A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer

Miklós Csűrös[1] and István Miklós[2]

[1] Department of Computer Science and Operations Research, Université de Montréal
C.P. 6128, succ. Centre-Ville, Montréal, Québec, Canada, H3C 3J7
csuros@iro.umontreal.ca
[2] Department of Plant Taxonomy and Ecology, Eötvös Lóránd University,
1117 Budapest, Pázmány Péter Sétány 1/c, Hungary.
miklosi@ramet.elte.hu

Abstract. We introduce a Markov model for the evolution of a gene family along a phylogeny. The model includes parameters for the rates of horizontal gene transfer, gene duplication, and gene loss, in addition to branch lengths in the phylogeny. The likelihood for the changes in the size of a gene family across different organisms can be calculated in $O(N + hM^2)$ time and $O(N + M^2)$ space, where N is the number of organisms, h is the height of the phylogeny, and M is the sum of family sizes. We apply the model to the evolution of gene content in Proteobacteria using the gene families in the COG (Clusters of Orthologous Groups) database.

1 Introduction

At this time, 294 microbial genomes have been sequenced, and that figure is expected to soon double (this in addition to 19 complete eukaryotic genomes, see `http://www.ncbi.nlm.nih.gov/Genomes/`). These numbers continue to grow exponentially with advances in technology and expertise [1]. The wealth of genome sequence data has already caused a revolution in molecular evolution methods [2, 3]. A few years ago, scientific studies had to focus on nucleotide-level differences between orthologous genes, mainly because of the technical and financial limitations on DNA sequence collection. With increasing amounts of whole genome information, however, it becomes possible to analyze genome-scale differences between organisms, and to identify the evolutionary forces responsible for these changes. In particular, sizes of gene families can be compared, allowing us to better understand adaptive evolutionary mechanisms and organismal phylogeny. Several studies suggest that gene content may carry sufficient phylogenetic signal for the construction of evolutionary trees [4–13]. Comparative analyses of genome-wide protein domain content [7, 14, 15] have also provided important insights into evolution. Gene content and similar features have been used to construct viral [16, 17], microbial [4, 5, 12], and universal trees [6, 14, 18]. Comparative gene content analysis is also used to estimate ancestral genome composition [19, 20]. The presence-absence pattern of homologs in different organisms, the so-called phyletic pattern [21, 22], provides clues about gene function [23] and the evolution of metabolic pathways [20].

A number of processes shape the gene content of an organism. New genes may be created by duplication of an existing gene, horizontal transfer from a different lineage, and rarer events such as gene fusion and fission [19]. It has been widely debated how the extent of horizontal gene transfer (HGT) compares to vertical inheritance [18, 19, 24–28]. It is clear that horizontal gene transfer plays a major role in microbial evolution [29], but there is still need for adequate mathematical models in which that role can be measured.

We introduce a probabilistic model for the evolution of gene content along a phylogeny. Our model accounts for gene duplication, gene loss and horizontal transfer. We consider the evolution of the size of a gene family, where the different processes add new genes to the family or erase members of it, and arrive at the family sizes observed at the terminal taxa. We describe an algorithm that calculates the likelihood of gene family sizes in different organisms, given an evolutionary tree. The algorithm computes the likelihood of family sizes in $O(N+M^2h)$ time where M is the total number of genes in the family, N is the number of genomes, and h is the height of the tree. Note that the tree height is at most linear in N, and on average, it is $O(\sqrt{N})$ or $O(\log N)$ for uniform or Yule-Harding distribution of random trees.

To our knowledge, no tractable stochastic model has yet been introduced that simultaneously accounts for horizontal transfer, gene loss, and duplication. These processes cannot be modeled by using only two parameters: whereas the intensity of gene loss and duplication depend on the size of a gene family, the rate of horizontal transfer has a constant component. Among other applications, a model that accounts for duplication and transfer is useful for analyzing the evolution of metabolic networks [30]: do new paths evolve by gene duplication and adaptive selection, or by accommodating genes with new functions via horizontal gene transfer?

A few probabilistic models were proposed for gene content evolution, which are less general than ours. Most studies use stochastic models with two parameters. Huson and Steel [11] analyzed a two-parameter model that accounts for gene loss and horizontal transfer but not for gene duplication. They derived a distance measure based on gene family sizes using likelihood maximization arguments. They further showed that traditional scores for shared gene content [5] are not as suitable for phylogeny reconstruction as either Dollo parsimony or their own distance function. Gu and Zhang [12] relied on a model that includes gene loss and gene duplication but no other modes of gene genesis, and assumes identical rates across different branches. They showed how gene family sizes can be used to define additive distances in such a model. Interestingly enough, the data can be reduced to a three-letter alphabet for the purposes of distance calculations: only 0, 1 or "many" homologs per family need to be counted. The distance metric relies on estimates of the rate parameters, which are obtained through likelihood optimization. Hahn et al. [31] developed an alternative likelihood-based approach for the same two-parameter model with constant rates across lineages. Karev et al. developed a rich probabilistic model of gene content evolution in a series of papers [32–34]. The model explains the distribution of gene

family sizes found in different organisms. It is, however, too general for exact detailed calculations, and for likelihood computations in particular. Our likelihood algorithm is also notable for its computational efficiency. For instance, the likelihood calculations of [31] in a two-parameter model take cubic time in M, and involve the evaluation of infinite sums that are truncated heuristically.

Not all comparative studies of gene content rely on gene family sizes. A frequently employed approach is to measure shared gene content [5, 6, 8–10] by identifying orthologs between each pair of genomes. Pairwise scores of shared gene content can be analyzed using distance-based methods of phylogeny construction or other clustering techniques. Lake and Rivera [13] proposed an improved technique of assessing shared gene content: for each genome, the presence and absence of homologs are marked with respect to genes of a reference genome. The presence-absence marks are encoded in a binary sequence for every genome. The sequences are used to compute a pairwise distance matrix using standard methods of phylogeny construction. Finally, a number of studies rely on families of homologous genes across many organisms, and record the absence or presence of each family in the genomes [4, 7, 24, 35]. The resulting absence-presence data are further analyzed with traditional parsimony or distance-based methods. Some specialized parsimony methods were purposely devised to analyze absence-presence data [20, 36] for gene families. Our work is concerned with the actual numbers of paralogs within the gene families, which give an even richer signal for evolutionary analyses [11, 19, 31].

The paper is organized in the following manner. Section 2 introduces our stochastic model of gene content evolution, and describes formulas for computing various associated probabilities, including likelihood. The formulas are used in an algorithm described in Section 3. Section 4 describes our initial experiments in modeling gene content evolution in 51 proteobacteria and 3555 gene families from the database of Clusters of Orthologous Groups (COGs) [22]. Section 5 concludes the paper.

2 Mathematical model

Let T be a phylogenetic tree over a set of organisms S. The tree T is a rooted tree with node set $V(T)$ and edge set $E(T)$, in which leaves are bijectively labeled with elements of S. Non-leaf nodes have at least two children. Every edge e has a length $t_e > 0$. We are interested in modeling the evolution of a gene family. The family size changes along the edges: genes may be duplicated, lost, or gained from an unknown source. We model the evolution of *gene counts* (family size) at the tree nodes: the gene count at every node $u \in V(T)$ is a random variable $\chi(u)$ that can take non-negative integer values. In addition to its length, each edge is equipped with a *duplication rate* λ, a *loss rate* μ, and a *transfer rate* κ. The loss rate accounts for all possible mechanisms of gene loss, including deletion and pseudogenization. The transfer rate accounts for processes of gene genesis, including HGT from another lineage in the same tree, or HGT from an unknown

organism. The tree topology, the edge lengths and rates determine the joint distribution of the gene counts.

In our model, the evolution of the gene counts on a branch follows a linear birth-and-death process [37] parametrized by λ, κ, and μ. Let $\{X(t) : t \geq 0\}$ denote the continuous-time Markov process formed by the gene counts along an edge uv: $\chi(u) = X(0)$ and $\chi(v) = X(t_{uv})$. The transition probabilities of the process are the following:

$$\begin{aligned}
\mathbb{P}\Big\{ X(t+\epsilon) = n+1 \;\Big|\; X(t) = n \Big\} &= \Big(\kappa + n\lambda\Big)\epsilon + o(\epsilon) \\
\mathbb{P}\Big\{ X(t+\epsilon) = n-1 \;\Big|\; X(t) = n \Big\} &= n\mu\epsilon + o(\epsilon) \\
\mathbb{P}\Big\{ |X(t+\epsilon) - n| > 1 \;\Big|\; X(t) = n \Big\} &= o(\epsilon).
\end{aligned}$$

In other words, every existing gene produces an offspring through duplication with an intensity of λ, or disappears with an intensity of μ, and new genes are acquired with an intensity of κ, independently from the number of existing genes.

Remark. For simplicity of notation, we impose the same rates across all edges throughout the paper. Nevertheless, the presented method accommodates branch-dependent rates in a straightforward manner.

The histories of individual genes on an edge form a *Galton-Watson* forest, see Figure 1. The figure illustrates a scenario where the gene count changes from three to five. The gene count at the child node is the result of many duplication, transfer and loss events. The change involves three horizontally transferred genes, from among which one survives, another one does not, and the third one produces two surviving paralogs.

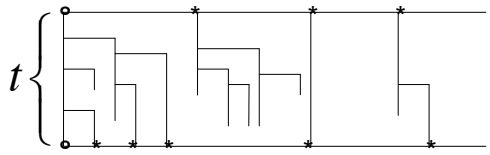


Fig. 1. Galton-Watson forest showing the evolution of genes in the same family along a tree edge. The top line represents the ancestral genome with three genes; the bottom line represents the descendant genome, in which there are five family members. Symbol ∘ represents the source from which genes might be transferred horizontally, symbols ⋆ represent paralogous genes in the genome at the beginning and the end of the investigated time span t. Each ∘ or ⋆ in the ancestral genome is the root of a Galton-Watson tree. Note that the physical order of genes is immaterial: here they are simply drawn next to each other for clarity.

While it is not too difficult to calculate the probabilities for any particular gene count on a branch (see §2.1), the likelihood L of observed gene counts at

the leaves involves an infinite number of possible gene counts at intermediate nodes:

$$L = \sum_{\langle m_x :\, x \in V(T)\rangle} \gamma(m_{\mathsf{root}}) \prod_{xy \in E(T)} \mathbb{P}\Big\{\chi(y) = m_y \;\Big|\; \chi(x) = m_x\Big\}, \tag{1}$$

where $\gamma(\cdot)$ defines the gene count distribution at the root, and the summation over the $\langle m_x \rangle$ vectors takes all values in agreement with the gene counts at the leaves in the input data. Our main technique for computing the likelihood is to restrict the computation to genes that have at least one surviving descendant at the leaves. In what follows we develop the formulas to compute the likelihood.

2.1 Basic transition probabilities

First we analyze the *blocks* of homologs at a node comprising genes of common origin. A *xenolog* block consists of the genes that trace back to a horizontal transfer event on the branch from the parent. For every gene at the parent, its descendants form an *inparalog* block. (Our terminology follows [38].) The homologs in Figure 1 belong to four blocks: a xenolog block of size three, an inparalog block of size zero for the deceased parental gene, and two inparalog blocks of size one. The independent birth-and-death processes associated with the blocks have been analyzed in the statistical literature.

Definition 1. *Define the following* basic transition probabilities *for gene count evolution on a branch. Let $h_t(n)$ denote the probability that there are n genes of foreign origin after time t. Let $g_t(n)$ denote the probability that a single gene has n copies after time t.*

In other words, $h_t(n)$ is the probability mass function for the number of xenologs at time t, and $g_t(n)$ defines the size distribution of an inparalog block at time t.

Theorem 1. *The basic transition probabilities can be written as follows.*

$$h_t(n) = \binom{\frac{\kappa}{\lambda} + n - 1}{n} \big(1 - \lambda\beta(t)\big)^{\frac{\kappa}{\lambda}} \big(\lambda\beta(t)\big)^n \tag{2}$$

where $\beta(t) = \frac{1-e^{-(\mu-\lambda)t}}{\mu - \lambda e^{-(\mu-\lambda)t}}$, *and*

$$\binom{\frac{\kappa}{\lambda} + n - 1}{n} = \begin{cases} 1 & \text{if } n = 0; \\ \frac{\left(\frac{\kappa}{\lambda}\right)\left(\frac{\kappa}{\lambda}+1\right)\cdots\left(\frac{\kappa}{\lambda}+n-1\right)}{n!} & \text{if } n > 0. \end{cases}$$

Furthermore,

$$g_t(n) = \begin{cases} \mu\beta(t) & \text{if } n = 0; \\ \big(1 - \mu\beta(t)\big)\big(1 - \lambda\beta(t)\big)\big(\lambda\beta(t)\big)^{n-1} & \text{if } n > 0. \end{cases} \tag{3}$$

Proof. The size of the xenolog block follows a birth-and-death process with a constant immigration rate κ and no emigration. The transition probabilities of (2) for such a process were analyzed by Karlin and McGregor [39]. An inparalog block evolves by a simple birth-and-death process: the transition probabilities of (3) are derived in, e.g., [37]. □

2.2 Gene extinction and survival

Definition 2. *A* surviving *gene at a node x is such that it has at least one modern descendant at the leaves below x.*

Let D_x denote the probability that a gene present at node x is not surviving, i.e., that it has no modern descendants.

Lemma 1. *The extinction probability D_x can be calculated as follows. If x is a leaf, then $D_x = 0$. Otherwise, let x be the parent of $x_1, x_2, \ldots, x_d$.*

$$D_x = \prod_{j=1}^{d} \Big(\mu\beta(t_j) + \big(1 - \mu\beta(t_j)\big)\big(1 - \lambda\beta(t_j)\big) \frac{D_{x_j}}{1 - \lambda\beta(t_j) D_{x_j}} \Big) \tag{4}$$

where t_j is the length of the branch leading from x to x_j.

Proof. For leaves, the statement is trivial. When x is not a leaf, condition on the gene counts at the children:

$$D_x = \prod_{j=1}^{d} \sum_{m=0}^{\infty} g_{t_j}(m) \big(D_{x_j}\big)^m.$$

Plugging in $g_t(m)$ from Eq. (3) and replacing the infinite series with a closed form gives (4). □

2.3 Effective transition probabilities

We introduce two new probabilities, denoted by $H_x(n)$ and $G_x(n)$, for having n surviving genes in a block at node x. The effective transition probabilities are related to $h_t(n)$, and $g_t(n)$, but take into consideration eventual extinction below node x. A formal definition follows.

Definition 3. *Let y be a non-root node. Define the following* effective transition probabilities. *Let $H_y(n)$ denote the probability that the xenolog block at node y contains n surviving genes. Let $G_y(n)$ denote the probability that an inparalog block at node n contains n surviving genes.*

Lemma 2. *Let y be a non-root node, let x be its ancestor, and let t be the length of the edge xy. The effective transition probabilities can be written as follows.*

$$H_y(n) = \binom{\frac{\kappa}{\lambda} + n - 1}{n} \left(\frac{1 - \lambda\beta(t)}{1 - D_y\lambda\beta(t)} \right)^{\frac{\kappa}{\lambda}} \left(\frac{(1 - D_y)\lambda\beta(t)}{1 - D_y\lambda\beta(t)} \right)^n \tag{5}$$

$$G_y(0) = 1 - \frac{\big(1 - \mu\beta(t)\big)(1 - D_y)}{1 - D_y\lambda\beta(t)}; \tag{6a}$$

$$G_y(n) = \frac{\big(1 - \mu\beta(t)\big)\big(1 - \lambda\beta(t)\big)}{\big(\lambda\beta(t)\big)(1 - D_y\lambda\beta(t))} \left(\frac{(1 - D_y)\lambda\beta(t)}{1 - D_y\lambda\beta(t)} \right)^n, \qquad n > 0. \tag{6b}$$

Proof. We condition on the number of xenologs at y (whether or not they survive).

$$H_y(n) = \sum_{i=0}^{\infty} \binom{n+i}{i} h_t(n+i) \big(D_y\big)^i \big(1-D_y\big)^n.$$

Using Eq. (2) leads to an infinite series that can be simplified to get (5). Similarly, write

$$G_y(n) = \sum_{i=0}^{\infty} \binom{n+i}{i} g_t(n+i) \big(D_y\big)^i \big(1-D_y\big)^n.$$

Taking the values of $g_t(n+i)$ from Eq. (3) and simplifying the resulting infinite series yields (6). □

2.4 Number of surviving genes on a branch

Definition 4. *Let y be a non-root node, and let x be its ancestor. Let $p_y(m|n)$ denote the* survival probability *defined as the probability of the event that there are m surviving genes at node y under the condition that there are n genes at node x (not necessarily surviving).*

Lemma 3. *The survival probabilities can be computed as follows.*

$$\begin{aligned}
p_y(m|0) &= H_y(m) && && \text{(7a)}\\
p_y(0|n) &= H_y(0)\big(G_y(0)\big)^n && 0<n && \text{(7b)}\\
p_y(1|n) &= G_y(0)p_y(1|n-1) + G_y(1)p_y(0|n-1) && 0<n && \text{(7c)}\\
p_y(m|n) &= \quad \alpha p_y(m-1|n) && 0<n, 1<m && \text{(7d)}\\
&\quad +\big(G_y(1)-\alpha G_y(0)\big)p_y(m-1|n-1)\\
&\quad +G_y(0)p_y(m|n-1)
\end{aligned}$$

where

$$\alpha = \frac{(1-D_y)\lambda\beta(t)}{1-D_y\lambda\beta(t)}. \tag{8}$$

Proof. For $p_y(m|0)$ and $p_y(0|n)$, the equations are straightforward. Otherwise, we condition on the surviving copies of a single gene at y:

$$p_y(m|n) = \sum_{i=0}^{m} G_y(i)p_y(m-i|n-1). \tag{9}$$

Now, using that $G_y(i+1) = \alpha G_y(i)$ whenever $i > 0$, and comparing (9) for $p_y(m|n)$ and $p_y(m-1|n)$, we can write $p_y(m|n)$ in a recursive form as shown. □

2.5 Conditional likelihoods

Definition 5. *Let x be a node in the tree. Define the* conditional likelihood *$L_x(n)$ for all n as the probability of having the observed gene counts at the leaves in the subtree rooted at x, under the condition that there are n surviving genes at x.*

Theorem 2. *The conditional likelihoods can be calculated as follows. In the case when x is a leaf, $L_x(n) = 1$ if n is the observed gene count at x, otherwise the likelihood is 0. If x is not a leaf, and has children $x_1, x_2, \ldots, x_d$, then the following recursions hold.*

$$L_x(0) = \prod_{j=1}^{d} \sum_{m=0}^{M_j} p_{x_j}(m|0) L_{x_j}(m); \tag{10a}$$

$$L_x(n) = (1 - D_x)^{-n} \Bigg(\prod_{j=1}^{d} \sum_{m=0}^{M_j} p_{x_j}(m|n) L_{x_j}(m) - \sum_{i=0}^{n-1} \binom{n}{i} (D_x)^{n-i} (1 - D_x)^i L_x(i) \Bigg); \qquad 0 < n \leq \sum_{j=1}^{d} M_j, \tag{10b}$$

where M_j is the sum of gene counts at the leaves in the subtree rooted at x_j. If $n > \sum_{j=1}^{d} M_j$, then $L_x(n) = 0$.

Proof. For a leaf node, or for $n > \sum_{j=1}^{d} M_j$, the theorem is trivial. Otherwise, consider the likelihood $\ell_x(n)$ of the observed gene counts at the leaves in the subtree rooted at x, conditioned on the event that there are n genes present at x, which may or may not survive. We write the likelihood in two ways. First, by conditioning on the number of surviving genes at the children,

$$\ell_x(n) = \prod_{j=1}^{d} \sum_{m=0}^{M_j} p_{x_j}(m|n) L_{x_j}(m). \tag{11}$$

Secondly, by conditioning on the number of surviving genes at x,

$$\ell_x(n) = \sum_{i=0}^{n} \binom{n}{i} \big(D_x\big)^{n-i} \big(1 - D_x\big)^i L_x(i). \tag{12}$$

Now, rearranging the equality of the two right-hand sides gives the desired result. □

Remark. Clearly, the gene counts M_x of Theorem 2 are easily computed for all x. If $m(x)$ is the gene count for every leaf x then

$$M_x = \begin{cases} m(x) & \text{if } x \text{ is a leaf;} \\ \sum_{j=1}^{d} M_{x_j} & \text{if } x_1, \ldots, x_k \text{ are the children of } x. \end{cases} \tag{13}$$

2.6 Likelihood

It is assumed that the family size at the root is distributed according to the equilibrium probabilities:

$$\gamma(n) = h_\infty(n) = \binom{\frac{\kappa}{\lambda} + n - 1}{n} \left(1 - \frac{\lambda}{\mu}\right)^{\frac{\kappa}{\lambda}} \left(\frac{\lambda}{\mu}\right)^n. \tag{14}$$

Theorem 3. *Let M be the total number of genes at the leaves. The likelihood of the observed gene counts equals*

$$L = \sum_{n=0}^{M} L_{\mathsf{root}}(n) \frac{\binom{\frac{\kappa}{\lambda}+n-1}{n} \left(1 - \frac{\lambda}{\mu}\right)^{\frac{\kappa}{\lambda}} \left((1 - D_{\mathsf{root}})\frac{\lambda}{\mu}\right)^n}{\left(1 - \frac{\lambda}{\mu} D_{\mathsf{root}}\right)^{\frac{\kappa}{\lambda}+n}}. \tag{15}$$

Proof. By summing the likelihoods conditioned on the surviving genes at the root,

$$L = \sum_{n=0}^{M} L_{\mathsf{root}}(n) \sum_{i=0}^{\infty} \gamma(n+i) \binom{n+i}{i} (D_{\mathsf{root}})^i (1 - D_{\mathsf{root}})^n. \tag{16}$$

Now, plugging in the values of $\gamma(\cdot)$ from Eq. (14) and replacing the infinite series by a closed form gives the theorem's formula. □

Remark. In place of the equilibrium probabilities of (14), many other prior distributions can be accommodated by the summation in (16).

3 Algorithm

This section employs the formulas of Section 2 in a dynamic programming algorithm to compute the likelihood exactly. More precisely, the algorithm computes the likelihood of gene counts at the tree leaves, given the duplication rate λ, the transfer rate κ, and the loss rate μ. Algorithm ComputeLikelihood below proceeds by a depth-first traversal; the necessary variables are calculated from the leaves towards the root. Let $m(u)$ denote the gene count at every leaf u.

```
ComputeLikelihood
Input λ, κ, μ, T, gene counts m(u): u is a leaf of T
Output likelihood of the m(·) values
1  for each node x ∈ V(t) in a depth-first traversal
2     Compute D_x using Eq. (4).
3     Compute the sum of gene counts M_x by Eq. (13).
4     if x is not the root then
5        Let y be the parent of x.
6        for n = 0, ..., M_y do
7           for m = 0, ..., M_x do compute p_x(m|n) by Eq. (7).
8     for n = 0, ..., M_x do compute L_x(n) by Eq. (10).
9  Compute the likelihood L at the root using Eq. (15).
10 return L.
```

Theorem 4 below analyzes the algorithm's complexity in terms of the topology of T. In particular, it uses the notions of *height of a node* x, defined as the number of edges on the path leading from the root to x, *levels* of nodes, which are sets of nodes with the same height, and *height of the tree*, which is the maximum of the leaf heights.

Theorem 4. *Let h be the height of T in Algorithm* ComputeLikelihood*, let N be the number of its leaves, and let $M = M_{\mathsf{root}}$ be the sum of gene counts. The algorithm can be implemented in such a way that it uses $O(N + M^2)$ space and runs in $O(N + hM^2)$ time.*

Proof. Computing D_x and M_x takes $O(1)$ time when x is a leaf, or $O(d)$ for an inner node with d children. There are $O(N)$ nodes in the tree and, thus, computing D_x and M_x for all x is done in $O(N)$ time. The computed values are stored in $O(N)$ space.

In order to analyze the computations in Lines 4–8, we consider nodes at the same level. Line 8 computes $L_x(n)$ for all $n = 0, \ldots, M_x$ in $O((M_x + 1)(M_x + d_x))$ total time where d_x is the number of children of node x. Lines 5–7 compute $p_x(m|n)$ for $(M_x + 1)(M_y + 1)$ pairs of n, m values. (Notice that $H_y(m)$ can be computed in $O(1)$ time for each m in the iteration over m using that $H_y(m) = \alpha\frac{m+\kappa/\lambda-1}{m}H_y(m-1)$ with the α of Eq. (8).) For the children $x_1, \ldots, x_{d_y}$ of the same node y, the total time spent in Lines 5–7 is $O((M_y + 1)(M_y + d_y))$. Terms of the type $O(d_x)$ sum up to $O(N)$ in the tree. Considering all nodes at the same level k, other terms' contribution to the running time is

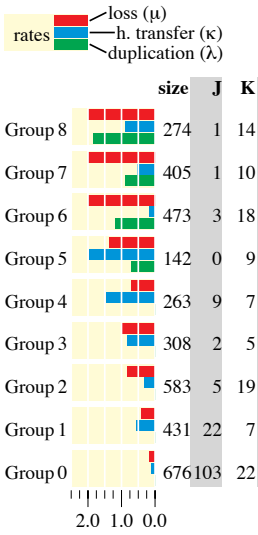
$$O\Big(\sum_{\text{all } y \text{ at level } k-1}(M_y^2 + dM_y) + \sum_{\text{all } x \text{ at level } k}(M_x^2 + dM_x)\Big),$$

where d is the maximum number of children. Clearly, $\sum_x M_x \leq M$ if the summation goes over x for which their subtrees do not overlap, such as nodes at the same level. Now, $\sum_x M_x^2 \leq (\sum_x M_x)^2 \leq M^2$, and, thus, $O(M^2 + Md)$ time is spent on each level. Therefore, the total time spent in the loop of Line 4 is $O\big(N + h(M^2 + Md)\big)$. Line 9 takes $O(M)$ time. Ignoring degenerate cases with $M \ll d$, the theorem's claim follows.

In order to obtain the space complexity result, notice that at the end of the loop in Line 8 the computed variables for the children of x are not needed anymore. Therefore, the nodes for which $p_x(\cdot|\cdot)$ is needed are such that their subtrees do not overlap. By the same type of argument as with time spent on a level, the number of variables that need to be kept in memory is $O(M^2)$. □

4 Gene content evolution in Proteobacteria

Proteobacteria form one of the most diverse groups of prokaryotes. Proteobacteria provide an excellent case study for gene content evolution: they include pathogens, endosymbionts, and free-living organisms. Genome sizes vary tenfold within this group, and horizontal transfer is abundant [25]. Their phylogeny



	size	J	K	L	D	V	T	M	N	U	O	C	G	E	F	H	I	P	Q	R	S
Group 8	274	1	14	22	2	0	18	6	3	13	7	12	22	21	1	2	4	10	13	55	61
Group 7	405	1	10	8	1	8	13	15	7	14	8	30	28	13	7	2	8	28	18	79	117
Group 6	473	3	18	23	2	6	13	13	4	1	8	22	30	14	3	3	9	9	18	81	208
Group 5	142	0	9	8	0	5	10	7	9	6	1	15	11	16	1	4	4	11	3	30	14
Group 4	263	9	7	9	2	1	5	22	22	15	20	25	10	30	12	14	10	25	5	32	16
Group 3	308	2	5	6	2	2	7	17	4	15	14	18	31	25	5	13	3	41	4	44	64
Group 2	583	5	19	17	2	5	18	19	11	7	22	31	27	19	15	23	8	27	9	98	220
Group 1	431	22	7	19	8	1	16	18	15	14	32	33	17	44	13	40	9	34	6	45	66
Group 0	676	103	22	48	19	5	9	36	4	15	23	27	12	40	31	48	17	7	1	58	168

Fig. 2. Rates in different groups and the distribution of COG functional categories. The functional categories are: J–translation, K–transcription, L–replication and repair, D–cell cycle control and mitosis, V—defense mechanisms, T–signal transduction, M–cell wall/membrane/envelope biogenesis, N–cell motility, U–intracellular trafficking and secretion, O–posttranslational modification, protein turnover and chaperones, C–energy production and conversion, G–carbohydrate transport and metabolism, E–amino acid transport and metabolism, F–nucleotide transport and metabolism, H–coenzyme transport and metabolism, I–lipid transport and metabolism, P–inorganic ion transport and metabolism, Q–secondary metabolites biosynthesis, transport and catabolism, R–general function prediction only, S–function unknown. The "size" columns gives the number of COGs in each rate group. (The numbers in a row do not always add up to the value in the "size" column because some COGs have more than one functional assignment.)

is still not resolved to satisfaction [40–43]. We used 51 proteobacteria in the first application of our likelihood method. Gene counts were based on the newer version [22] of the COG database. Each COG is a manually curated protein family of homologs. The COGs are classified into 23 functional categories. (For each of the 51 proteobacteria, the number of genes in each COG family was established by Pál et al. [30]. There are 3555 COG families that have at least one member in the organisms. The organisms and the phylogeny are shown at `http://www.iro.umontreal.ca/~csuros/gene_content/`.) The purpose of applying the likelihood method was not to carry out in-depth data analysis, but rather to get a first impression of our method's performance on realistic data.

First we optimized the branch lengths and the λ, κ parameters while keeping $\mu = 1.0$ to fix the scaling of edge lengths. In a second pass, we clustered the COG families with different rates in different groups. The groups were established in

several iterations of Expectation Maximization: in an E-step, each family was assigned to the best group (the one whose rates give the highest likelihood), in an M-step, rates were optimized within each group separately to maximize the likelihood of the COG gene counts within the group's families. Figure 2 shows the rates in different groups (Groups 0–8), as well as the distribution of COG functional classes across clusters. The picture shows that various rate groups are needed to describe the evolution of the families. While the results and the methodology still need a thorough critical assessment, some interesting patterns already emerge. About 19% of the families are very stable (Group 0), including the large majority of genes involved in translation (category J) such as tRNA synthetases and ribosomal proteins, and cell cycle control (category D). About one in nine families fall into groups with large horizontal transfer rates (Groups 4 and 5), while one in three families are in groups with very low transfer rates. In some categories duplication plays only a minor role: the evolution of cell motility (category N), and various metabolic functions (F,H,I) seem to be shaped mainly by horizontal transfer and loss.

5 Conclusion

We presented the first three-parameter model of gene content evolution, along with a fast algorithm for computing likelihoods. We implemented parameter optimization and a gene family clustering method and carried out a pilot experiment using COG family sizes in 51 Proteobacteria.

We modeled gene family evolution by a birth-and-death process. It was shown that birth-and-death processes of various complexity explain the observed power-law behavior of gene family sizes [32–34, 44]. In order to develop a truly realistic likelihood model, rate variation must be permitted across lineages and families. Our formulas can be readily adapted to branch-dependent rates. The challenge lies rather in the parametrization: introducing four parameters (three rates and branch length) for every tree edge and every family will lead to overfitting. A possible solution is to work with two sets of parameters: a branch-specific and a family-specific set. We are now working on developing adequate rate-variation models along these lines. In another related inquiry, we are investigating the possibility of pairing this model with sequence evolution models, to achieve a more nuanced modeling of homologies than simple counts. Incorporating gene similarity will undoubtedly lead to an improved likelihood model of gene content evolution.

This paper focuses on the core algorithmic problems of likelihood computations in a biologically realistic model of gene content evolution. The presented likelihood algorithm can be utilized in a number of contexts. The computations can be used in parameter optimization to estimate duplication, loss, and transfer rates in different gene families. By comparing the maximum likelihood values achieved with different evolutionary tree topologies, organismal phylogeny can be derived from gene content. "Unusual" branches with excess transfer, loss, etc., can be identified by examining the likelihoods, adapting an idea of [31].

The conditional likelihoods of §2.5 can be used in likelihood-based computations of ancestral gene content, similarly to standard methods employed in case of molecular sequences [45] and introns [46]. The likelihood computation allows for the sampling of different trees in a Bayesian Markov Chain Monte Carlo method. We believe that our approach — the efficient computation of exact likelihoods in a three-parameter model — will find many important applications in comparative gene content analysis.

Acknowledgments

We would like to thank Eugene Koonin, Hervé Philippe and Yuri Wolf for useful discussions concerning gene content evolution, as well as Csaba Pál and Martin Lercher for providing us with pre-publication data. This work was supported in part by the e-Science Regional Knowledge Center at Eötvös Lóránd University, Budapest, sponsored by the Hungarian National Office for Research and Technology (NKTH). M.Cs. is supported by grants from the Natural Sciences and Engineering Research Council of Canada and the *Fonds québecois de la recherche sur la nature et les technologies*. I.M. is supported by a Békésy György postdoctoral fellowship.

References

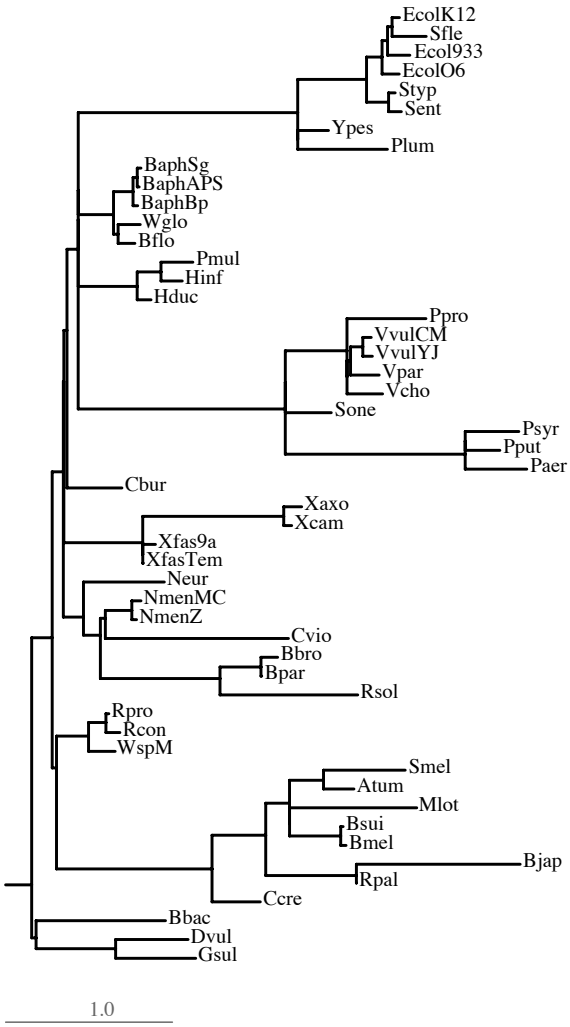
1. Green, E.D.: Strategies for the systematic sequencing of complex genomes. Nature Reviews Genetics **2** (2001) 573–583
2. Wolfe, K.H., Li, W.H.: Molecular evolution meets the genomic revolution. Nature Genetics **33** (2003) 255–265
3. Delsuc, F., Brinkmann, H., Philippe, H.: Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics **6** (2005) 361–375
4. Fitz-Gibbon, S.T., House, C.H.: Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Research **27** (1999) 4218–4222
5. Snel, B., Bork, P., Huynen, M.A.: Genome phylogeny based on gene content. Nature Genetics **21** (1999) 108–110
6. Tekaia, F., Lazcano, A., Dujon, B.: The genomic tree as revealed from whole proteome comparisons. Genome Research **9** (1999) 550–557
7. Lin, J., Gerstein, M.: Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Research **10** (2000) 808–818
8. Clarke, G.D.P., Beiko, R.G., Ragan, M.A., Charlebois, R.L.: Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. Journal of Bacteriology **184** (2002) 2072–2080
9. Korbel, J.O., Snel, B., Huynen, M.A., Bork, P.: SHOT: a web server for the construction of genome phylogenies. Trends in Genetics **18** (2002) 158–162
10. Dutilh, B.E., Huynen, M.A., Bruno, W.J., Snel, B.: The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. Journal of Molecular Evolution **58** (2004) 527–539

11. Huson, D.H., Steel, M.: Phylogenetic trees based on gene content. Bioinformatics **20** (2004) 2044–2049
12. Gu, X., Zhang, H.: Genome phylogenetic analysis based on extended gene contents. Molecular Biology and Evolution **21** (2004) 1401–1408
13. Lake, J.A., Rivera, M.C.: Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. Molecular Biology and Evolution **21** (2004) 681–690
14. Yang, S., Doolittle, R.F., Bourne, P.E.: Phylogeny determined by protein domain content. Proceedings of the National Academy of Sciences of the USA **102** (2005) 373–378
15. Deeds, E.J., Hennessey, H., Shakhnovich, E.I.: Prokaryotic phylogenies inferred from protein structural domains. Genome Research **15** (2005) 393–402
16. Montague, M.G., Hutchison III, C.A.: Gene content phylogeny of herpesviruses. Proceedings of the National Academy of Sciences of the USA **97** (2000) 5334–5339
17. Herniou, E.A., Luque, T., Chen, X., Vlak, J.M., Winstanley, D., Cory, J.S., O'Reilly, D.R.: Use of whole genome sequence data to infer baculovirus phylogeny. Journal of Virology **75** (2001) 8117–8126
18. Simonson, A.B., Servin, J.A., Skophammer, R.G., Herbold, C.W., Rivera, M.C., Lake, J.A.: Decoding the genomic tree of life. Proceedings of the National Academy of Sciences of the USA **102** (2005) 6608–6613
19. Snel, B., Bork, P., Huynen, M.A.: Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Research **12** (2002) 17–25
20. Mirkin, B.G., Fenner, T.I., Galperin, M.Y., Koonin, E.V.: Algorithms for computing 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** (2003) 2
21. Koonin, E.V., Galperin, M.Y.: Sequence-Evolution-Function: Computational Approaches in Comparative Genomics. Kluwer Academic Publishers, New York (2002)
22. 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., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A.: The COG database: an updated version includes eukaryotes. BMC Bioinformatics **4** (2003) 441
23. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., Yeates, T.O.: Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proceedings of the National Academy of Sciences of the USA **96** (1999) 4285–4288
24. Jordan, I.K., Makarova, K.S., Spouge, J.L., Wolf, Y.I., Koonin, E.V.: Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Research **11** (2001) 555–565
25. Gogarten, J.P., Doolittle, W.F., Lawrence, J.G.: Prokaryotic evolution in light of gene transfer. Molecular Biology and Evolution **19** (2002) 2226–2238
26. Kurland, C.G., Canback, B., Berg, O.G.: Horizontal gene transfer: a critical view. Proceedings of the National Academy of Sciences of the USA **100** (2003) 9658–9662
27. Kunin, V., Goldovsky, L., Darzentas, N., Ouzounis, C.A.: The net of life: reconstructing the microbial phylogenetic network. Genome Research **15** (2005) 954–959
28. Ge, F., Wang, L.S., Kim, J.: The cobweb of life revealed by genome-scale estimates of horizontal gene transfer. PLoS Biology **3** (2005) e316
29. Boucher, Y., Douady, C.J., Papke, R.T., Walsh, D.A., Boudreau, M.E.R., Nesbø, C.L., Case, R.J., Doolittle, W.F.: Lateral gene transfer and the origin of prokaryotic groups. Annual Review of Genetics **37** (2003) 283–328

30. Pál, C., Papp, B., Lercher, M.: Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nature Genetics **37** (2005) 1372–1375
31. Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C., Cristianini, N.: Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Research **15** (2005) 1153–1160
32. Karev, G.P., Wolf, Y.I., Rzhetsky, A.Y., Berezovskaya, F.S., Koonin, E.V.: Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evolutionary Biology **2** (2002) 18
33. Karev, G.P., Wolf, Y.I., Koonin, E.V.: Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? Bioinformatics **19** (2003) 1889–1900
34. Karev, G.P., Wolf, Y.I., Berezovskaya, F.S., Koonin, E.V.: Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. BMC Evolutionary Biology **4** (2004) 32
35. Wolf, Y.I., Rogozin, I.B., Grishin, N.V., Tatusov, R.L., Koonin, E.V.: Genome trees constructed by five different approaches suggest new major bacterial clades. BMC Evolutionary Biology **1** (2001) 8
36. Kunin, V., Ouzounis, C.A.: GeneTRACE-reconstruction of gene content of ancestral species. Bioinformatics **19** (2003) 1412–1416
37. Feller, W.: An Introduction to Probability Theory and Its Applications. Wiley & Sons (1950)
38. Sonnhammer, E.L.L., Koonin, E.V.: Orthology, paralogy and proposed classification for paralog subtypes. Trends in Genetics **18** (2002) 619–620
39. Karlin, S., McGregor, J.: Linear growth, birth, and death processes. Journal of Mathematics and Mechanics **7** (1958) 643–662
40. Lerat, E., Daubin, V., Moran, N.A.: From gene trees to organismal phylogeny in Prokaryotes: the case of the γ-Proteobacteria. PLoS Biology **1** (2003) E19
41. Boussau, B., Karlberg, E.O., Frank, A.C., Legault, B.A., Andersson, S.G.E.: Computational inference of scenarios for α-proteobacterial genome evolution. Proceedings of the National Academy of Sciences of the USA **101** (2004) 9722–9727
42. Herbeck, J.T., Degnan, P.H., Wernegren, J.J.: Nonhomogeneous model of sequence evolution indicates independent origins of endosymbionts within the Enterobacteriales (γ-Proteobacteria). Molecular Biology and Evolution **22** (2005) 520–532
43. Belda, E., Moya, A., Silva, F.J.: Genome rearrangement distances and gene order phylogeny in γ-Proteobacteria. Molecular Biology and Evolution **22** (2005) 1456–1467
44. Reed, W.J., Hughes, B.D.: A model explaining the size distribution of gene families. Mathematical Biosciences **189** (2004) 97–102
45. Pupko, T., Pe'er, I., Shamir, R., Graur, D.: A fast algorithm for joint reconstruction of ancestral amino acid sequences. Molecular Biology and Evolution **17** (2000) 890–896
46. Csűrös, M.: Likely scenarios of intron evolution. In McLysaght, A., Huson, D.H., eds.: Comparative Genomics. Volume 3678 of LNBI., Heidelberg, Springer-Verlag (2005) 47–60

Appendix: organisms in the data set

The picture below shows the organisms and the phylogeny in the experiments of Section 4. Branch lengths are already optimized to maximize the likelihood. Notice that branch lengths are not easy to interpret: scaling is defined in such a way that the rate $\mu = 1$ in Group 0, a modestly dynamic group (cf. Fig. 2). Long branches indicate major changes in gene content.



Abbreviations: EcolK12 –*Escherichia coli* K12, Sfle –*Shigella flexneri* 2a str. 2457T, Ecol933 –*Escherichia coli* O157:H7 str. EDL933, EcolO6 –*Escherichia*

coli O6, Styp –*Salmonella typhimurium* LT2, Sent –*Salmonella enterica* subsp. *enterica serovar Typhi* str. CT18, Ypes –*Yersinia pestis biovar Medievalis* str. 91001, Plum –*Photorhabdus luminescens* subsp. *laumondii* TTO1, BaphSg –*Buchnera aphidicola* str. Sg, BaphAPS –*Buchnera aphidicola* str. APS, BaphBp –*Buchnera aphidicola* str. Bp, Wglo –*Wigglesworthia glossinidia* endosymbiont of *Glossina brevipalpis*, Bflo –*[Candidatus] Blochmannia floridanus*, Pmul –*Pasteurella multocida* subsp. *multocida* str. Pm70, Hinf –*Haemophilus influenzae* Rd KW20, Hduc –*Haemophilus ducreyi* 35000HP, Ppro –*Photobacterium profundum* SS9, VvulCM –*Vibrio vulnificus* CMCP6, VvulYJ –*Vibrio vulnificus* YJ016, Vpar –*Vibrio parahaemolyticus* RIMD 2210633, Vcho –*Vibrio cholerae* O1 *biovar eltor* str. N16961, Sone –*Shewanella oneidensis* MR-1, Psyr –*Pseudomonas syringae* pv. tomato str. DC3000, Pput –*Pseudomonas putida* KT2440, Paer –*Pseudomonas aeruginosa* PAO1, Cbur –*Coxiella burnetii* RSA 493, Xaxo –*Xanthomonas axonopodis* pv. *citri* str. 306, Xcam –*Xanthomonas campestris* pv. *campestris* str. ATCC 33913, Xfas9a –*Xylella fastidiosa* 9a5c, XfasTem –*Xylella fastidiosa* Temecula1, Neur –*Nitrosomonas europaea* ATCC 19718, NmenMC –*Neisseria meningitidis* MC58, NmenZ –*Neisseria meningitidis* Z2491, Cvio –*Chromobacterium violaceum* ATCC 12472, Bbro –*Bordetella bronchiseptica* RB50, Bpar –*Bordetella parapertussis* 12822, Rsol –*Ralstonia solanacearum* GMI1000, Rpro –*Rickettsia prowazekii* str. Madrid E, Rcon –*Rickettsia conorii* str. Malish 7, WspM –*Wolbachia* endosymbiont of *Drosophila melanogaster*, Smel –*Sinorhizobium meliloti* 1021, Atum –*Agrobacterium tumefaciens* str. C58, Mlot –*Mesorhizobium loti* MAFF303099, Bsui –*Brucella suis* 1330, Bmel –*Brucella melitensis* 16M, Bjap –*Bradyrhizobium japonicum* USDA 110, Rpal –*Rhodopseudomonas palustris* CGA009, Ccre –*Caulobacter crescentus* CB15, Bbac –*Bdellovibrio bacteriovorus* HD100, Dvul –*Desulfovibrio vulgaris* subsp. *vulgaris* str. Hildenborough, Gsul –*Geobacter sulfurreducens* PCA.