Phylogenies based on gene content

Matthew Spencer,
Department of Mathematics and Statistics and
Department of Biochemistry and Molecular Biology,
Dalhousie University, Halifax, Nova Scotia, B3H 3J5, Canada.
matts@mathstat.dal.ca

April 4, 2005

Abstract

I review methods that can be used to estimate phylogenies from gene content data. These methods may be useful for inferring deep phylogenetic relationships, where sequence data can be misleading due to saturation, paralogy, and lateral gene transfer. The two major problems with gene content data are the difficulty of observing the absence of gene families and the possibility of multiple changes in gene content along an evolutionary pathway. I discuss parsimony, naive distance methods, the SHOT web server, paralinear distances, and model-based methods developed by Huson & Steel and Gu & Zhang. I suggest some possible improvements to these methods, and conclude with recommendations and areas for future research.

1 Why use gene content to reconstruct phylogeny?

These are many problems with deep phylogenetic reconstruction from nucleotide and amino acid sequences. Practical difficulties include phylogenetic artefacts such as long branch attraction, and the effects of rate variation over time, or heterotachy (Gribaldo and Philippe, 2002). These could in principle be solved by better models of sequence evolution. There are also two major problems that cannot be resolved in this way.

First, homologous nucleotide and amino acid sequences may be saturated with changes (Meyer et al., 1986). Figure 1 illustrates saturation at a single nucleotide. If the amount of evolutionary change (rate of change $\times$ time) is sufficiently small, closely related species will tend to have the same nucleotide states (Figure 1a). With more time or a higher rate of change, there may have been so many substitutions that nucleotide states are no longer informative about evolutionary history (Figure 1b).

Second, genomes gain genes by duplication and lateral transfer, and lose genes by deletion. Deletion results in distantly-related genomes having few homologous genes. For example, among approximately 100 genomes that had been

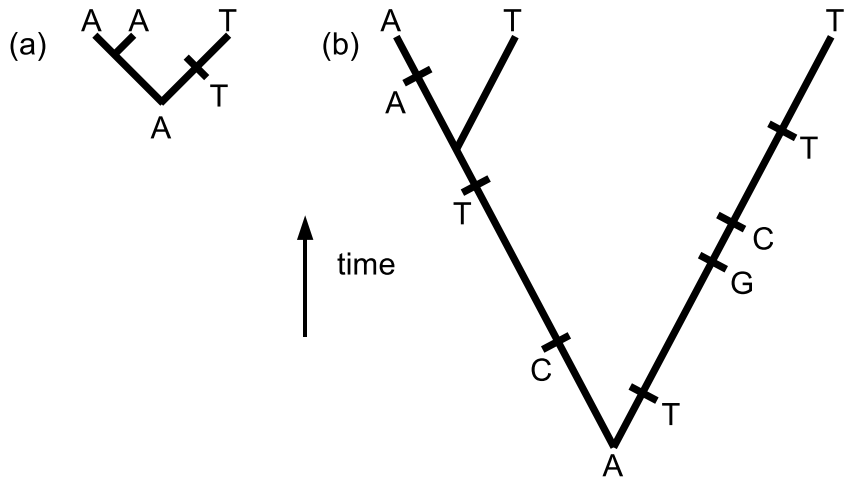

Figure 1: (a) Over a small amount of evolutionary time, nucleotide states are likely to contain useful information about phylogeny. (b) Over a larger amount of time, information on history may be lost. Labels are hypothetical nucleotide states at a single site on a tree leading from a common ancestor to extant species.

sequenced by 2003, there were only about 60 ubiquitous genes (Koonin, 2003). The scarcity of ubiquitous genes can make it difficult to find suitable sequences for deep phylogenetic reconstruction. Duplication can result in multiple copies of a sequence. If different copies are then deleted in different lineages, the relationships among the remaining copies may not reflect the species phylogeny. Given the generally high rates of gene deletion, this may be a common situation (Martin and Burg, 2002). Figure 2 shows an example. An ancestral duplication is followed by three speciation events, with different gene copies lost in the lineages leading to different extant taxa. The two surviving versions of copy B are more closely related to each other than to the two surviving versions of copy A. A and B are described as paralogous, which means they are related by a gene duplication event rather than a speciation event (Page and Holmes, 1998, page 31). In contrast, the two surviving versions of copy A are orthologous: they are related by a speciation event. If we did not recognize that A and B were paralogous rather than orthologous, we would incorrectly group the two taxa containing copy B together in a sequence-based phylogeny.

Because of lateral transfer, different genes may have different evolutionary histories (Doolittle et al., 2003). For example, in a situation like Figure 3, a phylogeny based on gene A would group taxa 1 and 2 together, but a phylogeny based on gene B would group taxon 2 with 3.

Gene duplication, deletion and transfer events are problems for sequence-based phylogenetics, but they also give rise to patterns of gene content across taxa. Here, I will discuss methods that use gene content data to infer phylogenies. Changes in gene content may be less prone to saturation than sequence data. Because we use information from the whole genome, we may be able to get a summary phylogeny that represents the dominant patterns of information flow. Furthermore, we can also attempt to estimate biological properties such as the gene content of an ancestral taxon, or the relative rates of duplication, deletion and transfer.

The idea that whole-genome phylogenies might be a good idea is not new. For example, Fitch and Margoliash (1967) wrote "Biochemists have attempted to use quantitative estimates of variance between substances obtained from different species to construct phylogenetic trees ... These methods have not been completely satisfactory because (i) the portion of the genome examined was often very restricted ..." It is only in the last decade that the data to produce whole-genome phylogenies have become available.

2 Introductory example: why is gene content phylogenetics difficult?

In this section, I will use some simple examples to illustrate the two major problems with using gene content data in phylogenetics. The rest of the material will be mainly concerned with different ways of solving these problems.

Consider the two *E. coli* strains K12 and 0157:H7 EDL933. I downloaded

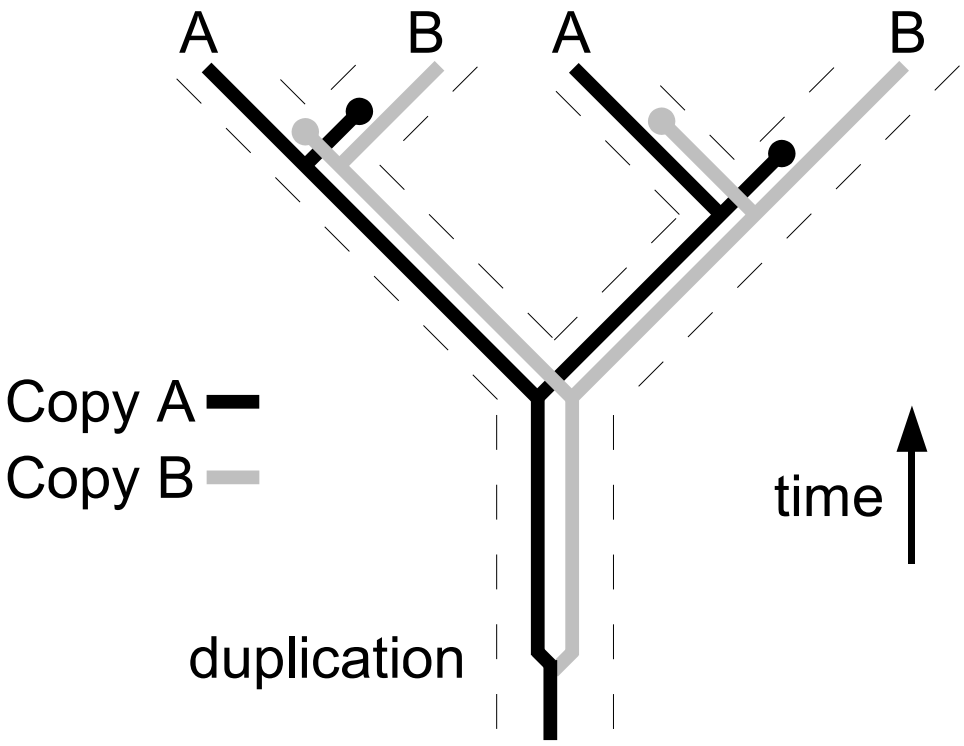

Figure 2: Deletion of different paralogs in different lineages. A single gene is duplicated to form two copies (black: copy A; grey: copy B). Three speciation events then give rise to four extant taxa (the light dashed lines indicate the species phylogeny). Copy A is lost from two taxa and copy B from the other two. Labels at the top of the tree indicate the copy that is present in each extant taxon. Circles indicate deletions.

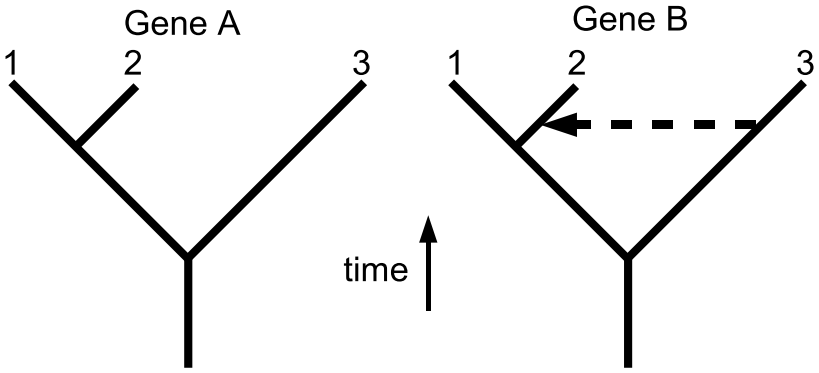

Figure 3: Lateral gene transfer results in different genes having different evolutionary histories. Numbers are labels for extant taxa. The horizontal dashed line indicates a lateral transfer of gene B from an ancestor of taxon 3 to an ancestor of taxon 2.

Table 1: Presence and absence of gene families in the COG database for two *E. coli* strains, K12 and 0157:H7 EDL933.

	0157:H7 EDL933 absent	0157:H7 EDL933 present
K12 absent	2622	120
K12 present	61	2070

data on these strains from the COG database (Tatusov et al., 2003). There are 4873 gene families in the version of COG I used. K12 has at least one gene from 2131 of these families, while 0157:H7 EDL933 has at least one gene from 2190 families. Table 1 shows the number of occurrences of each pattern of presence and absence of gene families in these strains.

There are only $120 + 61 = 181$ cases out of 4873 where one of the strains has no members of a gene family and the other has at least one member. Table 2 shows similar data for a pair of distantly-related species, *Archaeoglobus fulgidus* (Archaea) and *Bacillus subtilis* (gram positive bacteria). Here, there are $1181 + 654 = 1835$ cases where the two species have different presence/absence states. It looks as though distantly-related taxa tend to have different presence/absence states more often than closely-related taxa.

This suggests a simple measure of gene content distance d_{ij} between two

Table 2: Presence and absence of gene families in the COG database for *Archaeoglobus fulgidus* (Archaea) and *Bacillus subtilis* (gram positive bacteria).

	B. subtilis absent	*B. subtilis* present
A. fulgidus absent	2448	1181
A. fulgidus present	654	590

taxa i and j

$$d_{ij} = (n_{AP} + n_{PA})/N \tag{1}$$

where n_{AP} is the number of gene families absent from taxon i but present in taxon j, n_{PA} is the number of gene families present in taxon i but absent from taxon j, and N is the total number of gene families. The minimum possible value is 0, the maximum possible value is 1, and $d_{ij} = d_{ji}$. The two *E. coli* strains have $d_{ij} = 181/4873 = 0.04$, while *A. fulgidus* and *B. subtilis* have $d_{ij} = 1835/4873 = 0.38$.

If we had a set of m taxa of interest, we could compute the $m \times m$ matrix of pairwise distances between taxa. There are several good methods we could then use to estimate a phylogeny for the taxa, such as neighbor-joining (NJ) (Saitou and Nei, 1987; Studier and Keppler, 1988) and least-squares (Fitch and Margoliash, 1967). Both methods are available in standard phylogenetic software packages such as PAUP* (Swofford, 2003) and PHYLIP (Felsenstein, 2004b). See Page and Holmes (1998, section 6.2) for a brief introduction, or Felsenstein (2004a, chapter 11) for more detail. These methods will give the correct phylogeny if the true phylogeny is a tree and we know the true evolutionary distances.

Unfortunately, there are two problems with this approach, unobservable data and multiple changes.

2.1 Unobservable data

One fundamental difference between sequence data and gene content data is that some gene content data are unobservable. For nucleotide data, the possible states are A, C, G and T, and there is no obvious bias that affects the recording of these states. For amino acid data, there are twenty possible states (the twenty amino acids), and again there is no obvious recording bias. For gene content data, the possible states are either {absent, present} or {absent, present in one copy, present in two copies...}. It is almost always the case that absences are less likely to be recorded than presences. If a gene family is absent from every genome in a database, we may not know that the family exists.

Suppose that we have two independent genomes x and y. In each genome, a gene is present with probability 1/2. Denote the states absent and present by A and P respectively. Let x_A be the event that a gene is present in genome x, and x_P be the event that a gene is absent. Because the genomes are independent, the probability of observing the pattern $\{x_i, y_i\}$ is $P(x_i) \times P(y_i)$. If we could

observe every case, equation 1 gives the distance

$$\begin{aligned} d_{xy} &= \frac{P(x_A) \times P(y_P) + P(x_P) \times P(y_A)}{P(x_A) \times P(y_A) + P(x_A) \times P(y_P) + P(x_P) \times P(y_A) + P(x_P) \times P(y_P)} \\ &= P(x_A) \times P(y_P) + P(x_P) \times P(y_A) \\ &= \frac{1}{4} + \frac{1}{4} \\ &= \frac{1}{2} \end{aligned} \tag{2}$$

(the denominator in the first line includes all possible outcomes, so it sums to 1).

Now suppose that we can always observe the patterns $\{A, P\}$, $\{P, A\}$ and $\{P, P\}$, but we can never observe the pattern $\{A, A\}$. The distance we estimate would be

$$\begin{aligned} d_{xy} &= \frac{P(x_A) \times P(y_P) + P(x_P) \times P(y_A)}{P(x_A) \times P(y_P) + P(x_P) \times P(y_A) + P(x_P) \times P(y_P)} \\ &= \left(\frac{1}{2}\right) / \left(\frac{3}{4}\right) \\ &= \frac{2}{3} \end{aligned} \tag{3}$$

Because we cannot observe the double-absence pattern, we get a overestimate of the distance. Furthermore, this bias depends on the number of genes that are present. We can do the same calculations for a pair of independent genomes v and w, each of which has probability 3/4 that a gene is present. Thus v and w are larger genomes than x and y. The distance is 3/8 if we can observe every pattern, but 2/5 if we cannot observe the double-absence pattern. The proportion by which we overestimated the distance is less for v and w than for x and y. If the distances we calculate do not reflect the true proportion of differences, and are wrong by different amounts for pairs of genomes of different sizes, we cannot expect to reliably reconstruct a phylogeny for genomes that differ in size. This is an important problem, because genomes do differ greatly in size. For example, genomes in the COG database vary by almost a factor of 10 in the number of gene families that are present.

In reality, the under-recording of {absent, absent} patterns may be more complex. For example, the data in tables 1 and 2 came from the COG database (Tatusov et al., 2003). The database is constructed from patterns of pairwise sequence similarity detected using BLAST (Tatusov et al., 1997). Because of the way this is done, a gene family will only appear in the database if it is present in at least three genomes. Nevertheless, the same principles apply.

2.2 Multiple changes

We can estimate the correct tree from a matrix of pairwise evolutionary distances. An evolutionary distance is the number of changes of state between

Table 3: Classification of methods for gene content phylogenetics by the kind of data they use, how they deal with unobservable data, and how they deal with multiple changes. 'ML' means Maximum Likelihood.

Method	Data	Unobservable data	Multiple changes
parsimony	usually presence/absence	not necessary	no attempt
naive distances	presence/absence	various	no attempt
SHOT	presence/absence	correction based on pattern frequencies	approximate evolutionary distances
paralinear distances	presence/absence	conditioning genome	tree-additive distances
Huson and Steel	presence/absence	not necessary	ML distances
Gu and Zhang	absent/1 copy/ > 1 copy	conditional likelihood	ML distances
multi-gene events	number of genes in family	LOWESS extrapolation	ML distances

two taxa. The simple counting method we used above (equation 1) uses the observed number of changes.

To see why this is an underestimate of evolutionary distance, consider presence/absence data for a single site. If there was only one change at a given location, we will observe this as a change from absent to present or present to absent. If two changes occurred, we will end up where we started (absent to present to absent, or vice versa), and will observe no changes. If the data are the number of members of gene families, two changes might give us something like $10 \rightarrow 9 \rightarrow 8$. Although we did not end up where we started, we may not be able to distinguish this from the single change $10 \rightarrow 8$.

3 Methods

In this section, I will review some of the methods for estimating phylogenies from gene content data. Table 3 classifies these methods by the kind of data they use, how they deal with unobservable data, and how they deal with multiple changes. The list is not exhaustive, but I have attempted to include examples of all the major methods.

3.1 Parsimony

Maximum parsimony estimates the phylogeny that minimizes the total number of changes (here, gene gains or losses) needed to explain the observed data. Figure 4 illustrates how we might count the total number of changes needed in a simple case. A gene family is absent in taxa a and b, but present in taxon c. If a and b are sister taxa, then the absence of the gene in their common ancestor i will minimize the number of changes in this part of the tree. If we assumed the gene was present in i, it would have been independently lost along the evolutionary paths leading to a and b, requiring two changes instead of none. The gene could be either present or absent in the common ancestor j of i and c. We would either need to assume loss of the gene on the path leading from j to i (if the gene was present in j) or gain of the gene on the path leading from j to c (if the gene was absent in j). Either of these reconstructions is equally parsimonious, requiring one change. To estimate a phylogeny using parsimony, we would have to search over the set of possible trees, doing reconstructions of this kind on each tree. We would choose the tree with the smallest number

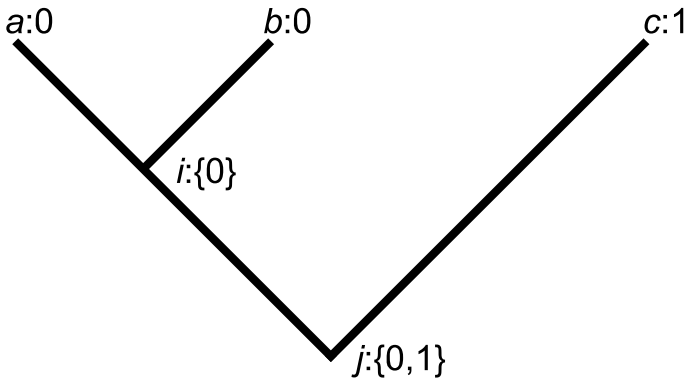

Figure 4: Inferred ancestral states in parsimony. Extant taxa are a, b and c. Internal nodes (hypothetical ancestors) are i and j.The labels at the top of the tree indicate observed states (0=absent, 1=present) in extant taxa. For example, a:0 means that extant taxon a has state 0. Labels in curly brackets indicate possible ancestral states.

of changes. For more information on parsimony, see Page and Holmes (1998, section 6.4) and Felsenstein (2004a, chapters 1, 6, 7, 9 and 10).

Parsimony was one of the first phylogenetic methods to be applied to gene content data (e.g. Fitz-Gibbon and House, 1999; Montague and Hutchinson III, 2000). Standard software such as PAUP* (Swofford, 2003) and PHYLIP (Felsenstein, 2004b) can be used. Parsimony has usually been applied to presence/absence data, but it could also be used with data on the number of members of gene families.

Unobservable data are not necessarily a problem for parsimony. If a gene family has the same state in every extant taxon, parsimony assumes that it did not change its state anywhere on the tree, no matter what topology the tree had. In most cases, most of the unobservable data are gene families that are absent from all extant taxa, and will therefore make no difference to the tree chosen by parsimony.

Parsimony does not use an explicit model for evolution. We assume changes are so rare that multiple changes along a single path never occur. This may be a problem if the rate of change is in fact quite high. As a result, there are situations in which parsimony is inconsistent: even with infinite data, it will not estimate the correct tree (Felsenstein, 2004a, pp. 107-122).

Another problem with the absence of an explicit model is that we do not know how to weight gains and losses. For example, if it were really the case that gains were rare but losses were common, then minimizing the number of gains would be more important than minimizing the number of losses. We should thus give higher weights to rare events. One possible source of external information on weights is the plausibility of reconstructed ancestral genomes, although this is rather subjective. Mirkin et al. (2003) reconstructed ancestral genomes for

taxa in the COG database, using parsimony on a fixed tree with a range of weights for gene gain and loss. They chose weights so that the reconstructed ancestors had complete metabolic pathways for essential functions. Boussau et al. (2004) performed a similar analysis for the α-proteobacteria, but used three different weights for gene duplication, deletion, and genesis (which includes lateral transfers). While Mirkin et al. (2003) suggested equal weights for gains and losses, Boussau et al. (2004) gave five times as much weight to genesis as to duplication or deletion. It is not clear how to reconcile these results.

It is sometimes assumed that the the transition from absence to presence of a gene can occur only once, but that a gene can be independently lost in different lineages. The version of parsimony that matches this assumption is known as Dollo parsimony, and has been applied to gene content data by Wolf et al. (2001) and Huson and Steel (2004). However, multiple transitions from absence to presence could occur by lateral transfer, in which case Dollo parsimony would not be appropriate.

3.2 Naive distances

By naive distance, I mean any distance measure which is not based on a proper evolutionary model for gene gain and loss. Some examples of naive distances that have been used with gene content are:

- Fitz-Gibbon and House (1999) coded gene presence/absence in 11 free-living microorganisms as a binary matrix, and used PAUP* (Swofford, 2003) to estimate a neighbor-joining tree. They do not describe the distance measure they used in detail, but it may be equation 1.

- Snel et al. (1999) used

$$d_{ij} = 1 - \frac{n_{PP}}{\min(a, b)} \tag{4}$$

  where a and b are the numbers of genes present in genomes i and j respectively, and n_{PP} is the number of gene families present in both genomes. Normalizing by the size of the smaller genome accounts for differences in genome size. If genome a contains 1000 genes and genome b contains 2000 genes, the maximum possible number of shared genes is 1000. Equation 4 will give a distance of 0 if all the genes in a are also in b. This seems like a good property.

- Wolf et al. (2002) used a distance measure based on the Jaccard coefficient:

$$d_{ij} = 1 - \frac{n_{PP}}{a + b - n_{PP}} \tag{5}$$

  In a case where all of the genes in a are also present in b, and there are additional genes in b but not a, equation 5 will give a distance greater than 0.

There are many other distances that could be tried. The examples here are easy to calculate, and standard software can be used to estimate phylogenies from the resulting distance matrices. However, we have no guarantee that any such distance will perform well. As discussed above (section 2.2), naive distances will underestimate evolutionary distances when the true number of changes is large. We are not certain to get the right tree, even if we have very large amounts of data. It therefore seems better to consider distances that take account of multiple changes.

3.3 SHOT

The SHOT web server:

`http://www.bork.embl-heidelberg.de/~korbel/SHOT_v2/`

has options for estimating phylogenies from gene presence/absence data. It uses either neighbor-joining or least-squares to estimate a phylogenetic tree from a matrix of approximate evolutionary distances. Unobservable data are dealt with using a correction based on pattern frequencies (Korbel et al., 2002). Because SHOT is one of the simpler model-based methods, it is worth looking at it in some detail.

The raw data are the presence/absence of families of orthologous genes. Orthologs are identified from the STRING database (von Mering et al., 2003), in which families are defined by bidirectional matches and triangles of reciprocal best matches. The similarity s in gene content between a pair of taxa is defined as the number of shared orthologs, normalized in a way that reflects genome size. The default measure of distance is then $-\log(s)$. To see why this is a good approximation to the evolutionary distance, we need to use a simple Markov model for the evolution of gene content.

Let A and $P = 1 - A$ be the probabilities of absence and presence of a gene family. Let q_{AP} be the instantaneous rate at which families are gained, and q_{PA} be the rate at which families are lost (both have dimensions of time^{-1}). Then a simple model for gene content evolution is the pair of differential equations

$$\begin{aligned} \frac{dA}{dt} &= -q_{AP}A + q_{PA}P \\ \frac{dP}{dt} &= q_{AP}A - q_{PA}P = -\frac{dA}{dt} \end{aligned} \tag{6}$$

Setting $\frac{dA}{dt} = 0$ and solving for A gives the stationary probability $\alpha = q_{PA}/(q_{PA} + q_{AP})$ of the absent state.

Let $\beta = \exp(-(q_{PA} + q_{AP})t)$, where t is the true evolutionary distance. For this model, the expected frequencies F of the patterns of absence and presence in a pair of taxa are

$$\begin{aligned} F_{AA} &= \alpha[1 - (1-\alpha)(1-\beta)] \\ F_{AP} &= \alpha[(1-\alpha)(1-\beta)] = F_{PA} \\ F_{PP} &= (1-\alpha)[1 - \alpha(1-\beta)] \end{aligned} \tag{7}$$

We obtain these frequencies by solving equation 6, but we do not show the details here. Assume that we can observe only those patterns in which a gene family is present in at least one of the taxa (this is not strictly true: a gene family will appear in STRING if it is present in at least two taxa, but these two taxa need not be the pair that we are interested in). The genome sizes a and b are

$$\begin{aligned} a &= n_{PA} + n_{PP} \\ b &= n_{AP} + n_{PP} \end{aligned} \tag{8}$$

Note that in this simple model, a and b are equal. We do not know n_{AA}, the number of gene families absent from both taxa. The total number of gene families N is therefore also unknown, but

$$\begin{aligned} a &= N(F_{PA} + F_{PP}) \\ &= N(1-\alpha)[\alpha(1-\beta) + 1 - \alpha(1-\beta)] \\ &= N(1-\alpha) \end{aligned} \tag{9}$$

Korbel et al. (2002) suggest the similarity measure s

$$\begin{aligned} s &= n_{PP}\frac{\sqrt{a^2+b^2}}{ab\sqrt{2}} \\ &= \frac{NF_{PP}}{N(1-\alpha)} \\ &= 1 - \alpha(1-\beta) \\ &= 1 - \alpha(1 - e^{-(q_{PA}+q_{AP})t}) \end{aligned} \tag{10}$$

We can directly estimate s from the first line of equation 10. Rearranging the last line of equation 10 shows us how to estimate t:

$$t = \frac{\log(s+\alpha-1) - \log(\alpha)}{-(q_{PA}+q_{AP})} \tag{11}$$

We do not know q_{PA} or q_{AP}, but the denominator is a constant if the parameters of the model do not change over time. It will shrink or grow every distance by the same amount. This will make no difference to the topology of the tree, so we can ignore it apart from its negative sign. Thus, to estimate a tree, the distance $\hat{t}$ is linearly related to t, and is just as good for our purposes:

$$\hat{t} = -\log(s+\alpha-1) + \log(\alpha) \tag{12}$$

We still cannot use $\hat{t}$, because we do not know α if we cannot observe the absence of a gene family in both taxa. We cannot estimate α either from the sizes of the two genomes, or from the number of times we observe each pattern. However, the observation that there are few ubiquitous gene families suggests that α (the stationary probability that a gene family is absent from a given genome) must be close to 1. If this is the case, then $\log(s+\alpha-1) \approx \log(s)$, and $\log(\alpha) \approx 0$. Then we can hope that t is approximately linearly related to $-\log(s)$.

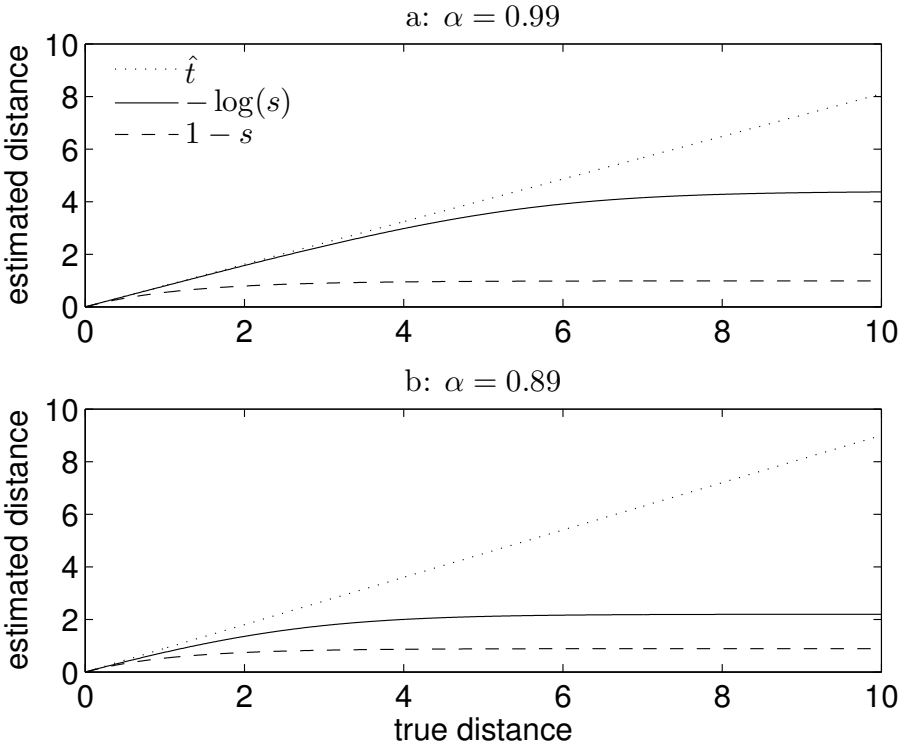

Figure 5: Relationship between true and estimated distances from the SHOT method under a simple Markov model of gene presence/absence, for two different stationary probabilities α of gene absence (a: $\alpha = 0.99$, b: $\alpha = 0.89$). The horizontal axis is the true evolutionary distance t (equation 11). The vertical axis is estimated distance, either $\hat{t}$ (equation 12, dotted line), $-\log(s)$ (the default estimate in SHOT, solid line), or $1-s$ (a naive distance measure, dashed line). Parameters: $q_{PA} = 0.8$, $q_{AP} = 0.01$ (a) or 0.1 (b).

Figure 5 shows that this is quite a good approximation for small evolutionary distances, so long as α is large. There is a linear relationship between true evolutionary distance t and $\hat{t}$. The default SHOT estimate $-\log(s)$ is almost linearly related to t when t is small, and is much better than the naive distance estimate $1 - s$. The nonlinearity in the relationship between t and $-\log(s)$ is more of a problem when α is smaller (Figure 5b).

3.4 Paralinear distances

Lake and Rivera (2004) suggested a different method to estimate pairwise distances, based on Markov models for presence/absence like the one discussed on section 3.3. Differences in the frequencies of the four nucleotides can cause

biases in sequence-based phylogeny estimation. For example, taxa with high GC content may be artefactually grouped together if analyzed using models that do not allow variation in nucleotide frequencies. Paralinear (also known as logdet) distances take account of such variation (Lake, 1994; Lockhart et al., 1994). There is substantial variation in genome size among organisms. For example, the smallest genome in the COG database (*Mycoplasma genitalium*) has members of only 362 gene families, while the largest (*Pseudomonas aeruginosa*) has members of 2243 gene families. Models like equation 6 assume the same equilibrium size for every genome, which is probably not realistic. Paralinear distances do not require this assumption, so it seems like they might be a good idea.

We still have to deal with the problem of unobservable states. Lake and Rivera (2004) suggested the use of a conditioning genome. Instead of considering all genes present in either genome i or genome j, they choose a conditioning genome c. They then calculate paralinear distances between each pair i and j using only those genes that are present in c. The choice of conditioning genome affects the distance we will estimate. Let t_k be the distance from the conditioning genome c to the common ancestor of i and j. Different values of t_k lead to different estimated distances (figure 6). If we use the same conditioning genome for many pairs of taxa, each pair will have a different value of t_k (figure 7) and the pairwise distances will be distorted.

When all states are observable, paralinear distances are tree-additive. This means that the distance between two taxa i and j that are connected by a path $i, x_1, x_2, \ldots, x_n, j$ is the sum of the distances between i and x_1, x_1 and x_2,...,x_n and j. Tree-additivity should ensure that we will get the correct tree if we have enough data. Given the effects of t_k discussed above, it is surprising that paralinear distances with a conditioning genome still seem to be tree additive. Nevertheless, this has not been formally proved as far as I know, and the choice of conditioning genome might be important in real situations with finite data. More work on this area is needed.

Paralinear distances with a conditioning genome have been applied to a set of ten prokaryotes and eukaryotes (Rivera and Lake, 2004). Rivera and Lake (2004) ensured that all taxa in their main analysis had similar-sized genomes. For additional analyses, they grouped taxa into sets with similar-sized genomes, and used a different conditioning genome for each set. The method is not yet implemented in standard software as far as I know.

3.5 Huson and Steel's method

Huson and Steel (2004) suggested a model for genome size evolution in which there is a constant rate of gene birth, and the rate of gene death is proportional to the number of genes in the genome:

$$\frac{dl}{dt} = \lambda - \mu l \tag{13}$$

where l is the number of genes in the genome, λ is the birth rate (dimensions

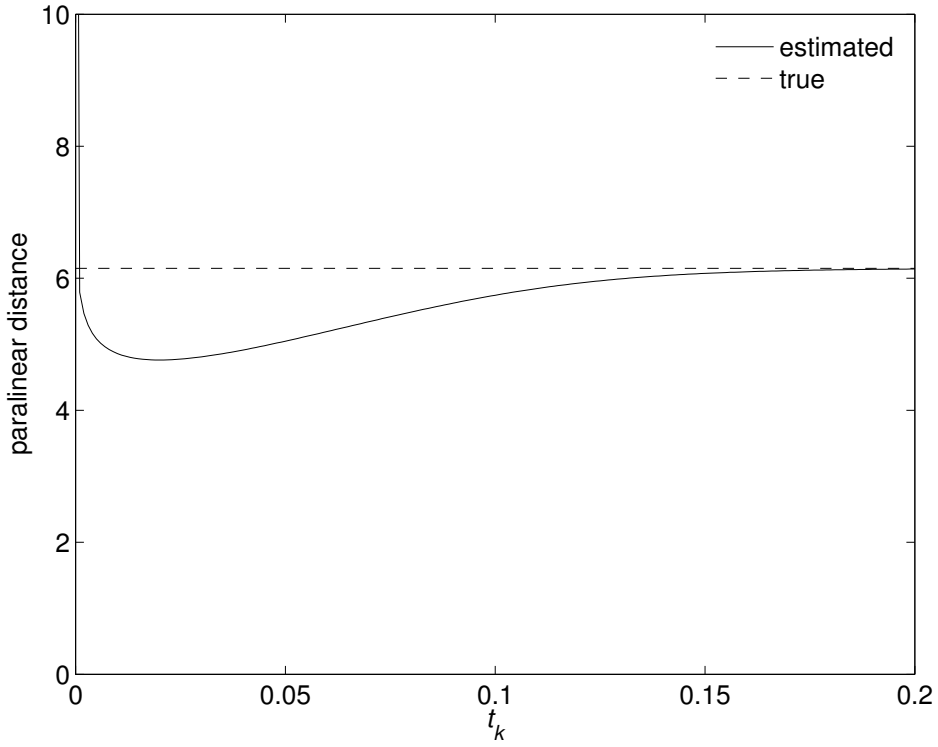

Figure 6: The evolutionary distance t_k from the common ancestor of two taxa i and j to the conditioning genome affects the estimated paralinear distance (solid line) between i and j. The true paralinear distance (dashed line) remains constant. Parameters: t_i (evolutionary distance from common ancestor to i)=0.1, t_j (evolutionary distance from common ancestor to j)=0.2, q_{AP} (rate of transitions from absent to present)=0.01, q_{PA} (rate of transitions from present to absent)=0.8.

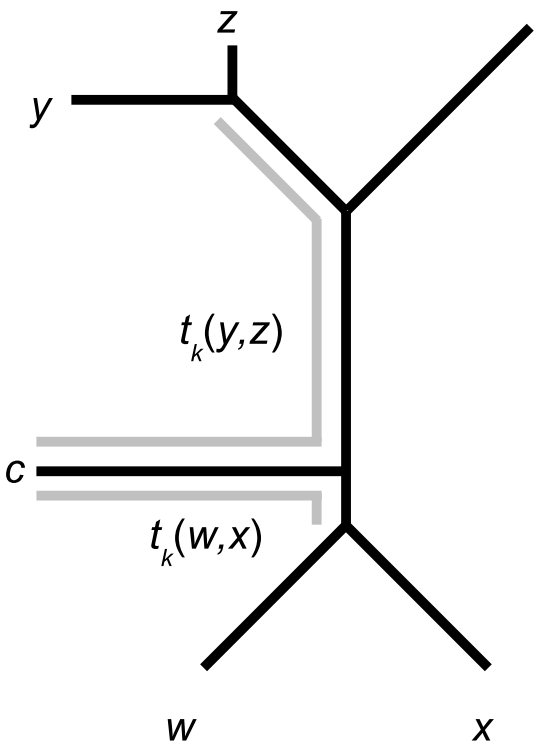

Figure 7: Distance to the conditioning genome depends on location of a pair of taxa. Here, c is the conditioning genome, and w, x and y, z are two pairs of taxa. The distances from the conditioning genome to the common ancestors are $t_k(w, x)$ and $t_k(y, z)$ respectively. Because these distances are different, we expect the estimated paralinear distances between w and x and between y and z to be different proportions of their true values (figure 6).

of genes $\times$ time^{-1}), and μ is the death rate (time^{-1}). Biologically, we can interpret this to mean that all genes have the same probability of being deleted, and that new genes arise by processes such as lateral transfer or evolution from non-coding sequences, rather than by duplication of existing genes. In reality, we might expect duplication to be an important source of new genes (e.g. Gevers et al., 2004). More sophisticated models for genome size could allow duplications as well (e.g. Karev et al., 2004).

Huson and Steel (2004) derive an evolutionary distance between two genomes i and j that evolve under the model specified by equation 13:

$$d_{ij} = -\log\left[\frac{1}{2}\left(u + \sqrt{u^2 + 4v_{12}}\right)\right] \tag{14}$$

where $u = 1 + v_{12} - v_1 - v_2$, $v_{12} = n_{PP}/m$, $v_1 = a/m$, $v_2 = b/m$, a and b are the sizes of the genomes i and j, and m is the expected number of genes in a genome, λ/μ. Huson and Steel (2004) suggested using the average number of genes per genome as an estimate of m. There is no problem with unobservable data, because the frequencies of absent genes do not appear in equation 14.

This method has been implemented in Splitstree 4:

`http://www-ab.informatik.uni-tuebingen.de/software/jsplits/welcome.html`

Simulations (Huson and Steel, 2004) suggest that it performs better than a naive distance (Equation 4), but not quite as well as Dollo parsimony. It has not yet been applied to real data as far as I know.

3.6 Gu and Zhang's method

One possible problem with Huson and Steel's method (section 3.5) is that the rate of gene gain does not depend on the number of genes in the genome, so the role of duplication is ignored. Gu and Zhang (2004) developed a birth-death model for the number of genes in a gene family, under the assumption that each gene is equally likely to be deleted or duplicated, with rates μ and λ respectively (dimensions of time^{-1}). In this model, the rate of change of the probability P_n of having n genes in a gene family is

$$\frac{dP_n}{dt} = \lambda(n-1)P_{n-1} + \mu(n+1)P_{n+1} - (\lambda+\mu)nP_n \tag{15}$$

In equation 15, the first term $\lambda(n-1)P_{n-1}$ is the rate at which we move from having $n-1$ genes to having n genes, by duplicating one gene. Similarly, $\mu(n+1)P_{n+1}$ is the rate at which we move from having $n+1$ genes to having n genes, by deleting one gene. The last term $(\lambda+\mu)nP_n$ is the sum of the rates at which we move from having n genes to having either $n+1$ or $n-1$ genes, by duplicating or deleting one gene. This model assumes that only one gene can be deleted or duplicated at a time.

Gu and Zhang (2004) showed that it is not possible to identify the parameters of this model if we consider only the presence or absence of a gene family. Data

on whether there are no genes, one gene, or more than one gene in the family are needed. Such data can be obtained from the COG database. Let A indicate no genes, P indicate one gene, and D indicate more than one gene in a family. Then we can use this model to calculate the likelihood of observing a given frequency of each combination of the states A, P and D in a pair of genomes separated by an evolutionary distance t. The estimate $\hat{t}$ that maximizes the likelihood is a maximum likelihood (ML) estimate of evolutionary distance.

This method is affected by unobservable data. Gu and Zhang (2004) overcame this problem using a conditional likelihood, in which the likelihood is normalized by the probability that the data are observable. They assumed that AA was the only unobservable pattern, and obtained a conditional ML distance estimate. This is a simplification, because a gene family will appear in the COG database only if it is present in at least three genomes. Thus AA in a pair of taxa is not guaranteed to be unobservable, and the other patterns are not guaranteed to be observable. They estimated conditional ML distances for a set of 35 microbial genomes from the COG database, and constructed a neighbor-joining tree. Software to do these analyses is available at

`http://xgu.zool.iastate.edu/software.html`

3.7 Models with multi-gene events

Thinking about the nature of the data and the biology of prokaryote genomes suggests some improvements to Gu and Zhang's method (section 3.6):

- The COG database identifies gene families based on triplets of reciprocal best matches (Tatusov et al., 2003), so a gene family must appear in at least three genomes before it will appear in the database. This means that the observability of a data pattern cannot be decided from pairwise criteria. Furthermore, the double absence pattern AA is usually the most commonly observed for pairs of taxa from the COG database (Tables 1 and 2). If we ignore this pattern in the cases where it can be observed, we will be discarding a large proportion of the data.

- The model allows only one gene to be deleted or duplicated at a time. There is empirical evidence that blocks of more than one gene can be duplicated (Gevers et al., 2004; Chen et al., 2003) or deleted (Ochman and Jones, 2000).

- Equation 15 shows that there is no way of leaving state $n = 0$ (the absence of a gene family). Only deletion and duplication are included in the model. Once a family has been lost, it is lost for ever, and there is no way to gain a family that was not initially present in a lineage. In reality, gene families may be gained by lateral transfer, and a new gene family could evolve from some other sequence (although this will probably be a rare event).

We developed a new method that addresses these problems (M. Spencer, E. Susko and A. J. Roger, unpublished). First, we estimated the proportions

of unobservable data for each of the patterns AA, AP, PA and PP for each pair of taxa in the COG database. Although our model works with the number of family members rather than presence/absence, the data were too sparse to estimate the proportion of unobservable data separately for each combination of i genes in a family in one taxon and j in another taxon. We estimated the proportion of unobservable data by extrapolation. Consider a pair of taxa. We know the number of times we observe the pattern AA in this pair of taxa, for gene families present in $3, 4, \ldots, m$ taxa (where m is the number of taxa in the COG database). We used this relationship to predict the number of times AA occurs in this pair, for gene families present in 0, 1 or 2 taxa in the database (the unobservable cases). We used a locally weighted least squares (LOWESS) regression (Cleveland, 1979), in which the slope of the prediction line at any given point on the horizontal axis is influenced most strongly by nearby points. This is an appropriate choice because we do not expect the relationship to have the same shape for all points on the horizontal axis or for all pairs of taxa. We performed similar analyses for the other patterns. Figure 8 shows examples of the extrapolations for two pairs of taxa. In most cases, there were many more unobservable data for the AA pattern than for any other. We then added estimates of the number of unobservable data to the observed counts of each combination of i genes in a family in one taxon and j in another. For example, we divided up unobservable data for the PP pattern in proportion to the number of times each combination of $i > 1, j > 1$.

To deal with gain of gene families by evolution from other sequences or by lateral transfer, we allowed a non-zero rate of transitions from 0 to 1 members of a gene family. To deal with duplications and deletions of multiple genes, we allowed transitions between i and j genes in a family, for any values of i and j. Because allowing a separate rate for each possible combination of i and j would involve too many parameters, we divided transitions up into blocks:

- Deletions of single genes
- Deletions of multiple genes that leave at least one member of a gene family present
- Deletions of entire gene families. We used a separate category here because deleting an entire gene family might increase the risk of losing some important function
- Duplications of single genes
- Duplications of multiple genes
- Transition from no members of a family to one member. This could occur by evolution from some other sequence (which might be rare) or by lateral gene transfer (which might be common).
- Lateral transfers of many genes (leading to larger increases in family size than could have occurred by duplication).

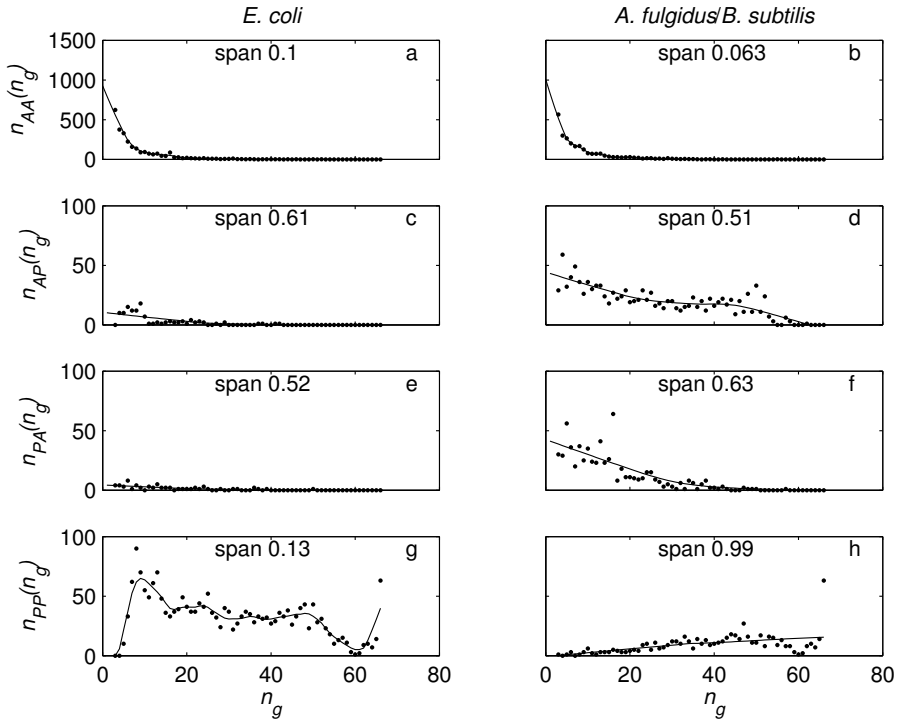

Figure 8: Relationship between number of genomes in which a gene family is found (horizontal axis, n_g) and number of observations of a category in the focal pair of genomes (vertical axis, $n_{..}(n_g)$), where $n_{..}$ is one of the categories AA (a, b), AP (c, d), PA (e, f) and PP (g, h). A indicates absent and P present in each member of the focal pair. Focal pairs are *E. coli* strains K12 and 0157:H7 EDL933 (a, c, e, g); *Archaeoglobus fulgidus* and *Bacillus subtilis* (b, d, f, h). Dots are observations, and solid lines are LOWESS curves with span (proportion of points used in each local regression, chosen by cross-validation) indicated on each panel. The vertical axis scale is fifteen times larger in a and b than in the other panels.

Within each block, we assumed that events operated independently and with equal probability on each possible unit (a single gene or a group of more than one gene). We estimated ML distances using this model for all 66 genomes in the COG database. We used likelihood ratio tests to compare the performance of this model with that of a birth-death model. For the majority of pairs of taxa, our model performed substantially better than the birth-death model. We then used least-squares to estimate a phylogeny based on the ML distances from our model (figure 9).

We have not yet implemented this method in easy-to-use software, and the large number of parameters means that large amounts of computer time are needed. Figure 9 has some good features. For example, the three kingdoms (archaea, bacteria, and eukaryotes) are clearly separated. Nevertheless, there are some obvious biological problems, most of which commonly occur with phylogenetic methods based on gene content.

First, the halophilic archaeon *Halobacterium* is placed near the root of the archaea, and the hyperthermophilic bacterium *Thermotoga* near the root of the bacteria. Both these results are probably artefacts arising from extensive lateral gene transfer. *Halobacterium* may have gained large numbers of genes from the halophilic bacteria (Ng et al., 2000; Kennedy et al., 2001; Brochier et al., 2004), and has therefore been displaced towards the bacteria. Similarly, *Thermotoga* may have gained many genes from the archaea (Nelson et al., 1999). Although our model includes lateral transfers, it will not give accurate tree reconstructions if large numbers of genes are transferred from a single source.

Second, a large group of parasitic and endosymbiotic bacteria have been wrongly grouped together: the parasitic α-proteobacteria *Rickettsia spp.*, chlamydiae (*Chlamydia trachomatis*, *Chlamydophila pneumoniae*), spirochaetes (*Treponema pallidum* and *Borrelia burgdorferi*), mycoplasmas (*Mycoplasma spp.* and *Ureaplasma urealyticum*), and the endosymbiont γ-proteobacterium *Buchnera*. This is probably a consequence of parallel loss of genes that are unnecessary for parasites (Wolf et al., 2001).

Fitting our models to data involves estimating the rates of gene gains and losses, and these rates can tell us about the biology of gene content evolution. We examined these rates in detail for two *E. coli* strains, and for *A. fulgidus* and *B. subtilis*. The estimated rate of lateral transfers of more genes than could be gained by duplication was not significantly different from zero in either case. Nevertheless, the rate of transitions from 0 to 1 members of a gene family was about one fifth of the rate of loss of entire gene families in both cases. If evolution of new gene families from other sequences is rare, this rate may mostly represent lateral transfers of single genes. We also estimated the residence time of a single gene, from when it appears (by duplication or transfer) to when it is deleted. The median residence time was of the same order of magnitude as the number of events separating *A. fulgidus* and *B. subtilis*. When using sequence data for deep phylogenetic reconstruction, it may therefore be better to focus on the subset of gene families with long residence times, rather than using all possible genes.

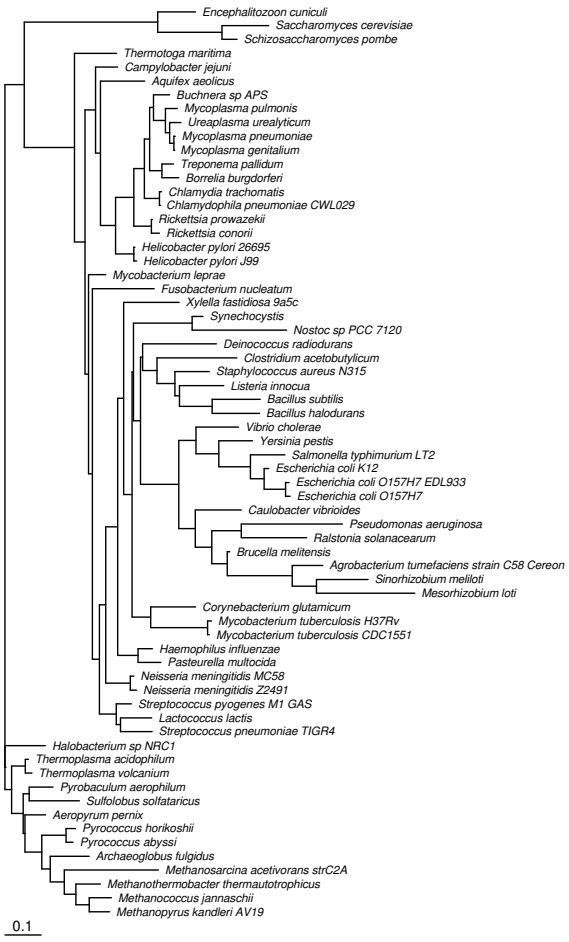

Figure 9: Phylogeny based on maximum likelihood distances for all 66 genomes in the COG database, estimated by least-squares with inverse square weighting. Distances are from the model described in M. Spencer, E. Susko and A. J. Roger (unpublished). The tree is rooted with the archaea as an outgroup. Edge lengths are expected numbers of gene events per gene family.

4 Conclusions

I conclude with some recommendations and some ideas for future work.

Which existing method is best? It is probably not sensible to rely on a single method. All are quite new, and are much less well developed than the corresponding methods for sequence-based phylogeny. Nevertheless, it is probably safe to suggest:

- Parsimony is unlikely to be the best method, because we do not know how to weight gene gains and losses. Furthermore, parsimony has well-known problems with multiple changes.
- Naive distance measures are unlikely to be the best method, because they do not account for multiple changes. The distance measure used in the SHOT web server (Korbel et al., 2002) is no more difficult to calculate, can be justified by simple models of gene presence/absence evolution, and is likely to perform better than naive distances (section 3.3). Therefore, it seems like a good choice for a quick analysis.
- Paralinear distances (Lake and Rivera, 2004; Rivera and Lake, 2004) are the only current method that may be unaffected by substantial variations in genome size in different parts of the tree. Nevertheless, the use of a conditioning genome to deal with unobservable data leads to some curious properties that have not yet been thoroughly investigated (section 3.4).
- The more sophisticated models (Huson and Steel, 2004; Gu and Zhang, 2004, M. Spencer, E. Susko and A. J. Roger, unpublished) offer some advantages. The models they use may be more realistic, and their parameters may give useful biological information. For example, our model can give some information about the rates of lateral gene transfer relative to other processes that affect gene content, and can be used to estimate the length of time a single gene is likely to persist in a genome.

Areas in which more work is needed include:

- Better models of gene content. All the models I discussed make some major assumptions that are unlikely to be true. Some of these are almost unavoidable. For example, we have treated genes or gene families as if they were independent. In reality, events such as duplication and deletion will affect blocks of contiguous genes, which will not necessarily belong to the same family. We would then need to model the order of genes as well as the number of members in a family. This is probably impractical (Felsenstein, 2004a, page 515). Other problems may be more easily addressed. For example, we assumed equal rates of gain and loss for all gene families, but models that allowed some families to change faster than others might be better. Variation in rates in different parts of the tree would allow for variations in genome size.

- Full maximum likelihood inference of phylogenies. We discussed estimating pairwise distances using explicit models for evolution. In principle, we could use these models to estimate the entire tree in a maximum likelihood framework. This is one of the best methods for DNA and amino acid sequence data, because it has good statistical properties and allows explicit hypothesis testing (for example, comparisons between models). Page and Holmes (1998, section 6.5) give a brief introduction to maximum likelihood phylogenetics, and Felsenstein (2004a, chapter 16) provides more detail. See Huelsenbeck and Crandall (1997) for an introduction to hypothesis testing for phylogenetics. Maximum likelihood has been applied to gene content data, but only for very small numbers of taxa (Zhang and Gu, 2004).

- Extensive lateral gene transfers from a single source are likely to mislead us if we assume treelike evolution. Nevertheless, there are good methods for estimating evolutionary networks from pairwise distances (Huson, 1998; Bryant and Moulton, 2002). So far, these methods have not yet seen much use with gene content data.

Acknowledgements

I am grateful to TIGR for supporting this workshop, and to Andrew Roger for the invitation to participate. This work was funded by the Genome Atlantic/Genome Canada Prokaryotic Genome Evolution and Diversity Project. The unpublished work described here was done in collaboration with Andrew Roger and Ed Susko. Members of the Statistical Evolutionary Bioinformatics group at Dalhousie University made many helpful suggestions.

References

Boussau, B., Karlberg, E. O., Frank, A. C., Legault, B.-A., and Andersson, S. G. E. (2004). Computational inference of scenarios for α-proteobacterial genome evolution. *Proceedings of the National Academy of Sciences*, 101(26):9722–9727.

Brochier, C., Forterre, P., and Gribaldo, S. (2004). Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the Methanopyrus kandleri paradox. *Genome Biology*, 5:R 17.

Bryant, D. and Moulton, V. (2002). Neighbornet: an agglomerative method for the construction of planar phylogenetic networks. *Lecture Notes in Computer Science*, 2452:375–391.

Chen, C.-Y., Wu, K.-M., Chang, Y.-C., Chang, C.-H., Tsai, H.-C., Liao, T.-L., Liu, Y.-M., Chen, H.-J., Shen, A. B.-T., Li, J.-C., Su, T.-L., Shao, C.-P.,

Lee, C.-T., Hor, L.-I., and Tsai, S.-F. (2003). Comparative genome analysis of Vibrio vulnificus, a marine pathogen. *Genome Research*, 13:2577–2587.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Doolittle, W. F., Boucher, Y., Nesbø, C. L., Douady, C. J., Andersson, J. O., and Roger, A. J. (2003). How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philosophical Transactions of the Royal Society of London Series B*, 358:39–58.

Felsenstein, J. (2004a). *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland, Massachusetts.

Felsenstein, J. (2004b). PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Fitch, W. M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155(3760):279–284.

Fitz-Gibbon, S. T. and House, C. H. (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research*, 27(21):4218–4222.

Gevers, D., Vanderpoele, K., Simillion, C., and Van de Peer, Y. (2004). Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends in Microbiology*, 12(4):148–154.

Gribaldo, S. and Philippe, H. (2002). Ancient phylogenetic relationships. *Theoretical Population Biology*, 61:391–408.

Gu, X. and Zhang, H. (2004). Genome phylogenetic analysis based on extended gene contents. *Molecular Biology and Evolution*, 21(7):1401–1408.

Huelsenbeck, J. P. and Crandall, K. A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics*, 28:437–466.

Huson, D. H. (1998). Splitstree: analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73.

Huson, D. H. and Steel, M. (2004). Phylogenetic trees based on gene content. *Bioinformatics*, 20(13):2044–2049.

Karev, G. P., Wolf, Y. I., Berezovskaya, F. S., and Koonin, E. V. (2004). Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evolutionary Biology*, 4:32.

Kennedy, S. P., Ng, W. V., Salzberg, S. L., Hood, L., and DasSarma, S. (2001). Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Research*, 11(10):1641–1650.

Koonin, E. V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Reviews Microbiology*, 1(2):127–136.

Korbel, J. O., Snel, B., Huynen, M. A., and Bork, P. (2002). SHOT: a web server for the construction of genome phylogenies. *Trends in Genetics*, 18(3):158–162.

Lake, J. A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. *Proceedings of the National Academy of Sciences*, 91:1455–1459.

Lake, J. A. and Rivera, M. C. (2004). Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Molecular Biology and Evolution*, 21(4):681–690.

Lockhart, P. J., Steel, M. A., Hendy, M. D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Molecular Biology and Evolution*, 11(4):605–612.

Martin, A. P. and Burg, T. M. (2002). Perils of paralogy: using HSP70 genes for inferring organismal phylogenies. *Systematic Biology*, 51(4):570–587.

Meyer, T. E., Cusanovich, M. A., and Kamen, M. D. (1986). Evidence against use of bacterial amino acid sequence data for construction of all-inclusive phylogenetic trees. *Proceedings of the National Academy of Sciences*, 83:217–220.

Mirkin, B. G., Fenner, T. I., Galperin, M. Y., and Koonin, E. V. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology*, 3(2).

Montague, M. G. and Hutchinson III, C. A. (2000). Gene content phylogeny of herpesviruses. *Proceedings of the National Academy of Sciences*, 97(10):5334–5339.

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., McDonald, L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., White, O., Salzberg, S. L., Smith, H. O., Venter, J. C., and Fraser, C. M. (1999). Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. *Nature*, 399:323–329.

Ng, W. V., Kennedy, S. P., Mahairas, G. G., Berquist, B., Pan, M., Shukla, H. D., Lasky, S. R., Baliga, N. S., Thorsson, V., Sbrogna, J., Swartzell, S., Weir, D., Hall, J., Dahl, T. A., Welti, R., Goo, Y. A., Leithauser, B., Keller, K., Cruz, R., Danson, M. J., Hough, D. W., Maddocks, D. G., Jablonski, P. E., Krebs, M. P., Angevine, C. M., Dale, H., Isenbarger, T. A., Peck, R. F., Pohlschroder, M., Spudich, J. L., Jung, K.-H., Alam, M., Freitas, T., Hou, S., Daniels, C. J., Dennis, P. P., Omer, A. D., Ebhardt, H., Lowe, T. M., Liang, P., Riley, M., Hood, L., and DasSarma, S. (2000). Genome sequence of Halobacterium species NRC-1. *Proceedings of the National Academy of Sciences*, 97(22):12176–12181.

Ochman, H. and Jones, I. B. (2000). Evolutionary dynamics of full genome content in Escherichia coli. *EMBO Journal*, 19(24):6637–6643.

Page, R. D. M. and Holmes, E. C. (1998). *Molecular evolution: a phylogenetic approach*. Blackwell Science, Oxford.

Rivera, M. C. and Lake, J. A. (2004). The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, 431:152–155.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Snel, B., Bork, P., and Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nature Genetics*, 21:108–110.

Studier, J. A. and Keppler, K. J. (1988). A note on the neighbor-joining algorithm of saitou and nei. *Molecular Biology and Evolution*, 5(6):729–731.

Swofford, D. L. (2003). Paup*. phylogenetic analysis using parsimony (*and other methods). Version 4 beta 10.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. (2003). The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science*, 278:631–637.

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*, 31(1):258–261.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., and Koonin, E. V. (2002). Genome trees and the Tree of Life. *Trends in Genetics*, 18(9):472–479.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L., and Koonin, E. V. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary Biology*, 1(8).

Zhang, H. and Gu, X. (2004). Maximum likelihood for genome phylogeny on gene content. *Statistical applications in genetics and molecular biology*, 3(1):article 31.