## 1. Introduction

Early life on Earth has left a variety of traces that can be utilized to reconstruct the history of life, e.g., the fossil and geological records, and information retained in living organisms. This proposal focuses on how information can be gained from the molecular record, i.e. information about the history of life that is retained in structure and sequence of macromolecules found in extant organisms. The interpretation of the molecular record necessitates its calibration with respect to the geochemical and fossil records, and needs to consider and incorporate information about biochemical pathways and evolutionary theory.

Ever since Darwin [1], the tree of life has provided a framework to study the evolution of organisms. Early evolutionary biologists already had recognized that the macroscopic complex organisms had evolved from more simple single-celled ancestors early in Earth's history (e.g. [2]). With the introduction of ribosomal rRNA as a taxonomic marker [3, 4] it became possible to extend the natural classification of organisms (i.e., a taxonomy based on shared ancestry) to single-celled microscopic organisms (see the glossary for discussion and explanation of terms). Initially, it had been hoped that more sequence information would allow to more accurately pinpoint the location of organisms on an increasingly detailed tree of life (e.g., [5]); however, when more molecular sequences became available, it became clear that evolutionary mechanisms of prokaryotes are fundamentally different for those of multicellular animals (see [6] for a recent summary of the impact of gene transfer on microbial evolution). Single celled organisms frequently exchange genetic information across species boundaries (this seems to be true for pro- and eukaryotes [7-13]), thereby turning the tree of life into a web or net.

For some time it seemed that the concept of tree-like organismal history had failed at least with respect to microbial evolution [14, 15]. Different approaches to identify transferred genes appeared to reinforce each other in suggesting that horizontal gene transfer (HGT) even among divergent microorganisms was rampant throughout evolutionary history [6, 16, 17]. Comparisons of genomes from closely related organisms seemed to provide the final blow to fell the tree of life. For example, comparison of completely sequenced genomes of three different strains of *Escherichia coli* revealed that these genomes differ by 778 to 1860 genes from one another, and only have 39.2% of their combined set of genes in common [18]. However, several recent manuscripts appear to resurrect the organismal tree [19-23]. Concatenation of genes was shown to result in statistically well-supported phylogenies, and the resulting multigene phylogenies are similar to ribosomal RNA phylogenies. Analysis of genomes from closely related organisms indicate that most of the transferred genes persist in the recipient genome for only a short time [24]. These authors suggest that it is only a special group of genes, which they believe to be phage derived, that is transferred between organisms. While large number of genes continues to be exchanged between divergent single celled organisms, the emerging picture of genome evolution is not a completely random tangle, but rather a web into which major lines of descent are embedded. These major lines of descent are like ropes; individual genes, or group of genes collaborating in a physiological trait, represent fibers that leave one rope and join another, but the organismal line of descent represented by the rope undoubtedly has reality (*cf.* [25, 26]). The extent to which genome content derived phylogenies [27-33] reflect the organismal evolution remains an open question [25, 29, 32].

Even complex traits were transferred between divergent microorganisms, including those that characterize major groups of bacteria, and those which most dramatically have changed

Earth's ecology: e.g., the genes encoding the photosynthetic machinery were transferred between different bacterial phyla [34-36], and a photosynthetic gene cluster encoding the complete photosynthetic machinery was transferred between the alpha and the beta proteobacteria [37] and between different alpha proteobacteria (Lucas Mix, Harvard, pers. communication). While gene transfer makes it more difficult to trace characters on the tree of life [38], major biochemical innovations can nevertheless be traced relative to one another (e.g., respiration, different types of photosynthesis [39]) and it seems feasible that future research will be able to pinpoint the emergence of these traits within the organismal web/tree of life.

Comparative genome analyses have revealed genomes as mosaics where different parts have different histories. This is caused by the exchange of genes between organisms. However, gene transfer is not so rampant as to turn genomes into assemblies of randomly selected pieces. Rather, genes are usually exchanged between closely related organisms, and the exchange between distantly related organisms is rare. While the concept of a TREE of life might have to be abandoned in favor of a WEB of life, there is hope that the different parts of genomes, in particular metabolic pathways of geobiological interest can be traced though this web and can aid the reconstruction of Earth's early history. Some of the major collaterals already have been well documented and corroborated through analysis of the molecular data, e.g., the uptake of plastids into the eukaryotic host cell, which resulted in the evolution of eukaryotic algae and plants [40, 41]. Other events remain hypothetical, e.g., the alleged contribution of the cytoskeletal machinery to the eukaryotic cell from a now extinct lineage [42-44].

Here we propose to develop tools that will enable scientists to trace the history of different part of the cellular and metabolic machinery through time, thereby contributing to a better understanding of the early history of life on Earth.

## 2. Goal and significance

### 2.1. Goal and objective

**Goal**: Unravel the life's early history on Earth

Our analyses will focus on the molecular records, but their interpretation will occur in the context of morphological and chemical fossils and within the confines posed by the geological record.

**Objective:**

Develop and explore different methods of representation for multidimensional data matrices that contain information from genome-scale phylogenetic analyses.

We plan to use two types of data to represent phylogenetic information: bipartitions and quartets (see glossary), and we will explore the use of Self Organizing Maps (SOM[45]), Locally Linear Embedding (LLE[46]) and Principle Component Analysis (PCA[47, 48]) to reduce the dimensionality of the data. In addition, we will use modified Lento plots [49] for the analyses of bipartition data for comparison of moderately small number of genomes (see [50]).

The preference will be given to a method that satisfies all of the following:

a. Reduces number of dimensions to two to facilitate interpretation of the data

b. Highlights consensus relationship among the organisms
c. Identifies instances of significant deviation from a plurality consensus (putatively horizontally transferred genes)

## 2.2. Significance

The research proposed here aims to develop algorithms and tools that will allow dissecting the mosaic nature of genomes, to reconstruct the evolutionary history of individual traits in relation to other traits and to the plurality or majority consensus of genes. This will allow detection of co-evolving traits, and correlation of the molecular record with the fossil and geological records. The proposed approach to comparative genome analysis will be especially useful with respect to the early evolution of life and the evolution of metabolic pathways. The proposed work fits into the context of NASA's Origin theme *"Understand the origin and evolution of life on Earth"*, in particular, the addressed questions are central to NASA's Astrobiology program, in particular Goals 4 and 5 of the revised astrobiology roadmap *"Understand how past life on Earth interacted with its changing planetary and Solar System environment"* and "Understand the evolutionary mechanisms and environmental limits of life".

## 3. Details and Justification for the proposed approaches

### 3.1. Outline of data flow

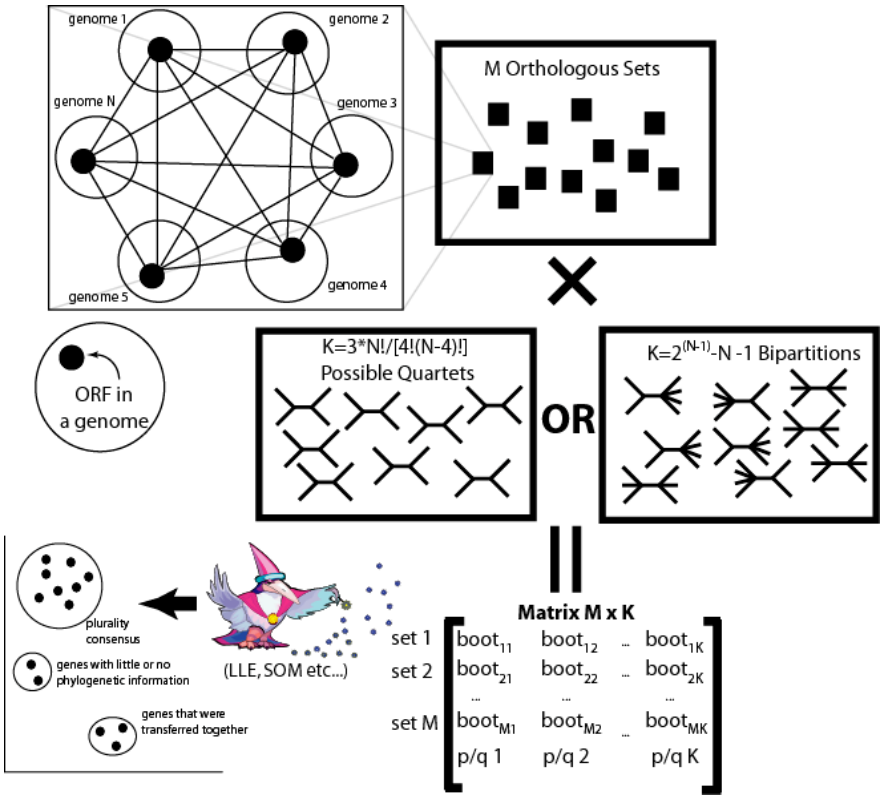



**Figure 1. Overview of the data flow.** This diagram illustrates the major steps in data analyses (see text for detailed description). Orthologous sets of amino acid sequences are depicted as black squares. The relationship between the genes within one orthologous set is shown in one example for 6 genomes: each genome is represented by one Open Reading Frame (ORF, black circle), and each gene picks up each other gene as the best hit in the BLAST search (represented as a line connecting two ORFs, see glossary for explanation of the best BLAST hit selection scheme). For each orthologous set either quartets or bipartitions are evaluated, and the data is compiled into a matrix. The matrix is further processed to produce a two dimensional representation of the multidimensional data matrix.

The overall data flow of the proposed analyses is depicted on the Figure 1. How do we compare *N* genomes among each other? We consider genes (more precisely, amino acid sequences coded by genes) in different genomes as collections of gene families, i.e. genes that are related to each other and shared a common ancestry in the past. To detect sets of orthologous genes (see glossary for the definition of orthology) we will utilize reciprocal best BLAST [51] hit relationships criterion (see glossary and section 3.2). Using this criterion our *N* genomes will be represented as *M* sets of orthologous genes for which we can perform phylogenetic analyses. If *N* is sufficiently large (*N*>=6), the number of possible tree topologies is high (see table 1), and it is computationally demanding to evaluate all possible tree topologies for each dataset. Instead, for each dataset all possible quartets and bipartitions will be evaluated (the evaluation criterion is the bootstrap support, the number of quartets/bipartitions is *K*). The data will be collected into a matrix (see section 3.3). The matrix is of very high dimension (*M* by *K*, where rows represent datasets, and columns give the bootstrap support values for the different partition or quartet). These data matrices will be analyzed with different methods that reduce dimensionality to a two-dimensional graphical representation (see section 3.4.). In these two-dimensional graphs gene families will be clustered into different groups: the plurality consensus group (genes that agree with each other on the relationships among genomes) will form one or several clusters, genes that strongly disagree with the plurality consensus will be located elsewhere and will be further examined for horizontal gene transfer. We also expect genes with no or little phylogenetic information to form a separate cluster. These diagrams will allow us to delineate the consensus signal present in the data, as well as to identify the potential horizontal gene transfer events and co-evolving genes.

### 3.2. Assembly of gene families

We will use double reciprocal BLAST [51] hits as criterion to assemble families of orthologous genes (see the glossary for definitions and explanations). This criterion requires that all members of a family pick each other as top scoring hits, when one is used as a query to search another genome. There is no recipe that guarantees the "correct" selection of orthologs. A commonly used approach is to use circular or reciprocal best BLAST hit relationships. For example, a circular BLAST hit scheme is employed by the Clusters of Orthologous Groups (COG) database [52]; it requires only unidirectional, circular best hit relationships for three of the reference genomes. The strict reciprocal best BLAST hit criterion employed, for example, by [53, 54] is more stringent, but not perfect. [55] reports an analysis of 353 quartets of orthologous genes assembled under the strict application of the reciprocal hit criterion. In only two instances was an unexpected phylogenetic relationship due to unrecognized paralogy. It is expected that the number of false positives will decline even further, when larger families of orthologous genes are assembled.

[19, 23] suggested that many of the claims for horizontal transfer might be due to the faulty selection of orthologous genes. Instead of a reciprocal hit criterion they used a single best-hit approach (non-reciprocal) but required as an additional criterion that no other hit above an arbitrary cut-off is present in the genome. This approach excludes all those families from analysis that underwent lineage specific gene duplications, and it also excludes ancient conserved paralogs. We repeated the analyses performed by Lerat and colleagues [56] using both ortholog selection schemes [25]. Under the reciprocal best hit criterion 54 additional gene families were assembled, but none of these additional families showed any indication of including paralogous sequences (i.e. their phylogenies agreed with the plurality consensus to the

same extent as the 207 gene families assembled under the alternative criterion proposed by [19, 23]. Clearly, the results observed by Lerat *et al*. are not due to an improved ortholog selection criterion, but rather due to the particular selection of genomes (probably to the inclusion of small genomes from symbiotic bacteria). Gene families assembled under either criterion result in the same conclusion. The reciprocal hit criterion appears to be slightly superior, since it detected more orthologs missed by the other scheme.

Both selection schemes err on the side of being overly restrictive and both produce a high number of false negatives (i.e., orthologs that are not detected and therefore excluded from the analyses). For example, genes that underwent lineage-specific amplification have a high chance of being excluded under both schemes, even though they are valid orthologs [57] and should be included. It is therefore important to consider approaches for the further analyses that are accommodating to missing data (see section 3.3.3. below).

Multiple sequence alignments will be calculated for each family of orthologs for the amino acid sequences. We have developed a program that compares two extreme paths through a multiple sequence alignment and removes ambiguously aligned positions form the alignment (JF Gogarten, O Zhaxybayeva, JP Gogarten, manuscript in preparation).

## 3.3. The data matrices

### 3.3.1. Bootstrap – the support measure of choice

To capture phylogenetic information contained in molecular data, it is important to include measures of statistical reliability. We have explored different measures in the past [50, 54, 58]: bootstrap support values, Bayesian posterior probabilities estimated either through Metropolis-coupled Markov Chain Monto Carlo (MCMCMC) exploration of tree space [59], or through empirical estimates [60]. The work described here we will use bootstrap support values for the following reasons:

- Both approaches to estimate posterior probabilities tend to over-estimate the reliability of phylogenies given a set of aligned sequences [54, 61-63], which can lead to spurious conflicting signals [50, 55].
- To avoid problems caused by limited taxon sampling [64-68], we will routinely add homologous sequences from the NCBI's non-redundant database. The resulting data sets often will have many more than 50 sequences. Under these conditions the currently available MCMCMC approaches frequently become stuck on local minima yielding unreliable support value estimates (unpublished data and Lucas Mix, pers. comm.).
- TREE-PUZZLE [69] provides a reasonable fast way to obtain maximum likelihood estimates for evolutionary parameters, allowing for the use of models that incorporate among site rate variation and complex substitution matrices. Using a non-parametric sampling of the multiple sequence alignments multiple distance matrices can be calculated and evaluated using neighbor joining[70, 71]. Especially in conjunction with the evaluation of embedded subtrees [55] this approach yields useful support measures.

### 3.3.2. Trees, Bipartitions or Quartets

***Trees:***

Initially, we and others (e.g.: [50, 56]) used completely resolved trees to capture phylogenetic information retained in molecular sequences. However, the number of trees increases dramatically with increasing number of species (with *n* taxa there are $(2n\text{-}5)!/[2^{(n-3)}(n\text{-}3)!]$ different unrooted tree topologies, *cf*. table 1). Furthermore, individual gene families are not likely to support a single completely resolved tree topology over its alternatives. A single unresolved internal branch already spreads the support evenly over three alternative topologies. While his approach is useful for the analysis of only a few genomes at a time (note: we solved the taxon sampling problem usually associated with analyzing a limited number of taxa only[58]), the described problems prevent the application of this approach to a larger number of genomes.

***Bipartitions:***

Provide a workable alternative for the comparative analysis of more than 5 genomes. The number of possible bipartitions grows much slower with an increasing number of genomes than the number of different trees (Number of internal bipartitions with *n* genomes $=2^{(n-1)} -n-1$). Even if the data cannot decide between some alternative bipartitions, other bipartitions might be strongly supported, and this strong support can be captured in data matrices based on bipartitions. For example, in the evolution of 13 genomes from selected gamma proteobacteria [56] the majority of gene families are non-conflicting, and the consensus signal can be captured analyzing bipartitions [50]. Furthermore, two bipartitions can be either compatible with one another (i.e. they might correspond to different branches in the same tree), or they can be incompatible (i.e. the same group is partitioned in different ways). Plotting the number of gene families that support and conflict with a bipartition is an excellent way to detect a week plurality phylogenetic signal [49, 50], and to detect those gene families that are in conflict with the consensus phylogeny assembled from the compatible consensus bipartitions (see [25] for examples).

However, in addition to these advantages, bipartitions also have limitations that restrict their usefulness:

- A single rogue-sequence reduces the support for many or all bipartitions. Assume that a gene family of six members supports a single completely resolved tree (genome 1-6 in Fig.2). If a seventh sequence is added to the alignment, that is not attracted to any of the other sequences in particular, the support for all the other bipartitions will decrease dramatically. If a single sequence moves from one end of a comb like phylogeny to the other, not a single bipartition is conserved between the two trees.
- It is difficult at best, to develop a scoring scheme that allows including gene families that do not have representatives in all genomes (see below). This greatly reduces the number of gene families that can be analyzed for a large number of genomes.
- The more sequences are included in an analysis, the shorter each individual branch becomes. Adding more sequences, especially if the additional sequences break up long branches, improves the reliability of the overall phylogeny [66, 72, 73]. However, the shorter a branch, the fewer substitution events occur along this branch, and the lower its bootstrap support becomes.

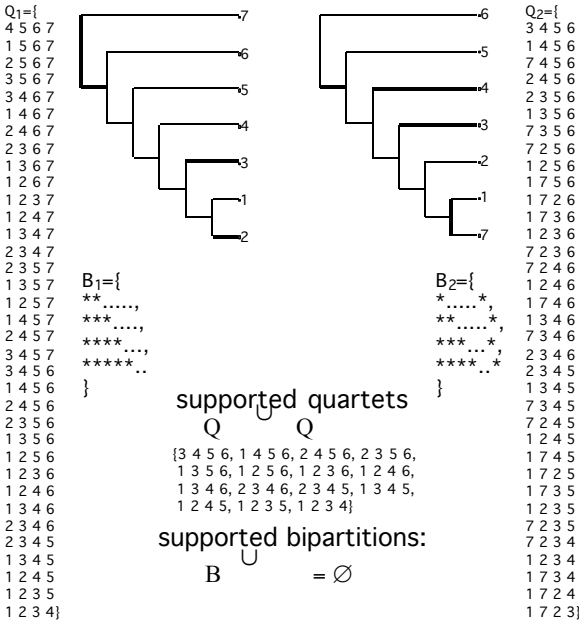



**Figure 2. Illustration of the topology case where quartet analyses are more useful than bipartition analyses.** Here we illustrate this special type of topology (so-called "hennigian comb") for tree with 7 taxa (named 1 through 7). The two trees differ only in position of one taxon (#7) and both trees should be considered unrooted. This hypothetical example can represent a real case where a gene from taxon #1 is horizontally transferred and replaced its homolog in taxon #7. All relationship among other taxa remain the same. Sets $Q_1$ and $Q_2$ list supported quartets by the trees on the left and on the right respectively. (Quartets are abbreviated as lists of 4 taxa (e.g., "4567") which stands for first two taxa grouping together (i.e., in the quartet "4567" taxa 4 and 5 go together, and 6 and 7 go together). Sets $B_1$ and $B_2$ list supported bipartitions for each of the trees (bipartitions are represented as a row of symbols, one symbol per taxon (either "." or "*"), dividing the dataset into two parts, see glossary for the definition of the bipartition). Intersection of the bipartition sets $B_1$ and $B_2$ is an empty set, i.e. there is no single bipartition shared between these two trees. In contrast, intersection of the sets $Q_1$ and $Q_2$ is not empty, and the similarities between two trees are captured in the quartet analyses. This demonstrates that in some cases quartets can retain more information than the bipartitions.

### *Quartets*

Quartets are formed from the analyses of four sequences at a time. For each quartet, three alternative tree topologies are possible (we consider the trees obtained from molecular data as unrooted). A multi-taxon tree contains many embedded quartet subtrees (see table 1 comparing number of trees, bipartitions and quartets for *N* genomes). For each quartet the support values for the three alternative topologies add up to 1 (or 100%). One problem posed by the analyses of data matrices based on quartets is that the data matrices are no longer sparse. Each of the possible quartets will have an associated support value triplet. And many of the quartets will have strong levels of support for one topology over the others. Note that we will NOT calculate the bootstrap support values from only four sequences at a time, rather we will proceed as detailed in [55]: we bootstrap the multiple sequence alignment, calculate multiple sequence phylogenies for each of the bootstrap samples, and then we parse each of the resulting phylogenies for the topology of the subtree containing the four sequences that are to be analyzed.

An advantage of using quartets is that the analysis is less affected by rogue sequences. Fig. 2 provides an example. The similarity between the two trees depicted in Fig. 2 is captured in the shared quartets. Considering the frequency with which a given topology for a quartet of four sequences is recovered in all bootstrap samples: additional sequences will increase the confidence with which the correct topology is recovered; however, on average the additional sequences will not reduce the support for one quartet topology over the two alternatives [55].

This feature suggests that quartets will be particular useful for the comparative analyses of a large numbers of genomes.

Table 1: Comparison of the number of unrooted tree topologies, internal bipartitions and quartets for different numbers of genomes.

| Number of genomes | Number of unrooted trees | Number of bipartitions | Number of quartets* |
|---|---|---|---|
| 4 | 3 | 3 | 1 |
| 6 | 105 | 25 | 15 |
| 8 | 10,395 | 119 | 70 |
| 10 | 2,027,025 | 501 | 210 |
| 13 | 1.37E+10 | 4,082 | 715 |
| 20 | 2.22E+20 | 5.24E+05 | 4,845 |
| 50 | 2.84E+74 | 5.63E+14 | 230,300 |
| 80 | 2.18E+137 | 6.04E+23 | 1,581,580 |

*: Each quartet has three possible alternative tree topologies

### 3.3.3. Comparison of datasets with varying number of taxa

In many instances the number of orthologous datasets for *N* genomes will be dramatically reduced if there is a small or very divergent genome included in the analyses. As the number of genomes increases, the number of orthologous genes tends to decrease (e.g., see [74] for review). It is therefore desirable to relax the selection scheme to allow datasets that do not have representative orthologs from all genomes. If the distribution of the gene is restricted to only a subset of genomes under consideration, it might nevertheless be important to consider the history of these genes, for example genes involved in photosynthesis will be absent in most genomes of non-photosynthetic organisms that one might want to use as an outgroup for an analysis of a group of photosynthetic bacteria. In case of a data matrix based on quartets, missing data are easily accommodated. We will exclude genomes one by one and re-calculate the orthologous sets for smaller and smaller subsets of genomes. Each dataset will be assigned a "dataset order" – a number that depends on how many of *N* genomes were included in ortholog selection scheme (see figure 3). When the genes from higher order sets will be compiled into the matrix, there will be missing data due to absence of some quartets. Interpreting the support values in a Bayesian context, the support value vectors for the missing quartets will be assigned to 1/3, 1/3 and 1/3 for the three possible tree topologies, reflecting an equal prior probability for each topology that is not modified by any data.

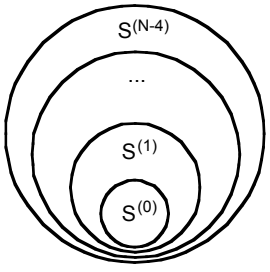



**Figure 3. Venn diagram of sets of orthologs that allow missing data.** Set $S^0$ allows no missing data, $S^1$ allows one genome representative sequence missing in the dataset, …, $S^{N-4}$ allows all but four sequences missing. The sets $S^0$, $S^1$ ,…, $S^{N-3}$ are proper subsets of $S^{N-4}$. We refer to the superscript of a set as "dataset order". Keeping track of the dataset orders will allow us to adjust how much missing data we will allow in analyses, as well as help us to re-use pre-calculated quartets from the database.

To summarize the advantages of quartets over bipartitions:

1) There are more compatible quartets between any two datasets than partitions, in particular quartets are more robust to the inclusion of divergent sequences that retain little or no phylogenetic information (see Fig. 2).
2) Easy to treat missing data
3) The quartet support values do not decrease as more sequences are added to the analyses
4) Quartets appear to represent information quanta of phylogenetic information

### 3.3.4. Storage of quartets data facilitate the analyses of genomes

If all quartets are analyzed as extended datasets, which are formed through addition of homologs from either a large number of reference genomes, or from the non-redundant data base at NCBI, then the results of already performed phylogenetic analyses can be stored in a database, and utilized in consecutive analysis that use sets of genomes that were already previously analyzed.

The information about the evaluated quartets will be stored in a database to speed up the data compilation for consecutive analyses. The information will be placed into one table. Fields of the table are described in Table 2.

Table 2. Proposed fields for table in the database of quartets

| Table Quartets | Description of the table fields |
|---|---|
| Quartet_ID | Unique ID number assigned to each quartet |
| GI1 | GenBank Identification number of the 1st member of the quartet |
| GI2 | GenBank Identification number of the 2nd member of the quartet |
| GI3 | GenBank Identification number of the 3rd member of the quartet |
| GI4 | GenBank Identification number of the 4th member of the quartet |
| Name1 | Name of the genome of the 1st member of the quartet |
| Name2 | Name of the genome of the 2nd member of the quartet |
| Name3 | Name of the genome of the 3rd member of the quartet |
| Name4 | Name of the genome of the 4th member of the quartet |
| Boot1 | Bootstrap support value for topology ((1,2),3,4) |
| Boot2 | Bootstrap support value for topology ((1,3),2,4) |
| Boot3 | Bootstrap support value for topology ((1,4),3,2) |
| P_description | Description of the protein function |
| COG | Functional category of the protein (following COG database notations) |
| order | Dataset order |

## 3.4. Algorithms to analyze the data matrices

We will explore different tools to analyze the data matrices (see section 3.3). The tools considered for this research map the high dimensional data space represented by these matrices into a lower dimensional space while preserving the phylogenetically useful information present in the data. Of the many different tools available today for dimension reduction we will particularly concentrate on two.

First we will consider the self-organizing map (SOM) algorithm [45]. This neural network-based algorithm attempts to detect the essential structure of the input data based on the

similarity between the points in the high dimensional space. The points (each representing a gene family in our case) in high-dimensional space are embedded in a two dimensional map in such a way that points similar to each other appear to be close together on the two dimensional map, points which are dissimilar are far apart. The geometric interpretation being that points that are similar to each other tend to be clustered in the high-dimensional space and SOM preserves this clustering on the 2-dimensional map. Next, we will consider the locally linear embedding algorithm (LLE) [46], which maps each point in the high dimensional space into a point in two-dimensional space via an encoding based on a covariance matrix of the distances to a selected set of nearest neighbors. This tends to preserve local neighborhood structures. In general, this algorithm performs well in providing low dimensional projections of high dimensional data retaining the natural structure of the original data. However, the interpretation that similar points are clustered together is not as straightforward as in the case of SOM, since the algorithm also takes the local geometric structure of the high dimensional data space into account. The LLE algorithm itself is eigenvector-based and computes an optimal embedding of the high-dimensional data in the low dimensional projection. Therefore, this algorithm does not suffer from the potential local minimum problem of the SOM algorithm. On the other hand, the straightforward interpretation of the maps provided by SOM makes SOM an ideal exploratory tool during the initial phases of our research. Both algorithms map high dimensional data into a single global coordinate system of lower dimension (typically of dimensionality two) and are considered to be non-linear dimension reduction schemes.

In addition to the above approaches we will also investigate the following: principal component analysis (PCA) [47, 48], multidimensional scaling (MDS) [75], support vector based kernel PCA [76], and ISOMAP [77] to produce informative two-dimensional maps depicting phylogenetic relationships among genomes. Some of these algorithms are considered linear dimension reduction schemes, in particular PCA and MDS. They are well-understood and studied schemes and it is important to include them even if only to underscore the fact that the relationships among the gene histories are indeed non-linear. We will develop tools to make gene types and tree topologies readily recognizable in each map.

Criterion for successful mapping is our aim to produce diagrams that depict gene families with similar histories as neighboring, whereas genes with different histories as falling into separate clusters. We will explore the utility of the different approaches using well-studied test cases: the eukaryotic genome where currently three large groups of genes are recognized based on their different evolutionary history (those from the host genome and those from the endosymbionts that evolved into mitochondria and plastid), and selected bacterial genomes where genes involved in the same function of metabolic pathway often have the same evolutionary history.

### 3.4.1. Supertrees versus Maps

Supertrees [78, 79] currently are the *dernier crie* of the "Tree of Life" community, comprised mainly of researchers funded through NSF's Tree of Life initiative. In contrast, the emphasis of the proposed research on maps and plots rather than trees or supertrees might seem quaint and a little old fashioned. We prefer maps for the following reasons: to-date the genome-wide analyses indicate the opposite: the evolution of genomes is not tree-like. One can easily take genome scale data and calculate trees that reflect the observed levels of similarity, but the fact that one can obtain a tree should not be confused the demonstration that the data were

generated in a process that is best described by a tree. At present, every indication from the data is that genome evolution is not a steadily furcating process, especially among single celled microorganisms. Supertrees might be more appropriate for multicellular eukaryotes, which appear to have become the main focus of NSF's Tree of Life initiative.

If one so desires, one can easily calculate supertrees from the unprocessed data matrices that we will have compiled. In case of the bipartition matrices, one can combine the compatible plurality bipartitions. The gene families supporting conflicting bipartitions are immediately available to illustrate the "non-tree" component of genome evolution. In case of data matrices based on quartets (which is more interesting because they accommodate missing data) one can use the quartet puzzling [80] to repeatedly build trees from randomly selected quartets, taking into consideration the support the that the different topologies receive through the data. Repeated assembly of complete trees from randomly selected quartets will give rise to trees that are slightly different from one another. These differences can be utilized calculate quartet puzzling support values [80] for individual branches.

However, the conversion of quartets or bipartitions into trees will be more illuminating, if applied to those gene families that map close to each other in the obtained maps, and that therefore are assumed to share evolutionary history.

### 3.4.2. Development of interactive tools to explore the calculated maps

The produced two-dimensional maps contain layers of information. Interpreting these maps requires summarizing the phylogenetic signal of families clustered together and the extraction of features that distinguish between different clusters. We plan to develop an interactive tool to automate these tasks.

**Proposed interface features for the map exploration program:**
Once the map is generated, user will have an interactive exploration tool that aids the interpretation of the map. The tool will have the following features:

- Identification of genes in a query cluster (those genes referred from now on as "cluster members"). The genes will be identified by name and also could be further queried for GI numbers and functional categories (according to COG database)
- Identify bipartitions/quartets supported by each member of the cluster
- Identify bipartitions/quartets supported by all members of the cluster
- Identify bipartitions/quartets that separate a query cluster from neighboring clusters
- Combine bipartitions/quartets from plurality clusters into a plurality consensus diagram
- Highlight clusters that contain only one member and are well-separated from other clusters
- Transform the scoring scale or introduce thresholds below which support values are set to zero, followed by recalculation of the map
- Interactively merge clusters (i.e. place constrains and recalculate the map)

## 3.5. Preliminary results: A case study of thirteen gamma-proteobacterial genomes

Thirteen completely sequenced gamma-proteobacterial genomes were recently analyzed by [56] and [25] (see also section 3.2.). This group of bacteria provides an excellent example to test approaches to detect genes deviating from the majority consensus. The analyses in [56] and [25] revealed that these genomes show a strong plurality phylogenetic signal (fig. 4). The lack of

obvious transfer is probably due to many of these bacteria living in close association with multicellular eukaryotes, and due to the inclusion of highly reduced genomes. Of the 252 families analyzed only three families strongly contradict with the consensus (fig. 4).

An analysis of significantly supported bipartitions is depicted in fig. 4. Eight bipartitions are supported with more than 99% bootstrap support by more than 100 gene families each. In addition to this strong majority signal, there are hardly any strongly supported bipartitions that are incompatible with these majority bipartitions. This analysis illustrates the power of bipartitions analysis to extract a phylogenetic signal shared by many gene families, and it illustrates that a plot of conflicting bipartitions can reliably pinpoint gene families whose members underwent gene transfer between different lineages. We analyzed a data matrix listing strong bootstrap support (>99%) for all bipartitions for the 252 gene families that had a representative in each of the thirteen genomes. When the gene families were clustered in bipartition space using SOM (fig. 5) the three genes that had been identified as having undergone horizontal gene transfer grouped separate from the other gene families. The major clusters formed correspond to genes that share the majority bipartitions to different extent (i.e. one cluster contains only gene families that share all of these bipartitions, one contains gene families that do not significantly support only one of the majority bipartitions, etc.). This example illustrates that SOM can be used to map gene families in a way that reveals gene families with atypical phylogenetic information content.

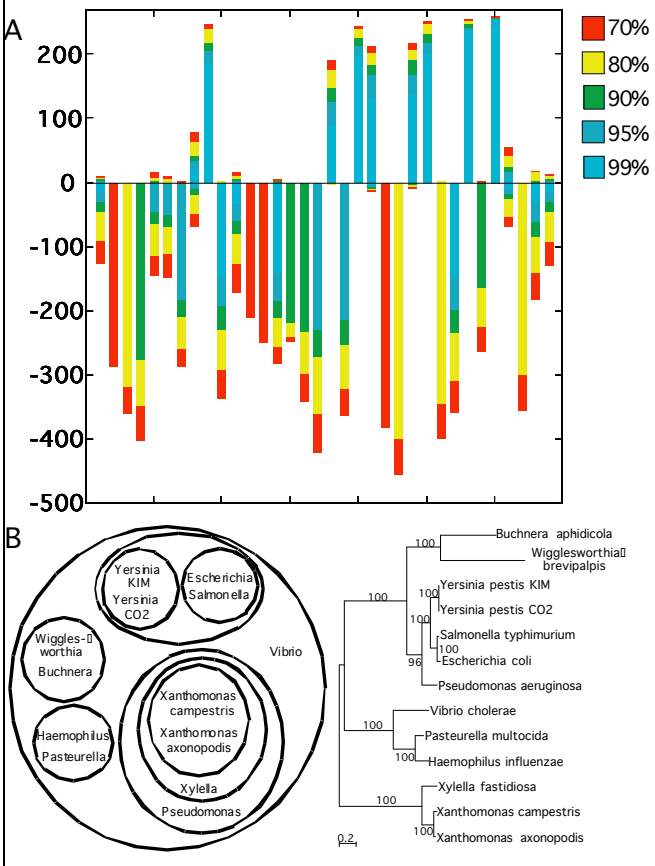



**Figure 4. Analyses of bipartitions for gamma-proteobacteria.** A modified Lento-plot (after [49]) for partitions with at least 70% bootstrap support is depicted in panel A. Each column represents the number of datasets that support or conflict (downwards pointing columns) a bipartition. The level of support is color-coded. The support for partitions was extracted from the output of the CONSENSE program [71] that was ran on 100 Neighbor Joining trees calculated using NEIGHBOR [71] from TREE-PUZZLE [69] distances determined from 100 bootstrap samples. The partitions supported by the majority of the datasets are non-conflicting with one another, and therefore can be combined into nested consensus clusters depicted in the panel B. Only three datasets strongly conflict with these bipartitions, one of which, virulence factor homologs (mviN), is depicted in panel C. This topology is not among the 13 topologies analyzed in [56]. The tree was calculated using NEIGHBOR [71] from distances calculated with TREE-PUZZLE v.5.1 [69] under the auto-detected substitution model with among site rate variation taken into account (estimated alpha =0.94). The support values are bootstrap support values from 100 bootstrap samples.

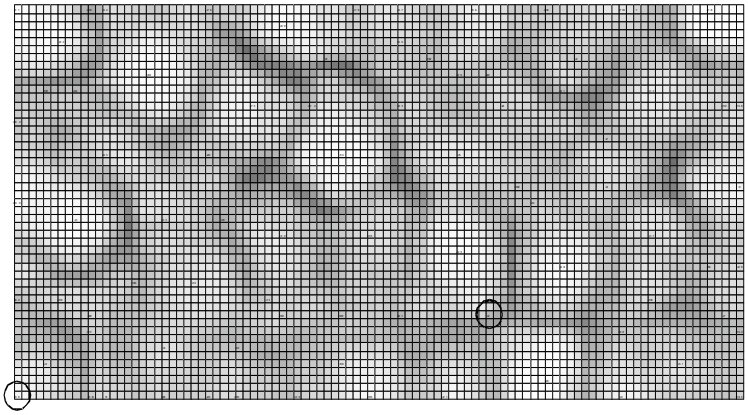


**Figure 5. SOM map of genes in bipartition space for the test case of thirteen gamma proteobacteria.** In this map white areas indicate groups of gene families whose phylogenetic signal is distinctly different from the surrounding genes. Genes that are in conflict with plurality bipartitions (encircled) group separately from the clusters containing genes conforming to the to plurality phylogeny. Proposed exploration tools will allow interactive extraction of information from the map.

### 3.6. Timeline and addressed questions

**1st year:**

- Implement different algorithms to generate maps.
- Using gamma proteo- and cyanobacterial genomes [25] as test cases tune the parameters in the different methods to optimize visualization. For example the following questions will be addressed: What is the effect of applying cut-offs for support values? Does a transformation of the support value scale (compressing the lower end of the scale) improve resolution?
- Using the above-mentioned test cases, explore advantages of quartets *vs.* bipartitions.
- Determine the best way to encode the support values for the three alternative topologies for each quartet. The support for the three alternative topologies always adds up to one. Therefore, it should be sufficient to enter only two of the three values into the data matrix, possibly encoded as complex numbers, where the real and the imaginary part represent support for two of the three alternatives.
- Explore the maximum number of genomes that feasibly can be analyzed at one time using the different mapping approaches.
- Begin implementation of the quartet database and develop scripts that can utilize the quartet database to shorten computation times.

**2nd Year:**

- Gradually fill the quartet database with data.
- Design the program package for the future development by a graduate student:
    - Design class diagram to show the relationships between different classes of the program, as well as specify which methods and attributes belong in which class.
    - Define an ADT (Abstract Data Type) for each class
    - Generate interface documentation.
- Explore the possibility to include higher order characteristics (e.g. insertion/deletion in the individual sequences; structural features of proteins like the length of surface loops or domain fusions; morphological, metabolic and ecological characteristics of the different

species). The combined data matrices might be useful to study the co-evolution of traits and they might add resolution to the inference of organismal lineages.

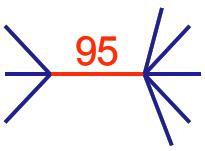
- Follow the evolution of operons, metabolic pathways and physiological characteristics in the maps and through the inferred putative organismal history. The following are examples for the questions that will be explored: In which biological context did sulfate reduction emerge [29, 81, 82]? How did the enzymes evolve that catalyze (and utilize) iron oxidation in reducing environments as electron donors in photosynthesis [83]? Could these enzymes be responsible for the iron deposition in the banded iron formation? Can the enzymes that protect against superoxide stress [84], or enzymes like squalene epoxidase that use molecular oxygen [85] be used to relate the molecular record to the origin of oxygen producing photosynthesis and the rise of atmospheric oxygen levels?

**3rd Year:**

- Develop a stand-alone user-friendly program that would perform the analyses (this task will be performed by a graduate student enrolled in the M.S. program of the Computer Science and Statistics Department at the University of Rhode Island). The code will be released to the research community under GNU Public License at the end of the year 3.
- Continue to apply the program to analyze the available completely sequenced genomes.
- Make inferences about early evolution from the analyses of genomes. Analyze genes inside the obtained clusters and tie these cluster specific histories to geological and fossil records.
- Prepare manuscripts for publication that report on the developed tools and techniques, and on the findings obtained using these tools.

### 3.8. Glossary:

- **Among Site Rate Variation (ASRV) –** Variation of the substitution rate in different parts of sequence (often due to functional constraints on the protein). Usually ASRV is approximated by "sorting" all sites in the dataset into several categories according to their rates. A commonly used distribution describing ASRV is a discrete approximation of the Gamma distribution [86].
- **Bipartition –** A division of a phylogenetic tree into two parts that are connected by a single internal branch. It divides a dataset into two groups, but it does not consider the relationships within each of the two groups. An example of a bipartition is shown in the figure. In our analyses we define the support for a bipartition (95 in this particular example) as the bootstrap support of the internal branch. The number of all possible bipartitions for $N$ genomes is equal to $(2^{(N-1)}-N-1)$.




- **Bootstrap Support** (in the context of this proposal) – A statistical measure to assess the significance of the branch as obtain from analyses of one dataset. Positions from the sequence alignment are re-sampled with replacement to generate pseudosamples. Each pseudosample is analyzed and the percentage of pseudosamples supporting the same branch constitutes bootstrap support of this branch [87].
- **Organismal Lineage –** Can be defined as the majority consensus of genes passed on over very short time intervals. Provided the time intervals are sufficiently short, this definition only fails in the rare event of two organisms making co-equal contributions to a new line of

descent. Gary Olsen (University of Illinois, Urbana-Champaign) used the metaphor of a rope to illustrate this concept. Not a single cellulose fiber (representing the genes) might persist throughout a rope (representing the organismal lineage) from beginning to end; nevertheless, the rope has continuity.

- **Ortholog selection schemes:** Let $A_1$, $A_2$, … , $A_n$ denote a set of orthologous genes from *n* genomes (one gene $A_i$ from each of *n* genomes). Let $A_i$->$A_j$ denote the best BLAST hit relationship among two genes $A_i$ and $A_j$, where gene $A_j$ from the genome *j* is the best hit in the BLAST search of gene $A_i$ against genome j. In **the circular BLAST hit ortholog selection scheme** for each set of selected putative orthologous genes $A_1$, $A_2$, …, $A_n$ members are connected through a "circular" unidirectional best BLAST hit relationships , i.e.: $A_1$ -> $A_2$ -> …->$A_n$->$A_1$. In **the reciprocal BLAST hit ortholog selection scheme** a set of genes is considered orthologous if and only if all genes in the set pick each other as a top BLAST hit, i.e. $A_1$<->$A_2$, $A_1$<->$A_3$,…, $A_1$<->$A_n$, $A_2$<->$A_3$,…,$A_2$<->$A_n$,…, $A_{n-1}$<->$A_n$. This selection scheme is more stringent than circular BLAST hit ortholog selection scheme. In **a single BLAST hit ortholog selection scheme** genes from one reference genome are used to search all other genomes, and top-scoring BLAST hits (above a preset cutoff) are merged into a dataset. Additional criteria need to be applied to eliminate paralogs. For example, one can exclude datasets that have more than one representative gene per genome.
- **Orthologs –** Genes in different species that are related to one another by speciation events.
- **Paralogs -** Genes in different or the same species that are related by a gene duplication event. Especially in conjunction with gene losses, paralogs can be mistaken for orthologs.
- **Quartet –** In this context, a quartet describes the phylogenetic relations between 4 sequences. If no molecular clock is assumed, the possible relationships are described by the topologies of three unrooted trees: ((1,2),(3,4)); ((1,3),(2,4)); ((1,4),(2,3)). Using either posterior probabilities or bootstrap support values the support for the three alternatives add up to 1. The quartet might be regarded as the smallest quantum of phylogenetic information
- **Unrooted Tree** – a tree that only specifies relationships between the taxa (determined by the tree topology), but does not make any assumptions about ancestors and descendants (i.e., the tree does not have direction in time). Usually trees calculated from molecular data are unrooted. Placing the root requires additional information[8].

## References


1. Darwin, C. (1859). On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. London: John Murray, Albemarle Street.
2. Haeckel, E. (1866). Generelle Morphologie der Organismen: Allgemeine Grundzüge der organischen Formen-Wissenschaft mechanisch begründet durch die von Charles Darvin reformierte Descendenz-Theorie. Berlin: Georg Riemer.
3. Woese, C. R., Fox, G. E. (1977). The concept of cellular evolution. J Mol Evol *10*, 1-6.
4. Woese, C. R., Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A *74*, 5088-5090.
5. Woese, C. R. (1987). Bacterial evolution. Microbiol Rev *51*, 221-271.
6. Gogarten, J. P., Doolittle, W. F., Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. Mol Biol Evol *19*, 2226-2238.
7. Gogarten, J. P. (1995). The early evolution of cellular life. Trends in Ecology and Evolution *10*, 147-151.
8. Gogarten, J. P., Taiz, L. (1992). Evolution of proton pumping ATPases: Rooting the tree of life. Photosynthesis Research *33*, 137-146.
9. Gogarten, J. P. (2003). Gene transfer: gene swapping craze reaches eukaryotes. Curr Biol *13*, R53-54.
10. Doolittle, W. F. (1998). You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet *14*, 307-311.
11. Bapteste, E., Moreira, D., Philippe, H. (2003). Rampant horizontal gene transfer and phospho-donor change in the evolution of the phosphofructokinase. Gene *318*, 185-191.
12. Koonin, E., Makarova, K., Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol *55*, 709 - 742.
13. Koonin, E. V. (2003). Horizontal gene transfer: the path to maturity. Mol Microbiol *50*, 725-727.
14. Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. Science *284*, 2124-2129.
15. Hilario, E., Gogarten, J. P. (1993). Horizontal transfer of ATPase genes--the tree of life becomes a net of life. Biosystems *31*, 111-119.
16. Lawrence, J. G., Hendrickson, H. (2003). Lateral gene transfer: when will adolescence end? Mol Microbiol *50*, 739-749.
17. Ragan, M. A. (2002). Reconciling the many faces of lateral gene transfer. Trends in Microbiology *10*, 4.
18. Welch, R. A., Burland, V., Plunkett, G., 3rd, Redford, P., Roesch, P., Rasko, D., Buckles, E. L.*, et al.* (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci U S A *99*, 17020-17024.

19. Lerat, E., Daubin, V., Moran, N. A. (2003). From Gene Trees to Organismal Phylogeny in Prokaryotes:The Case of the gamma-Proteobacteria. PLoS Biol *1*, 19.

20. Brown, J. R., Italia, M. J., Douady, C., Stanhope, M. J.: Horizontal Gene Transfer and the Universal Tree of Life. In: *Horizontal Gene Transfer* Edited by Syvanen M, Kado CI, 2nd ed. pp. 305-349: Academic Press; 2002: 305-349.

21. Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. Nat Genet *28*, 281-285.

22. Brochier, C., Bapteste, E., Moreira, D., Philippe, H. (2002). Eubacterial phylogeny based on translational apparatus proteins. Trends in Genetics *18*, 1-5.

23. Daubin, V., Moran, N. A., Ochman, H. (2003). Phylogenetics and the cohesion of bacterial genomes. Science *301*, 829-832.

24. Daubin, V., Lerat, E., Perriere, G. (2003). The source of laterally transferred genes in bacterial genomes. Genome Biol *4*, R57.

25. Zhaxybayeva, O., Lapierre, P., Gogarten, J. P. (2004). Genome mosaicism and organismal lineages. Trends in Genetics *in press*.

26. Martin, W. (1999). Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. Bioessays *21*, 99-104.

27. Tekaia, F., Lazcano, A., Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. Genome Res *9*, 550-557.

28. Fitz-Gibbon, S. T., House, C. H. (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res *27*, 4218-4222.

29. House, C. H., Runnegar, B., Fitz-Gibbon, S. T. (2003). Geobiological analysis using whole genome-based tree building applied to the Bacteria, Archaea, and Eukarya. Geobiology *1*, 15-26.

30. Snel, B., Bork, P., Huynen, M. A. (1999). Genome phylogeny based on gene content. Nat Genet *21*, 108-110.

31. Korbel, J. O., Snel, B., Huynen, M. A., Bork, P. (2002). SHOT: a web server for the construction of genome phylogenies. Trends in Genetics *18*, 158-162.

32. Clarke, G. D. P., Beiko, R. G., Ragan, M. A., Charlebois, R. L. (2002). Inferring Genome Trees by Using a Filter To Eliminate Phylogenetically Discordant Sequences and a Distance Matrix Based on Mean Normalized BLASTP Scores. J. Bacteriol. *184*, 2072-2080.

33. Lin, J., Gerstein, M. (2000). Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res *10*, 808-818.

34. Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y., Blankenship, R. E. (2002). Whole-genome analysis of photosynthetic prokaryotes. Science *298*, 1616-1620.

35. Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Blankenship, R. E. (2003). Evolution of photosynthetic prokaryotes: a maximum-likelihood mapping approach. Philos Trans R Soc Lond B Biol Sci *358*, 223-230.

36. Xiong, J., Bauer, C. E. (2002). Complex evolution of photosynthesis. Annu Rev Plant Biol *53*, 503-521.

37. Igarashi, N., Harada, J., Nagashima, S., Matsuura, K., Shimada, K., Nagashima, K. V. (2001). Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. J Mol Evol *52*, 333-341.

38. Zhaxybayeva, O., Gogarten, J. P. (2004). Cladogenesis, Coalescence and the Evolution of the Three Domains of Life. Trends in Genetics *20*, 182-187.

39. Xiong, J., Bauer, C. E. (2002). A cytochrome b origin of photosynthetic reaction centers: an evolutionary link between respiration and photosynthesis. J Mol Biol *322*, 1025-1037.

40. Margulis, L. (1995). Symbiosis in Cell Evolution: Microbial Communities in the Archean and Proterozoic Eons, 2nd edn: W H Freeman & Co.

41. Palmer, J. D. (2003). The symbiotic birth and spread of plastids: How many times and whodunit? Journal of Phycology *39*, 4–11.

42. Sogin, M. L. (1991). Early evolution and the origin of eukaryotes. Curr Opin Genet Dev *1*, 457-463.

43. Hartman, H., Fedorov, A. (2002). The origin of the eukaryotic cell: a genomic investigation. Proc Natl Acad Sci U S A *99*, 1420-1425.

44. Hartman, H. (1984). The origin of the eukaryotic cell. Speculations Sci Technol *7*, 77-81.

45. Kohonen, T. (2001). Self-organizing maps, 3rd edn. Berlin ; New York: Springer.

46. Roweis, S. T., Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science *290*, 2323-2326.

47. Jolliffe, I. T. (2002). Principal component analysis, 2nd edn. New York: Springer-Verlag.

48. Pearson, K. (1901). On lines and planes of closest fit to points in space. Philosophical Magazine *2*, 559-572.

49. Lento, G., Hickson, R., Chambers, G., Penny, D. (1995). Use of spectral analysis to test hypotheses on the origin of pinnipeds. Mol Biol Evol *12*, 28 - 52.

50. Zhaxybayeva, O., Hamel, L., Raymond, J., Gogarten, J. (2004). Visualization of the phylogenetic content of five genomes using dekapentagonal maps. Genome Biology *5*, R20.

51. Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389 - 3402.

52. Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B.*, et al.* (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res *29*, 22-28.

53. Montague, M. G., Hutchison, C. A., 3rd (2000). Gene content phylogeny of herpesviruses. Proc Natl Acad Sci U S A *97*, 5334-5339.

54. Zhaxybayeva, O., Gogarten, J. (2002). Bootstrap, Bayesian probability and maximum likelihood mapping: Exploring new tools for comparative genome analyses. BMC Genomics *3*, 4.

55. Zhaxybayeva, O., Peter Gogarten, J. (2003). An Improved Probability Mapping Approach to Assess Genome Mosaicism. BMC Genomics *4*, 37.

56. Lerat, E., Daubin, V., Moran, N. A. (2003). From Gene Trees to Organismal Phylogeny in Prokaryotes:The Case of the gamma-Proteobacteria. PLoS Biol *1*, E19.

57. Fitch, W. (2000). Homology a personal view on some of the problems. Trends Genet *16*, 227 - 231.

58. Zhaxybayeva, O., Gogarten, J. P. (2003). An improved probability mapping approach to assess genome mosaicism. BMC Genomics *4*, 37.

59. Huelsenbeck, J., Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics *17*, 754 - 755.

60. Strimmer, K., Goldman, N., Von Haeseler, A. (1997). Bayesian probabilities and quartet puzzling. Molecular Biology and Evolution *14*, 210-211.

61. Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F., Douzery, E. J. (2003). Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol Biol Evol *20*, 248-254.

62. Erixon, P., Svennblad, B., Britton, T., Oxelman, B. (2003). Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst Biol *52*, 665-673.

63. Suzuki, Y., Glazko, G. V., Nei, M. (2002). Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc Natl Acad Sci U S A *99*, 16138-16143.

64. Philippe, H., Douzery, E. (1994). The pitfalls of molecular phylogeny based on four species as illustrated by the Cetacea/Artiodactyla relationships. Journal of Mammalian Evolution *2*, 133-152.

65. Adachi, J., Hasegawa, M. (1996). Instability of quartet analyses of molecular sequence data by the maximum likelihood method: the Cetacea/Artiodactyla relationships. Mol Phylogenet Evol *6*, 72-76.

66. Hillis, D., Pollock, D., McGuire, J., Zwickl, D. (2003). Is sparse taxon sampling a problem for phylogenetic inference? Syst Biol *52*, 124 - 126.

67. Graybeal, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol *47*, 9 - 17.

68. Rosenberg, M., Kumar, S. (2003). Taxon sampling, bioinformatics, and phylogenomics. Syst Biol *52*, 119 - 124.

69. Schmidt, H., Strimmer, K., Vingron, M., von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics *18*, 502 - 504.

70. Saitou, N., Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol *4*, 406-425.

71. Felsenstein, J. (1993). PHYLIP (Phylogeny Inference Package) version 3.6 Distributed by the author. Department of Genetics, University of Washington, Seattle.

72. Graybeal, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem? Syst Biol *47*, 9-17.

73. Rosenberg, M. S., Kumar, S. (2003). Taxon sampling, bioinformatics, and phylogenomics. Syst Biol *52*, 119-124.

74. Koonin, E. V. (2000). How many genes can make a cell: the minimal-gene-set concept. Annu Rev Genomics Hum Genet *1*, 99-116.

75. Cox, T. F., Cox, M. A. A. (2001). Multidimensional scaling. In: *Book* Multidimensional scaling (Editor ed.^eds.). City: Chapman & Hall/CRC.

76. Schölkopf, B., Smola, A. J. (2002). Learning with kernels : support vector machines, regularization, optimization, and beyond. Cambridge, Mass.: MIT Press.

77. Tenenbaum, J. B., Silva, V. d., Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science *290*, 2319-2323.

78. Sanderson, M. J., Purvis, A., Henze, C. (1998). Phylogenetic supertrees: assembling the trees of life. Trends in Ecology & Evolution *13*, 105-109.

79. Bininda-Emonds, O. R. P., Gittleman, J. L., Steel, M. A. (2002). THE (SUPER)TREE OF LIFE: Procedures, Problems, and Prospects. Annual Review of Ecology and Systematics *33*, 265-289.

80. Strimmer, K., von Haeseler, A. (1996). Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, 964-969.

81. Habicht, K. S., Gade, M., Thamdrup, B., Berg, P., Canfield, D. E. (2002). Calibration of sulfate levels in the archean ocean. Science *298*, 2372-2374.

82. Vasconcelos, C., McKenzie, J. A. (2000). Biogeochemistry. Sulfate reducers--dominant players in a low-oxygen world? Science *290*, 1711-1712.

83. Ehrenreich, A., Widdel, F. (1994). Anaerobic oxidation of ferrous iron by purple bacteria, a new type of phototrophic metabolism. Appl Environ Microbiol *60*, 4517-4526.

84. Imlay, J. A. (2003). Pathways of oxidative damage. Annual Review of Microbiology *57*, 395-418.

85. Brocks, J. J., Logan, G. A., Buick, R., Summons, R. E. (1999). Archean molecular fossils and the early rise of eukaryotes. Science *285*, 1033-1036.

86. Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol *39*, 306-314.

87. Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution *39*, 783-791.