than in *Clostridia*, *Bacteroidetes*, or *Betaproteobacteria*.

Based on E_{gd}, measures of sampling density and group relatedness of each group can be defined (see above). The groups with the highest density are the *Bacillales* and *Lactobacillales*. The *Gammaproteobacteria* contains more sequences than those groups, but their density is lower (Figure 2-10). Relatedness of all groups is greater than 0.5. The lowest relatedness is observed in the *Verrucomicrobia* and *Planctomycetes* and the highest is the *Gammaproteobacteria* (Figure 2-11). Relatedness showed higher relation to the number of leaves than the density but the *Nitrospirae* group showed higher relatedness even though it contains only 3 sequences, less than the *Planctomycetes* and *Verrucomicrobia*.

The size of the tree, as measured by E_{gd}, varies along an alignment depending on which sequences are represented at a particular position (i.e., on the presence of gaps). Using weights of all branches, E_{gd} was calculated position by position. The E_{gd} of a position is the sum of weights of sequences having any amino acid at the position. Among 72,700 positions, 23,934 positions have E_{gd} less than 5 (Figure 2-12). The positions having low E_{gd} values seem to not have enough positions to derive conserved

amino acids within a group so a threshold for E_{gd} of a group should be applied on all of groups in a tree.

Finding Conserved Amino Acid Type for a Group on a Position

Once the weights of all terminal leaves in the ML supertree are calculated, sequences on each position are divided into sub-alignments corresponding to the groups. The size of a group (E_{gd}) is calculated as the sum of weights of all branches except for gapped branches and the weight of the most frequent amino acid types is the sum of leaves with the most frequent amino acid type.

For each position, E_{gd} is calculated by summing weights of the leaves having any amino acid type and belonging to a group. ACD for a group is calculated as follows. Among all amino acid types shown in the position, only the leaves with the most frequent amino acid type are chosen and summed then divided by E_{gd}.

After applying the two thresholds to all groups on each position, each group would have a gap or one amino acid at the position in its conserved amino acid sequence (CAS). All the positions are concatenated into a big alignment over different protein families. The positions are classified into three different types based on the number of amino acid types and contribution to generation of the tree: 1) phylogenetic-informative positions - at least two or more amino acid types shared with two or more groups, 2) not-informative positions – at least two or more amino acid types but not informative, 3) conserved positions between groups – only one amino acid type and no gaps are shown. Only phylogenetic informative sites could contribute to reconstruct trees.

Deep-Branching Trees Based on CAS Alignments

Two thresholds are applied to get conserved amino acids for each group on each sequence: (1) group size by E_{gd} and (2) amino acid conservation degree (ACD) which means the fraction of the weight of the most frequent amino acid type in the group. Due to different distribution of gaps at different positions, group size depends on position, and E_{gd} of a group is calculated for each position as the sum of weights of non-gapped sequences. If both thresholds are satisfied for a particular amino acid type within a group at a given position, the corresponding Conserved Amino Acid sequence is assigned that amino acid type; otherwise it is assigned a gap.

In this study, a total of 30 combinations of two thresholds were used: 10 E_{gd} thresholds from 1.5 to 6.0 increased by 0.5 and 0.8, 0.9, and 1.0 as ACD in the group. With different sets of E_{gd} threshold and ACD, a different number of groups were included in the CAS alignment. As thresholds become more stringent, the number of phyla/positions of big enough size, and the number of conserved positions decrease. All groups could be satisfied when E_{gd} is lowered onto 1.5 ($E_{gd} < 1.5$).

The ML trees from the CAS alignments at different threshold were generated. Firstly a draft tree is estimated using RAxML v.7.3.2 with GTR+Γ+I (command= raxmlHPC-PTHREADS-SSE3 -n lg.all.rxl -m PROTGTRI -c 4 -f a -N 1 -T 8), estimation of GTR matrix, gamma-distributed rates and proportion of invariants. To find substitution pattern and rates between amino acids, a GTR matrix is estimated during the RAxML process. PhyML optimizes the ML tree and calculates aLRT support values.

The trees with ACD=1.0 and different E_{gd} were shown in Figure 2-13. All of the trees showed that five classes of the *Proteobacteria* formed the *Proteobacteria* clade,

and the *Lactobacillales*, *Bacillales* and *Clostridia* grouped well into a clade of the *Firmicutes.*

All trees showed three big clusters: (1) the *Bacillales, Lactobacillales*, and *Clostridia* (three classes of the *Firmicutes), Actinobacteria, Cyanobacteria, Tenericutes, Thermotogae, Deinococcus-Thermus, Fusobacteria* and *Synergists* (called the "F cluster") (2) the *Proteobacteria, Acidobacteria, Deferribacteres, Nitrospirae*, and *Aquificae* (called the "P cluster"), and (3) the *Bacteroidetes, Chlorobi, Planctomycetes*, and *Chlamydiae* (called the "B cluster"). Except for the *Spirochaetes*, three clusters were maintained over different trees as long as the groups were contained in the tree (Figure 2-13). While most of groups showed consistent patterns of grouping over different thresholds, the location of the *Spirochaetes* was changed.

Substitution patterns of conserved amino acids might be different from represented amino acids in the alignment. Instead of using a fixed amino acid substitution model, GTR matrices were derived from CAS alignments and used for constructing trees. When compared to LG matrix (Figure 2-14), GTR matrices showed positive correlation except for some rates of amino acids, which were not seen in the data set. This trend was maintained with thresholds changed. This means that substitution pattern of amino acids in deep branching might be similar to the patterns of terminal branches.

CAS Alignments of Informational and Not-informational Protein Families

The positions on a CAS alignment were divided according to the functional category of protein families each position belongs to: (1) Informational alignment with 118 families (Table 2-4) (2) Non-informational alignment with 51 families (Table 2-5). With the same procedure and set of thresholds applied on the informational alignment

and the non-informational alignment, CAS alignments with different number of positions were generated (Table 2-8 for the informational families and Table 2-9 for the non-informational families). The ML trees with aLRT support values were generated (Figure 2-15). In both informational and non-informational alignments, most groups except for the *Aquificae*, *Epsilonproteobacteria*, and *Spirochaetes* were clustered to the three large clusters: the B cluster (the *Bacteroidetes* and others), F cluster (the *Firmicutes* and others) and P cluster (the *Proteobacteria* and others). The placement of the *Epsilonproteobacteria* was found to be dependent on the collection of families. Informational alignments showed that the *Epsilonproteobacteria* was grouped with other proteobacteria while it is located outside of the *Proteobacteria* clade in the Non-informational tree. Compared to the tree based on the alignment with all families, the *Aquificae, Epsilonproteobacteria,* and *Spirochaetes* are more dependent to the collection of protein families and the others are more consistently grouped independent from the functions of protein families.

Estimation of the Amount of Homoplasy in Position-Specific Manner

It is well known that bacterial genomes have shown homoplasy frequently due to horizontal gene transfer, transposon elements, and convergent evolution. There are two classical measures to detect Homoplasy based on parsimony, Consistency Index and Retention Index (Farris 1989). Retention index is scaled to [0,1], indicating that higher RI means the lower homoplasy, so it is more appropriate to compare different sets of data. The average Retention Index is calculated on positions for CAS alignments as well as the complete alignment.

Retention Index is calculated by

$$RI = g - s/g - m \quad (2\text{-}16)$$

where *g* = maximum steps to reconstruct the tree, *s* = number of steps to reconstruct the tree under investigation, and *m* = the minimum number of steps to reconstruct the tree. To estimate the number of steps to reconstruct the tree, the Fitch algorithm (Fitch 1971) a parsimony method is used. In the Fitch algorithm (Fitch 1971), the possible states of internal node *i* having the nodes, *j* and *k* as its children can be obtained by

$$S_i = \begin{cases} S_j \cup S_k & if\ S_j \cap S_k \neq \emptyset \\ S_j \cap S_k & elsewhere \end{cases} \quad (2\text{-}17)$$

where S_w is the set of amino acids at the node *w*. All terminal leaves are assigned as corresponding amino acids and then possible states of internal nodes are calculated, traversing from terminal leaves to root. The number of steps (changes) to reconstruct the tree, *s*, is the number of union operations used while traversing the tree. The minimum number of steps for the tree, *m*, is the number of amino acid types -1. In unordered multiple states like amino acid types, maximum steps (changes) for the tree, *g*, can be (the number of leaves with any amino acid type) – (the number of leaves with the most frequent amino acid type) (Steel and Penny 2006; Cuthill et al. 2010).

The RI of a CAS alignment was increased as the threshold was increased, meaning that more strict thresholds showed lower homoplasy level (Figure 2-16). The alignments with the lower group-thresholds contain more groups with smaller number of sequences, e.g. the *Nitrospirae* (3 sequences) and *Fusobacteria* (4 sequences). According to the assumptions of predicting the ancestral state (amino acid type), only existing sequences should be considered to get the ancestral state of the “small” groups so that the resulting amino acid type might not be noise rather than “truly” the ancestral states. The possibility of Homoplasy is also increased due to the increased noise in the ancestral sequence.

Even though the complete tree was expected to have higher homoplasy, the RI of the complete alignment is 0.831 while most of CAS alignments showed RI less than 0.81. From the formulation of RI, the number of leaves in the tree can affect the maximum number of steps to construct the tree. That is, when the number of leaves of the tree is extremely bigger than the number of character states, the maximum number of steps could be extremely high and then RI should be high no matter what RI is. Therefore, I concluded it is difficult to compare the amount of homoplasy of trees with extremely different numbers of leaves such as the complete tree (1814 leaves) and the CAS tree (16~25 leaves) based on RI.

Table 2-1. Distribution of groups over protein families in the dataset.

Group	Number of sequences
Acidobacteria	8
Actinobacteria	206
Aquificae	10
Bacteroidetes	78
Chlamydiae	39
Chlorobi	11
Chloroflexi	16
Cyanobacteria	41
Deferribacteres	4
Deinococcus-Thermus	17
Fusobacteria	5
Nitrospirae	3
Planctomycetes	5
Spirochaetes	45
Synergistetes	4
Tenericutes	58
Thermotogae	15
Verrucomicrobia	4
Firmicutes	
Bacillales	134
Clostridia	108
Lactobacillus	149
Proteobacteria	
Alpha-proteobacteria	205
Beta-proteobacteria	117
Gamma-proteobacteria	415
Delta-proteobacteria	45
Epsilon-proteobacteria	72
Total	1814

Table 2-2. Group/Species reported as target of HGT (Wolf et al. 1999; Ciccarelli et al. 2006; Kanhere and Vingron 2009).

COG	Groups
COG0008	Proteobacteria, Spirochaetes, Chlamydiae
COG0018	Spirochaetes, Proteobacteria, Cyanobacteria, Chlamydia
COG0024	Rhodopirellula baltica, Epsilonproteobacteria
COG0030	Rhodopirellula baltica, Epsilonproteobacteria
COG0048	Clostridia
COG0060	Spirochaetes, Chlamydia, Mycobacteria,Clostridium acetobutylicum
COG0084	Agrobacterium tumefaciens strain C58 (Cereon)
COG0088	Aquifex aeolicus
COG0099	Aquifex aeolicus
COG0103	Aquifex aeolicus
COG0124	Spirochaetes, Helicobacter, Aquifex, Synechocystis, Bacillus
COG0143	Spirochaetes, Chlamydia, Proteobacteria
COG0162	Escherichia
COG0172	Leptospiracea, Spirochaetes
COG0186	Clostridia
COG0195	Rhodopirellula baltica, Epsilonproteobacteria
COG0197	Clostridia, Thermotogae
COG0201	Clostridia
COG0202	Aquifex aeolicus
COG0216	Rhodopirellula baltica
COG0216	Epsilonprobacteria
COG0227	Deinococcus radiodurans
COG0242	Aquifex aeolicus
COG0359	Aquifex aeolicus
COG0442	Borrelia, Mycoplasma, Spirochates, Borrelia burgdorferi, Deinococcus radiodurans, Mycobacterium leprae
COG0482	Aquifex aeolicus
COG0495	Bacteriovorus
COG0522	Thermotoga maritima
COG0525	Leptospiraceae, Rickettsiaceae
COG0536	Aquifex aeolicus

Table 2-3. Sizes and distribution of groups over protein families in the dataset.

Group	Number of sequences	Number of involved families
Acidobacteria	8	137
Actinobacteria	206	167
Alphaproteobacteria	205	166
Aquificae	10	166
Bacillales	134	167
Bacteroidetes	78	156
Betaproteobacteria	117	165
Chlamydiae	39	159
Chlorobi	11	129
Chloroflexi	16	150
Clostridia	108	162
Cyanobacteria	41	164
Deferribacteres	4	130
Deinococcus	17	162
Deltaproteobacteria	45	161
Epsilonproteobacteria	72	161
Fusobacteria	5	156
Gammaproteobacteria	415	156
Lactobacillus	149	163
Nitrospirae	3	135
Planctomycetes	5	127
Spirochaetes	45	159
Synergistetes	4	130
Tenericutes	58	164
Thermotogae	15	161
Verrucomicrobia	4	144

Table 2-4. COG functional categories of informational protein families.

Category	Number of Families	Description
Information storage and processing		
J	89	Translation, ribosomal structure and biogenesis
A	0	RNA processing and modification
K	8	Transcription
L	21	Replication, recombination and repair
B	0	Chromatin structure and dynamics
Total	118 families	

Table 2-5. COG functional categories of non-informational protein families.

Category	Number of Families	Division
Cellular processes and signaling		
D	1	Cell cycle control, cell division, chromosome partitioning
Y	0	Nuclear structure
V	1	Defense mechanisms
T	0	Signal transduction mechanisms
M	4	Cell wall/membrane/envelope biogenesis
N	0	Cell motility
Z	0	Cytoskeleton
W	0	Extracellular structures
U	6	Intracellular trafficking, secretion, and vesicular transport
O	13	Posttranslational modification, protein turnover, chaperones
Metabolism		
C	3	Energy production and conversion
G	2	Carbohydrate transport and metabolism
E	3	Amino acid transport and metabolism
F	5	Nucleotide transport and metabolism
H	3	Coenzyme transport and metabolism
I	3	Lipid transport and metabolism
P	1	Inorganic ion transport and metabolism
Q	2	Secondary metabolites biosynthesis, transport and catabolism
Poorly characterized		
R	9	General function prediction only
S	1	Function unknown
Total	51 families	

Table 2-6. E_{gd} of each group derived from the ML supertree.

Group	E_{gd}	Number of sequences
Acidobacteria	2.582	8
Actinobacteria	23.078	206
Alphaproteobacteria	23.694	205
Aquificae	2.975	10
Bacillales	7.057	134
Bacteroidetes	16.383	78
Betaproteobacteria	10.063	117
Chlamydiae	2.922	39
Chlorobi	2.310	11
Chloroflexi	4.048	16
Clostridia	21.707	108
Cyanobacteria	6.449	41
Deferribacteres	1.544	4
Deinococcus	3.324	17
Deltaproteobacteria	12.762	45
Epsilonproteobacteria	6.573	72
Fusobacteria	1.944	6
Gammaproteobacteria	28.292	415
Lactobacillus	8.892	149
Nitrospirae	1.583	3
Planctomycetes	2.742	5
Spirochaetes	7.471	45
Synergistetes	1.733	4
Tenericutes	12.114	58
Thermotogae	3.887	15
Verrucomicrobia	2.181	4

Table 2-7. Number of positions and alignment size when different set of thresholds, E_{gd} and ACD, are applied to each position to obtain conserved amino acid alignments.

E_{gd} of Group	Amino acid conservation degree	Number of positions			sum
		informative	not informative	conserved	
1.5	0.8	9729	15321	22135	47185
1.5	0.9	5659	12942	25098	43699
1.5	1.0	3544	11054	26473	41071
2.0	0.8	7321	14416	23000	44737
2.0	0.9	3864	11116	25666	40646
2.0	1.0	1998	8306	26250	36554
2.5	0.8	5334	13008	24160	42502
2.5	0.9	2620	9234	26094	37948
2.5	1.0	1102	5978	25552	32632
3.0	0.8	3528	11187	25718	40433
3.0	0.9	1637	7471	26434	35542
3.0	1.0	510	3769	24155	28434
3.5	0.8	2925	10154	25988	39067
3.5	0.9	1332	6564	26204	34100
3.5	1.0	387	2990	23008	26385
4.0	0.8	2399	9360	26413	38172
4.0	0.9	1072	5875	26113	33060
4.0	1.0	279	2449	22248	24976
4.5	0.8	2037	8830	26700	37567
4.5	0.9	893	5404	26084	32381
4.5	1.0	212	2109	21710	24031
5.0	0.8	1774	8245	26880	36899
5.0	0.9	741	4952	25973	31666
5.0	1.0	155	1799	21285	23239
5.5	0.8	1515	7526	26954	35995
5.5	0.9	614	4434	25632	30680
5.5	1.0	114	1498	20491	22103
6.0	0.8	1331	6835	26678	34844
6.0	0.9	540	3967	25066	29573
6.0	1.0	93	1318	19639	21050

Table 2-8. Number of positions and alignment size when different set of thresholds, E_{gd} and ACD, are applied to each position of informational protein alignment to identify conserved amino acids.

E_{gd} of Group	Amino acid conservation degree	Number of positions			sum
		informative	not informative	conserved	
1.5	0.8	6867	10408	15438	32713
1.5	0.9	4018	8903	17444	30365
1.5	1.0	2523	7685	18412	28620
2.0	0.8	5360	10030	15687	31077
2.0	0.9	2869	7869	17597	28335
2.0	1.0	1500	6001	18200	25701
2.5	0.8	3922	9133	16436	29491
2.5	0.9	1936	6549	17966	26451
2.5	1.0	821	4328	17808	22957
3.0	0.8	2609	7936	17579	28124
3.0	0.9	1204	5393	18265	24862
3.0	1.0	379	2776	16942	20097
3.5	0.8	2163	7226	17807	27196
3.5	0.9	982	4734	18169	23885
3.5	1.0	281	2195	16192	18668
4.0	0.8	1761	6695	18137	26593
4.0	0.9	781	4237	18113	23131
4.0	1.0	207	1788	15654	17649
4.5	0.8	1511	6385	18354	26250
4.5	0.9	648	3946	18151	22745
4.5	1.0	154	1554	15391	17099
5.0	0.8	1340	5919	18470	25729
5.0	0.9	558	3575	18056	22189
5.0	1.0	120	1309	15056	16485
5.5	0.8	1128	5359	18586	25073
5.5	0.9	451	3162	17843	21456
5.5	1.0	83	1059	14489	15631
6.0	0.8	989	4804	18565	24358
6.0	0.9	395	2792	17539	20726
6.0	1.0	66	914	13859	14839

Table 2-9. Number of positions and alignment size of the non-informational protein alignment when different set of thresholds, E_{gd} of Group and Amino acid conservation degree, are applied to each position.

E_{gd} of Group	Amino acid conservation degree	Number of positions			sum
		informative	not informative	conserved	
1.5	0.8	2862	4913	6697	14472
1.5	0.9	1641	4039	7654	13334
1.5	1.0	1021	3369	8061	12451
2.0	0.8	1961	4386	7313	13660
2.0	0.9	995	3247	8069	12311
2.0	1.0	498	2305	8050	10853
2.5	0.8	1412	3875	7724	13011
2.5	0.9	684	2685	8128	11497
2.5	1.0	281	1650	7744	9675
3.0	0.8	919	3251	8139	12309
3.0	0.9	433	2078	8169	10680
3.0	1.0	131	993	7213	8337
3.5	0.8	762	2928	8181	11871
3.5	0.9	350	1830	8035	10215
3.5	1.0	106	795	6816	7717
4.0	0.8	638	2665	8276	11579
4.0	0.9	291	1638	8000	9929
4.0	1.0	72	661	6594	7327
4.5	0.8	526	2445	8346	11317
4.5	0.9	245	1458	7933	9636
4.5	1.0	58	555	6319	6932
5.0	0.8	434	2326	8410	11170
5.0	0.9	183	1377	7917	9477
5.0	1.0	35	490	6229	6754
5.5	0.8	387	2167	8368	10922
5.5	0.9	163	1272	7789	9224
5.5	1.0	31	439	6002	6472
6.0	0.8	343	2030	8113	10486
6.0	0.9	145	1175	7527	8847
6.0	1.0	27	404	5780	6211

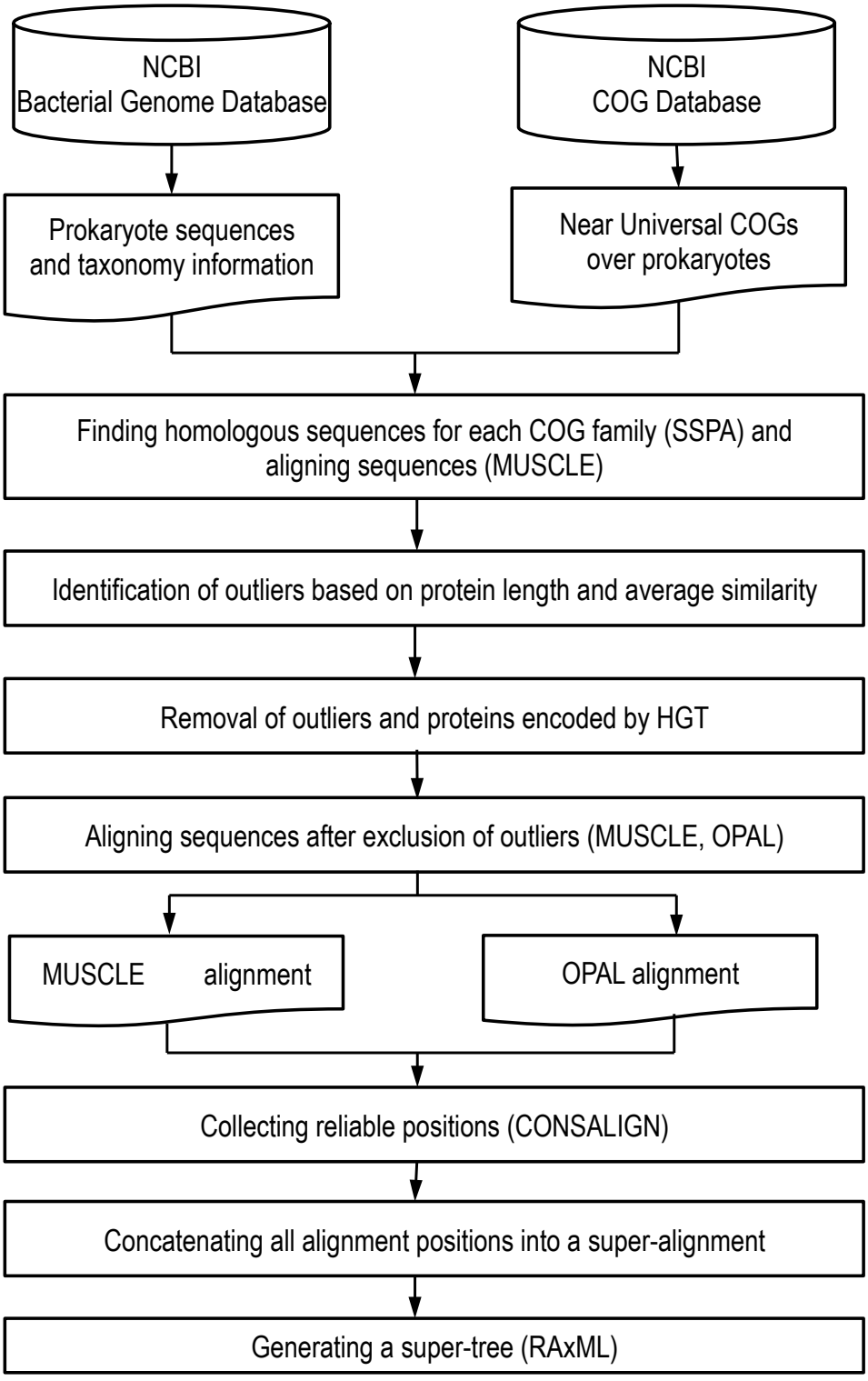


Figure 2-1. A workflow for data construction and generation of super tree.

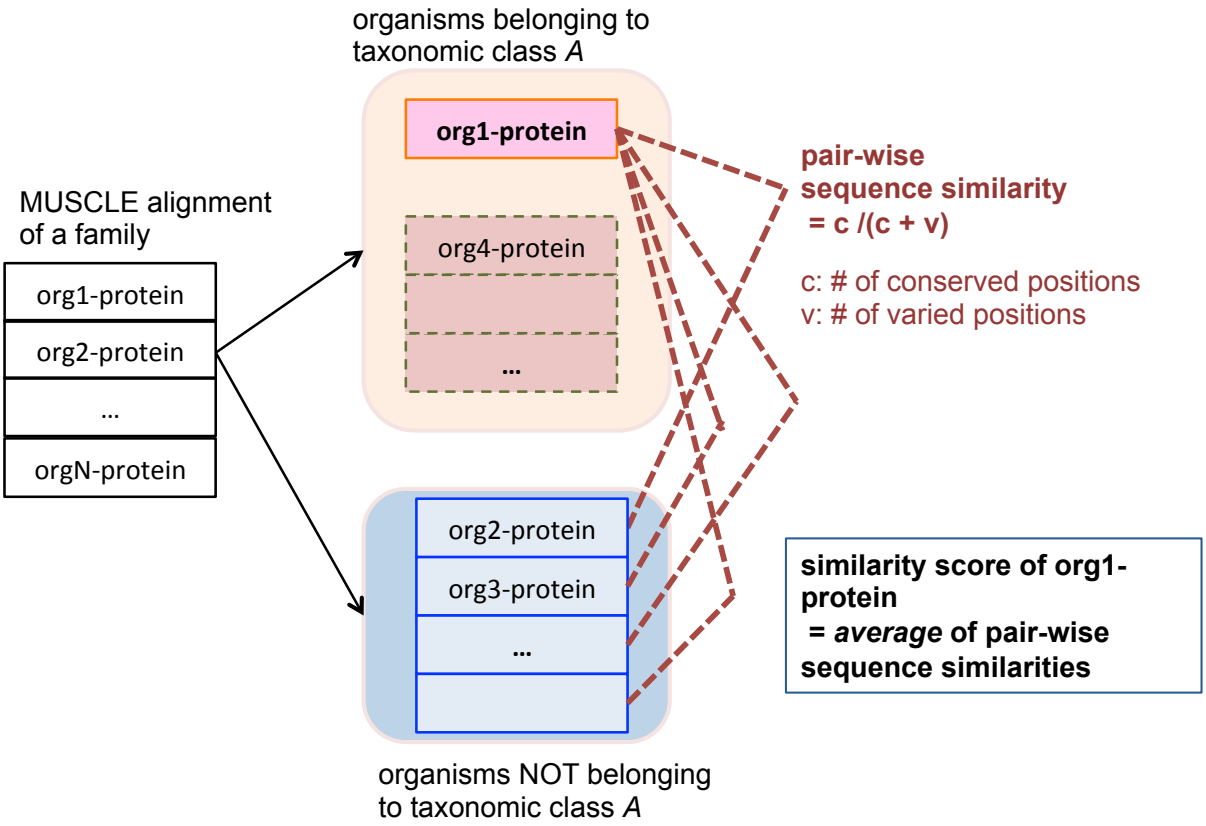


Figure 2-2. An example of calculating similarity score of a protein based on average of pairwise sequence similarity.

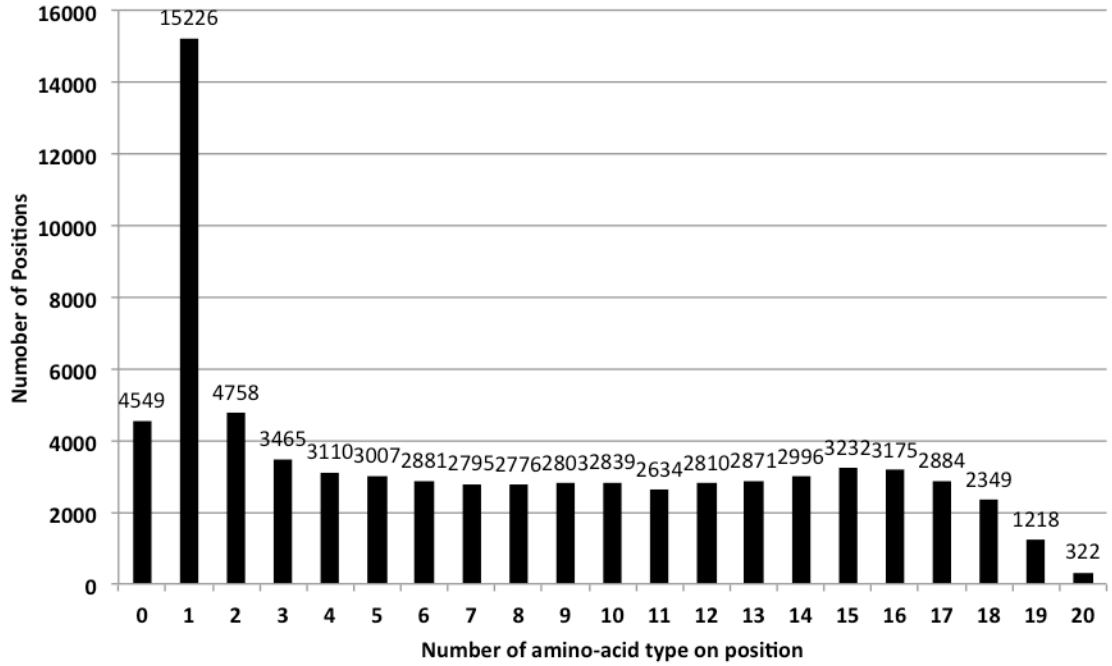


Figure 2-3. Histogram showing distribution of amino acid subset size over 72,700 positions in the complete alignment.

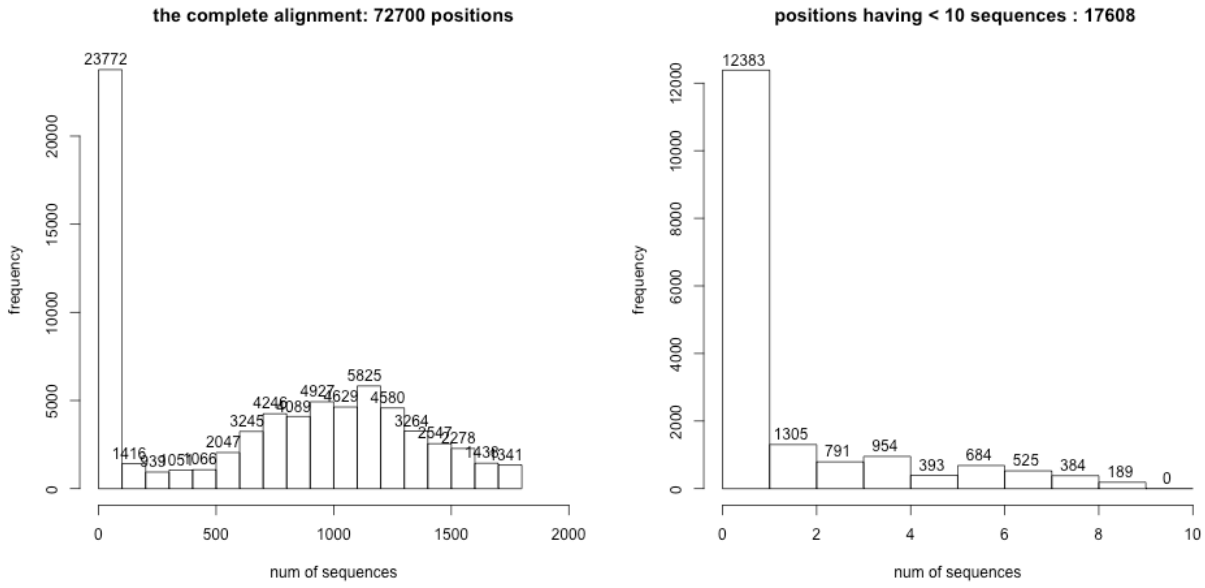


Figure 2-4. Distribution of number of sequences represented on each position.

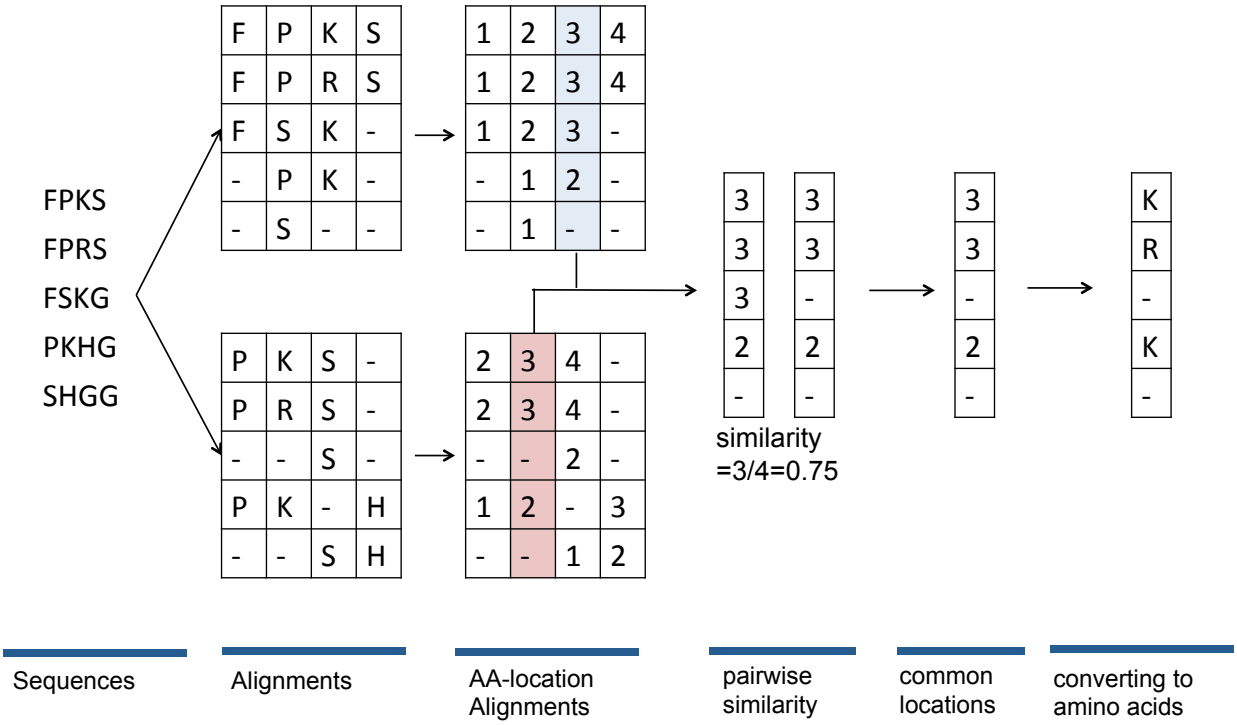


Figure 2-5. An example of obtaining reliable position from alignments of different softwares.

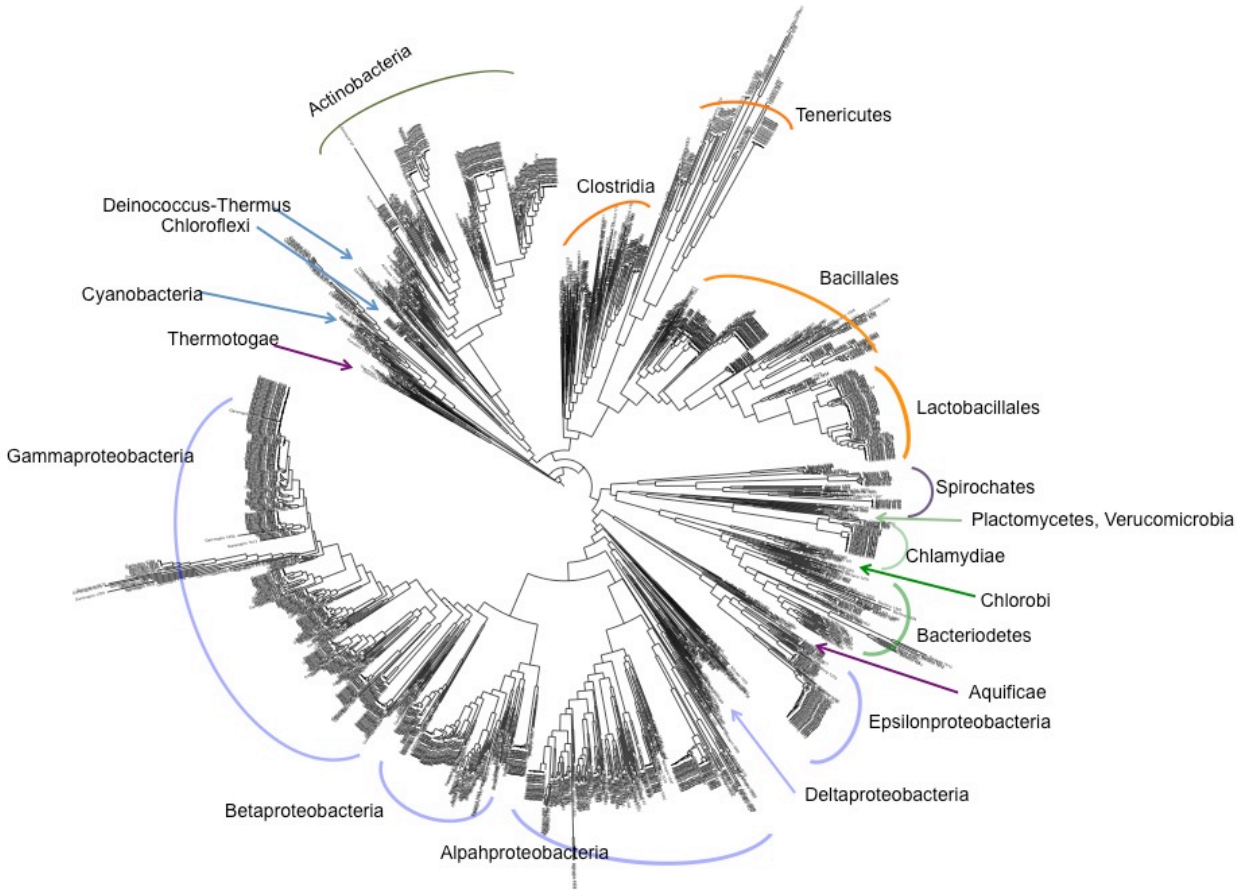


Figure 2-6. A maximum likelihood tree of bacterial sequences is obtained using RAxML (Stamatakis 2006). The tree was visualized by a tree viewer, iTOL (Letunic and Bork 2007).

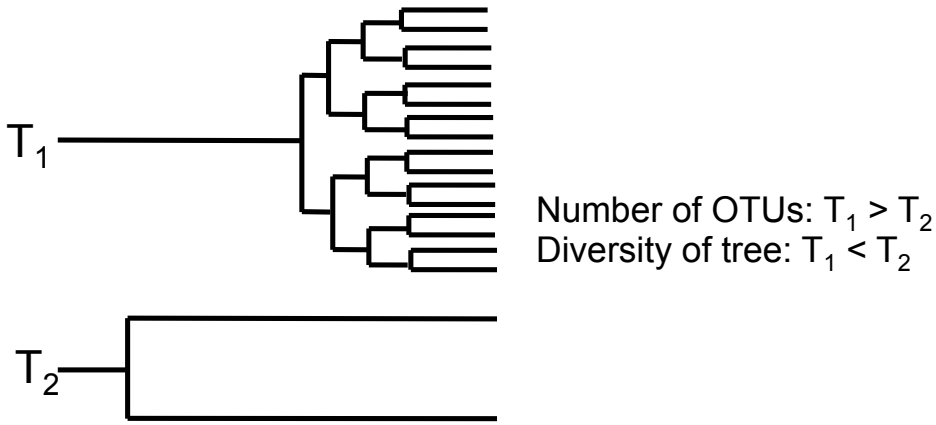


Figure 2-7. Diversity of a tree depending on the tree topology. A tree having leaves connected to each other by very short branches may correspond to a smaller number of "independent" observations than a tree with fewer leaves connected by very long branches.

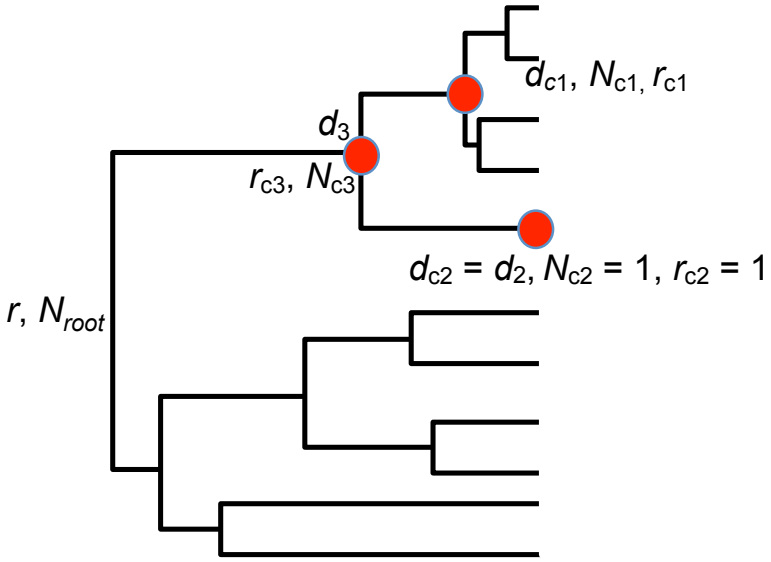


Figure 2-8. Calculation of E_{gd} with traversing the tree from leaves to the root.

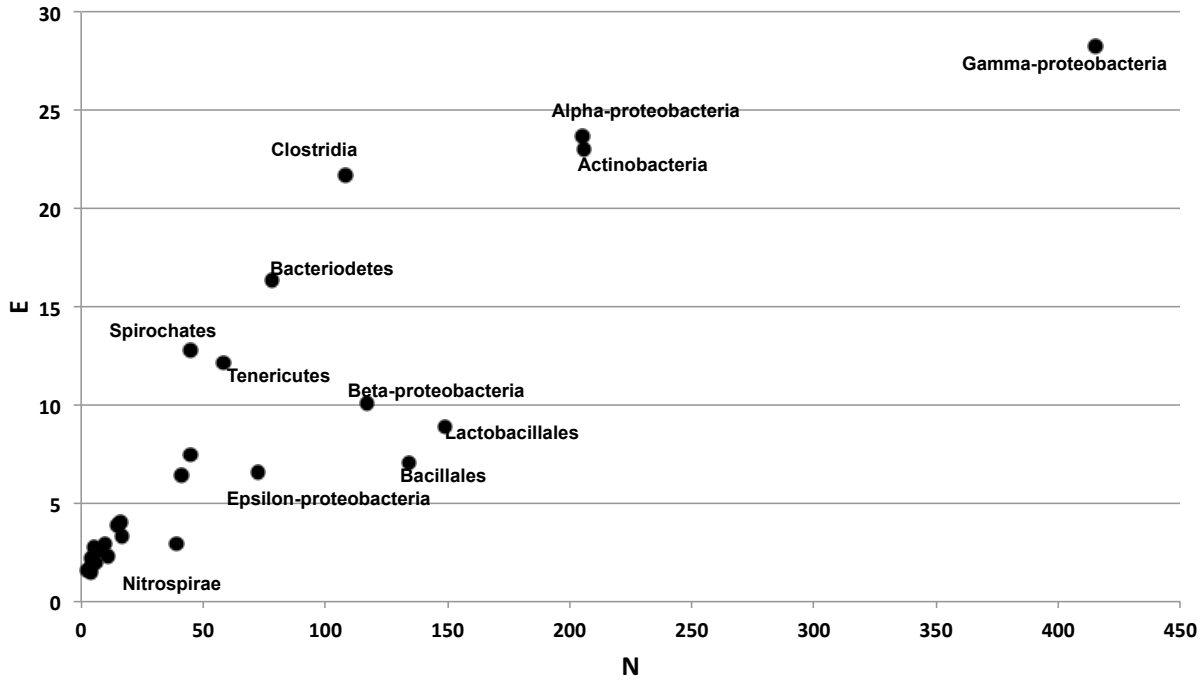


Figure 2-9. Plot of the number of sequences and E_{gd} for each group.

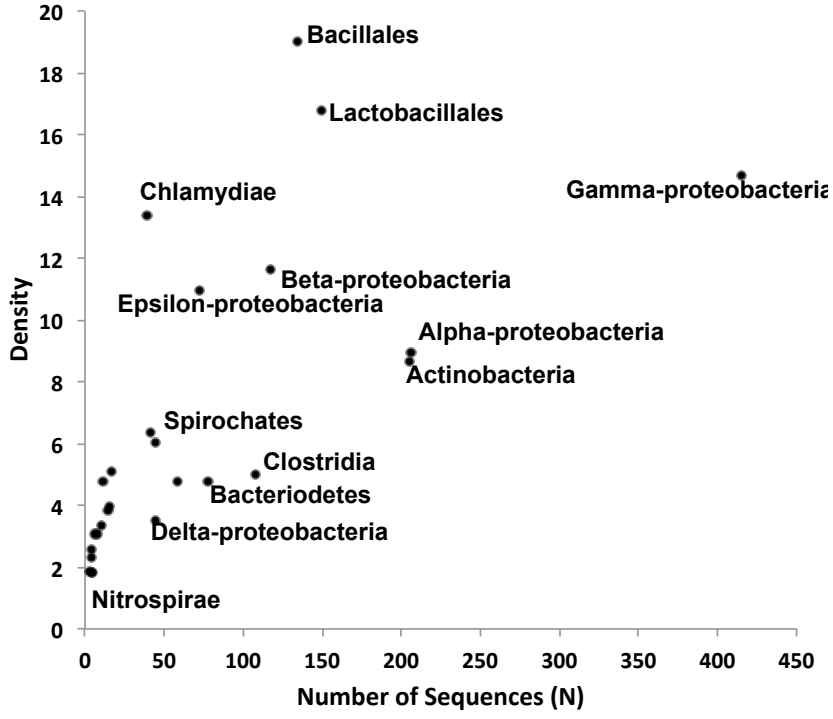
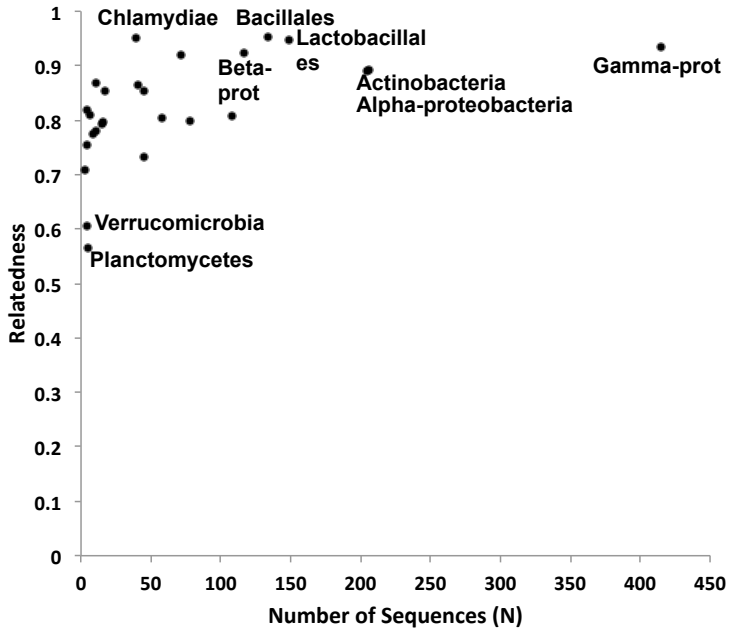


Figure 2-10. A plot of density and number of sequences of group.



Group	Number of sequences	Relatedness
Acidobacteria	8	0.774
Actinobacteria	206	0.892
Alphaproteobacteria	205	0.889
Aquificae	10	0.781
Bacillales	134	0.954
Bacteroidetes	78	0.800
Betaproteobacteria	117	0.922
Chlamydiae	39	0.949
Chlorobi	11	0.869
Chloroflexi	16	0.797
Clostridia	108	0.806
Cyanobacteria	41	0.864
Deferribacteres	4	0.819
Deinococcus	17	0.855
Deltaproteobacteria	45	0.733
Epsilonproteobacteria	72	0.922
Fusobacteria	6	0.811
Gammaproteobacteria	415	0.934
Lactobacillus	149	0.947
Nitrospirae	3	0.709
Planctomycetes	5	0.565
Spirochaetes	45	0.853
Synergistetes	4	0.756
Tenericutes	58	0.805
Thermotogae	15	0.794
Verrucomicrobia	4	0.606

Figure 2-11. A plot and table of relatedness and number of sequences of groups.

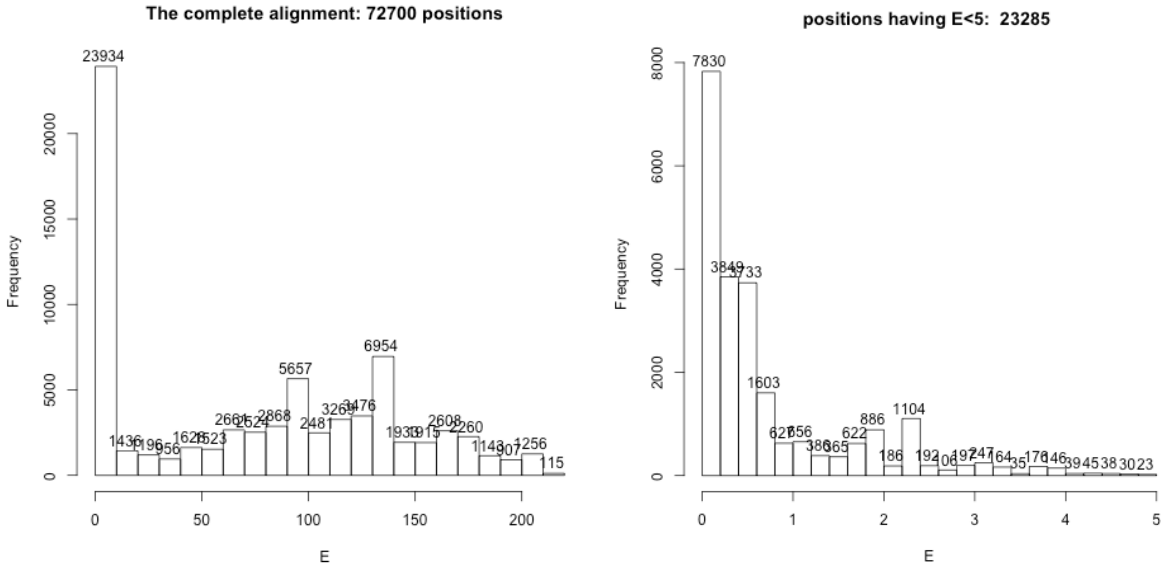


Figure 2-12. Histogram of E_{gd} over all positions. Left is E_{gd} of all positions and right is E_{gd} of positions having E_{gd} <5.

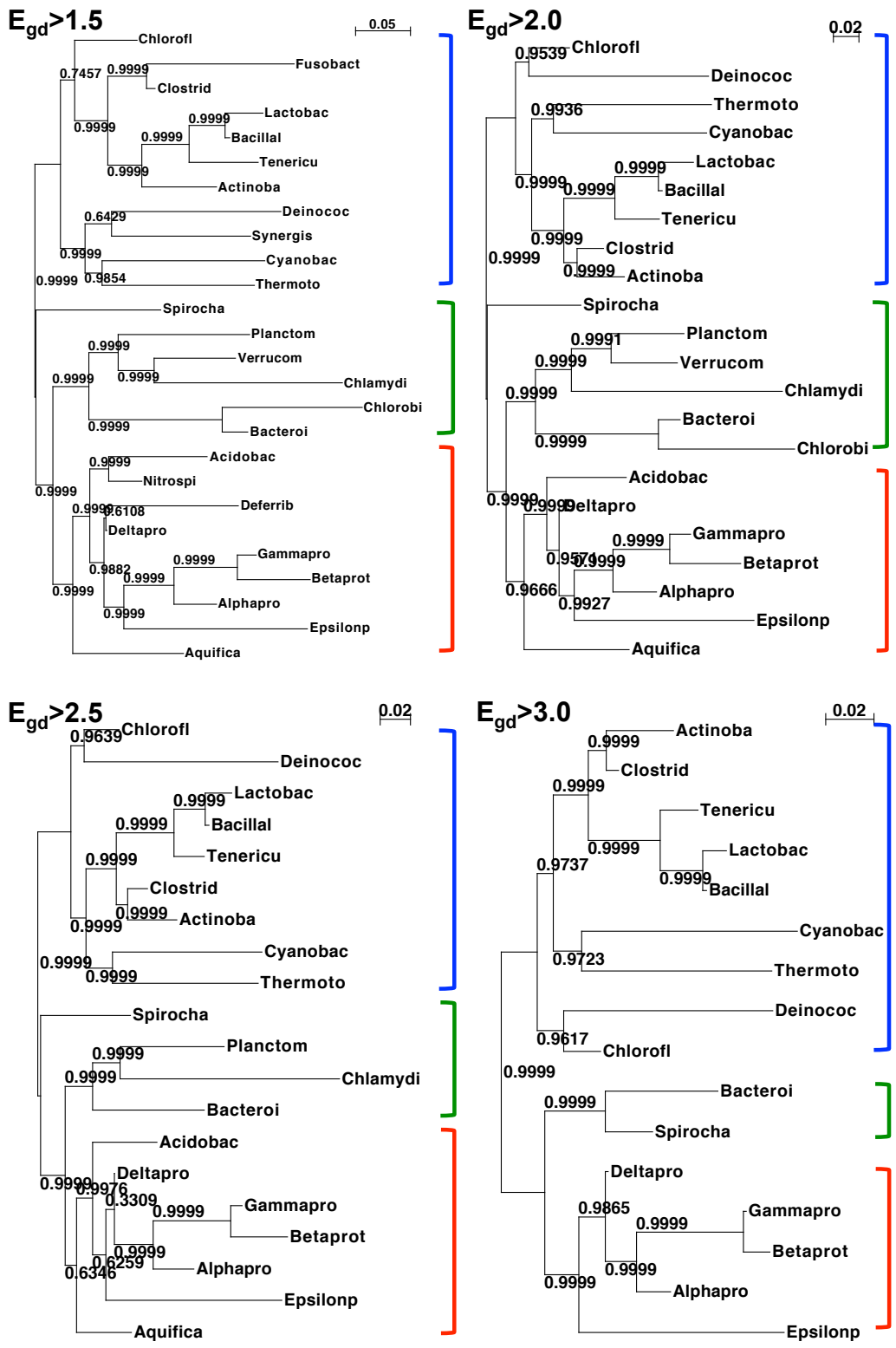


Figure 2-13. ML trees from conserved amino acid alignments when different threshold of E_{gd} is applied with ACD =1.0. The aLRT supports on most branches were shown very high (>0.99).

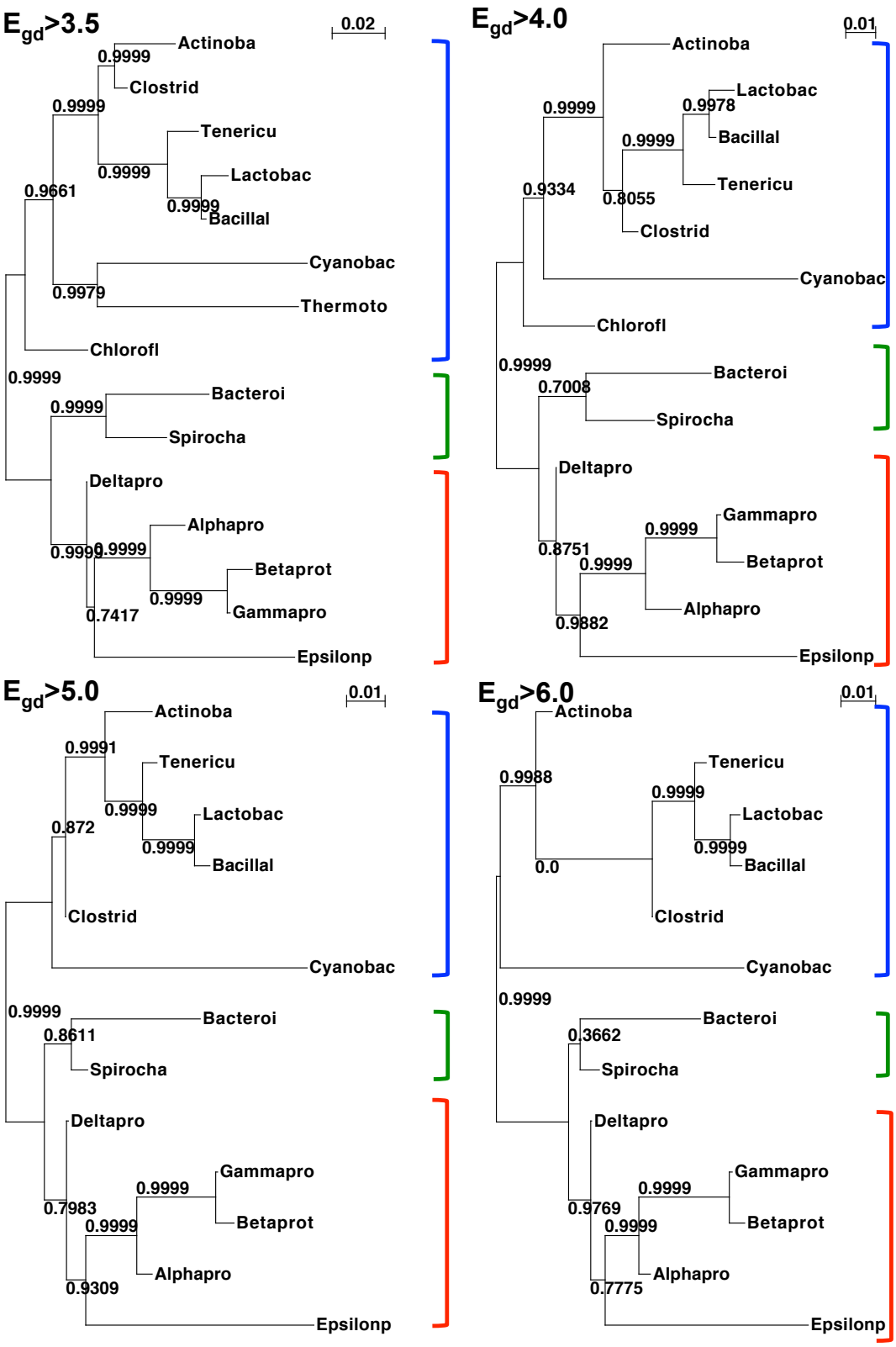


Figure 2-13. Continued.

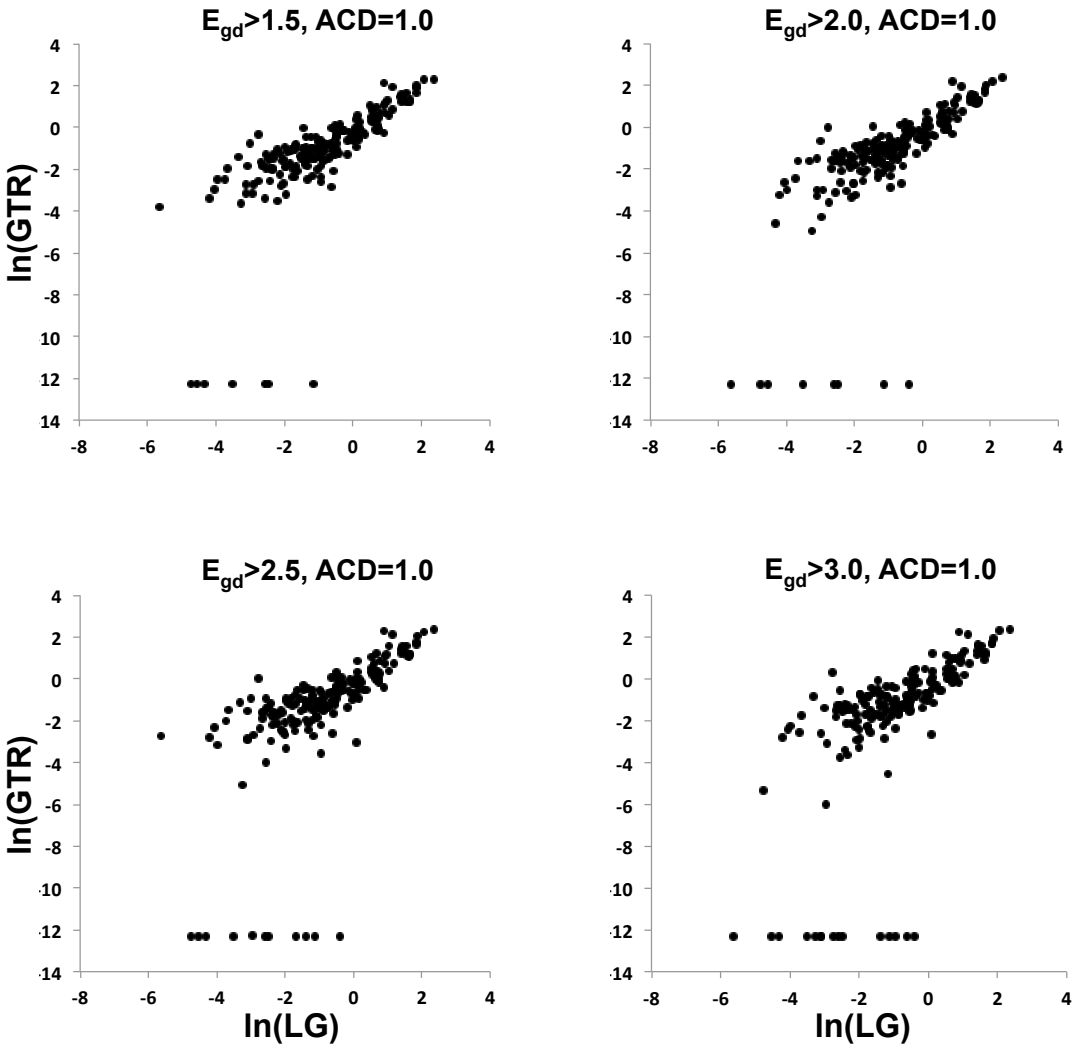


Figure 2-14. Comparison of GTR matrices from CAS alignments to LG matrix. The points in the bottom means pairs of amino acid types which are not seen in the alignment. The thresholds of E_{gd} and ACD are shown above plots in the same order of Figure 2-13.

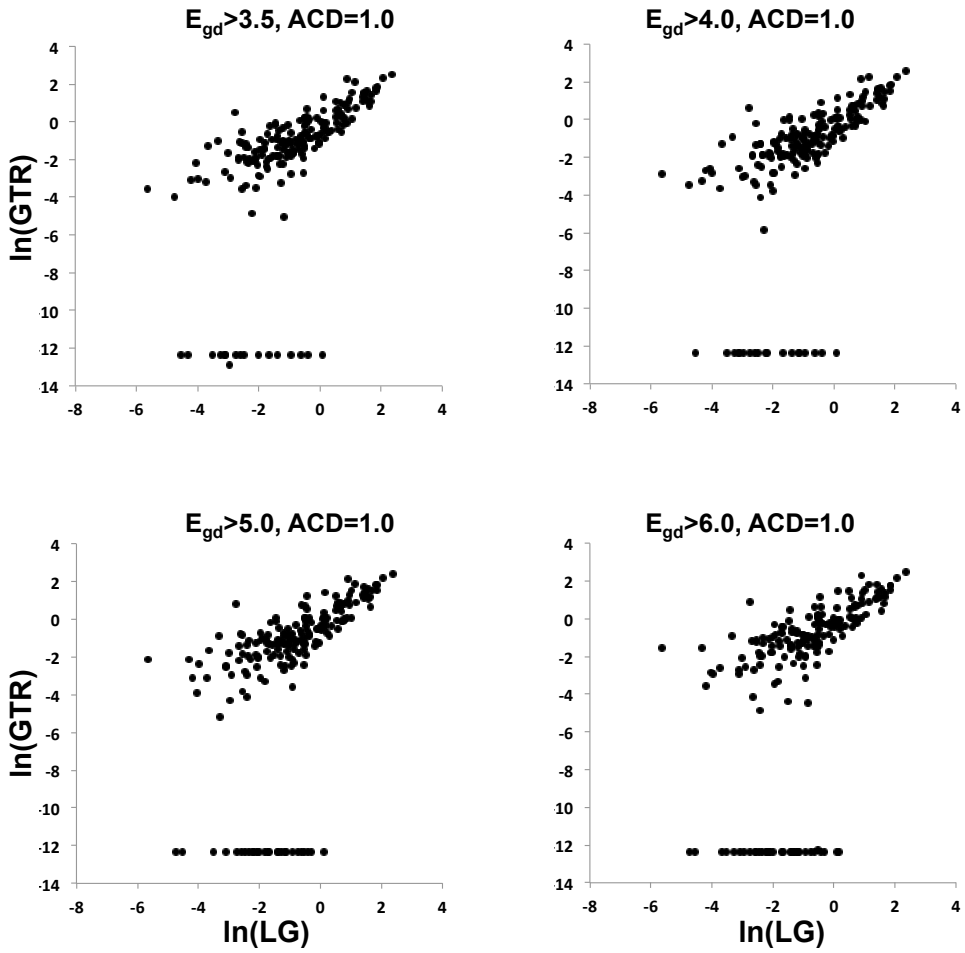


Figure 2-14. Continued.

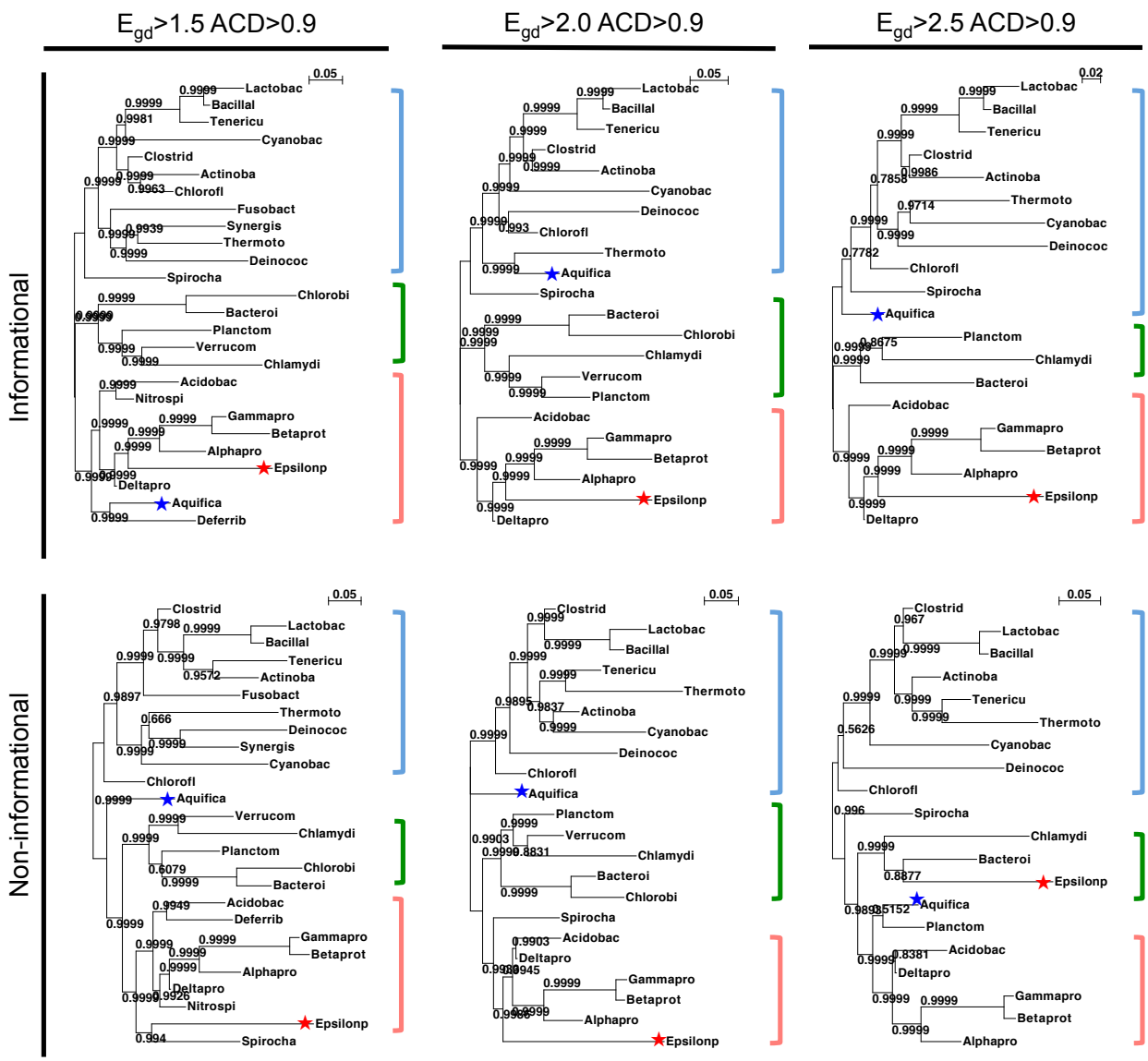


Figure 2-15. Comparison of ML trees from informational-protein alignment and non-informational alignment. The Epsilon-proteobacteria (red star) is not clustered with other proteobacteria in non-informational alignment. The placement of the Aquificae (blue star) is changed over trees not depending on the functional category.

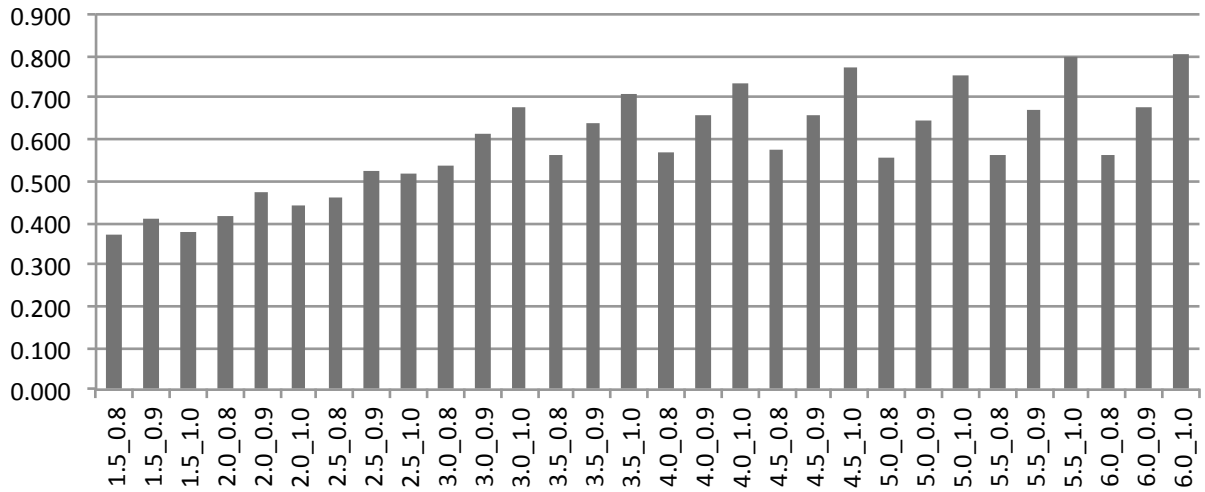


Figure 2-16. A plot of Retention Index (RI) of conserved amino acid alignments with different set of thresholds (E_{gd} and ACD for each alignment are shown as the first and second number in the x-axis). As thresholds, E_{gd} and ACD, are increased and stricter up to E_{gd}=6.0 and ACD=1.0, the RI is increased indicating lower homoplasy.

CHAPTER 3
MODEL DEVELOPMENT AND APPLICATION

Constrained Neutral Model of Amino Acid Replacements

Model of Amino Acid Substitutions With Equal Probability

The properties of the constrained neutral model (Brocchieri 2001), referred to "CNM", was analyzed quantitatively with assuming that amino acids in the subset are equally substitutable. Without taking into account the codon structures of different amino acid types, this simplified model is useful to identify and interpret several important features observed in the differentiation and phylogenetic tree reconstructions of real protein families. In this model, the concept of position-specific "background distribution" of amino acid types is substituted with the concept of "subset" of amino acids.

Suppose that each position of an amino acid sequences can accept a subset of amino acid types with no effect on fitness and that all other amino acids at that position cannot be accepted. Hence, at each position all amino acid types represented in the subset are equivalent. A subset for each position can be defined by, for example, listing amino acid types observed at that position (Figure 3-1). This way it is assumed that a protein, as a functional unit, can be equivalently represented by any of the sequences permitted by the vector of subsets and that all members of a protein family are drawn from the vectors of subsets.

In this simplified model, amino acid "mutations" to any other amino acid type are assumed to occur with equal probability at all positions, mimicking randomness in the mutational process. However, such "mutations" are accepted only if they conform to the profile. As a consequence the rate of substitution of amino acid types at a given position depends on and only on the *size* of the subset of amino acid types allowed at that

position. For example, if at a particular position 4 different amino acid types are allowed, only 3 out of 19 proposed amino acid substitutions will be accepted; if 7 amino acid types are allowed at a different position, 6 out of 19 proposed substitutions will be accepted. In the first case the expected rate of substitution (3/19) will be ½ the rate of substitution at the second site (6/19). That is, the evolution rate will be proportional to the number of allowed amino acid types. Hence, the model: (i) suggests that different sites evolve at different rates depending on how many different amino acid types they can accept; (ii) predicts that a protein will evolve slowly or fast depending on the average size of its subsets; and (iii) generates for each protein a position-specific distribution of substitution rates that depends on the distribution of subset sizes across positions and can be modeled, for example, by a gamma distribution (Ota and Nei 1994; Grishin 1995).

If *r* substitutions occur at a site where *a* amino acid types are allowed, the expected sequence dissimilarity between the original and the mutated sequence at this position is:

$$p(A_r \neq A_0) = 1 + \sum_{i=1}^{r-1} (\frac{1}{1-a})^i \quad (3\text{-}1)$$

Hence, the expected sequence dissimilarity (expected p-distance) between two sequences at a particular position converges to *(a-1)/a* when *a* > 2 as mutations accumulate at the position (Figure 3-2) whereas when *a* = 2 dissimilarity will fluctuate between 0 (for even number of mutations) and 1 (for odd number of mutations) averaging 0.5 over time (Figure 3-2).

The probability of each mutation being accepted at any of N_a sites accepting *a* amino acid types is:

$$p_a = \frac{N_a(a-1)}{\sum_{j=2}^{20} N_j(j-1)}, \quad (3\text{-}2)$$

Based on this, an exact relation of sequence dissimilarity (*p*-distance) with evolutionary distance (number of mutations per site, *R/N*) for a protein family can be computed by:

$$E(p-distance) = \frac{1}{N}\sum_{a=2}^{20} N_a \sum_{R_a=1}^{R} p(R_a) \sum_{R'_a=1}^{R_a} p(R'_a|R_a)\left[1 + \sum_{i=1}^{R'_a-1}(\frac{1}{1-a})^i\right] \quad (3\text{-}3)$$

where R = total number of mutations in N positions (the evolutionary distance is defined as R/N), R_a = number of mutations falling in all N positions that allow a different amino acid types, and R'_a = number of mutations falling in one specific position among them = number of mutations falling in one specific position among them.

The relation between sequence (dis)similarity and evolutionary distance in the toy model can also be approximated using the basic Markov model of sequence evolution used for likelihood analyses (e.g. Le and Gascuel 2008). From example of figure 3-1, the first position is observed to have a subset [A, R, K]. As the size of subset is 3, the expected sequence dissimilarity will converge to 2/3. As another way to examine the relation of sequence similarity and evolutionary distance based on the subset, the general procedures to get transition probability matrix (P) of amino acid interpreting molecular evolution as a continuous process can be used. The subset determines amino acid replacement matrix (R) (Figure 3-3). R is used to derive the amino acid instantaneous substitution matrix Q in conjunction with the vector π of stationary amino acid frequencies at a particular site. In our example π will have 0 value in all positions except those corresponding to amino acids A, K and R, where $\pi_A = \pi_K = \pi_R = 1/3$. Q is constructed as $q_{ij} = r_{ij}\pi_j$ $(i \neq j)$ and $q_{ii} = -\sum_{i=j} q_{ij}$. And Q is normalized so that a time unit (t=1) corresponds to 1.0 expected substitution per site by $Q = \{q_{ij} / C\}$ with

normalization factor $C = -\sum q_{ii}\pi_i$. Then the transition probability matrix during evolutionary time t, $P(t) = \{p_{ij}\}$, where p_{ij} is the probability of observing a change from amino acid i to j, is calculated by $P(t) = (p_{ij}(t)) = e^{Qt}$. The expected sequence similarity between sequences separated by an evolutionary time t can be calculated with π and probabilities of observing amino acids unchanged from P matrix, $p_{ii}(t)$. In this way, once relation between evolutionary distance and sequence similarity can be derived the evolutionary distance with sequence similarity observed on the sequences can be estimated.

The average size of the subsets also predicts the asymptotic value towards which sequence similarity converges as evolutionary distance increases. For example, given a subset where the average size of the position-specific subsets is a = 3, as evolutionary distance increases the sequence similarity converges to 1/a = 0.333 (Figure 3-4). Hence, the toy model also predicts that: (iv) an asymptotic sequence similarity value between two sequences corresponds under the model to the point of loss of phylogenetic information (mutational saturation) around which evolutionary distance cannot be estimated; and (v) the level of similarity corresponding to mutational saturation is proportional to the inverse of the rate of evolution of the protein. In other words, under the model proteins that evolve more slowly reach saturation at higher levels of similarity.

Application of CNM with Equal Probability to a Conserved Protein Family

Using this formulation, our model is applied to a family of human *Hsp70* proteins, which is one of the conserved protein families, building a profile based on the amino acid variants observed in their sequences (Figure 3-5). For this protein family our model predicts loss of phylogenetic information (evolutionary distance not predictable from the

sequence) at high similarity values. Comparing the predictions of the model with estimates of evolutionary distances suggested by the gamma distance method with different alpha parameters (Ota and Nei 1994; Grishin 1995) (Figure 3-5), it is observed that evolutionary distance predictions by these methods can severely under-estimate true evolutionary distances if the assumptions of our model are correct and: (vi) a model of constrained neutral substitution provides an interpretation for the phenomenon of "long branch attraction" generated by traditional tree-reconstruction methods as the result of their progressive under-estimation of longer evolutionary distances.

Consequently, a model of constrained neutral evolution by predicting asymptotic values of protein similarity may produce a more reliable tree than a detailed probabilistic model describing more accurately the dynamics of the evolutionary process but over-estimating site-specific boundaries of differentiation. Especially, the constrained neutral model may be more powerful for highly conserved protein family, for which it suggests lowest site-specific variability and possible loss of phylogenetic information even at high sequence similarity. In this respect, the neutral constrained model indicates that slowly evolving positions do not necessarily provide more phylogenetic information.

A General Model of Protein Evolution

Our analyses are based on the assumption of the following general model of protein evolution, which can be considered a reformulation of similar models envisioned in different contexts by Crooks and Brenner (2004) and by Blanquart and Lartillot (2006). In this model a protein can be functionally represented as a vector π of position-specific profiles. At each position, the profile $\pi^{(i)}$ is a position-specific vector π of amino acid equilibrium frequency that defines the probability of observing each amino acid

type at a particular position *i*. Hence, the size of $\pi^{(i)}$ is *L* – the length of the protein, the size of $\pi^{(i)}$ is 20 – the number of different amino acid types. The profile $\pi^{(i)}$ conceptually result from the combination of a binary "occupancy vector" $b^{(i)}$ of size 20, whose elements define the subsets and are 1 if the corresponding amino acid type is allowed at position *i* and 0 otherwise, and relative frequencies calculated from a DNA mutational process and codon structure as detailed below.

Deriving Position-Specific Profile Based on the Model with Incorporating Nucleotide Substitution Model and Codon Structures

In the toy model, a subset of amino acids allowed at a position determined position-specific profile and evolutionary rates. Here, a mechanistic approach is used to develop a model of neutral constrained by incorporating a nucleotide substitution model and codon structures so that position-specific profiles can reflect nucleotide-level mutational bias such as nucleotide compositional bias as well as amino acid-level neutral evolution processes. The procedure to derive a position-specific profile based on the nucleotide substitution model and the codon structure of amino acids in the subset is as follows. An example of a position characterized by a subset [D, E] of amino acids and parameters is shown in the Figure 3-6.

Deriving Nucleotide Substitution Matrix

A nucleotide substitution model can be derived with Kimura's two parameter model where α is the transition rate and β is the transversion rate and transition-transversion ratio, $\kappa = \alpha/\beta$, can be calculated. For each family or each sequence, α $\hat{\alpha}$and β are the observed values and κ can be calculated from those values (Kimura 1980). When $\kappa = 1$ there is no rate difference between transitions and transversions. Alternatively, transition and transversion substitution rates can be determined, arbitrarily

scaled, by assuming a value for $\kappa = \alpha/\beta$. Using transition and transversion rates, a replacement matrix of nucleotides, R_{DNA}, can be constructed.

$$R_{DNA} = \{r_{ij}\} = \begin{pmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{pmatrix} \quad (3\text{-}4)$$

$where\ i,j\ \in \{A,G,C,T\}$

Using the resulting R matrix, nucleotide mutational biases can be derived from equilibrium frequencies of nucleotides, π, which result in an instantaneous substitution matrix of nucleotides, $Q_{DNA} = \{q_{ij}\}$ as:

$$q_{ij} = \pi_j r_{i \leftrightarrow j} \quad (3\text{-}5)$$

$$q_{ii} = -\sum_{j \neq i} q_{ii} \quad (3\text{-}6)$$

Deriving a General Codon-substitution Rate Matrix and Position-specific Codon Substitution Rate Matrices based on Nucleotide Substitution Model

An instantaneous 61 x 61 Q_{codon} matrix is derived from Q_{DNA} in time *dt* assuming independence. Its formulation is based on the general formulation for estimating transition probabilities. A transition probability matrix, $P(t) = \{p_{ij}\}$ where p_{ij} is the probability of changing amino acid *i* to *j*, is determined by Q and evolutionary time *t*.

$$P(t) = \{p_{ij}\} = e^{Qt} \quad (3\text{-}7)$$

The probability of changing from nucleotide *i* to *j* ($i \neq j$) during a short time span d*t*, $q_{ij}dt$, is equal to $q_{ij}dt$, whereas the probability of observing amino acid *i* unchanged, $p_{ii}dt$ is equal to $1+ q_{ii}dt$ (Le and Gascuel 2008). Based on these probabilities, I can approximate the substitution rate of codons as the product of the corresponding nucleotide substitution probabilities. For example, the substitution rate of change from codon GAC to codon GAA can be calculated as:

$$q_{GAC \to GAA} = (1 + q_{G \to G}) * (1 + q_{A \to A}) * q_{C \to A}. \quad (3\text{-}8)$$

Finally, given a position-specific occupancy vector a *position-specific codon-substitution matrix* is obtained by resetting to 0 all positions of the general matrix corresponding by row or column to codons (amino acid types) not allowed at that position and normalized by recalculated diagonal terms, $q_{ii} = -\sum_{j \neq i} q_{ij}$.

Notice that this procedure reproduces a distribution of position-specific evolutionary rates, as in our toy model, where rates are reduced as excluding substitutions toward amino acid types excluded by the occupancy vectors.

Deriving an Amino-Acid Substitution Rate Matrix Based on a Codon-Substitution Rate Matrix

To explain how to calculate codon substitution rate for a given site, take a position, as an example, with the occupancy vector where only aspartic acid (D) and glutamic acid (E) are allowed. Codons encoding Aspartic acid and Glutamic acid are [GAC, GAT] and [GAA, GAG], respectively. In this case, substitution rates of four pairs of codons encoding the amino acids can be calculated and only the corresponding cells in the position-specific codon substitution matrix will have non-zero values. For a given subset of amino acid types and corresponding Q_{codon}, an instantaneous 20 x 20 Q_{AA} matrix where cells of amino acids in the subset have only non-zero values. Substitution rates of changing amino acid *i* to an other amino acid, *j,* can be calculated by weight-averaging corresponding codon substitution rates with relative frequencies of codons encoding the amino acid *i.* Relative frequencies (*w*) of codons encoding an amino acid is determined by Q_{DNA} and when there is no compositional bias of DNA it will be determined by the number of codons encoding the amino acids. Given equilibrium frequency of DNA [0.25, 0.25, 0.25, 0.25], indicating no compositional bias, the relative

codon frequencies of amino acids D and E will be [0.5, 0.5]. To construct Q_{AA}, substitution rate of change from D to E, for example, the following equation is used:

$$Q_{D\to E} = (q_{GAC\to GAA} + q_{GAC\to GAG)}) * w_{GAC} + (q_{GAT\to GAA} + q_{GAT\to GAG)}) * w_{GAT} \quad (3\text{-}9)$$

Deriving Equilibrium-Frequency Profiles From an Amino-Acid Substitution Rate Matrix

If the initial distribution of frequencies of amino acids is $\pi^{(0)} = [\pi_j^{(0)}]_{j=1\ldots20}$ where $\sum_j \pi_j = 1$, the distribution after duration of evolutionary time t, $\pi^{(t)} = [\pi_j^{(t)}]_{j=1\ldots20}$, can be given as

$$\pi^{(t)} = \pi^{(0)}P(t). \quad (3\text{-}10)$$

If the initial distribution is no longer changed after time t, it can be said the distribution is at equilibrium and π at this time is the equilibrium distribution. Therefore the equilibrium distribution is given by

$$\pi P(t) = \pi. \quad (3\text{-}11)$$

This is equivalent to

$$\pi(I + Q) = \pi \quad (3\text{-}12)$$

$$\pi Q = 0. \quad (3\text{-}13)$$

Therefore, equilibrium frequencies of amino acids can be derived from Q using this equation (Yang 2006). These equilibrium frequencies constitute the profile $\pi^{(k)}$ of amino acid equilibrium frequency at position k. After getting profiles of all positions, a vector π of position-specific profiles, $\pi = [\pi^{(k)}]_{k=1\ldots L}$, is obtained.

In deriving a position-specific profile, the DNA equilibrium frequency is used as a parameter. That is the DNA equilibrium frequency can be set depending on the specific characteristics of the gene families or genomes of interest. Estimating DNA equilibrium frequency is not a trivial problem, since estimation of DNA-base equilibrium frequencies

is needed before selection at the amino acid level. DNA equilibrium frequency can be simply approximated based on the frequency of each base in the coding sequences. Or genome sequences of organisms can be used that appear in the alignment to estimate DNA equilibrium frequencies. As a third alternative, it can be estimated from “silent” sites, i.e., from the third codon-position of 4-fold degenerate codons. Finally, equilibrium frequencies can be estimated by constructing a model based on (i) the relation between overall base composition and base composition at the three codon positions; (ii) estimates of the codon-position specific proportion of variable and constrained sites; and (iii) assumptions on the determinants of base usage as either biases in the mutational process or genome-wide selective processes at the nucleotide level.

Test of Lumpability

Lumpability

To derive an amino acid substitution probability matrix from codon and nucleotide substitutions in terms of a Markov process, the codon matrix should be “lumpable” into an amino acid matrix. Lumpability of a finite Markov chain is defined, as “the process resulting from the observation of the Markov chain by masking out transitions and states inside a partition group is also a Markov chain” (Buchholz 1994). “Exact Lumpability” of a partition, one derivation of Lumpability, is defined that a partition is exactly lumpable if the sums of transition probabilities from states of a fixed partition group into each state of another or the same partition group are all equal (column sum criterion) (Schweitzer 1984).

When the state space of a finite Markov chain can be partitioned into subsets, each of which can be treated as a single state in a smaller chain that retains the Markov property, the original chain is called “lumpable” with respect to that partition (Figure 3-7).

Disjoint subsets can be considered as amino acids and states in each subset as codons encoding the amino acid. To make sure an amino acid substitution probability is also a Markov process when it is derived from codon substitution probability matrix, the codon substitution matrix should be lumpable into an amino acid substitution probability matrix.

Test of Ordinary/Exact Lumpability

The Lumpability Test is well described in Buchholz's paper (Buchholz, 1994). Buchholz's equations for the Lumpability Test is applied to check if the 61 by 61 codon transition probability matrix, P_{codon}, can be lumpable into 20 by 20 amino acid transition probability matrix, P_{aa_B} with several assumptions to make it simple.

- All nucleotides are equally substitutable. Therefore the equilibrium frequency of nucleotides is [0.25,0.25,0.25,0.25]
- Kimura model is used with two parameters: nucleotide substitution rate (p) = 0.001 and transition-transversion ratio (k) = 2.
- Relative equilibrium frequencies of all codons encoding an amino acid are dependent on the nucleotide substitution probability and equally substitutable between codons encoding the same amino acid.

These assumptions result in equilibrium frequencies for each codon = 0.016. Based on these assumptions Q_{codon} is generated and used to derive P_{codon} (t) by $P(t) = e^{Qt}$. And then P_{codon} is lumped into P_{aa_B} by the Buchholz's method.

For test of "Strict Lumpability", set a Markov chain, P_{codon}, having equilibrium frequency, π, state space, S, and transition probability matrix, P_{aa_B}. and Partitions C(I) constituting A.

$$C(I) \subseteq A, C(I) \neq \emptyset, C(I) \cap C(J) = \emptyset, \ \cup_{I=1}^{20} C(I) = A \qquad (3\text{-}14)$$

for $I, J \in \{1, \dots, N\}$. N_I is the number of states in partition group $C(I)$.

The transition matrix of the aggregated Markov chain P_{aa_B} is generated using a collector matrix V and a distributor matrix W as shown in the following equation

$$\underline{P_{aa\_B}} = \underline{WP_{aa\_B}V} \tag{3-15}$$

where $\underline{V} \in \mathbb{R}^{n \times N}, V\ (i,I) = \ 1\ if\ i\ \in P_{codon}(I)$ and 0 otherwise. $\underline{V} = \underline{D}^{-1}\left(\underline{V}\right)^{T}, D = diag\left(\underline{aV}\right) > \underline{0}$, where $diag\left(\underline{\alpha}\right)$ is a diagonal matrix with $\alpha(i)$ in position *i*. When P_{aa_B} is the irreducible transition matrix of a finite Markov chain.

- P_{codon} is ordinary lumpable, iff for all $I\ \in \{1, N\}$ and all $i, j\ \in C(I)$:
  $$(\underline{\epsilon_i} - \underline{\epsilon_j})\underline{P_{aa\_B}V} = \underline{0} \tag{3-16}$$
- P_{codon} is exactly lumpable, iff for all $I\ \in \{1, N\}$ and all $i, j\ \in C(I)$:
  $$(\underline{\epsilon_i} - \underline{\epsilon_j})\underline{P_{\boldsymbol{aa\_B}}^{T}V} = \underline{0} \tag{3-17}$$
- P_{codon} is strictly lumpable, iff it is ordinary and exactly lumpable

where $\underline{\epsilon_i}$ is a row vector with 1.0 in position I and 0 elsewhere.

As a result of testing lumpability of P_{codon} based on Buchholz's equations, P_{codon} is neither ordinary nor exactly lumpable. But several sub-matrices of P_{codon} corresponding to subsets of amino acids (e.g. [DENK], [AGPTV], and [MW]) are exactly lumpable.

Test of Near Lumpability

Because P_{codon} is not strictly lumpable, it is tested whether or not P_{codon} is "near lumpable" by the Buchholz formula for the near lumpability (Buchholz 1994) with the same notations and definitions as in the test for strict lumpability.

- P_{codon} is nearly ordinary lumpable, iff for all $I\ \in \{1, N\}$ and all $i, j\ \in C(I)$:
  $$\left(\underline{e_i} - \underline{e_j}\right)\underline{P_{aa\_B}V} < \underline{\epsilon e} \tag{3-18}$$
- P_{codon} is nearly exactly lumpable, iff for all $I\ \in \{1, N\}$and all $i, j\ \in C(I)$:
  $$\underline{P_{aa\_B}^{T}V} < \underline{\epsilon e} \tag{3-19}$$
- P_{codon} is nearly strictly lumpable, iff it is ordinary and exactly lumpable

for $0 < \epsilon << 1$.

Test of "near lumpability" on P_{codon} from the subsets of amino acids showed that ϵ is distributed in the range of 0.0000001 to 0.0001 such that P_{oodon} is not strictly lumpable but nearly strictly lumpable. Having an extremely low value of ϵ provides enough support to continue to use the reduction of Q_{codon} to derive an amino acid substitution model from the nucleotide model.

Comparison of the Lumped P_{codon} to the P_{aa} derived from Q_{aa}

To visualize the difference between the lumped P_{codon}, $P_{aa\_B}(t)$, and the $P_{aa}(t)$ derived from Q_{aa} a plot was generated showing the sequence similarity and evolutionary distance. Sequence similarity given *t* is calculated by

$$sequence\ similarity\ (t) = \sum_i p_{ii}(t) * \pi_i \ \text{ where } t \text{ is evolutionary time} \qquad (3\text{-}20)$$

As shown previously when deriving Q_{aa} in the mechanistic protein evolution model, Q_{aa} is derived from Q_{codon} by summing the rates of substitution between all pairs of codons encoding an amino acid, *A*, and another amino acid, *B*, with weighting relative frequencies of codons of amino acid *A*.

When 20 amino acids are used to construct Q_{codon}, the curves of the lumped Q_{aa_B} , P_{aa_B}, and the P_{aa} derived from Q_{aa} are very close in the plot (Figure 3-8) but not superimposed to each other, which means P_{codon} with 20 amino acids is not "strictly" lumpable into P_{aa_B}. When amino acid subsets [DENK], [AGPTV], or [MW], two curves are identical and overlapped because P_{codon} based on the subset is "strictly" lumpable to P_{aa_B}. (Figure 3-9).

As shown in the test of near lumpability, it is "nearly lumpable" and in the plot of evolutionary distance and sequence similarity the curve of the lumped P_{codon} is very close to P_{aa}, Therefore, the proposed way of converting Q_{codon} to Q_{aa} in this study is

thought to be reasonable and practical despite P_{codon} (t) not being strictly lumpable into $P_{aa}(t)$.

Profile Analysis by Conditional Likelihood

Based on our application of a constrained neutral model (CNM), position-specific Q and profile depend on which amino acids are observed on the position. That is, the model assumes that amino acids on the positions are appeared due to evolutionary events. The assumption, however, is not valid when amino acids are randomly distributed, or a particular amino acid or set of amino acids was not observed due to sampling or other reason. The complete alignment and individual alignment of 3 protein families were chosen for this analysis: (1) DnaK (COG 0443, 771 positions) (2) Chaperonin GroEL (*HSP60* family) (COG 0359, 543 positions) and (3) RecA/RadA recombinase (COG 0468, 461 positions).

Pruning Algorithm and Conditional Likelihood on an Alignment Position

Conditional Likelihood Analysis is used to calculate the probability of occurrence of each amino acid on the root of the tree (Felsenstein 1981). The likelihood of the tree on a particular position was calculated as follows.

Conditional likelihoods at a node w are a 20-dimensional vector, $L_w = \{L_w(A), \ldots, L_w(Y)\}$, where $L_w(x)$ is the likelihood of the observed data in the subtree rooted by the node w given state x at node w. Let a node w be related to two leaves α and β. When the likelihood of the nodes α and β are $L_\alpha = \{L_\alpha(A), \ldots, L_\alpha(Y)\}$ and $L_\beta = \{L_\beta(A), \ldots, L_\beta(Y)\}$, the conditional likelihood of amino acid k_w at node w is calculated as

$$L_w(k_w) = \Pr(k_\alpha | k_w, d_\alpha) \Pr\left(k_\beta \middle| k_w, d_\beta\right) \quad (3\text{-}21)$$

where k_m and d_m are the observed amino acid and brand length at the node m. This formula can be converted into:

$$L_w(k_w) = \sum_x \Pr(x|k_w, d_\alpha)\, L_\alpha(x)\ \sum_y \Pr\left(y\middle|k_w, d_\beta\right) d_\beta(y) \quad (3\text{-}22)$$

where x, y are amino acid type.

The overall position-specific likelihood of the tree is

$$L_{root} = \sum_x \pi_x L_{root}(x) \quad (3\text{-}23)$$

The conditional likelihood calculation of the tree could be initialized by assigning the initial likelihood onto terminal nodes depending on the amino acid on the position. For example, the initial likelihood of amino acid D is L = {$L(A)$, $L(C)$, $L(D)$, … , $L(Y)$} = { 0,0,1,0,..,0 }.

Based on the neutral constrained model, the initial likelihood depends on the subset of amino acids that can be substituted on the position not a single amino acid. For a particular position, terminal nodes in a given tree are labeled with corresponding amino acid or gaps. Branches with amino acids are only selected and gapped branches are removed. To calculate the likelihood of each node, amino acid substitution probability matrix, P, equilibrium frequency, π, and likelihoods of children nodes are needed. For terminal nodes the initial likelihood was assigned. And then based on the subset of amino acids observed in the position, Q and π are derived and used to drive P given a branch length.

Test on Random Distribution of Amino Acids

To test if amino acids are randomly occurred and distributed, an alternative was questioned: for an amino acid type in the subset, when a certain cluster contains all the amino acid type on the tree, what is the probability of not seeing the amino acid type

outside of the cluster? To calculate the probability of each amino acid based on the position-specific profile and Q, I perform conditional likelihood analysis based on a pruning algorithm (Felsenstein 1981) as described below.

Let one amino acid type in the subset be aa_i (Figure 3-10). The minimal cluster including all leaves with amino acid aa_i is chosen. The likelihoods of leaves are initialized depending on their belonging or not to the clusters and on the subset of amino acid types observed across all sequences at that position. For all leaves within the cluster, their conditional likelihoods $\{l_j\}$ are $l_j = 1$ if j is in the subset, otherwise, $l_j = 0$ if j is not in the subset. For any leaf outside of the cluster, the likelihoods are initialized as $l_j = 1$ if j is in the subset but $j \neq i$, and $l_j = 0$ otherwise. The likelihood of each node is calculated recursively traversing from the root to the terminals and then the overall likelihood is calculated as the likelihood of the amino acid.

The hypothesis is that an amino acid type is seen in a cluster of the most recent common ancestor of sequences showing the amino acid type but it is not seen outside of the cluster. Therefore, if the probability of the amino acid were less than p-value, the hypothesis would be rejected. That is, this amino acid can be said as randomly distributed. This conditional likelihood analysis is applied to the complete alignment and three protein families, DnaK, GroEL and RacA proteins. Among them, results from the complete alignment is shown in Table 3-1 and those for DnaK, GroEL and RacA were shown in Table A-2, Table A-3, and Table A-4 in Appendix. On the complete alignment, average cluster size is 406.92 and average standard deviation is 335.69. The test results with two p-values, 0.01 and 0.001, are shown in Table 3-1. On average, 5% and 1% of positions having each amino acid type shows the amino acid type is randomly

distributes over the tree when $p < 0.01$ and $p < 0.001$, respectively. The results from three protein families, DnaK, GroEl and RacA proteins, are not significantly different. The fraction of amino acids showing significantly low probability is less than 10% (Table A-2, A-3, and A-4).

If there is only one sequence showing a particular amino acid type on the tree, the probability of not seeing the amino acid type outside of the cluster is very low. About 50% of amino amid types showing $p < 0.01$ have one-sequence clusters. This means that most of amino acid types are not randomly distributed and just small number of amino acid types might be randomly distributed.

Test on Change of Amino Acid Subset With Tree

It might be happen a particular amino acid or set of amino acids was not observed due to sampling or other reason leading that the profile would not allow it to be substitutable on the position. Therefore it should be measured by how likely the amino acid types could be seen even though they are not seen on the position. To accomplish this, for an amino acid type not in the subset, the probability of seeing the amino acid type on the tree assuming it could be substituted was calculated with conditional likelihood analysis (Figure 3-11).

Based on the position-specific profile, Q, and a corresponding tree, conditional likelihood analysis was performed to measure the probability of seeing an amino acid type not in the subset in the three protein families when it is assumed to be substitutable. Let an amino acid type not in the subset be aa_i. Position-specific Q and π can be derived using amino acid subset including aa_i. Because aa_i is not seen anywhere in the tree, the initial likelihood is $\{l_j\}$ where $l_j = 1$ if j is in the subset otherwise $l_j = 0$ if j is not in the subset. Using Q, π, branch length and initial likelihood, the

likelihood for every internal node can be calculated as described as in previous section of “test on random distribution of amino acids”. After calculating likelihood on the root of the tree, overall likelihood of the amino acid type is calculated.

This approach is applied to the complete alignment and three protein families as used in previous conditional likelihood analysis. Among them, the result from the complete alignment is shown in Table 3-2. About 42% of amino acid types showing $p < 0.05$ rejected the hypothesis, which means half of amino acid types with probability less than 0.05 cannot be seen even when it is not zero in the profile. It could be interpreted that for the positions having any amino acid type(s) with higher probabilities profiles might be reliable but with weak support. The results from three protein families, DnaK, GroEl and RacA proteins, are shown in Table A-5, A-6, and A-7, respectively. The results from three protein families and the complete alignment are not significantly different each other. Therefore these findings are not dependent on the selection of data set and could be observed on the overall dataset.

Phylogenetic Tree based on Position Specific Profiles

A Complete Tree Based on Position-Specific Q and Profile

In most of softwares commonly used for tree reconstruction, a position-specific Q matrix and profile cannot be used as the only tools for phylogenetic tree reconstruction. The following is a tree construction strategy using a collection of position-specific Q matrix and profile from an alignment (Figure 3-12).

On each position, a position-specific Q and profile are generated. Q is divided by the average of normalizing factors on all positions where the normalizing factor (C) is calculated as $C = -\sum q_{ii}\pi_i$. Using them, P matrices are generated with a series of

evolutionary distance *t* (from 0.001 to 6.000 incrementing by 0.001). Expected sequence similarity can be calculated by

$$s^k(t) = \sum_i p_{ii}^k(t) * \pi_i^k \ . \quad (3\text{-}24)$$

where *t* is evolutionary time on a particular position *k*. After repeating the calculation of expected sequence similarity based on position-specific Q and profile on all positions, an average of sequence similarity given *t* over all positions is calculated.

$$average\ of\ sequence\ similarity(t) = \sum_k^K s^k(t) \quad (3\text{-}25)$$

Therefore, a relationship between sequence similarity and evolutionary distance based on position-specific Q and profile is generated.

To obtain pairwise evolutionary distance for a distance matrix of all sequences that can be used in neighbor-joining method, pairwise sequence similarity of any two sequences was calculated as p-distance. Given a pair of sequences, sequence similarity is calculated on each position and then averaged for all positions counted.

- If any sequence has a gap, that position does not count.
- If both sequences have the same amino acid, similarity is 1.
- If each sequence has an amino acid but they are different, similarity is 0.

The p-distance between two sequences is sum of similarity divided by (number of alignment columns - gapped alignment columns), considered as sequence similarity of the pair of sequences. The expected evolution rate per position is calculated as the average of evolution rate of all positions where evolution rate, *r,* can be calculated as $r = -\sum q_{ii}\pi_i$ from position-specific Q and profile.

The evolutionary distance corresponding to the sequence similarity between the pair of sequences can be found from the relationship of evolutionary distance and

sequence similarity estimated from position-specific Q and profile. For 1,814 sequences, pairwise similarity of 1,644,391 pairs of sequences were calculated and used for a distance matrix to generate a NJ tree (Figure 3-13). The NJ tree was generated using QuickTree 1.1 (Howe et al. 2002) by submitting a distance matrix in the Phylip format.

Most of sequences corresponding to various main phyla or classes of Bacteria formed distinct clades and they were generally well matched to their taxonomy groups. *Lactobacillales*, *Bacillales* and *Clostridia* sequences were grouped by class as well as formed their phylum-level clade, the *Firmicutes*. The *Actinobacteria* sequences are clustered well and their closest group is the *Firmicutes*. The *Firmicutes* and *Actinobacteria* are shown to have the closest association. Sequences belonging to alpha-, beta-, gamma-, delta-, and epsilon-proteobacteria are closely clustered and formed a clade of the *Proteobacteria*. When the topologies of the CAS trees were compared to those of the ML supertree, closeness within the group is higher and there is not any extremely long branch or outlier branch located outside of the group cluster. Computation time to construct a profile-based NJ tree with 1,814 sequences and 72,700 positions was less than 2 days on a personal computer as while it takes 15 days with 16 nodes of a computer cluster in UF high performance computing center to obtain a ML tree.

Comparison with a ML tree based on LG+Γ+I

To compare the CNM with a generalized substitution matrix, LG, the relation of evolutionary distance and sequence similarity inferred by the CNM and LG+Γ+I model was compared. The plot (Figure 3-14) generated to compare the evolutionary distances

predicted by the CNM evolutionary model and those predicted by on the LG+Γ+I model, base on pairwise sequence similarity, with the corresponding pairwise evolutionary distances inferred from the ML tree (patristic distances) based on the complete alignment. To compute evolutionary distance expectations based on the LG+Γ+I model, the four classes of rates corresponding to gamma distributed rates with alpha = 0.90 and invariant=0.026 (as estimated from RAxML), were applied to the LG matrix. Sequence similarity given evolutionary distance then was calculated for each class and averaged over 4 classes. Furthermore, maximum-likelihood distances were calculated for each pair of sequences individually based on their alignment and using the same parameter-value for alpha and *I* identified from the complete alignment. It was expected that if the ML tree correctly and consistently reflected all information in the data, the patristic distances of each pair of sequences would correspond to the evolutionary distances predicted by the model.

Maximum-likelihood estimates of evolutionary distances calculated for each pair of sequences matched very well the theoretical expectations based on the corresponding sequence similarity (except for the least similar sequences, Figure 3-14). However, it was found that the patristic pairwise distances calculated from the ML tree did not correspond to the distances estimated for each individual pair of sequences based on the same model parameters. To test how parameter estimation might have affected the fitting of the data to the model, expected evolutionary distances were computed based on different values of invariants (Figure 3-14) and parameter alpha (Figure A-1). None of the expected distances based on LG with/without gamma matched the ML tree patristic distances. As invariants increased, the curve of expected

distances approached closer to the distribution of patristic distances from the ML tree, but it never fitted it satisfactorily. Then expected evolutionary distances were estimated using the position-specific Q matrices calculated with the CNM model, shown in Figure 3-14. It was found that even the CNM model would not provide a good fit to the distribution of patristic distances, and although it resulted in estimates of greater evolutionary distances for less similar sequences than predicted my the LG+Γ+I model, these would still under-estimate the corresponding LG+Γ+I model patristic distances.

I finally evaluated the effect of sequence sampling density on the estimated patristic distances. Sampling from each bacterial phylum one sequence at random, a ML tree using the LG+Γ+I model was obtained independently and patristic distances based on this tree were calculated. This sparser sampling resulted in a distribution of patristic distances much closer to the expectations from the model, but yet still separated from it. This result suggested that as more details on sequence evolution are provided by denser sampling, more clearly the inadequacy of the model is revealed.

Table 3-1. Test on random distribution of each amino acid type (AA) in the complete alignment (72,700 positions).

AA	Num. of positions having AA	Num. of positions having random-distributed AA		Size of the cluster containing AA		Size of the rest on the Tree	
		P< .01	P< .001	Average	Stdev	Average	Stdev
A	56618	2872	991	486.58	469.52	135.17	427.55
C	28762	362	106	257.91	329.65	415.30	337.35
D	40283	494	136	441.11	325.40	163.92	272.49
E	43831	364	76	451.68	323.09	150.21	263.87
F	38526	1294	362	359.60	335.29	277.12	319.07
G	43536	4954	1336	406.21	336.08	212.80	301.86
H	38232	456	149	368.55	337.38	274.61	316.79
I	48235	3792	698	448.10	332.64	186.65	284.72
K	45197	609	132	466.11	322.53	147.16	258.97
L	51084	5377	2158	459.09	328.41	167.89	275.95
M	42354	800	404	375.25	339.07	273.32	317.89
N	44362	1202	339	429.92	328.72	194.21	286.85
P	30338	3516	940	352.12	332.39	233.55	307.46
Q	44939	326	115	454.31	327.07	172.02	277.32
R	43617	5368	2904	451.49	329.64	166.78	275.25
S	53130	6802	3904	469.43	324.89	156.05	268.14
T	51975	4715	1676	450.70	329.02	179.16	280.68
V	52031	3304	658	470.50	328.46	160.99	269.98
W	22356	18	5	226.80	307.18	411.69	329.51
Y	35217	158	37	312.88	327.44	326.71	323.99
SUM		26995	13183				

Table 3-2. Test on candidacy of each amino acid type (AA) in position-specific profile in the complete alignment.

AA	Num. of positions not having AA	[a]Num. of positions having significant AA		
		P<.01	P<.05	P<.5
A	16082	5242	7222	11880
C	43938	4495	13361	33772
D	32417	3700	11445	25848
E	28869	3218	8364	22083
F	34174	6550	13677	26076
G	29164	13229	18217	25153
H	34468	3112	8844	25486
I	24465	9090	14113	19534
K	27503	2556	6357	20219
L	21616	10573	13475	19320
M	30346	2539	6242	20649
N	28338	5008	12130	22365
P	42362	26182	32210	38534
Q	27761	1994	4726	17944
R	29083	17416	21140	27613
S	19570	12972	13825	18811
T	20725	8664	11440	17279
V	20669	7424	10756	16808
W	50344	1377	3939	39492
Y	37483	2314	7203	26422

[a]Num. of positions having significant AA means number of positions where the likelihood of the amino acid is less than p-value.

Position	Sequences	Profile
1.	KKRKRKKKRKKK	[K,R]
2.	QSSSQNNDNAAK	[Q,S,N,D,A,K]
3.	FFFFFYYYFFFF	[F,Y]
4.	GGGGGGGGGGGG	[G]
5.	ESKDDDDEHAEQ	[A,D,E,H,K,Q,S]
6.	DQDNREKRAESK	[A,D,E,K,N,R,S]
7.	VLIVIIIVVVIV	[V,L,I]
8.	EEEEEEEEEEEE	[E]

Figure 3-1. An example of how to define a subset on each homologous position under the constrained neutral model.

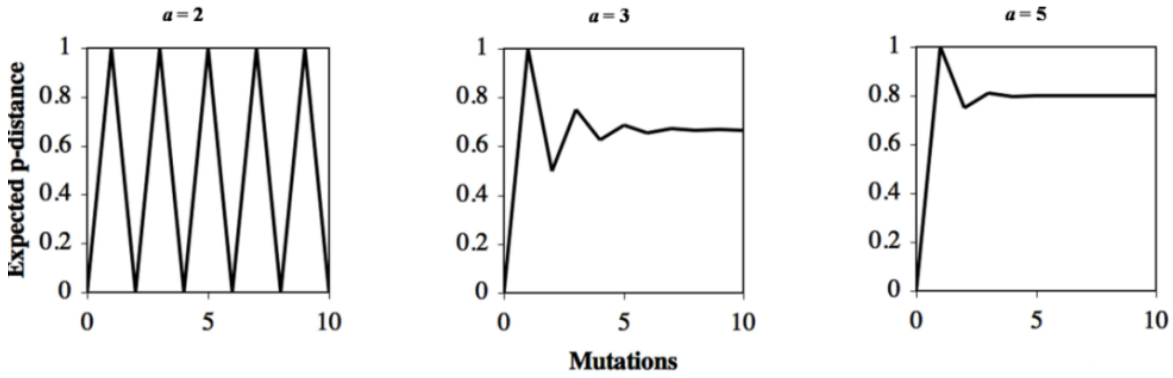


Figure 3-2. Relation of expected evolutionary distance and mutations at different number of amino acid type allowed (*a*). The evolutionary distance between the two sequences on a position is converged to (*a*-1)/*a* only when *a* > 2 as mutations occur on the position.

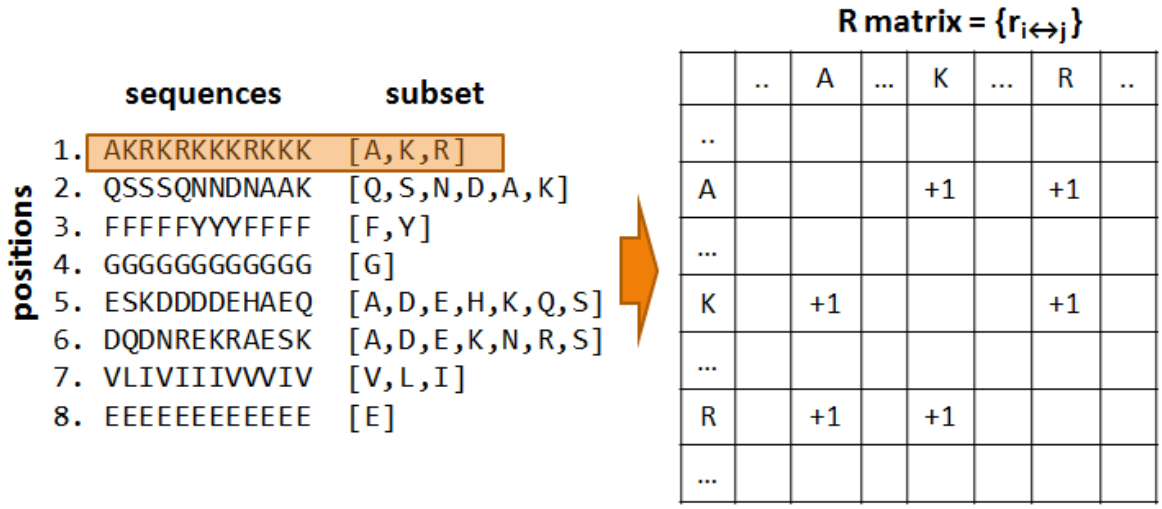


Figure 3-3. How to construct amino acid replacement matrix, R, with a subset [A, K, R]. Starting the first position with subset [A, K, R], cells corresponding to A, K and R unchanged and cells for change of amino acids in the subset are added by 1 and R matrix.

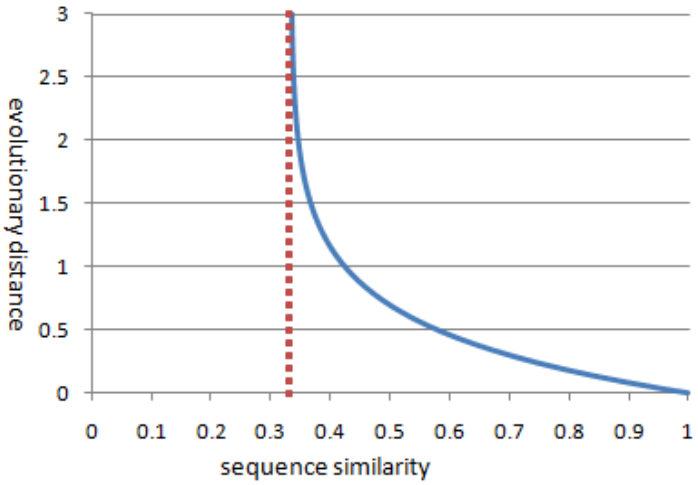


Figure 3-4. A plot showing the relation between evolutionary distance and sequence similarity when subset size is 3 (blue line). As evolutionary distance increases, sequence similarity is converged to a positive limit which is 0.333 here (dotted red line).

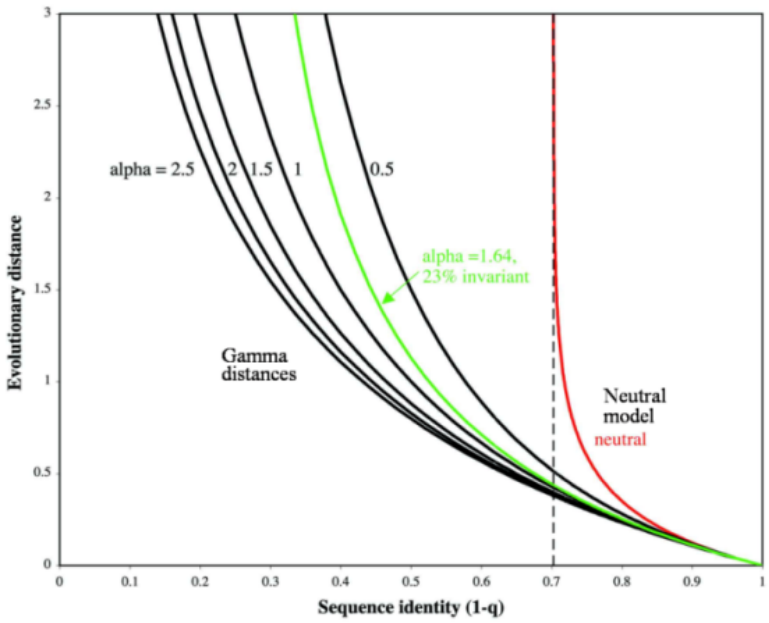


Figure 3-5. Comparison of the neutral model and gamma distance using relation between evolutionary distance and sequence similarity from Hsp70 protein family.

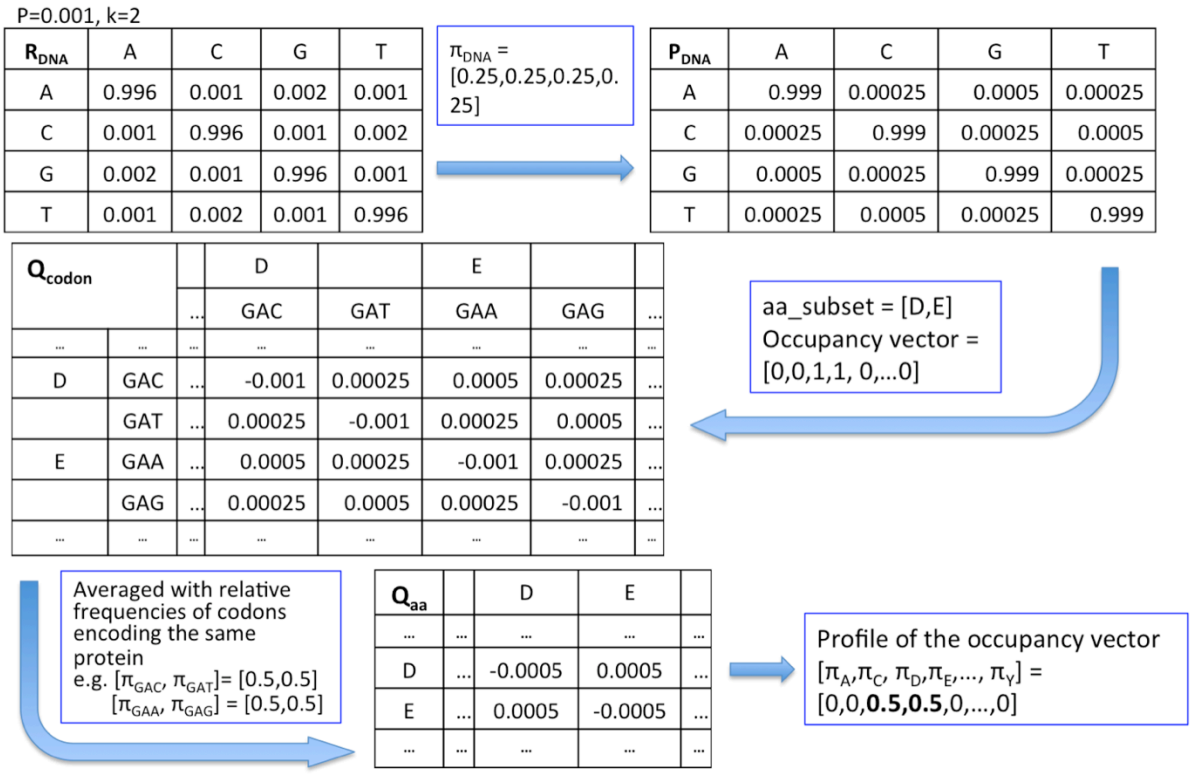


Figure 3-6. A procedure to drive amino acid substitution rate matrix, Q, and profile on the position representing amino acids D and E.

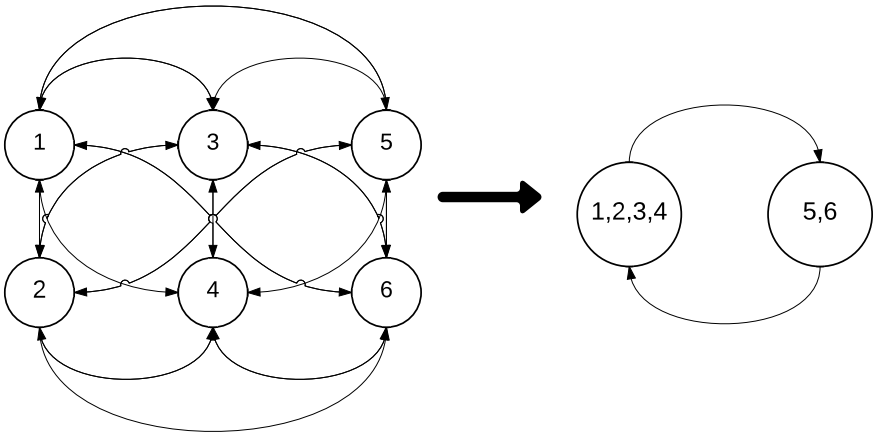


Figure 3-7. Lumpability of Markov states 1 to 6 into the lumped states [1,2,3,4] and [5,6].

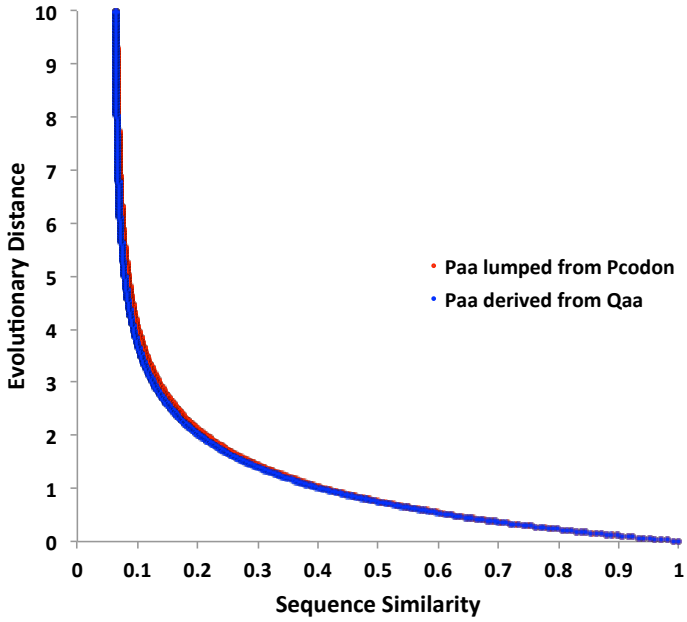


Figure 3-8. Plot of sequence similarity vs. evolutionary distance based on P_{aa} derived from Q_{aa} in blue and P_{aa} lumped from P_{codon} by Buchholz's way (1994) in red when 20 amino acids are assumed to be substitutable. Two curves are very close but not the same because P_{codon} given t is not strictly lumpable into $P_{aa}(t)$.

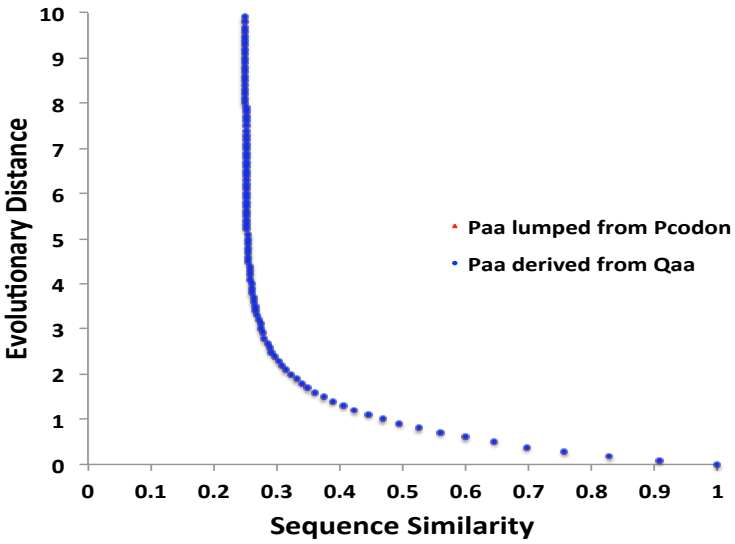
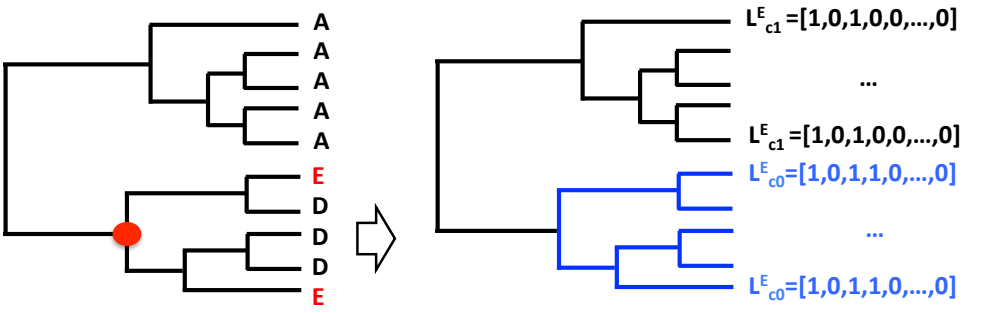


Figure 3-9. Plot of sequence similarity vs. evolutionary distance based on P_{aa} derived from Q_{aa} in blue and P_{aa} lumped from P_{codon} by Buchholz's way (1994) in red when a subset of amino acids [D,E,N,K] are assumed. P_{codon} is strictly lumpable so the curves are superimposed.



Step 1. ***Q*** and π are generated with aa_subset = [A,D,E]
and used to derive ***P*** given branch length.

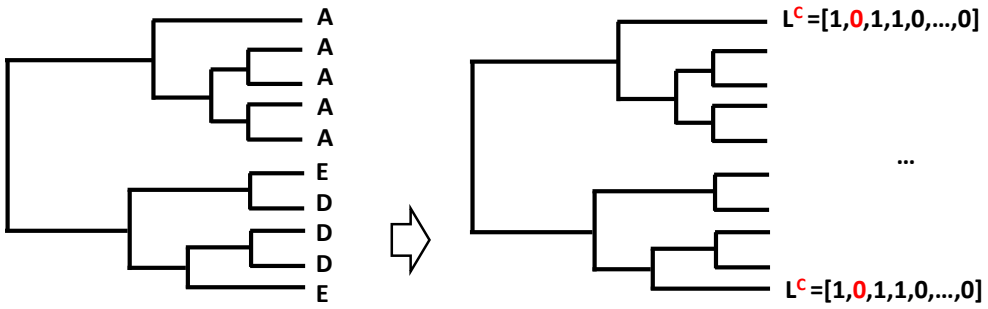
Step 2. the minimal cluster containing all amino acid E is chosen.

Step 3. Initialize conditional likelihood.
$L^i_{c1} = \{l_j\}$ where $l_j = 1$ if j is in aa_subset for all j, otherwise, $l_j = 0$
$L^i_{c0} = \{l_j\}$ where $l_j = 1$ if j is in aa_subset snf $j \neq i$, otherwise, $l_j = 0$

Step 4. calculate overall likelihood of the node and likelihood on nodes.

$$L^{(i)} = \sum_x \pi_x L_o^{(i)}(x)$$

Figure 3-10. An example of calculation of conditional likelihood in order to calculate probability of not seeing an amino acid outside of the cluster where all the leaves with amino acids are included.



Step 1. Drive Q and π based on the amino acid subsets.

When aa_subset = [A,D,E], for amino acid **C**, Q and π are generated on the subset **[A, C, D, E]** and used to derive P given branch length.

Step 2. Assign a initial likelihood onto leaves.

For amino acid *i* which is not in the aa_subset
$L^i = \{l_j\}$ where l_j is 1 if amino acid j is in the aa_subset for all j

Step 3. Calculate likelihood on nodes and overall likelihood of the node.

$$L^{(i)} = \sum_x \pi_x L_o^{(i)}(x)$$

Figure 3-11. An example of conditional likelihood calculation to estimate probability of seeing an amino acid NOT in the subset.

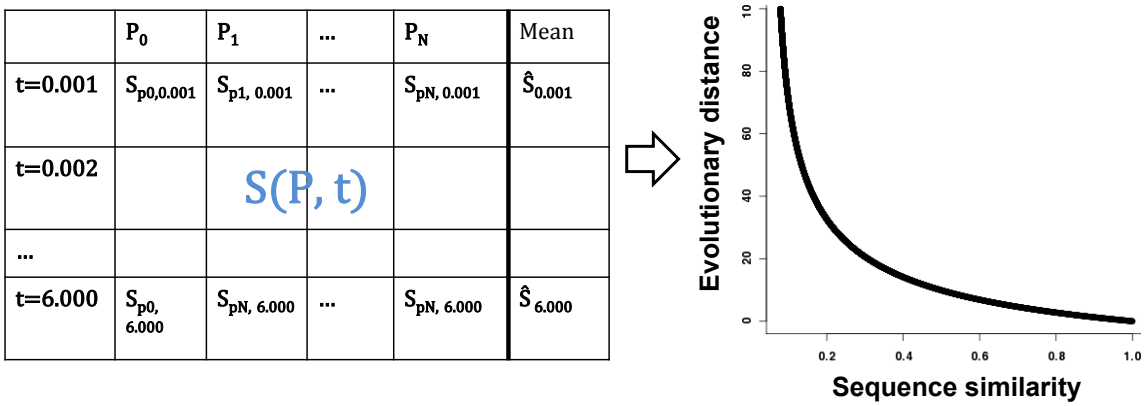


Figure 3-12. Derivation of relation between sequence similarity and evolutionary distance from position-specific Q and profile. The substitution probability matrix P is derived from position-specific Q and profile given evolutionary distance t. Sequence similarity at given t is calculated using P and profile. Average of sequence similarity over positions given t generates the relation of evolutionary time and sequence similarity.

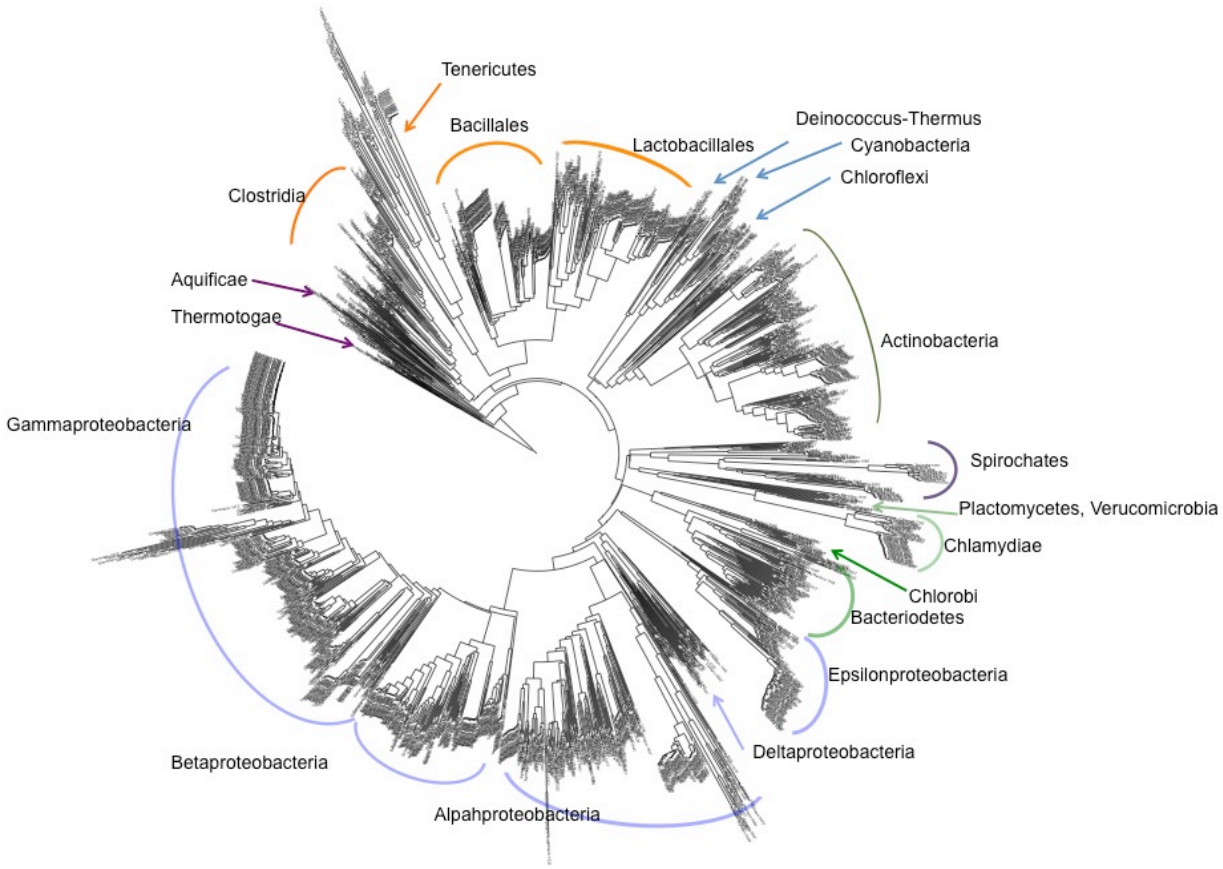


Figure 3-13. Neighbor Joining tree (unrooted) based on Q matrices and profiles under the CNM from the complete alignment. The tree was visualized by a tree viewer, iTOL (Letunic and Bork 2007).

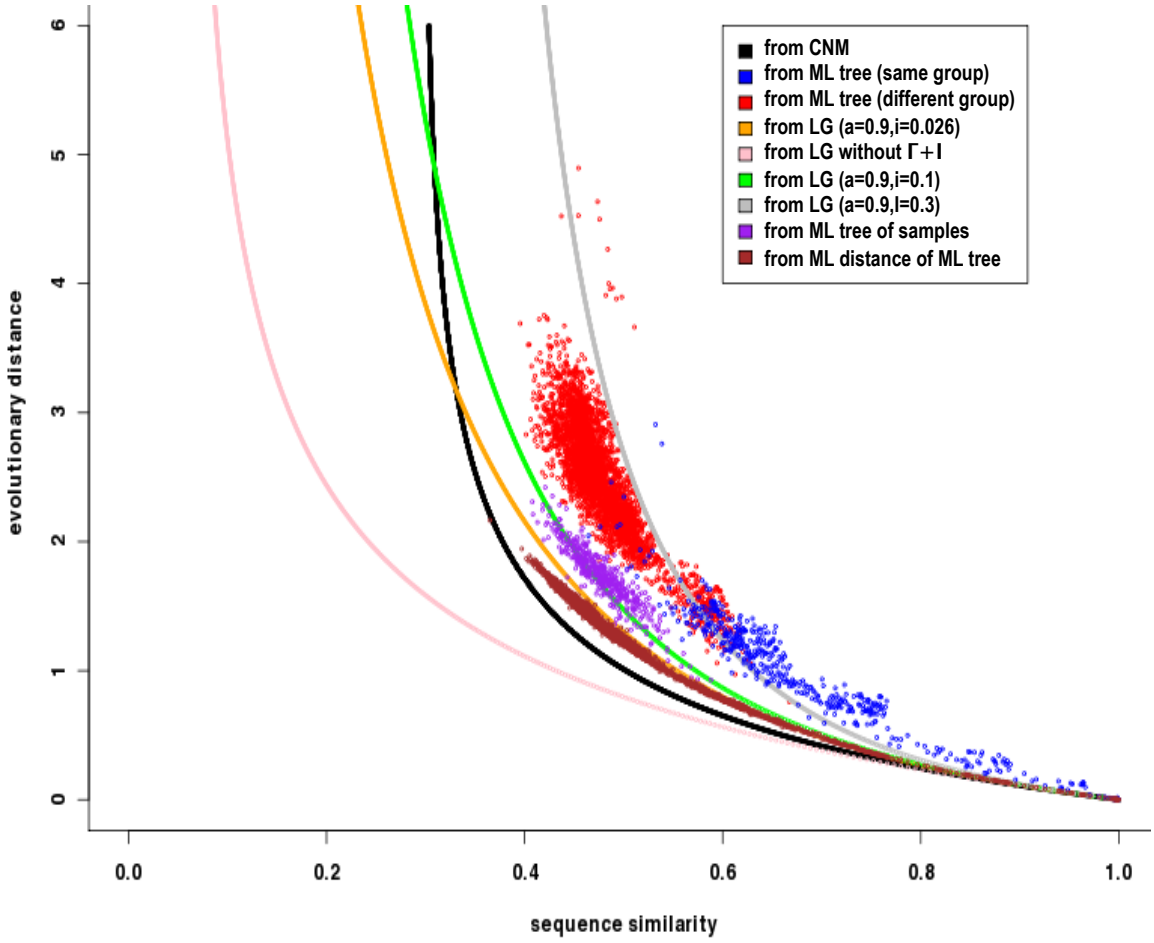


Figure 3-14. Comparison of LG model and the CNM model (position-specific Q) when applied to the bacterial alignment. Red and blue points are patristic distance on the ML tree of each pair of sequences. The LG model with the same parameters used to derive the ML tree, as well as CNM do not provide good fitting to the distribution of patristic distances of the ML tree.

CHAPTER 4
DISCUSSION

In this study, evolutionary relations of bacterial phyla were investigated using three different approaches: (1) a complete bacterial phylogeny first was derived by a conventional ML method using the LG transition matrix and gamma distributed evolutionary rates (LG+Γ+I model), based on the consensus (robust) alignment of a large number of sequences belonging to conserved protein families, (2) then the relation of bacterial phyla (deep-branch patterns) based uniquely on the evolution of amino acids conserved within groups was investigated using a ML approach with a generalized time-reversible matrix and gamma-distributed rates (GTR+Γ+I model), and (3) a complete bacterial phylogeny using the constrained neutral evolution model (CNM) was derived.

I based my analyses on a dataset of 169 protein families classified in the COG database and conserved across most of 1814 bacterial species. A concatenated alignment of reliable positions and characters was created using a newly-developed procedure (CONSALIGN) based on selection of alignment positions consistently aligned by different alignment methods. The corresponding ML tree based on LG+Γ+I model showed clustering of organisms consistent with taxonomic classification and with similar trees previously obtained from large collections of sequences, where individual bacterial phyla were well-supported but were connected by short, poorly supported branches, resulting in a “bush-like” phylogenetic tree.

A Deep-Branching Bacterial Phylogeny Based on Conserved Amino Acids

To evaluate if the bush-like shape of the tree obtained using a site-independent transition matrix and amino acid position characters evolving a different rates could be a

consequence of the inadequacy of the model and of mutational saturation of fast-evolving characters, I reconstructed a deep-branching bacterial phylogeny based only on state evolution of slowly evolving characters, identified as alignment positions conserved within phyla [Conserved Amino acid Sequence (CAS) analysis]. To identify amino acid types conserved within a group, I developed a new index of phylogenetic cluster size, E_{gd}, which takes into consideration of the phylogenetic relations among sequences. This measure was necessary to distinguish clusters where character-state conservation could be ascribed to a selective constraint, from conservation resulting from close evolutionary distances among sequences, as would occur, for example, when multiple strains of the same species are represented among the sequences. The E_{gd} index estimates the expected number of leaf-states not identical by descent represented in a cluster of *N* leaves, and it is bounded between 1.0, when all sequences are connected by branches of length 0, and *N*, when all sequences are connected by branches of infinite length. In connection with E_{gd}, I also defined leaf-weights (≤ 1.0), which measured the "independence" of each leaf from other leaves in the tree, whose sum equaled E_{gd}. Having defined cluster size E_{gd} and leaf weights, we used E_{gd} thresholds for considering conservation within a group, and thresholds of amino acid conservation within a group (ACD), based on the fractional weight of conserved positions within the group. These thresholds were used to select positions conserved within phylogenetic groups representing slowly evolving sites. Rarely evolving characters are expected to provide more reliable information of ancient evolutionary events and I expected them to resolve possible issues of mutational saturation,

evolutionary distance underestimation, and long-branch attraction that could have bias phylogenetic results obtained using complete alignments.

Maximum likelihood trees obtained based on conserved characters resulted in consistent clustering over trees obtained with different stringency in the definition of conserved characters, of the same bacterial phyla. These clusters were reproduced with high statistical support, as measured by the approximate Likelihood Ratio Test (aLRT). Furthermore, the trees confirmed the bush-like shape of previous evolutionary trees (Koonin 2007; Puigbò et al. 2009), suggesting that this indeed reflected the natural history of the evolution of bacterial phyla. Three major clusters, named P, B and F clusters, were consistently reproduced over trees with different thresholds of E_{gd} and ACD. Most of groupings of phyla were consistent with literature based on different type of analysis e.g. signature indels and proteins (Figure 4-1), as detailed below.

The P Cluster

The P cluster consisted of the *Proteobacteria, Acidobacteria, Deferribacteres, Nitrospirae*, and *Aquificae* clades, whose clustering consistently appeared in trees obtained with different thresholds of E_{gd} and ACD.

Nitrospirae is a group represented by relatively few genomes (E_{gd} < 2.0) and thus its phylogenetic position could be estimated only using a permissive size threshold (E ≥ 1.5). In these trees *Nitrospirae* were placed close to the *Proteobacteria*, confirming the results of a previous study based on gene order data from completely sequenced genomes (Kunisawa 2010). A study for the phylogenetic position of the *Deferribacteres* suggested that its proximal phylum is the *Proteobacteria* and *Nitrospirae* (Kunisawa 2011) . In trees containing the *Deferribacteres*, the *Proteobacteria* and *Nitrospirae* were closer to the *Deferribacteres* than others except for the *Acidobacteria*. The *Aquificae*

seemed to be closely related to the *Epsilonproteobacteria* as proposed in other studies (Cavalier-Smith 2002; Deeds et al. 2005; Oshima et al. 2012). The *Acidobacteria* is a major phylum whose cohesiveness is highly supported based on 16s rRNA analysis, but its relation with other phyla is still debated (Ward et al. 2009). Two recent studies based on two acidobacterial isolates suggested that the *Acidobacteria* is a sister group of the *Deltaproteobacteria* (Ciccarelli et al. 2006; Wu and Eisen 2008). These results suggested instead that *Acidobacteria* and *Nitrospirae* are sister groups that diverged from *Proteobacteria* before they were separated into different subgroups (Alpha-, Beta-, Gamma-, Delta-, Epsilon-proteobacteria).

The B Cluster

The B cluster, consistently shown over trees with different thresholds of E_{gd} and ACD, is composed of the *Bacteroidetes, Chlorobi, Planctomycetes*, and *Chlamydiae.*

The PVC clade, composed of *Planctomycetes, Verrucomicrobia* and *Chlamydiae* was obtained consistently with previous analyses (Wagner and Horn 2006; Kamneva et al. 2010). The *Bacteroidetes* and *Chlorobi* clustering was consistent with the signature analysis and formed with the *Fibrobacteres* a superphylum, FCB group (Gupta 2004; Gupta and Lorenzini 2007). In contrast to other clades, the position of *Spirochaetes* was not consistently reproduced by different trees. The *Spirochaetes* was placed outside of other clades at lower thresholds, but at higher thresholds ($E_{gd} > 3.0$) it was consistently grouped with the *Bacteroidetes*. The latter grouping is supported by a signature sequence analysis, which suggested that the *Spirochaetes* are relatives of the FCB group (Gupta 2004). The grouping of *Chlamydiae* with *Spirochaetes* is in agreement with the conclusions of Wolf el al. (2001), based on five different approaches to reconstruct genome trees.

The F Cluster

The *Bacillales, Lactobacillales*, and *Clostridia* (classified as *Firmicutes*) consistently clustered in a clade of *Firmicutes*. The *Tenericutes* were closely associated with the *Firmicutes*, consistently with there previous classification (Garrity et al. 2004) before their classification as a separate group (Ludwig et al. 2009). A study based on a housekeeping gene *pgk* also put the Mollicutes as a sister of other classes of the *Firmicutes* (Wolf et al. 2004). The *Actinobacteria* was clustered in close relation with the *Firmicutes*, often as a sister group of the *Clostridia*, suggesting that all gram-positive bacteria are monophyletic. The *Deinococcus-Thermus* and *Cyanobacteria* (gram-negatives) were closely related to the group of gram-positives. This grouping was also suggested in another study (Wolf et al. 2001) which used five different approaches.

The *Firmicutes*, *Actinobacteria*, *Cyanobacteria*, *Deinococcus-Thermus* and *Mollicutes* (*Tenericutes*) in the F cluster correspond to the group of "Terrabacteria," which was suggested as a major clade of prokaryotes arising from an early colonization of land, and possessing important terrestrial adaptations such as resistance to environmental hazard and oxygenic photosynthesis (Battistuzzi et al. 2004; Battistuzzi and Hedges 2009).

Informational vs. Not-informational Families

CAS analysis uniquely based on protein families associated with Informational processes ("informational"), compared to equivalent analysis based on all other protein families ("non-informational"), resulted in trees with topologies more consistently reproduced using different parameters, and more consistent trees with the topologies resulting from the analysis of all families. A first likely explanation for this result is that the number of informational families is almost twice the number of non-informational

families. The number of informative conserved characters and positions in the informational alignment is also larger than the corresponding number across non-informational proteins. A second possible reason could be that the evolutionary information provided by non-informational proteins, such as metabolism- and structure-related proteins, could be affected by convergent evolution due to adaptations to similar environmental factors, and might be confused by higher proneness to horizontal gene transfer between different species (Jain et al. 1999).

Patterns of Amino Acid Substitution in CAS Alignments

At each position of the CAS alignment, conservation of amino acid types over the time of radiation of individual species within a phylum can be explained by strong negative selection resulting in very slow evolutionary rates. Therefore, conserved amino acids sites should be more likely to contain information on deep evolutionary relations. Furthermore, widespread conservation across individual species would provide more reliable information on the state of the common ancestral sequence eliminating uncertainty resulting from further differentiation.

Changes ("switches") in conserved amino acid types between phyla can be explained by phylum-specific functional changes affecting the fitness of the different amino acid types at homologous positions. Conserved amino acids in a protein family have been used to predict functional residues in a protein in various ways (Polacco and Babbitt 2006; Capra and Singh 2007; Manning et al. 2008; Kalinina et al. 2009). Amino acids conserved within a subfamily of sequences but different between sequence-groups were used to predict functional sites such as residues responsible for determining ligand specificity (Kalinina et al. 2009). Besides of conservation of amino acids, amino acid substitution pattern was used to find functional residues and

understand evolutionary processes. Hanada et al. (2007) suggested 4 classes of amino acid types based on the synonymous/nonsynonymous substitution ratio (Table 4-1), and radical changes in the classification of amino acids were related to the purifying selection of the protein (Hanada et al. 2007). Radical substitution of amino acids was used to find the evolutionary origins of *Trichomonas* hydrogenosomes, where 6 classes of amino acid types were defined and used for phylogenetic analysis (Hrdy et al. 2004).

In other way, amino acid substitution could be occurred by epistatic effect. Epistasis means that impacts of a mutation in one gene depend on the allele states at other loci (Phillips 2008; Breen et al. 2012). The “sign epistasis” is used to define a specific case of epistasis where beneficial mutation on a gene in a species might be deleterious in other species (Weinreich et al. 2005). If different amino acid types were observed in different sequence families at a orthologous site, it could be explained by epistasis where amino acid substitutions that were beneficial or neutral in one species should often be deleterious in another (Breen et al. 2012).

The pairwise substitution rates in the GTR matrices generated from the CAS alignments showed a strong positive correlation with the rates in the LG matrix. That is, among switch positions, substitutions between amino acid with similar physico-chemical properties are prevalent. Hence, the patterns of substitution between conserved amino acid types prevalently involved preservation of the functional/structural properties of the protein sites involved, in contrast to the radical changes in physic-chemical properties mostly associated with functionally differentiated sites (Hanada 2007). This result suggests that the slow rate of differentiation of these characters was affected by epistasis more than by functional differentiation.

Reconstruction of a Bacterial Phylogeny Based on the Constrained Neutral Model

By using slowly evolving amino acid characters ("switch" positions) I sought to use a particular class of rare genomic changes that could provide unambiguous information on the deep phylogenetic relations of bacterial phyla. The ML trees obtained from these data were in fact generally characterized by deep branches with very high support values. Furthermore, they identified and supported many of the phylogenetic relations proposed in the literatures and identified using alternative sources of information. One of the features of the bacterial tree that I thought might be modified by using switch-position information was the bush-like shape of the tree. I hypothesized that the bush-like appearance of the tree could not reflect the natural history of bacterial evolution but rather a bias imposed by under-estimation of long evolutionary distances, resulting in compression of the length of deep branches. In fact, I showed (Figure 3-13) that the patristic pairwise distances in the ML tree obtained from the complete set of reliably aligned sequence positions with the LG+Γ+I evolutionary model, were considerably longer than the pairwise distances predicted by the same model, either as expectations or as maximum-likelihood estimates. This result showed the inadequacy of the model in fitting alignment data. The tree resulting from only using switch-position information, however, had a similar bush-like appearance, suggesting that rapid radiation has been a real feature of early bacterial evolution.

Seeking for an evolutionary model that would better fit the details of the evolution of amino acid characters during bacterial phylogenies, I envisioned a mechanistic neutral model of protein evolution constrained by the definition of position-specific subsets of amino acid types that limit the evolvability of each position, in terms of both rate of evolution and of which amino acid types are allowed at a particular position. The

rationale of applying such model is that, similarly to the CAT model (Lartillot and Philippe 2004), it will generally result in increased estimates of evolutionary distance between sequences with low similarity (distantly related) than with an unspecific transition matrix (e.g., LG) and it will correspond to higher asymptotic similarity as evolutionary distances tend to go to infinity. These properties seemed to be consistent with the relation between evolutionary distance and sequence similarity suggested by the pairwise patristic distances calculated from the LG+Γ+I tree of bacterial sequences. In the application of a constrained neutral model (CNM), I derived position-specific amino acid usage profiles, substitution matrices, and expected evolutionary rates, based on nucleotide and codon substitution models, and on alignment data. When compared to the ML tree from the complete alignment in terms of relationship between phyla, a NJ tree based on CNM is more close to CAS trees and another ML tree generated by Wu and Eisen (2008) (Figure 4-2).

Conclusions and Future Perspectives

Application of different evolutionary models and results from selection of amino acid sites characterized by rare substitution events, suggested that the bush-like shape of reconstructions of the evolution of bacterial phyla is not due to biases in estimating evolutionary distances but reflects the natural history of bacteria evolution. In contrast to the weak support that characterizes tree reconstructions based on complete sequence alignments, selection of slow evolving “switch position” sites resulted in robust reconstructions of the relations among bacterial phyla, which was mostly consistent with the results found in the literature on specific taxonomic groups. Characterization of which rare substitution events occur more frequently at slowly evolving protein sites, suggested that epistasis may play a greater role than functional differentiation in

affecting the rate of evolution of these sites. However, neither application of a general transition rate model to all data or to slowly evolving sites, nor introduction of position-specific profiles of equilibrium amino acid frequencies resulted in predictions of evolutionary distances that satisfactorily fit the data. Furthermore, the results suggested that the introduction of position-specific constraints on amino acid allowable types and equilibrium frequencies is also not sufficient to fit the relation between sequence similarities and evolutionary distances inferred from the resulting phylogenetic tree. I concluded that a different class, possibly lineage-specific, class of evolutionary events must be considered to describe both short-range and long-range evolution of bacterial proteins. Exploratory analyses suggest that introduction of families of models where parameters are defined as a function of evolutionary distance can result in a much better fit of model predictions to the data and I predict that pursuing this new interesting and uncharted territory will produce significant insight on the dynamics of protein evolution.

Table 4-1 Classes of amino acid types based on synonymous and nonsynonymous ratio (Hanada et al. 2007)

Classes	Amino acid type
Hydrophilic and small (MW: 75–146)	A N C G P S T
Hydrophobic and small (MW: 117-149)	I L M V
Negatively charged (MW: 133–147)	D E
Positively charged, aromatic and relatively large (MW: 146-204)	R Q H K F W Y

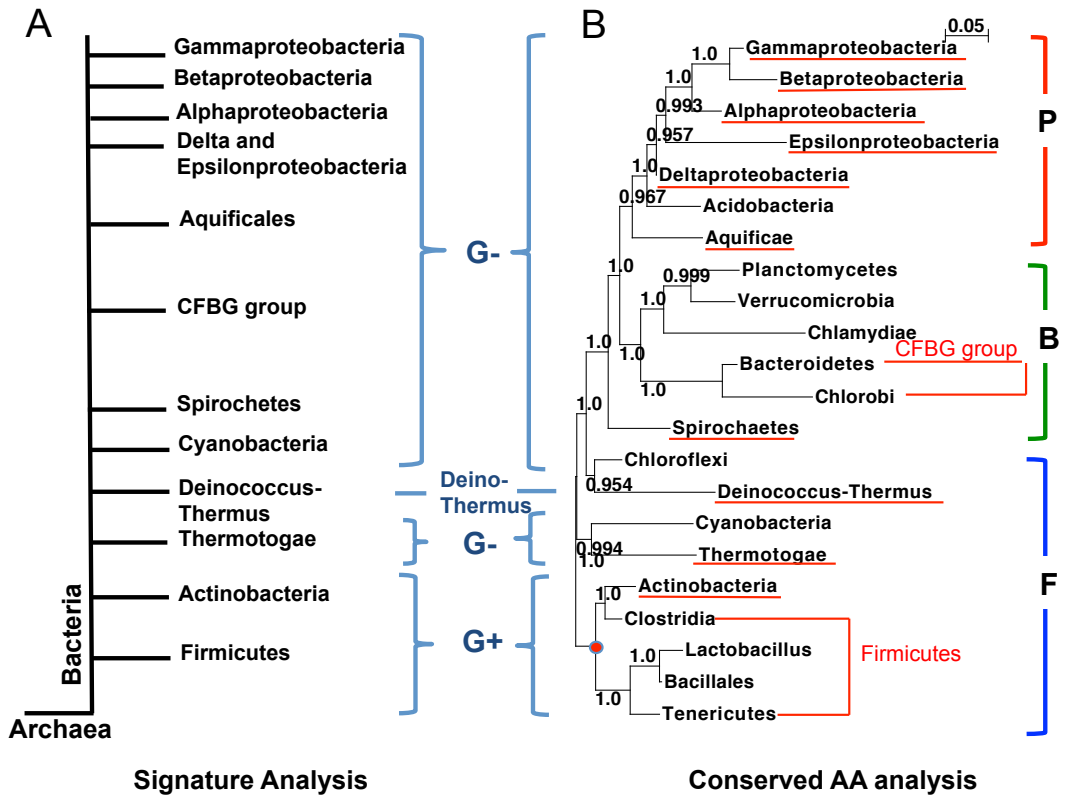


Figure 4-1. (A) Phylogenies based on signature indels and proteins (Gupta, 2005) and (B) A phylogenetic tree from conserved amino acid analysis (E_{gd}>2.0, ACD=1.0, rooted by the *Firmicutes* clade in red point). The groups from the conserved amino acid analysis were marked as red when they were in the same order as signature analysis. G+ and G- indicate gram-positive and gram-negative, respectively.

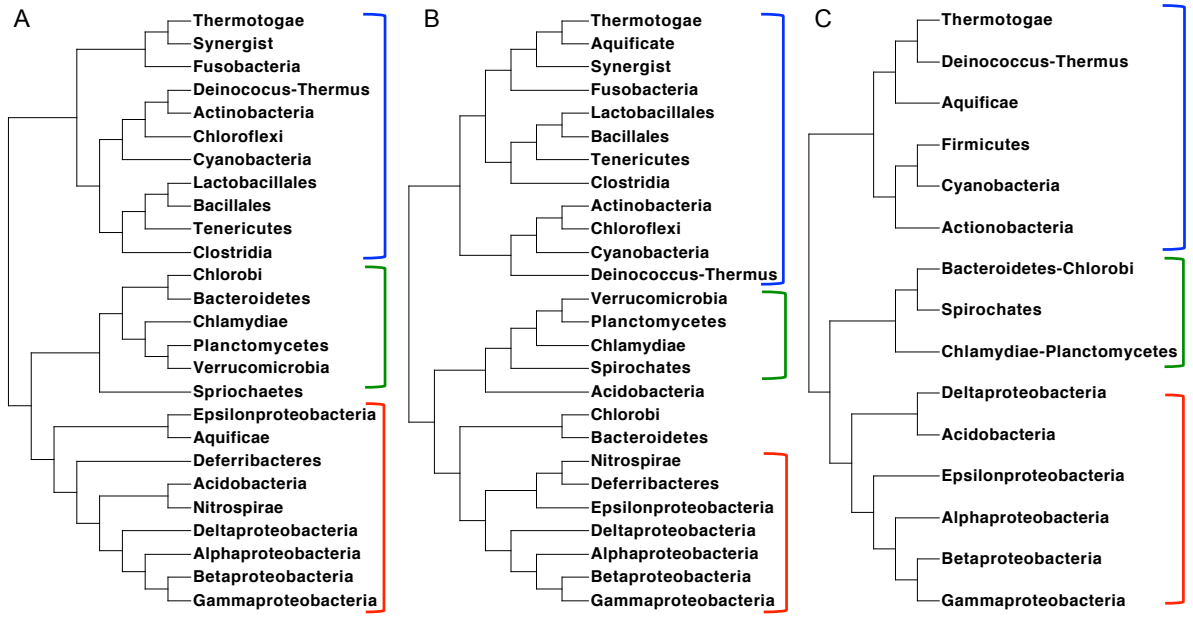


Figure 4-2. Comparison of topologies of three bacterial phylogenetic trees. (A) NJ tree and (B) ML tree from this study and (C) ML tree reproduced from Figure A-2 (Wu and Eisen, 2008). All are unrooted trees. Groups of three major clusters were indicated by blue (F), green (B), and red (P).

APPENDIX
SUPPLEMENTS

Table A-1. Nearly universal COG families used for the dataset. See Tables 2-4 and 2-5 for description of functional category.

COG	Functional Category	Description
COG0006	E	Xaa-Pro aminopeptidase
COG0008	J	Glutamyl- and glutaminyl-tRNA synthetases
COG0012	J	Predicted GTPase, probable translation factor
COG0013	J	Alanyl-tRNA synthetase
COG0016	J	Phenylalanyl-tRNA synthetase alpha subunit
COG0018	J	Arginyl-tRNA synthetase
COG0024	J	Methionine aminopeptidase
COG0030	J	Dimethyladenosine transferase (rRNA methylation)
COG0037	D	Predicted ATPase of the PP-loop superfamily implicated in cell cycle control
COG0048	J	Ribosomal protein S12
COG0049	J	Ribosomal protein S7
COG0050	J	GTPases - translation elongation factors
COG0051	J	Ribosomal protein S10
COG0052	J	Ribosomal protein S2
COG0060	J	Isoleucyl-tRNA synthetase
COG0072	J	Phenylalanyl-tRNA synthetase beta subunit
COG0073	R	EMAP domain
COG0080	J	Ribosomal protein L11
COG0081	J	Ribosomal protein L1
COG0084	L	Mg-dependent DNase
COG0085	K	DNA-directed RNA polymerase, beta subunit/140 kD subunit
COG0086	K	DNA-directed RNA polymerase, beta subunit/160 kD subunit
COG0087	J	Ribosomal protein L3
COG0088	J	Ribosomal protein L4
COG0089	J	Ribosomal protein L23
COG0090	J	Ribosomal protein L2
COG0091	J	Ribosomal protein L22
COG0092	J	Ribosomal protein S3
COG0093	J	Ribosomal protein L14
COG0094	J	Ribosomal protein L5
COG0096	J	Ribosomal protein S8
COG0097	J	Ribosomal protein L6P/L9E
COG0098	J	Ribosomal protein S5

Table A-1. Continued.

COG	Functional Category	Description
COG0099	J	Ribosomal protein S13
COG0100	J	Ribosomal protein S11
COG0101	J	Pseudouridylate synthase
COG0102	J	Ribosomal protein L13
COG0103	J	Ribosomal protein S9
COG0112	E	Glycine/serine hydroxymethyltransferase
COG0124	J	Histidyl-tRNA synthetase
COG0125	F	Thymidylate kinase
COG0130	J	Pseudouridine synthase
COG0143	J	Methionyl-tRNA synthetase
COG0162	J	Tyrosyl-tRNA synthetase
COG0172	J	Seryl-tRNA synthetase
COG0173	J	Aspartyl-tRNA synthetase
COG0178	L	Excinuclease ATPase subunit
COG0180	J	Tryptophanyl-tRNA synthetase
COG0184	J	Ribosomal protein S15P/S13E
COG0185	J	Ribosomal protein S19
COG0186	J	Ribosomal protein S17
COG0187	L	Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV), B subunit
COG0188	L	Type IIA topoisomerase (DNA gyrase/topo II, topoisomerase IV), A subunit
COG0190	H	5,10-methylene-tetrahydrofolate dehydrogenase/Methenyl tetrahydrofolate cyclohydrolase
COG0193	J	Peptidyl-tRNA hydrolase
COG0195	K	Transcription elongation factor
COG0197	J	Ribosomal protein L16/L10E
COG0198	J	Ribosomal protein L24
COG0199	J	Ribosomal protein S14
COG0200	J	Ribosomal protein L15
COG0201	U	Preprotein translocase subunit SecY
COG0202	K	DNA-directed RNA polymerase, alpha subunit/40 kD subunit
COG0203	J	Ribosomal protein L17
COG0204	I	1-acyl-sn-glycerol-3-phosphate acyltransferase
COG0209	F	Ribonucleotide reductase, alpha subunit
COG0210	L	Superfamily I DNA and RNA helicases
COG0211	J	Ribosomal protein L27
COG0215	J	Cysteinyl-tRNA synthetase
COG0216	J	Protein chain release factor A

Table A-1. Continued.

COG	Functional Category	Description
COG0217	S	Uncharacterized conserved protein
COG0220	R	Predicted S-adenosylmethionine-dependent methyltransferase
COG0222	J	Ribosomal protein L7/L12
COG0223	J	Methionyl-tRNA formyltransferase
COG0227	J	Ribosomal protein L28
COG0228	J	Ribosomal protein S16
COG0230	J	Ribosomal protein L34
COG0231	J	Translation elongation factor P (EF-P)/translation initiation factor 5A (eIF-5A)
COG0233	J	Ribosome recycling factor
COG0234	O	Co-chaperonin GroES (HSP10)
COG0236	IQ	Acyl carrier protein
COG0237	H	Dephospho-CoA kinase
COG0238	J	Ribosomal protein S18
COG0240	C	Glycerol-3-phosphate dehydrogenase
COG0242	J	N-formylmethionyl-tRNA deformylase
COG0244	J	Ribosomal protein L10
COG0250	K	Transcription antiterminator
COG0254	J	Ribosomal protein L31
COG0255	J	Ribosomal protein L29
COG0256	J	Ribosomal protein L18
COG0257	J	Ribosomal protein L36
COG0258	L	5-3 exonuclease (including N-terminal domain of PolI)
COG0261	J	Ribosomal protein L21
COG0264	J	Translation elongation factor Ts
COG0267	J	Ribosomal protein L33
COG0268	J	Ribosomal protein S20
COG0272	L	NAD-dependent DNA ligase (contains BRCT domain type II)
COG0275	M	Predicted S-adenosylmethionine-dependent methyltransferase involved in cell envelope biogenesis
COG0290	J	Translation initiation factor 3 (IF-3)
COG0291	J	Ribosomal protein L35
COG0292	J	Ribosomal protein L20
COG0305	L	Replicative DNA helicase
COG0313	R	Predicted methyltransferases
COG0319	R	Predicted metal-dependent hydrolase
COG0322	L	Nuclease subunit of the excinuclease complex
COG0335	J	Ribosomal protein L19

Table A-1. Continued.

COG	Functional Category	Description
COG0336	J	tRNA-(guanine-N1)-methyltransferase
COG0358	L	DNA primase (bacterial type)
COG0359	J	Ribosomal protein L9
COG0360	J	Ribosomal protein S6
COG0361	J	Translation initiation factor 1 (IF-1)
COG0441	J	Threonyl-tRNA synthetase
COG0442	J	Prolyl-tRNA synthetase
COG0443	O	Molecular chaperone
COG0459	O	Chaperonin GroEL (HSP60 family)
COG0465	O	ATP-dependent Zn proteases
COG0468	L	RecA/RadA recombinase
COG0477	GEPR	Permeases of the major facilitator superfamily
COG0480	J	Translation elongation factors (GTPases)
COG0481	M	Membrane GTPase LepA
COG0482	J	Predicted tRNA(5-methylaminomethyl-2-thiouridylate) methyltransferase, contains the PP-loop ATPase domain
COG0484	O	DnaJ-class molecular chaperone with C-terminal Zn finger domain
COG0492	O	Thioredoxin reductase
COG0495	J	Leucyl-tRNA synthetase
COG0500	QR	SAM-dependent methyltransferases
COG0504	F	CTP synthase (UTP-ammonia lyase)
COG0522	J	Ribosomal protein S4 and related proteins
COG0525	J	Valyl-tRNA synthetase
COG0526	OC	Thiol-disulfide isomerase and thioredoxins
COG0528	F	Uridylate kinase
COG0532	J	Translation initiation factor 2 (IF-2; GTPase)
COG0533	O	Metal-dependent proteases with possible chaperone activity
COG0536	R	Predicted GTPase
COG0541	U	Signal recognition particle GTPase
COG0542	O	ATPases with chaperone activity, ATP-binding subunit
COG0544	O	FKBP-type peptidyl-prolyl cis-trans isomerase (trigger factor)
COG0550	L	Topoisomerase IA
COG0552	U	Signal recognition particle GTPase
COG0556	L	Helicase subunit of the DNA excision repair complex
COG0563	F	Adenylate kinase and related kinases
COG0564	J	Pseudouridylate synthases, 23S RNA-specific

Table A-1. Continued

COG	Functional Category	Description
COG0566	J	rRNA methylases
COG0568	K	DNA-directed RNA polymerase, sigma subunit (sigma70/sigma32)
COG0571	K	dsRNA-specific ribonuclease
COG0575	I	CDP-diglyceride synthetase
COG0576	O	Molecular chaperone GrpE (heat shock protein)
COG0582	L	Integrase
COG0587	L	DNA polymerase III, alpha subunit
COG0592	L	DNA polymerase sliding clamp subunit (PCNA homolog)
COG0593	L	ATPase involved in DNA replication initiation
COG0597	MU	Lipoprotein signal peptidase
COG0629	L	Single-stranded DNA-binding protein
COG0635	H	Coproporphyrinogen III oxidase and related Fe-S oxidoreductases
COG0636	C	F0F1-type ATP synthase, subunit c/Archaeal/vacuolar-type H+-ATPase, subunit K
COG0653	U	Preprotein translocase subunit SecA (ATPase, RNA helicase)
COG0682	M	Prolipoprotein diacylglyceryltransferase
COG0690	U	Preprotein translocase subunit SecE
COG0691	O	tmRNA-binding protein
COG0706	U	Preprotein translocase subunit YidC
COG0776	L	Bacterial nucleoid DNA-binding protein
COG0781	K	Transcription termination factor
COG0858	J	Ribosome-binding factor A
COG1109	G	Phosphomannomutase
COG1136	V	ABC-type antimicrobial peptide transport system, ATPase component
COG1160	R	Predicted GTPases
COG1214	O	Inactive homolog of metal-dependent proteases, putative molecular chaperone
COG1253	R	Hemolysins and related proteins containing CBS domains
COG1466	L	DNA polymerase III, delta subunit
COG2812	L	DNA polymerase III, gamma/tau subunits
COG2890	J	Methylase of polypeptide chain release factors

Table A-2. Test on random distribution of each amino acid type (AA) in DnaK protein family (771 positions).

AA	Num. of positions having AA	Num. of positions having random-distributed AA		Size of the cluster containing AA		Size of the rest on the Tree	
		P< .01	P< .001	Average	Stdev	Average	Stdev
A	496	50	22	346.69	224.89	144.09	215.57
C	179	3	0	155.70	213.77	361.14	212.24
D	361	11	1	347.43	214.44	132.78	200.84
E	404	9	3	355.68	211.21	130.45	200.86
F	245	8	3	242.86	238.02	257.72	237.13
G	333	45	19	293.73	221.76	177.95	220.01
H	269	8	4	183.63	217.38	303.28	223.59
I	388	47	7	317.74	232.83	183.61	226.31
K	408	6	1	358.97	210.24	115.21	189.85
L	362	45	17	312.52	234.28	180.73	226.22
M	318	15	10	270.30	237.21	227.03	235.53
N	398	16	6	316.98	218.67	163.44	214.58
P	220	25	6	273.64	235.20	199.45	227.48
Q	414	6	0	314.90	225.28	169.62	216.19
R	352	61	35	312.21	231.38	163.76	213.95
S	475	76	40	345.25	217.33	133.89	197.50
T	442	59	21	331.16	225.69	158.07	213.79
V	413	40	12	341.26	222.19	170.29	214.99
W	151	0	0	131.25	201.73	358.94	199.65
Y	219	1	0	230.10	227.30	273.58	228.17
SUM	6847	531	207				

Table A-3. Test on random distribution of each amino acid type (AA) in GroEL protein family (583 positions).

AA	Num. of positions having AA	Num. of positions having random-distributed AA		Size of the cluster containing AA		Size of the rest on the Tree	
		P< .01	P< .001	Average	Stdev	Average	Stdev
A	403	46	24	581.16	329.73	209.71	325.42
C	151	4	2	270.90	352.67	525.00	349.30
D	250	13	7	591.66	321.06	177.91	308.04
E	278	4	1	607.53	305.36	167.50	293.34
F	185	20	7	331.88	353.50	435.44	350.83
G	265	59	35	458.63	365.45	312.84	364.14
H	208	6	2	351.65	360.22	437.80	365.49
I	317	44	15	567.50	335.10	225.15	327.42
K	277	16	4	560.78	332.55	222.80	324.90
L	312	61	37	507.50	354.67	273.17	345.62
M	247	22	12	406.98	362.06	388.06	360.85
N	265	26	8	518.01	341.87	252.69	335.54
P	139	13	5	440.97	351.78	293.95	344.99
Q	278	6	4	494.20	352.67	298.38	351.38
R	248	53	41	493.38	358.84	286.43	354.33
S	378	61	41	555.01	331.41	229.19	327.05
T	339	47	37	546.75	345.57	241.50	341.83
V	356	38	18	575.25	329.17	211.74	321.79
W	99	0	0	179.94	298.18	610.61	294.24
Y	165	3	0	333.55	336.16	441.14	342.30
SUM		287	212				

Table A-4. Test on random distribution of each amino acid type (AA) in RecA protein family (408 positions).

AA	Num. of positions having AA	Num. of positions having random-distributed AA		Size of the cluster containing AA		Size of the rest on the Tree	
		P< .01	P< .001	Average	Stdev	Average	Stdev
A	23	23	14	644.30	346.82	223.92	349.33
C	2	2	1	355.13	393.85	523.21	393.21
D	4	4	2	621.18	350.76	228.26	343.39
E	6	6	2	602.18	364.00	243.31	359.21
F	11	11	4	438.89	409.05	429.93	403.35
G	32	32	10	530.13	389.04	331.92	387.60
H	2	2	0	486.33	384.12	369.72	381.16
I	20	20	7	645.76	349.96	215.96	342.80
K	4	4	2	634.54	344.51	220.56	345.40
L	35	35	17	586.28	369.58	282.61	367.27
M	11	11	3	487.36	391.68	379.72	394.31
N	12	12	5	613.79	352.88	249.57	357.58
P	16	16	9	492.09	382.82	354.70	396.93
Q	2	2	1	620.75	349.03	241.74	353.51
R	28	28	19	593.24	367.38	266.29	362.80
S	21	21	13	643.36	345.83	219.11	343.66
T	25	25	16	590.15	348.62	267.33	347.13
V	25	25	7	626.32	354.99	232.43	351.07
W	0	0	0	225.56	375.78	652.31	375.70
Y	1	1	0	417.68	405.82	442.69	401.55
SUM		163	101				

Table A-5. Test on candidacy of each amino acid type (AA) in position-specific profile in DnaK protein family.

AA	Num. of positions not having AA	[a]Num. of positions having significant AA		
		P<.01	P<.05	P<.5
A	222	119	138	188
C	539	55	141	389
D	357	65	147	296
E	314	57	135	252
F	473	118	183	384
G	385	249	295	352
H	449	66	159	354
I	330	136	229	298
K	310	76	118	253
L	356	185	267	344
M	400	63	106	287
N	320	96	192	281
P	498	342	387	459
Q	304	30	78	178
R	366	252	288	364
S	243	190	205	243
T	276	180	199	261
V	305	140	194	272
W	567	9	52	441
Y	499	49	117	364

[a]Num. of positions having significant AA means number of positions where the likelihood of the amino acid is less than p-value.

Table A-6. Test on candidacy of each amino acid type (AA) in position-specific profile in GroEL protein family.

AA	Num. of positions not having AA	[a]Num. of positions having significant AA		
		P<.01	P<.05	P<.5
A	180	98	107	142
C	432	93	187	322
D	333	80	195	288
E	305	79	175	261
F	398	155	228	329
G	318	231	259	293
H	375	73	156	273
I	266	150	203	232
K	306	63	132	250
L	271	180	207	255
M	336	70	142	256
N	318	122	224	283
P	444	358	383	421
Q	305	48	95	196
R	335	257	288	331
S	205	172	177	202
T	244	178	195	223
V	227	141	159	201
W	484	25	90	414
Y	418	60	142	306

[a]Num. of positions having significant AA means number of positions where the likelihood of the amino acid is less than p-value.

Table A-7. Test on candidacy of each amino acid type (AA) in position-specific profile in RacA protein family.

AA	Num. of positions not having AA	[a]Num. of positions having significant AA		
		P<.01	P<.05	P<.5
A	118	61	73	102
C	251	60	125	206
D	210	48	123	187
E	194	40	98	168
F	221	86	134	186
G	184	124	144	170
H	229	41	98	186
I	148	83	113	136
K	181	30	65	150
L	144	95	114	137
M	175	34	71	134
N	186	62	119	163
P	257	191	219	246
Q	183	26	52	133
R	205	158	183	205
S	130	114	116	130
T	133	82	99	121
V	125	63	79	110
W	288	22	54	263
Y	232	26	74	168

[a]Num. of positions having significant AA means number of positions where the likelihood of the amino acid is less than p-value.

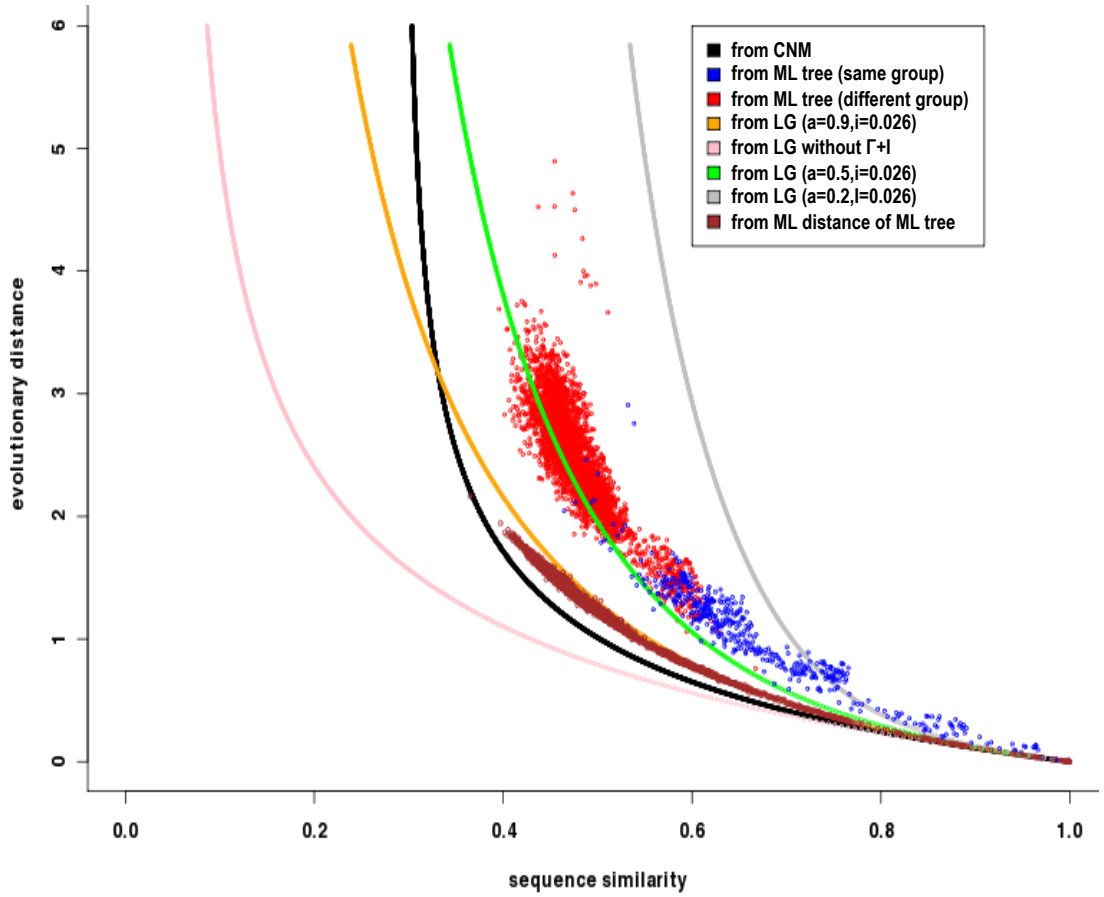


Figure A-1. Comparison of LG model (with different alpha parameters) and the CNM model (position-specific Q) when applied to the bacterial alignment. The expectations inferred by the LG model with different alpha parameters of gamma distribution (green and pink curves) do not provide good fitting to the distribution of patristic distances of the ML tree.

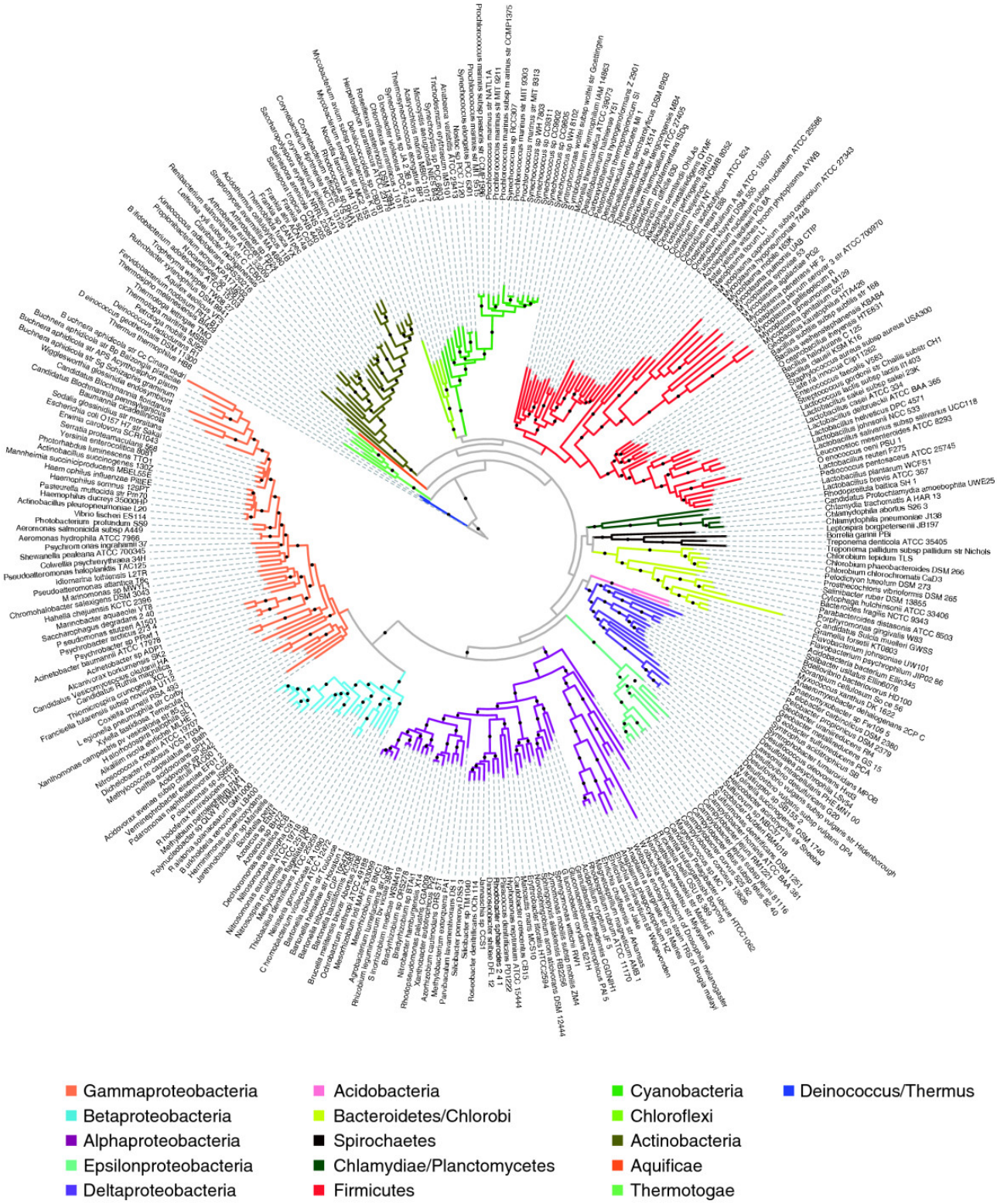


Figure A-2. A ML tree of bacterial phylogeny from Wu and Eisen (2008). The branches with bootstrap support>0.8 are indicated with black dots.

LIST OF REFERENCES

Bapteste E, Boucher Y, Leigh J, Doolittle WF. 2004. Phylogenetic reconstruction and lateral gene transfer. Trends in microbiology 12:406–411.

Bapteste E, Susko E, Leigh J, MacLeod D, Charlebois RL, Doolittle WF. 2005. Do orthologous gene phylogenies really support tree-thinking? BMC evolutionary biology 5:33.

Battistuzzi FU, Feijao A, Hedges SB. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC evolutionary biology 4:44.

Battistuzzi FU, Hedges SB. 2009. A major clade of prokaryotes with ancient adaptations to life on land. Molecular biology and evolution 26:335–343.

Blanquart S, Lartillot N. 2006. A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. Molecular biology and evolution 23:2058–2071.

Boussau B, Guéguen L, Gouy M. 2008. Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. BMC evolutionary biology 8:272.

Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA. 2012. Epistasis as the primary factor in molecular evolution. Nature 490:535–538.

Brocchieri L, Karlin S. 1998. A symmetric-iterated multiple alignment of protein sequences. Journal of molecular biology 276:249–264.

Brocchieri L. 2001. Phylogenetic inferences from molecular sequences: review and critique. Theoretical population biology 59:27–40.

Brochier C, Bapteste E, Moreira D, Philippe H. 2002. Eubacterial phylogeny based on translational apparatus proteins. Trends in genetics : TIG 18:1–5.

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ. 2001. Universal trees based on large combined protein sequence data sets. Nature genetics 28:281–285.

Buchholz P. 1994. Exact and ordinary lumpability in finite Markov chains. Journal of applied probability 31:59.

Cadotte MW, Jonathan Davies T, Regetz J, Kembel SW, Cleland E, Oakley TH. 2010. Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. Ecology letters 13:96–105.

Capra JA, Singh M. 2007. Predicting functionally important residues from sequence conservation. Bioinformatics (Oxford, England) 23:1875–1882.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution 17:540–552.

Cavalier-Smith T. 2002. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. International journal of systematic and evolutionary microbiology 52:7–76.

Choi I-G, Kim S-H. 2007. Global extent of horizontal gene transfer. Proceedings of the National Academy of Sciences of the United States of America 104:4489–4494.

Churchill GA, Von Haeseler A, Navidi WC. 1992. Sample size for a phylogenetic inference. Molecular biology and evolution 9:753–769.

Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311:1283–1287.

Crooks GE, Brenner SE. 2005. An alternative model of amino acid replacement. Bioinformatics 21:975–980.

Cuthill JFH, Braddy SJ, Donoghue PCJ. 2010. A formula for maximum possible steps in multistate characters: isolating matrix parameter effects on measures of evolutionary convergence. Cladistics 26:98–102.

Daubin V, Gouy M, Perrière G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome research 12:1080–1090.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff M, editor. Atlas of Protein Sequence and Structure. 5th ed. Wachington,DC: National Biomedical Research Foundation. p. 345–352.

Deeds EJ, Hennessey H, Shakhnovich EI. 2005. Prokaryotic phylogenies inferred from protein structural domains. Genome research 15:393–402.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nature reviews. Genetics 6:361–375.

Doolittle WF. 1999. Phylogenetic Classification and the Universal Tree. Science 284:2124–2128.

Duret L, Mouchiroud D. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Molecular biology and evolution 17:68–74.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research 32:1792–1797.

Faith DP. 1992. conservation evaluation and phylogenetic diversity. Biological Conservation 61:1–10.

Farris JS. 1989. The Retention index and the rescaled consistency index. Cladistics 5:417–419.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution 17:368–376.

Filzmoser P, Garrett RG, Reimann C. 2005. Multivariate outlier detection in exploration geochemistry. Computers & Geosciences 31:579–587.

Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology 20:406–416.

Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. Journal of molecular evolution 48:284–290.

Foster PG. 2004. Modeling compositional heterogeneity. Systematic biology 53:485–495.

Galtier N, Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. Proceedings of the National Academy of Sciences of the United States of America 92:11317–11321.

Gao B, Mohan R, Gupta RS. 2009. Phylogenomics and protein signatures elucidating the evolutionary relationships among the Gammaproteobacteria. International journal of systematic and evolutionary microbiology 59:234–247.

Garrity GM, Bell JA, Lilburn TG. 2004. Taxonomic Outline of the Prokaryotes. In: Bergey's Manual of Systematic Bacteriology. Second. Springer New York. p. 140–204.

Gogarten JP, Townsend JP. 2005. Horizontal gene transfer, genome innovation and evolution. Nature reviews. Microbiology 3:679–687.

Grishin NV. 1995. Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. Journal of molecular evolution 41:675–679.

Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. Molecular biology and evolution 16:1664–1674.

Gu X. 2006. A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. Molecular biology and evolution 23:1937–1945.

Gupta RS, Lorenzini E. 2007. Phylogeny and molecular signatures (conserved proteins and indels) that are specific for the Bacteroidetes and Chlorobi species. BMC evolutionary biology 7:71.

Gupta RS. 2004. The phylogeny and signature sequences characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes. Critical reviews in microbiology 30:123–143.

Hanada K, Shiu S-H, Li W-H. 2007. The nonsynonymous/synonymous substitution rate ratio versus the radical/conservative replacement rate ratio in the evolution of mammalian genes. Molecular biology and evolution 24:2235–2241.

Henikoff S, Henikoff JG. 1991. Automated assembly of protein blocks for database searching. Nucleic acids research 19:6565–6572.

Hillis DM, Moritz C, Mable BK. 1996. Molecular Systematics. second edi. Sunderland, MA, USA: Sinauer Associates, Inc.

Holder M, Lewis PO. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nature reviews. Genetics 4:275–284.

Howe K, Bateman A, Durbin R. 2002. QuickTree : building huge Neighbour-Joining trees of protein sequences. Bioinformatics 18:1546–1547.

Hrdy I, Hirt RP, Dolezal P, Bardonová L, Foster PG, Tachezy J, Embley TM. 2004. Trichomonas hydrogenosomes contain the NADH dehydrogenase module of mitochondrial complex I. Nature 432:618–622.

Isaac NJB, Turvey ST, Collen B, Waterman C, Baillie JEM. 2007. Mammals on the EDGE: conservation priorities based on threat and phylogeny. PloS one 2:e296.

Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proceedings of the National Academy of Sciences of the United States of America [Internet] 96:3801–3806.

Jermiin L, Ho SY, Ababneh F, Robinson J, Larkum AW. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Systematic biology 53:638–643.

Jin L, Nei M. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Molecular biology and evolution 7:82–102.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Computer applications in the biosciences : CABIOS 8:275–282.

Kalinina O V, Gelfand MS, Russell RB. 2009. Combining specificity determining and conserved residues improves functional site prediction. BMC bioinformatics 10:174.

Kamneva OK, Liberles D a, Ward NL. 2010. Genome-wide influence of indel Substitutions on evolution of bacteria of the PVC superphylum, revealed using a novel computational method. Genome biology and evolution 2:870–886.

Kanhere A, Vingron M. 2009. Horizontal Gene Transfers in prokaryotes show differential preferences for metabolic and translational genes. BMC evolutionary biology 9:9.

Karlin S, Brocchieri L. 1998. Heat shock protein 70 family: multiple sequence comparisons, function, and evolution. Journal of molecular evolution 47:565–577.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution 16:111–120.

Komar AA, Lesnik T, Reiss C. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS letters 462:387–391.

Koonin EV. 2007. The Biological Big Bang model for the major transitions in evolution. Biology direct 2:21.

Kosiol C, Holmes I, Goldman N. 2007. An empirical codon model for protein sequence evolution. Molecular biology and evolution 24:1464–1479.

Kuma K, Iwabe N, Miyata T. 1995. Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families. Molecular biology and evolution 12:123–130.

Kunisawa T. 2010. The phylogenetic placement of the non-phototrophic gram-positive thermophile Thermobaculum terrenum and branching orders within the phylum Chloroflexi inferred from gene order comparisons. International journal of systematic and evolutionary microbiology.

Kunisawa T. 2011. Inference of the phylogenetic position of the phylum Deferribacteres from gene order comparison. Antonie van Leeuwenhoek 99:417–422.

Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC evolutionary biology 7 Suppl 1:S4.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular biology and evolution 21:1095–1109.

Lartillot N, Philippe H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 363:1463–1472.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Molecular biology and evolution 25:1307–1320.

Lee HW, Brocchieri L. 2013. The Evolution of Fuzzy Proteins. J Phylogen Evolution Biol 1:e102.

Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics (Oxford, England) 23:127–128.

Lichtarge O, Bourne HR, Cohen FE. 1996. An evolutionary trace method defines binding surfaces common to protein families. Journal of molecular biology 257:342–358.

Liò P, Goldman N. 1998. Models of Molecular Evolution and Phylogeny Models of Molecular Evolution and Phylogeny. Genome Research:1233–1244.

Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AW. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. Journal of molecular evolution 34:153–162.

Ludwig W, Klenk H. 2001. Overview : A Phylogenetic Backbone and Taxonomic Framework for Procaryotic Systematics. In: D.R.Boone, R.W.Castenholz, editors. Bergey's Manual of Systematic Bacteriology. 2nd ed. Berlin, Germany: Springer-Verlag. p. 49–66.

Ludwig W, Schleifer K-H, Whitmans WB. 2009. Revised road map to the phylum Firmicutes. In: Bergey's Manual of Systematic Bacteriology. 2nd ed. New York: Springer-Verlag. p. 1–32.

Manning JR, Jefferson ER, Barton GJ. 2008. The contrasting properties of conservation and correlated phylogeny in protein functional residue prediction. BMC bioinformatics 9:51.

McInerney JO, Cotton JA, Pisani D. 2008. The prokaryotic tree of life: past, present... and future? Trends in ecology & evolution 23:276–281.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology 48:443–453.

Nei M, Chakraborty R, Fuerst PA. 1976. Infinite allele model with varying mutation rate. Proceedings of the National Academy of Sciences of the United States of America 73:4164–4168.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular biology and evolution 3:418–426.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology 302:205–217.

Oshima K, Chiba Y, Igarashi Y, Arai H, Ishii M. 2012. Phylogenetic position of aquificales based on the whole genome sequences of six aquificales species. International journal of evolutionary biology 2012:859264.

Ota T, Nei M. 1994. Estimation of the Number of Amino Acid Substitutions Per Site When the Substitution Rate Varies Among Sites. Journal of molecular evolution 38:642–643.

Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927–931.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. Nature reviews. Genetics 7:337–348.

Phillips PC. 2008. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. Nature reviews. Genetics 9:855–867.

Polacco BJ, Babbitt PC. 2006. Automated discovery of 3D motifs for protein function annotation. Bioinformatics (Oxford, England) 22:723–730.

Puigbò P, Wolf YI, Koonin EV. 2009. Search for a “Tree of Life” in the thicket of the phylogenetic forest. Journal of biology 8:59.

Quang le S, Gascuel O, Lartillot N. 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24:2317–2323.

Quang le S, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 363:3965–3976.

Reeves JH. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. Journal of molecular evolution 35:17–31.

Rodríguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Systematic biology 56:389–399.

Rogozin IB, Thomson K, Csürös M, Carmel L, Koonin EV. 2008. Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series. Biology direct 3:7.

Rogozin IB, Wolf YI, Carmel L, Koonin EV. 2007. Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. Molecular biology and evolution 24:1080–1090.

Rokas A, Carroll SB. 2008. Frequent and widespread parallel evolution of protein sequences. Molecular biology and evolution 25:1943–1953.

Rokas A, Holland P. 2000. Rare genomic changes as a tool for phylogenetics. Trends in ecology & evolution 15:454–459.

Sauna ZE, Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease. Nature reviews. Genetics 12:683–691.

Schneider A, Cannarozzi GM, Gonnet GH. 2005. Empirical codon substitution matrix. BMC bioinformatics 6:134.

Schweitzer PJ. 1984. Aggregation Methods for Large Markov Chains. Proceedings of the International Workshop on Computer Performance and Reliability:275–286.

Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Computer applications in the biosciences : CABIOS 12:327–345.

Sleator RD. 2011. Phylogenetics. Archives of microbiology 193:235–239.

Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456–463.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics (Oxford, England) 22:2688–2690.

Steel M, Penny D. 2006. Maximum parsimony and the phylogenetic information in multistate characters. In: Albert VA, editor. Parsimony, Phylgeny, and Genomics. Oxford (UK): Oxford University Press. p. 163–180.

Swithers KS, Gogarten JP, Fournier GP. 2009. Trees in the web of life. Journal of biology 8:54.

Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular biology and evolution 10:512–526.

Tarrío R, Rodríguez-Trelles F, Ayala FJ. 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Molecular biology and evolution 18:1464–1473.

Tatusov RL, Fedorova ND, Jackson JD, et al. 2003. The COG database: an updated version includes eukaryotes. BMC bioinformatics:41.

Tatusov RL. 1997. A Genomic Perspective on Protein Families. Science 278:631–637.

Teichmann SA, Mitchison G. 1999. Is there a phylogenetic signal in prokaryote proteins? Journal of molecular evolution 49:98–107.

Thorne JL. 2000. Models of protein sequence evolution and their applications. Current opinion in genetics & development 10:602–605.

Uzzell T, Corbin KW. 1971. Fitting discrete probability distributions to evolutionary events. Science 172:1089–1096.

Vane-Wright RI, Humphries CJ, Williams PH. 1991. What to Protect? Systematics and the Agony of Choice. Biological Conservation 55:235–254.

Vellend M, Cornwell WK, Magnuson-Ford K, Mooers AØ. 2010. Measuring phylogenetic biodiversity. In: Magurran E, McGil J, editors. Biological diversity: Frontiers in measurement and assessment. A. E.Magur. Oxford (UK): Oxford University Press. p. 194–207.

Wagner M, Horn M. 2006. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. Current opinion in biotechnology 17:241–249.

Wang H-C, Li K, Susko E, Roger AJ. 2008. A class frequency mixture model that adjusts for site-specific amino acid frequencies and improves inference of protein phylogeny. BMC evolutionary biology 8:331.

Warwick RM, Clarke KR. 1995. New “biodiversity” measures reveal a decrease in taxonomic distinctness with increasing stress. Marine Ecology Progress Series 129:301–305.

Webb CO, Ackerly DD, McPeek MA, Donoghue MJ. 2002. Phylogenies and Community Ecology. Annual Review of Ecology and Systematics 33:475–505.

Webb CO. 2000. Exploring the Phylogenetic Structure of Ecological Communities: An Example for Rain Forest Trees. The American Naturalist 156:145–155.

Weinreich DM, Watson RA, Chao L. 2005. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evolution; international journal of organic evolution 59:1165–1174.

Wheeler TJ, Kececioglu JD. 2007. Multiple alignment by aligning alignments. Bioinformatics 23:i559–68.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular biology and evolution 18:691–699.

Whelan S, Liò P, Goldman N. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends in genetics : TIG 17:262–272.

Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America 74:5088–5090.

Woese CR, Stackebrandt E, Macke TJ, Fox GE. 1985. A phylogenetic definition of the major eubacterial taxa. Systematic and applied microbiology 6:143–151.

Woese CR. 1987. Bacterial evolution. Microbiological reviews 51:221–271.

Wolf M, Müller T, Dandekar T, Pollack JD. 2004. Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. International journal of systematic and evolutionary microbiology 54:871–875.

Wolf YI, Aravind L, Grishin NV, Koonin EV. 1999. Evolution of Aminoacyl-tRNA Synthetases -- Analysis of Unique Domain Architectures and Phylogenetic Trees Reveals a Complex History of Horizontal Gene Transfer Events. Genome Research 9:689–710.

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. BMC evolutionary biology 1:8.

Wróbel B. 2008. Statistical measures of uncertainty for branches in phylogenetic trees inferred from molecular sequences by using model-based methods. Journal of applied genetics 49:49–67.

Wu M, Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome biology 9:R151.

Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular biology and evolution 10:1396–1401.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of molecular evolution 39:306–314.

Yang Z. 2006. Models of amino acid and codon substitution. In: Computational Molecular Evolution. New York: Oxford University Press. p. 40–70.

BIOGRAPHICAL SKETCH

Hye Won Lee earned a Bachelor of Science in Chemical Engineering and Biotechnology from Ajou University in South Korea. She earned a Master of Medical Science in Molecular Genomic Medicine from Seoul National University in South Korea. In 2007, she joined Genetics and Genomics Graduate Program of the University of Florida Genetics Institute.