

## Phylogenetic Trees Based on Gene Content

Daniel H. Huson<sup>1</sup> and Mike Steel<sup>2</sup>

<sup>1</sup> Center for Bioinformatics (ZBIT), Tübingen University, Sand 14, 72076 Tübingen, Germany (To whom correspondence should be addressed).

<sup>2</sup> Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

**Abstract.** Comparing gene content between species can be a useful approach for reconstructing phylogenetic trees. In this paper we derive a maximum likelihood estimation of evolutionary distance between species under a simple model of gene genesis and gene loss. Using simulated data on a biological tree with 107 taxa (and on a number of randomly generated trees) we compare the accuracy of tree reconstruction using this ML distance measure to an earlier ad-hoc distance. We then compare these distance-based approaches to a character-based tree reconstruction method (Dollo parsimony) which seems well-suited to the analysis of gene content data. To simplify simulations, we give a formal proof of the well known “fact” that the Dollo parsimony score is independent of the choice of root. Our results show a consistent trend, with the character-based method and ML distance measure outperforming the earlier ad-hoc distance method.

*Keywords:* Gene content phylogeny, maximum likelihood estimator, Dollo parsimony

*Running head:* Gene content phylogeny

## 1 Introduction

As more and more whole genome sequences become available, there is growing interest in new methods that infer phylogenies from whole genome data [17], either directly from DNA comparisons [7], from comparisons of gene content [15, 6], from conserved gene synteny [13], or from other genomic features [11].

In [15], Snel *et al.* present a phylogenetic tree for 13 genomes based on a simple measure of shared gene content that corresponds quite well to other phylogenies based on the comparison of 16S rRNA sequence data. The accuracy of their method was subsequently investigated by [9]. Simultaneously, a tree was published based on a parsimony analysis of the gene content of 11 genomes [6].

The aims of our paper are three-fold. First, we investigate a simple model of gene content evolution involving gene genesis and gene loss, but not gene transfer, and derive from it a new distance measure that is the maximum likelihood estimator under the given model.

Secondly, we discuss a specific flavor of maximum parsimony tree construction, *Dollo parsimony*, that is an appropriate approach under this model [18]. In this context we prove that the Dollo parsimony score is independent of the choice of root, a result that is useful beyond the context of gene content evolution.

Thirdly, based on this simple model of gene content evolution, we have undertaken an experimental study to compare the performance of all three tree construction methods mentioned above.

The result of this study is that Dollo parsimony is generally the most accurate method, but it is often very closely matched by the maximum-likelihood distance estimator, which consistently outperforms the simple gene content distance.

Implementations of our simulator and tree construction software are freely available from: [www-ab.informatik.uni-tuebingen.de/software/genecontent/welcome\\_en.html](http://www-ab.informatik.uni-tuebingen.de/software/genecontent/welcome_en.html).

## 2 Methods

### 2.1 A Simple Model of Gene Content Evolution

Consider the following model. A *genome*  $G$  will consist of a set of genes, which we will regard as formal labels (gene identifiers). We assume that genomes evolve according to a constant-birth, proportional-death Markov process. That is, at each instant each gene in  $G$  can be independently deleted, at intensity rate  $\mu$ , or  $G$  can acquire a new gene (gene genesis) independently, at intensity rate  $\lambda > 0$ . This model has some similar features to the model described in [9]. However, we do not model horizontal gene transfer or selection pressure. As our goal is mathematical tractability, the model that we discuss is less sophisticated than, for example, a *Birth, Death and Innovation Model* that explicitly models domain duplication, deletion and innovation [8].

Under the constant birth, proportional-death model, let  $G(t)$  denote the random genome that results from this process after duration  $t$ , and let  $l(t) = |G(t)|$ . Let  $m := \frac{\lambda}{\mu}$ . From the theory of Markov processes ([4], (p. 414)) we have that:

- (i) Independent of  $G(0)$ , the genome size  $l(t)$  converges to a Poisson distribution with mean  $m$  as  $t$  grows.
- (ii) If  $l(0)$  is chosen according to this equilibrium Poisson distribution, then the process  $G(t)$  is a time-reversible Markov process.

As usual, a *phylogenetic tree*  $\mathcal{T}$  on a set  $X$  of taxa is a tree, with *vertex set*  $V(\mathcal{T})$  and *edge set*  $E(\mathcal{T})$ , whose internal vertices all have degree at least 3, leaves are bijectively labeled by elements of  $X$  and each edge  $e$  is labeled by a positive number  $\tau(e)$ . Such a tree  $\mathcal{T}$  is called *rooted*, if there is a specified *root*  $\rho$  for  $\mathcal{T}$ , which can be any internal vertex of  $\mathcal{T}$  or the midpoint of some edge of  $\mathcal{T}$ . If the edges of  $\mathcal{T}$  are assigned lengths we obtain a *additive* distance on  $X$  by setting the distance between any pair  $x$  and  $y$  to be the sum of the lengths of the edges on the path in  $\mathcal{T}$  connecting  $x$  and  $y$  (for further background see [14]).

We assume that the genomes assigned to each vertex of  $\mathcal{T}$  are the result of the Markov process (in equilibrium) described above. This model can be described by just three parameters  $(\mathcal{T}, \tau, m)$  where  $\mathcal{T}$  is a rooted (or unrooted) phylogenetic tree,  $\tau$  is a set of edge lengths for  $\mathcal{T}$  and  $m (= \lambda/\mu)$  is the expected number of genes at any vertex of  $\mathcal{T}$ .

Such a model  $M = (\mathcal{T}, \tau, m)$  assigns a set  $G(v)$  of genes to every vertex  $v$  of  $\mathcal{T}$  as follows: First  $G(\rho)$  is some set of chosen genes whose size is taken from the equilibrium distribution for the model - namely a Poisson distribution with mean  $m$ .

In a depth-first traverse of the tree, let  $v$  denote a parent vertex,  $w$  a child vertex and  $e$  the edge connecting the two. Assume that  $G(v)$  has already been assigned and we want to assign  $G(w)$ . Then  $G(w)$  is chosen according to the above model as  $G(t)$  conditional on  $G(0) = G(v)$ . That is, each gene  $g \in G(v)$  survives along  $e$  and is present in  $G(w)$  with probability  $e^{-\mu\tau(e)}$ . Furthermore, new genes (not present anywhere else in the tree) are born along  $e$  at rate  $\lambda$  and for duration  $\tau(e)$ .

Note that, in particular, we thereby assign a set of genes to each taxon  $x \in X$  and define  $G(x) = G(\nu(x))$ , where  $\nu(x)$  is the leaf with label  $x$ .

## 2.2 Maximum Likelihood distance estimation

Now suppose we have two genomes  $G_1$  and  $G_2$  which have evolved from an ancestral genome,  $G(0)$  according to this process, which acted for duration  $t_1$  to form  $G_1$  and  $t_2$  to form  $G_2$ . We assume that the process is in equilibrium (and thus a reversible Markov process). We wish to calculate  $t_{ML} := t_1 + t_2$  to maximize the joint probability of observing  $G_1$  and  $G_2$  under this model. Since the genes are essentially abstract markers in this model, all that is relevant in any probability calculations for estimating  $t_{ML}$  are the number of genes in  $G_1$  and  $G_2$  and the number of genes common to both  $G_1$  and  $G_2$ . Thus, for  $i = 1, 2$ , let  $l_i = |G_i|$ , and let  $l_{12} = |G_1 \cap G_2|$ .

**Theorem 1.**

$$t_{ML} = -\frac{1}{\mu} \log \left( \frac{\beta + \sqrt{\beta^2 + 4\alpha_{12}}}{2} \right),$$

where  $\alpha_i := \frac{l_i}{m}$ ,  $\alpha_{12} = \frac{l_{12}}{m}$ , and  $\beta := 1 + \alpha_{12} - \alpha_1 - \alpha_2$ .

*Proof.* Recalling that  $l(t) = |G(t)|$ , let

$$p_{ij}(t) = \mathbb{P}(l(t+s) = j | l(s) = i).$$

By the Markov assumption this quantity does not depend on  $s$ . Also, let  $p_j(t) = p_{0j}(t)$ , which is the probability that  $j$  new genes will arise in a genome over duration  $t$ . The following expression for  $p_{0j}(t)$  is given by standard Markov process theory - see for example [4] (equation (11.10) as described on pp. 434-435) - which shows that  $p_j(t)$  has a Poisson distribution with mean  $m(1 - e^{-\mu t})$ , that is:

$$p_j(t) = \frac{m^j}{j!} (1 - e^{-\mu t})^j \exp(-m(1 - e^{-\mu t})). \quad (1)$$

Let us write  $\mathbb{P}(G_1, G_2|T = t)$  to denote the joint probability of generating genomes of length  $l_1$  and  $l_2$ , that share  $l_{12}$  genes, if the genomes have been separated by duration  $t$ . Similarly, we write  $\mathbb{P}(G_2|G_1, T = t)$  to denote the associated conditional probability, and  $\mathbb{P}(G_1)$  to denote the probability of a genome of length  $l_1$ . Then, by elementary probability theory,

$$\mathbb{P}(G_1, G_2|T = t) = \mathbb{P}(G_2|G_1, T = t) \cdot \mathbb{P}(G_1|T = t).$$

Now,  $\mathbb{P}(G_1|T = t) = \mathbb{P}(G_1) = \frac{m^{l_1}}{l_1!} e^{-m}$  and so

$$\mathbb{P}(G_1, G_2|T = t) = \mathbb{P}(G_2|G_1, T = t) \cdot \frac{m^{l_1}}{l_1!} e^{-m} \quad (2)$$

To calculate  $\mathbb{P}(G_2|G_1, T = t)$  we use the fact that the model is a reversible Markov process and so we can consider  $G_2$  as evolving from  $G_1$ . To obtain  $G_2$  from  $G_1$  over duration  $t$  requires that (i) precisely  $l_1 - l_{12}$  of the  $l_1$  genes in  $G_1$  must be eliminated and (ii) one must have precisely  $l_2 - l_{12}$  additional genes created. Since the genes are treated independently and according to an identical process, the probability of event (i) is given by a binomial distribution with  $l_1$  independent trials, and for which the probability of success on each trial is  $1 - e^{-\mu t}$ . Furthermore, the probability of event (ii) is, by definition,  $p_{l_2 - l_{12}}(t)$ . Consequently,

$$\mathbb{P}(G_2|G_1, T = t) = \frac{l_1!}{l_{12}!(l_1 - l_{12})!} (e^{-\mu t})^{l_{12}} (1 - e^{-\mu t})^{l_1 - l_{12}} \cdot p_{l_2 - l_{12}}(t). \quad (3)$$

Thus if we let

$$L(t) := e^{-\mu t l_{12}} \cdot (1 - e^{-\mu t})^{l_1 + l_2 - 2l_{12}} \exp(-m(1 - e^{-\mu t}))$$

then, combining (1), (2), and (3), we have:

$$\mathbb{P}(G_1, G_2|T = t) = c \cdot L(t)$$

for a positive constant  $c (= \frac{m^{l_1 + l_2 + l_{12}} e^{-m}}{l_{12}!(l_1 - l_{12})!(l_2 - l_{12})!})$  that is independent of  $t$ . Maximizing  $\mathbb{P}(G_1, G_2|T = t)$  corresponds to maximizing  $L(t)$ . Solving the equation  $\frac{dL(t)}{dt} = 0$  for  $t = T_{ML}$  by routine techniques from differential calculus gives the claimed solution.  $\square$

Theorem 1 allows us to define an additive evolutionary distance between genomes by setting:

$$d_{G_1, G_2} = -\log \left( \frac{\beta + \sqrt{\beta^2 + 4\alpha_{12}}}{2} \right). \quad (4)$$

Note that this distance estimate does not require knowledge of  $\lambda$  or  $\mu$  separately, and although it does involve  $m$  (the expected number of genes in a genome) in practice this may be estimated by averaging the number of genes across the genomes of the taxa being compared.

### 2.3 Dollo Parsimony

Dollo Parsimony is a method for reconstructing phylogenetic trees from binary sequences. Named after Louis Dollo, who argued that it is harder to evolve a complex feature than it is to lose it, the approach was suggested in [10] and further analyzed in [3]. Although the method has traditionally

been applied to more classical types of evolutionary data (such as morphological characters), it has been applied to molecular studies involving restriction site data [2].

The method assumes that the transition event  $0 \rightarrow 1$  (which in our setting corresponds to gene genesis) can occur at most once in the rooted phylogenetic tree  $\mathcal{T}$  that describes the evolution of the genomes. Gene losses – represented by the transition  $1 \rightarrow 0$  – can occur multiple times. The rooted tree  $\mathcal{T}$  is scored by the total number of transitions that occur (0 to 1 and 1 to 0) in the tree, across all the genes. The ancestral gene compositions (at the interior vertices and root of  $\mathcal{T}$ ) are chosen so as to minimize this score. Finally the tree that minimizes this Dollo score is selected (we define this more precisely shortly).

We pause to describe the binary sequences that are naturally associated with gene content data. Suppose we have a set of taxa  $X$  and a set  $G(x)$  of genes for each taxon  $x \in X$ . Let  $\mathcal{G} = \cup_{x \in X} G(x)$  be the set of all mentioned genes and assume that  $\mathcal{G} = \{g_1, \dots, g_k\}$ . To each taxon  $x \in X$  we assign a  $\{0, 1\}$ -sequence  $S_x$  of length  $k$  by setting  $S_x[i] = 1$  if and only if  $i \in G(x)$ . Thus, each gene  $g_i \in \mathcal{G}$  can be regarded as a *binary character*, that is a function  $\chi : X \rightarrow \{0, 1\}$  defined by setting  $\chi(x) = 1$  if  $g_i \in G(x)$ , otherwise  $\chi(x) = 0$ .

Returning to the general setting of binary characters, given  $\chi : X \rightarrow \{0, 1\}$  together with a rooted phylogenetic  $X$ -tree, consider extensions  $\bar{\chi} : V(\mathcal{T}) \rightarrow \{0, 1\}$  for which (i) there is at most one edge  $(u, v)$  in  $\mathcal{T}$  with  $\bar{\chi}(u) = 0$  and  $\bar{\chi}(v) = 1$ , and (ii) which minimizes the number  $m$  of edges  $(u, v)$  for which  $\chi(u) \neq \chi(v)$  - this value of  $m$  we call the *DP-score* of  $\chi$  on  $\mathcal{T}$ , and any such extension  $\bar{\chi}$  we call a *minimal DP-extension* of  $\chi$  on  $\mathcal{T}$ . Given a sequence of characters a *DP-tree* is any rooted phylogenetic  $X$ -tree that minimizes the sum of the DP-scores of the characters in the sequence.

It might be expected that the Dollo scoring criterion would favor certain placements of the root in a tree. However, it is part of the “folklore” (see for example [16]) that this is not so. Thus the method effectively constructs an unrooted phylogenetic  $X$ -tree which makes it comparable with the distance-based tree reconstruction methods in our study. We now provide a formal proof of this result, which also shows that the DP-score of a character on a rooted phylogenetic tree can be very easily described and computed.

Given an unrooted phylogenetic  $X$ -tree  $\mathcal{T}$  and  $x, y \in X$ , let  $p(x, y) := p_{\mathcal{T}}(x, y)$  denote the set of vertices on the path in  $\mathcal{T}$  connecting  $x$  and  $y$ . For a character  $\chi : X \rightarrow \{0, 1\}$ , and a phylogenetic  $X$ -tree  $\mathcal{T}$ , let

$$V(\chi, \mathcal{T}) = \{v \in V(\mathcal{T}) : \exists x, y \in X : \chi(x) = \chi(y) = 1, v \in p(x, y)\},$$

and let

$$\Delta(\chi, \mathcal{T}) = |\{\{u, v\} \in E(\mathcal{T}) : |\{u, v\} \cap V(\chi, \mathcal{T})| = 1\}|.$$

Given a rooted phylogenetic  $X$ -tree,  $\mathcal{T}$ , let  $\mathcal{T}^{-\rho}$  denote the phylogenetic  $X$ -tree obtained from  $\mathcal{T}$  by *suppressing* the root vertex  $\rho$ , that is, if  $\rho$  has degree two, then we delete it from the tree and join the two adjacent vertices by a new edge, whereas, if  $\rho$  has degree more than two, then we simply stop regarding  $\rho$  as the root.

Let  $l_{DP}(\chi, \mathcal{T})$  be the DP-score of  $\chi$  on  $\mathcal{T}$ .

**Theorem 2.** *For a rooted phylogenetic  $X$ -tree  $\mathcal{T}$  and a character  $\chi : X \rightarrow \{0, 1\}$ , we have*

$$l_{DP}(\chi, \mathcal{T}) = \Delta(\chi, \mathcal{T}^{-\rho}).$$

Thus  $l_{DP}(\chi, \mathcal{T})$  is independent of the placement of a root. Furthermore, there are at most two minimal DP-extensions of  $\chi$  and these differ only on the assignment of states to the root vertex. For every other vertex  $v$  of  $\mathcal{T}$ , the optimal Dollo assignment to  $v$  is 1 if and only if  $v \in V(\chi, \mathcal{T}^{-\rho})$  and 0 otherwise.

*Proof.* Suppose  $\bar{\chi}$  is a minimal DP-extension of  $\chi$  on  $\mathcal{T}$ . We may regard  $V(\mathcal{T}^{-\rho})$  and  $V(\chi, \mathcal{T}^{-\rho})$  as subsets of  $V(\mathcal{T})$ . Let  $v$  be a vertex of  $V(\mathcal{T})$ . Then there are three possibilities, either:

- (i)  $v \in V(\chi, \mathcal{T}^{-\rho})$ , in which case we will show that  $\bar{\chi}(v) = 1$ , or
- (ii)  $v \in V(\mathcal{T}^{-\rho}) - V(\chi, \mathcal{T}^{-\rho})$ , in which case we will show that  $\bar{\chi}(v) = 0$ , or else
- (iii)  $v \notin V(\mathcal{T}^{-\rho})$ .

The claim accompanying case (i) follows from the requirement that the transition  $0 \rightarrow 1$  can occur at most once in  $\mathcal{T}$  under the Dollo requirement.

Regarding case (ii), to establish the accompanying claim, we apply induction on the edge distance  $d$  from  $v$  to its most distant descendant leaf. If  $d = 0$  then  $v$  is a leaf, and so  $\bar{\chi}(v) = \chi(v) = 0$ . For the induction step consider the vertices  $v_1, v_2, \dots, v_k$  ( $k \geq 2$ ) that are adjacent to, but descended from  $v$ . Note that all but at most one of these vertices must lie in  $V(\mathcal{T}^{-\rho}) - V(\chi, \mathcal{T}^{-\rho})$ , otherwise  $v$  would lie in  $V(\chi, \mathcal{T}^{-\rho})$ . If  $\{v_1, v_2, \dots, v_k\} \subseteq V(\mathcal{T}^{-\rho}) - V(\chi, \mathcal{T}^{-\rho})$  then by the induction hypothesis  $\bar{\chi}(v_i) = 0$  for all  $i \in \{1, \dots, k\}$  and so, since  $k \geq 2$ ,  $\bar{\chi}(v) = 0$  is the only possible assignment (as  $\bar{\chi}$  is a minimal DP-extension of  $\chi$ ). On the other hand, suppose one of the elements – say  $v_1$  – lies in  $V(\chi, \mathcal{T}^{-\rho})$ . In that case, if  $v = \rho$ , the root vertex of  $\mathcal{T}$ , then  $k \geq 3$  (if  $v$  had degree 2 then  $v \notin V(\mathcal{T}^{-\rho})$ ) and once again applying the induction hypothesis to  $v_1, \dots, v_k$  we have that  $\bar{\chi}(v) = 0$ . If  $v$  is not the root vertex, then there exists an adjacent vertex  $w$  that is an immediate ancestor of  $v$ . Consider the other subtree of  $\mathcal{T}$  that descends from  $w$ . None of the leaves of this tree can have a  $\chi$  value of 1 for otherwise this would force  $v \in V(\chi, \mathcal{T}^{-\rho})$ . Thus, the assignment  $\bar{\chi}(w) = \bar{\chi}(v) = 0$  is the only possible assignment since  $\bar{\chi}$  is a minimal DP-extension of  $\chi$ . In all these cases then the induction step holds.

Finally, consider Case (iii). In this case  $v = \rho$  and  $v$  has degree 2. Let  $v_1, v_2$  denote the two vertices of  $\mathcal{T}$  adjacent to  $\rho$ . If (a)  $v_1$  and  $v_2$  are both in  $V(\chi, \mathcal{T}^{-\rho})$  then  $\bar{\chi}(\rho) = 1$ . If (b) neither of  $v_1, v_2$  are in  $V(\chi, \mathcal{T}^{-\rho})$  then by part (ii),  $\bar{\chi}(v_1) = \bar{\chi}(v_2) = 0$  and so  $\bar{\chi}(\rho) = 0$ . If (c) exactly one of  $v_1, v_2$  are in  $V(\chi, \mathcal{T}^{-\rho})$  then by cases (i) and (ii)  $\bar{\chi}$  assigns state 0 to one vertex and 1 to the other, and so there are two equally parsimonious assignments to  $\rho$  – either 0 or 1. Note that in all three cases (a)–(c) the total number of edges between  $\rho, v_1$  and  $v_2$  on which there is a transition under  $\bar{\chi}$  is 1, when  $|\{v_1, v_2\} \cap V(\chi, \mathcal{T}^{-\rho})| = 1$ , otherwise the total is 0.

Summarizing, the value of  $\bar{\chi}$  is determined on all the vertices of  $\mathcal{T}$ , except perhaps the root, for which there are two possible assignments precisely when the root has exactly two adjacent vertices that receive different states by  $\bar{\chi}$ . Also, it is clear from considering the various cases above that the total number of edges that receive different states by  $\bar{\chi}$  is precisely  $\Delta(\chi, \mathcal{T})$  as claimed.  $\square$

### 3 Results

Given two genomes  $G_1$  and  $G_2$ . Snel *et al* [15] suggested using the following simple distance measure between genomes, based on the shared gene content for phylogenetic reconstruction

$$d_{G_1, G_2} = 1 - \frac{|G_1 \cap G_2|}{\min\{|G_1|, |G_2|\}},$$

which we will refer to as the *shared genes distance*. Applying this idea to the genomes of 13 different unicellular species, they obtained a Neighbor-Joining tree [12] which they argued is biologically reasonable and correlates with other published phylogenies.

Similarly, using the ML distance estimation described in Section 2.2, we can obtain a phylogenetic tree by applying a method such as Neighbor-Joining to the distance matrix.

Thus, we have described three different ways of obtaining a phylogeny from gene content data:

- by the *shared genes distance* approach due to Snel *et al.*,
- using our new *ML distance estimator* method (given by equation (4)) or
- applying *Dollo parsimony*.

Given the model of evolution described in Section 2.1, how well do these three different approaches perform? To address this question we have undertaken a simulation study in which we evolved sets of genes along a given tree and then applied the three methods on the sets of genes observed at the leaves of the tree, in an attempt to recover the original tree.

We ran simulations on 15 different trees, ranging in size from 50 – 107 taxa. All simulations produced similar results. In the following, we exemplify these simulations by describing the study done on a binary tree  $\mathcal{T}_0$  on  $n = 107$  taxa which comes from the biological literature [1].

The edge lengths on the tree ranged from 1 to 189, with a median of 21, and were interpreted as time. A discrete approximation of the continuous time Markov process was used. For this the probability  $p_{genesis}$  of gene genesis in one unit of time was varied from 0.05 – 0.80, whereas the probability  $p_{loss}$  for any given gene to be lost in one unit of time was set to  $\frac{p_{genesis}}{m}$ , where  $m$  is the average size of the (root) genome; this ensures that the process is in equilibrium and the expected genome size remains constant with time. We performed 50 runs from each choice of  $p_{genesis}$ . In each run, the genome size at the root of the tree was chosen from its equilibrium (Poisson) distribution with mean = 1000 and stddev =  $\sqrt{1000}$  (in practice a Gaussian distribution was used since the Poisson converges to this distribution for large  $m$ ).

For each of the three phylogenetic methods and for each of the values of  $p_{gain}$ , we computed the average *accuracy score* over  $N = 50$  runs as:

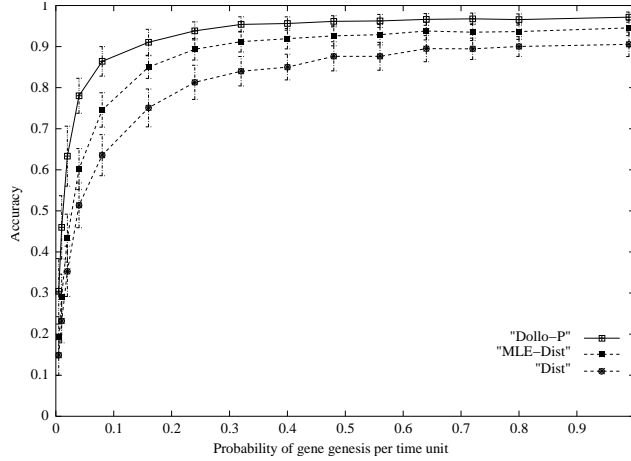
$$score = \frac{(n - 3) - \frac{FP}{N}}{n - 3},$$

where  $FP$  denotes the total number of false positive “splits” observed in all  $N$  runs, where a *split* is bipartition of the set of taxa corresponding to an edge of a tree. Note that this score is 1, if the tree topology is completely correct, and 0, if completely wrong.

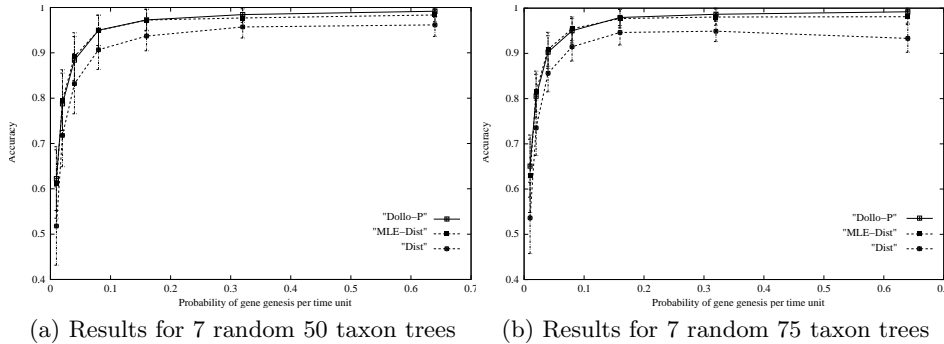
To perform this study, we developed our own simulator software and implementations of both distance methods, and we used the Phylip [5] implementations of Neighbor-Joining (the `neighbor` program) and Dollo parsimony (the `dollop` program), in both cases using the programs’ default settings.

As reported in Figure 1, our simulations on a 107 taxon tree indicated that Dollo parsimony and the ML distance estimator consistently shows higher accuracy than the shared genes distance approach of Snel *et al.* Although in Figure 1 Dollo parsimony is clearly performing better than the ML distance estimator, this trend was less evident in the simulations on the other 14 trees, see Figure 2.

In summary, while the work in [15] and [6], has established that shared gene content is useful for inferring phylogenies, our study indicates that using more sophisticated techniques such as Dollo



**Fig. 1.** Here we exemplify the performance of the three gene content-based phylogenetic methods on data simulated on a binary tree with 107 taxa. For all three methods “Dollo parsimony” (Dollo-P), “ML distance estimator method” (MLE-Dist) and “shared genes distance” approach (Dist) we plot the accuracy as a function of  $p_{gain}$ . Each data point represents the performance averaged over 50 independent runs and the error bars span one standard deviation below and above the mean.



**Fig. 2.** More simulation results comparing the performance of Dollo-P, MLE-Dist and Dist as a function of  $p_{gain}$ . We report the mean accuracy and standard deviation on seven randomly generated tree topologies, (a) on 50 taxa, and, (b) on 75 taxa, using 10 independent runs per tree.

parsimony or the ML distance estimator should provide more accurate trees. In a forthcoming paper we intend to demonstrate the utility of these new techniques when applied to a range of real data sets.

## 4 Acknowledgments

We thank the New Zealand Institute for Mathematics and its Applications (NZIMA) for support under the *Phylogenetic Genomics* programme.

## References

1. J.R. Cole, B. Chai, T.L. Marsh, R.J. Farris, Q. Wang, S.A. Kulam, S. Chandra, D.M. McGarrell, T.M. Schmidt, G.M. Garrity, and J.M. Tiedje. The ribosomal database project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res*, 31(1):442–3, 2003.

2. R.W. DeBry and N.A. Slade. Cladistic analysis of restriction endonuclease cleavage map data within a maximum likelihood framework. *Systematic Zoology*, 31:21–34, 1985.
3. J. S. Farris. Phylogenetic analysis under Dollo’s law. *Systematic Zoology*, 26:77–88, 1977.
4. W. Feller. *An introduction to probability theory and its applications*, volume 1. John Wiley and Sons, Inc., New York, 2nd edition, 1950.
5. J. Felsenstein. PHYLIP – phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.
6. S. T. Fitz-Gibbon and C. H. House. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research*, 27(21):4218–4222, 1999.
7. S. R. Henz, A. F. Auch, D. H. Huson, K. Nieselt-Struwe, and S. C. Schuster. Whole genome-based prokaryotic phylogeny. Short paper in ECCB, 2003.
8. G. P. Karev, Y. I. Wolf, and E. V. Koonin. Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics*, 19(15):1889–1900, 2003.
9. V. Kunin and C.A. Ouzounis. GenTRACE- reconstruction of gene content of ancestral species. *Bioinformatics*, 19(11):1412–1416, 2003.
10. W.J. Le Quesne. The uniquely evolved character and its cladistic application. *Systematic Zoology*, 23:513–517, 1974.
11. Antonis Rokas and Peter W.H. Holland. Rare genomic changes as a tool for phylogenetics. *Trends in Ecology and Evolution*, 15(11):454–459, 2000.
12. N. Saitou and M. Nei. The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.
13. D. Sankoff and J.H. Nadeau. Conserved synteny as a measure of genomic distance. *Discr. Appl. Math.*, 71:247–257, 1996.
14. C. Semple and M.A. Steel. *Phylogenetics*. Oxford University Press, 2003.
15. B. Snel, P. Bork, and M. A. Huynen. Genome phylogeny based on gene content. *Nature*, 21:108–110, 1999.
16. D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. Chapter 11: Phylogenetic inference. In D. M. Hillis, C. Moritz, and B. K. Mable, editors, *Molecular Systematics*, pages 407–514. Sinauer Associates, Inc., 2nd edition, 1996.
17. Y. I. Wolf, I. B. Rogozin, N. V. Grishin, and E. V. Koonin. Genome trees and the Tree of Life. *TRENDS in Genetics*, 18(9):472–479, 2002.
18. Y. I. Wolf, I. B. Rogozin, N. V. Grishin, R. L. Tatusov, and E. V. Koonin. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evolutionary Biology*, 1:8, 2001.