DIMACS Series in Discrete Mathematics
and Theoretical Computer Science

A Top-down Method for Building Genome Classification Trees with Linear Binary Hierarchies

Boris Mirkin and Eugene Koonin

ABSTRACT. With complete genome sequence data becoming available at an increasing rate, the problem of classification of the genomes on the basis of different criteria is becoming pressing. Here we present an approach that applies linear embedding of binary hierarchies to the analysis of the representation of genomes in clusters of orthologs. Rather than imposing an evolutionary postulate such as the additivity of tree metrics or the principle of parsimony of the character changes, this approach tries to approximate the observed patterns of presence/absence of individual genes in genomes. An individual dene is modeled by a cluster of orthologous groups of proteins from different genomes (COG). The developed method differs from other tree-building methods in that it works by sequentially splitting the entity set and, accordingly, modeling potential speciation events. This allows, first, to remove "worked-out" characters from further splits and, second, to stop the splitting process when results become unreliable. The linear binary hierarchy theory allows to do this and, also, to explicitly reformulate the criterion and the method in terms of a between-genome similarity index. The resulting classification trees have the potential of revealing previously unnoticed but potentially biologically relevant relationships between the corresponding organisms.

1. Introduction: Clusters of Orthologous Groups as characters

The availability of multiple, complete genome sequences of bacteria, archaea and eukaryotes has stimulated several attempts on construction of whole-genome evolutionary trees. Typically, these studies employ certain integral characteristics of the evolutionary process to investigate the relationships between genomes and, by inference, the corresponding life forms themselves. Probably the most obvious of such characteristics is the presence/absence of representatives of the analyzed species in orthologous groups of genes, and recently, several groups have used this approach to compare the sequenced prokaryotic genomes [**4, 14, 17**].

The concept and data of clusters of orthologous groups of proteins (COGs) have been developed in [**16**]; see also http://www.ncbi.nlm.nih.gov/COG/. A COG consists of genes from different species that are genome-specific best hits as identified by sequence similarity searches using BLAST. This criterion of orthology assumes primarily vertical inheritance of the members of each COG. However, as indicated by numerous comparative-genomic studies, the vertical inheritance pattern of evolution is confounded by numerous

2000 *Mathematics Subject Classification.* Primary 62H30, 92B10; Secondary 05C50.
Key words and phrases. Evolutionary tree, similarity, character-based tree, hierarchy.

non-canonical evolutionary events, such as lineage-specific gene loss and horizontal gene transfer. In particular, if a COG is assumed to be represented in the ancestor of all the species under consideration, the hypothesis of strict vertical evolution would imply that all the species must be represented in the COG. This pattern, the ubiquitous representation of a COG, is indeed seen, but only in a small fraction of the COGs, only 84 of the 3166 analyzed COGs. These observations suggest that (1) many COGs have not been inherited from the Last Universal Common Ancestor (LUCA) but rather evolved in more recent ancestors, and (2) COG members have been repeatedly lost in different lineages, even in the ancient COGs inherited from LUCA. Furthermore, the patterns of genome representation in COGs probably have been affected also by horizontal gene transfer.

These hypotheses can be tested if the species tree is known for the genomes in question. The topology of such a tree is a matter of contention, but an approximation has been derived by phylogenetic analysis of concatenated alignments of ribosomal proteins in [**19**]. It appears that most COGs in our collection have not strictly followed the vertical path of evolution; instead, their histories involved both losses and horizontal transfer events [**11**]. Combined with the topology of the (putative) species tree, the data on presence/absence of COGs in genomes (phyletic patterns of the COGs) can be exploited for constructing parsimonious evolutionary scenarios for COGs. Furthermore, this approach should allow one to estimate the amount of horizontal transfer and lineage-specific gene loss in each COG and to build gene presence-absence trees using only those COGs that show evidence of none or few of these events. Presence-absence trees constructed from COGs that evolved predominantly vertically are expected to approximate the species tree. Indeed, a COG that emerges on a particular ancestral node can be treated as a character that is present in this node and its descendents but is absent in the sibling(s) of the given ancestral node and its descendents. Under this approach, presence or absence of a COG can be considered a feature differentiating diverged branches of ancestral forms during evolution.

To derive a putative evolutionary tree on the basis of the COG presence/absence data, a tree building method should be employed that produces divergences at the ancestor branches in such a way that COGs, which are not inherited by (almost) all of the ancestor's offspring, are not randomly distributed among the offspring, but go to either of the descendent siblings. A method explicating this requirement should work top-down, i.e. from the tree root to its leaves, producing divergences at which clearly distinguishable COGs go to either of the children branches. These COGs should be considered divergence markers, the characters that diverged at the corresponding ancestor nodes, whereas other COGs, e.g. the 84 COGs universal COGs, are considered as inherited characters.

However, to our knowledge, among the numerous currently available tree building methods, none builds a tree top-down; these methods build a tree in the opposite direction, bottom-up, starting from leaves (individual genes or genomes) and merging them one by one (for recent compendia of tree-building methods see the phylogenetic packages PHYLIP [**3**] and PAUP [**15**]). Thus, the primary goal of this work is to develop a method for building an evolutionary tree in the top-down fashion and to apply it to the COG-to-genome presence/absence data. We propose such a method using the linear theory of binary hierarchies [**9, 10**]. This method combines two major approaches to tree building, the character-based approach and the distance-based approach. Also, this method evaluates the quality of a divergence event on the basis of the proportion of the data scatter taken into account by the given divergence. This allows one to assess the relative reliability of different tree parts and, in particular, to produce an unresolved subtree (multifurcation) by stopping

further divisions of unreliable tree parts. Moreover, the method leads to a new genome-to-genome similarity measure which has some obvious advantages over other measures such as the Jaccard coefficient used, for instance, by Snel et al. [**14**].

However, application of the method to the COG data produced an unrealistic tree (see Figure 3 later). Since similarly unrealistic trees have been produced with other methods applied to the same data [**19**], this makes us suggest that the presence/absence data are poorly suited for phylogenetic analysis because the phyletic profiles of many (if not most) of the COGs contain too little phylogenetic signal and too much history of repeated horizontal transfers and losses. If we could separate COGs with such complicated history and remove them from consideration, the remaining part of the data should lead to a more realistic tree. To this end, we applied a method for building a parsimonious scenario of a COG's evolutionary history given a hypothetical genome (species) tree, which we have developed for this and other purposes in [**11**]. Specifically, each of the COGs was assigned a so-called inconsistency value, which is the minimum number of losses or horizontal transfer events required to explain the COG's phyletic pattern. By removing most inconsistent COGs, we obtained a set of COGs that probably evolved exclusively or predominantly in the vertical mode, leading to a more realistic portrayal of the evolution (see Figure 4).

Since we employ a top-down method for building a tree, we can undertake a further adjustment in order to obtain a more realistic tree. This adjustment takes into account the following property: if a character has diverged in an ancestral node, it does not participate in the subsequent divisions. Thus, after each divergence step, we can remove those COGs that went, mainly, to one of the diverged branches and leave for further divisions only those COGs that have been more or less uniformly distributed between the children. Such a modification is described in section 5.3.

2. Linear theory of binary hierarchies

2.1. Basis for an hierarchy. Suppose I is a set of N genomes. The relationships between the genomes in I can be represented by a rooted tree, with the leaves corresponding to individual genomes. Each node in the tree represents the set of genomes corresponding to the leaves in the subtree rooted at this node. Thus the tree may be represented by the set of all such clusters. This set can, in turn, be represented by an orthonormal basis of the linear space of all N-dimensional centered vectors [**9**]. In this representation, a three-valued N-dimensional vector $\phi_w(i)$, $i \in I$, is defined for each interior node w of the tree. If the children of w are $w1$ and $w2$, then $\phi_w(i)$ is zero outside of cluster w, a_w within cluster $w1$, and $-b_w$ within cluster $w2$. We consider only binary trees, i.e. trees in which each internal node has two children, the normal situation in phylogenetic trees (although some subtrees may be unresolved leading to multifurcations). The values a_w and b_w are (essentially) uniquely defined by the condition that the vector ϕ_w is centered and normalized. This condition means that $a_w N_{w1} - b_w N_{w2} = 0$, where N_{w1} and N_{w2} are the cardinalities of the clusters $w1$ and $w2$, respectively; their sum is obviously N_w, the cardinality of the parent cluster w. Together with the condition of normalization, this implies that $a_w = \sqrt{N_{w2}/N_{w1}N_w}$ and $b_w = \sqrt{N_{w1}/N_{w2}N_w}$. It is not difficult to prove that all the vectors ϕ_w are pairwise mutually orthogonal, thus forming an orthonormal basis of the space of all N-dimensional centered vectors.

Generalizing to the case in which some lower clusters may be omitted, thus possibly leaving some (leaf) nodes unresolved, we refer to the set of nodes (clusters) as a binary hierarchy. For a binary hierarchy, the set of its interior nodes will be denoted by W, and

its cardinality $|W|$ by m. The $N \times m$ matrix whose columns are the vectors ϕ_w, for each $w \in W$, will be denoted by Φ.

2.2. Building a LBH for a data matrix. This formalism can be applied to the problem of deriving a tree from an $N \times n$ entity-to-feature data matrix Y (with the rows corresponding to the N genomes in I and columns to n pre-specified features – such as the presence/absence or frequency of certain proteins) as follows. Given a binary hierarchy with m interior nodes represented by the $N \times m$ matrix Φ, we can express Y in the following form:

$$Y = \Phi C + E, \tag{1}$$

where C is an $m \times n$ matrix of elements c_{wk} and $E = (e_{ik})$ is a matrix of residuals (the differences between the corresponding entries of Y and ΦC). The columns of Y are assumed to be centered (as are the columns of Φ). Then, given Φ, the problem is to determine C to minimize the residuals. Having done this, the next problem is to modify the hierarchy (i.e., Φ) to further reduce the residuals. (When the hierarchy is completely resolved, i.e., all N singleton clusters belong to the hierarchy, the matrix of differences E equals zero.)

The matrix C can be found using the least-squares approach, by minimizing the sum of the squared residuals,

$$D(\Phi, C) = \sum_{i,k} e_{ik}^2. \tag{2}$$

The optimal C is $C^* = \Phi^T Y$, that is, $c^*_{wk} = \sum_i y_{ik}\phi_{iw}$ for any w and k, and

$$D(\Phi, C^*) = Tr(Y^T Y) - \sum_{w \in W} \mu_w^2, \tag{3}$$

where

$$\mu_w^2 = \sum_{k=1}^{n} c_{wk}^2 = \frac{N_{w1}N_{w2}}{N_w} d^2(y_{w1}, y_{w2}), \tag{4}$$

the vectors y_{w1} and y_{w2} are the centers of gravity of the row-vectors in Y belonging to the clusters $w1$ and $w2$, respectively, and $d^2(y_1, y_2) = \sum_k (y_{1k} - y_{2k})^2$, is the square of Euclidean distance between any two n-dimensional vectors y_1 and y_2. Since $Tr(Y^T Y) = \sum_{i,k} y_{ik}^2$ is the data scatter (which is proportional to the data variance), equation (3) can be considered as a decomposition of the data scatter into the explained (by the hierarchy) and unexplained parts.

From (3) we see that any least-squares optimal hierarchy must maximize

$$\sum_{w \in W} \mu_w^2 = \sum_{w \in W} \frac{N_{w1}N_{w2}}{N_w} d^2(y_{w1}, y_{w2}).$$

Therefore, the problem of building a tree according to the least-squares criterion reduces to the combinatorial problem of finding $m = |W|$ successive splits of I which maximize the sum of the weighted squared between-centre distances $d^2(y_{w1}, y_{w2})$. The complexity of solving the problem can be reduced by exploiting the principle of sequential fitting [**10**], that is, by doing splits iteratively, one split at a time.

A tree building method, according to this approach, starts by splitting the entire set I into subsets S_1 and S_2 to maximize the criterion

$$\mu^2 = \frac{N_1 N_2}{N} d^2(y_1, y_2) \tag{5}$$

where N_1, N_2 and N are cardinalities of sets S_1, S_2 and I, respectively, and y_1 and y_2 are gravity centers of S_1 and S_2, that is, vectors of average within cluster values of variables $k = 1, ..., n$. Then the operation of splitting is reiterated each time applied to one of the found split parts, until they all become singletons. The value of μ^2 is the proportion of the data scatter, $\sum_{i,k} y_{ik}^2$, taken into account by the split.

2.3. Deriving a similarity index for LBH. Criterion (5) can be equivalently reformulated to the format of a square genome-to-genome similarity matrix $a = (a_{ij})$, $i, j \in I$. Indeed, μ^2 in (5) can be equivalently presented as $\mu^2 = \sum_{k=1}^{n} c_k^2$ where $c_k = \sum_{i \in I} \phi_i y_{ik}$ and ϕ_i is $\sqrt{1/N_1 - 1/N}$ if $i \in S_1$ or $\sqrt{1/N_2 - 1/N}$ if $i \in S_2$. This leads to $\mu^2 = \sum_{k=1}^{n} \sum_{i \in I} \phi i y_{ik} \sum_{j \in I} \phi j y_{jk} = \sum_{i \in I} \sum_{j \in I} \phi i \phi j \sum_{k=1}^{n} y_{ik} y_{jk}$. Finally, criterion (5) is expressed as

$$\mu^2 = \sum_{i \in I} \sum_{j \in I} a_{ij} \phi i \phi j \tag{6}$$

where

$$a_{ij} = \sum_{k=1}^{n} y_{ik} y_{jk} \tag{7}$$

is similarity between genomes i and j from I according to data matrix Y.

Matrix $a = (a_{ij})$ is obviously double-centered, that is, all within-column averages and all within-row averages in it are equal to zero.

Criterion (6) can be easily transformed into:

$$\mu^2 = \frac{N_1 N_2}{N}(a_{11} + a_{22} - 2a_{12}) \tag{8}$$

where

$$a_{ww'} = \frac{\sum_{i \in S_w} \sum_{j \in S_{w'}} a_{ij}}{N_w N_{w'}},$$

the average similarity between S_w and $S_{w'}$ or within S_w if $w = w'$, $w, w' = 1, 2$.

2.4. The splitting algorithm. The above result obviously can be extended to the general case of splitting a cluster S_w into non-overlapping parts S_{w1} and S_{w2}. When an element $i \in S_{w1}$ is moved into S_{w2}, the change $\Delta(i)$ of the criterion μ_w^2 (8) can be easily expressed through elements of matrix a and the average similarities.

The algorithm suggested for maximization of μ^2 in (8) is a modification of the traditional exchange algorithm:

1. Find a pair i^*, j^* maximizing $g(i, j) = a_{ii} + a_{jj} - 2a_{ij}$ over $i, j \in S$.

2. Create initial S_1 and S_2 by distributing each $i \in S$, $i \neq i^*, j^*$, either to i^* or j^*, according to the sign of an analogue to $\Delta(i)$, $3(a_{ii^*} - a_{ij^*}) - (a_{i^*i^*} - a_{j^*j^*})$.

3. For any $i \in S$ calculate $\Delta(i)$ and take i^* maximizing it.

4. If $\Delta(i^*)$ is not positive, stop the process and output current clusters S_1 and S_2 along with corresponding μ^2. If $\Delta(i*) > 0$, move $i*$ to the other cluster and go to the beginning of step 3.

It can be proven that, when a split is found with the above algorithm, all members of a cluster are more attracted to their own cluster than to the other cluster in the split. It is assumed, in this statement, that the measure of attraction of an element i to subset S is defined as $\alpha(i, S) = a(i, S) - a(S)/2$ where $a(i, S)$ is the average similarity between i and elements of S and $a(S)$ is the average similarity between elements of S.

3. Applying LBH to the data on the presence/absence of genomes in COGs

As follows from the above, the criterion employed in the LBH approach differs from other known criteria used for building trees in that the LBH tree approximates the data matrix entries instead of following any pre-specified phylogenetic assumption. This implies some other features that are emphasized in this section:

- reformulation of the criterion with regard to the specifics of the data;
- stopping the splitting process;
- identification of the COGs that make the greatest contribution to a split;
- derivation of a similarity measure for genome comparison.

3.1. Reformulation of the splitting criterion. Let us consider how criterion (2) can be applied in the case when all the variables are binary descriptors of presence/absence of COGs in genomes represented by zero-one columns (one for presence, zero for absence).

Let us compute the within-cluster average of a binary variable k. Since the variable has been centered initially, the entries $1-p_k$ and $-p_k$ stand for 1 and 0, respectively, where p_k denotes the relative frequency of 1's in the column k.

Thus, the average is $y_{wk} = (1-p_k)p_{wk}/p_w - p_k(1-p_{wk}/p_w)$ where p_{wk} is the frequency of simultaneously observing descriptor k and cluster S_w, and p_w is the frequency of S_w. This leads to:

$$y_{wk} = p_{wk}/p_w - p_k. \tag{9}$$

which implies

$$c^2_{wk} = \frac{N_{w1}N_{w2}}{N_w}\left(\frac{p_{w1k}}{p_{w1}} - \frac{p_{w2k}}{p_{w2}}\right)^2 \tag{10}$$

The first factor here takes care of the sizes of split parts: the more uniform the partition, the greater the factor, which corresponds to the information concepts of the search theory. The second factor maximizes the difference between the frequencies of 1's in the subclusters.

The criterion (2) in this case is simply the weighted distance between within-cluster probability profiles:

$$\mu^2_w = \frac{N_{w1}N_{w2}}{N_w} d^2(p(w1), p(w2)) \tag{11}$$

where $p(w)$ is the vector of (conditional) probabilities of categories k in cluster S_w. It should be noted that this measure closely resembles the so-called twoing rule used in the CART techniques for deriving decision trees [**1**].

The relative value of contribution of a COG (variable) k in the overall splitting criterion is c^2_{wk}/μ^2_w and does not depend on the split cardinalities nor on factor $N_{w1}N_{w2}/N_w$.

3.2. Contribution of a split and stopping the splitting process. The quantity μ^2_w in (11) is, in fact, the contribution of the split w to the total data scatter. It can be employed for deciding whether the process of splitting should be stopped at this point or continued. We suggest that the average value of the data scatter per genome, $sc = \sum_{i,k} y^2_{ik}/N$, can be considered the threshold: if μ^2_w becomes less than sc, the cluster should not be partitioned anymore because its contribution becomes too small and reflects the noise in the data rather than a meaningful signal.

3.3. Most contributing COGs. A feature's contribution to the data scatter is equal to the value of c_{wk}^2 in (10). Thus, features that make the greatest contribution to the split w can be found by sorting these quantities in the descending order. The features thus selected will maximally differ by their contents in the split parts, which may be exploited for the interpretation of the given split.

3.4. Similarity between genomes. Similarity between genomes i and j considered as sets (bags) of COGs can be estimated by using the Jaccard coefficient,

$$J = \frac{|ij|}{|i| + |j| - |ij|}$$

where $|i|$, $|j|$, and $|ij|$ are cardinalities of i, j and their overlap, respectively. This measure was applied in some studies on gene content (presence-absence) trees [**14**]. However, there appears to be an intrinsic flaw in the Jaccard coefficient in that it systematically underestimates the similarity between genomes.

When the sizes of sets i and j are about the same and their overlap is about half of the elements in each of them, the Jaccard coefficient is about 1/3, whereas, intuitively, one would expect the similarity score to be, in this case, about 1/2. To make the coefficient equal to 1/2, the overlap must contain about 2/3 of the elements from each of the sets, which intuitively should correspond to a score of 2/3.

The underestimate becomes even more striking when one set is much smaller than the other, as it frequently happens with genomes. Let, for instance, the size of i be 1/4 of the size of j, and i a subset of j. Then the value of the Jaccard coefficient will also be 1/4. This small value contradicts our predictions on the evolutionary relationships between genomes i and j because, in case they are closely related, the smaller one is likely to be a derivative of the larger one. To cope with this problem, another coefficient can be proposed:

$$M = (|ij|/|i| + |ij|/|j|)/2 = \frac{|i| + |j|}{2|i||j|}|ij|,$$

the average proportion of the overlap in the genomes. This index, which we refer to as the Maryland Bridge coefficient, has all the advantages of the Jaccard coefficient (such as the independence of the genome sizes) and, in fact, is co-monotone with the latter, but properly evaluates similarity in the cases described. The genome trees produced with these two coefficients are similar and coincide in the deep bifurcations which contribute the most to the data scatter.

The LBH theory suggests another similarity measure according to formula (7). Under this measure, the elements of the similarity matrix $A = YY^T$ are equal to

$$a_{ij} = |ij| - |i| - |j| + h(i) + h(j) + g \tag{12}$$

where $h(i)$ (or $h(j)$) is the total weight of COGs represented in i (or, in j). The weight of a COG k is defined as $1-p_k$ where p_k is the proportion of genomes whose proteins constitute the COG. The greater the number of genomes represented in a COG, the smaller is the weight. The value of $g = \sum_k p_k^2$ is simply an averaging constant. The LBH similarity index (12) is a linear analogue of the Maryland Bridge coefficient. However, it is adjusted according to the information content of the COGs in i and j.

4. Inconsistencies between COGs and an evolutionary tree

4.1. Loss and horizontal transfer events. To assess the consistency of a COG's phyletic profile with the evolution of the corresponding species, a species tree needs to

be specified. Such a tree, based on the phylogenetic analysis of concatenated alignments of ribosomal proteins in [**19**], is shown in Figure 1.

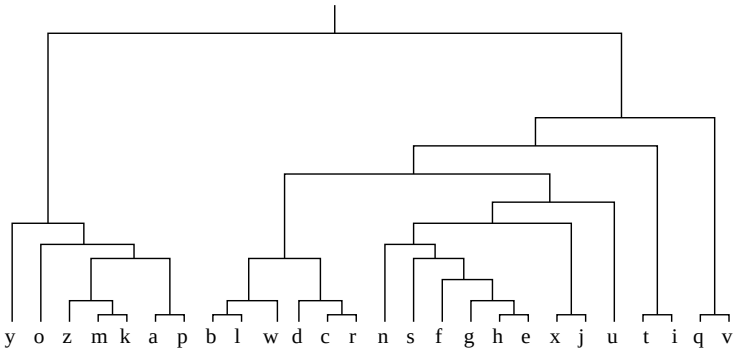

FIGURE 1. The inter-genome tree from [**19**]; the genomes are: y - *Saccharomyces cerevisiae* (and *Candida albicans*), o - *Halobacterium sp.NRC-1*, z - *Aeropyrum pernix*, m - *Methanococcus jannaschii*, (and *Methanobacterium thermoautotrophicum*), k - *Pyrococcus horikoshii* (and *Pyrococcus abyssi*), a - *Archaeoglobus fulgidus*, p - *Thermoplasma acidophilum* (and *Thermoplasma volcanium*), b - *Bacillus subtilis* (and *Bacillus halodurans*), l - *Lactococcus lactis* (and *Streptococcus pyogenes*), w - *Ureaplasma urealyticum* (and *Mycoplasma genitalium*, *Mycoplasma pneumoniae*), d - *Deinococcus radiodurans*, c - *Synechocystis*, r - *Mycobacterium tuberculosis* (and *Mycobacterium leprae*), n - *Neisseria meningitidis*, s - *Xylella fastidiosa*, f - *Pseudomonas aeruginosa*, g - *Vibrio cholera*, h - *Haemophilus influenzae* (and *Pasteurella multocida*), e - *Escherichia coli* (and *Buchnera sp. APS*), x - *Rickettsia prowazekii*, j - *Mesorhizobium loti* (and *Caulobacter crescentus*), u - *Helicobacter pylori* (and *Campylobacter jejuni*), t - *Treponema pallidum* (and *Borrelia burgdorferi*), i - *Chlamydia trachomatis* (and *Chlamydia pneumoniae*), q - *Aquifex aeolicus*, v - *Thermotoga maritima*.

The tree includes 26 entities, which are either individual species or pairs of closely related species (see legend to Figure 1; in the rest of this work we will refer to them as species) and represent the three primary kingdoms of life: bacteria (nineteen species constituting the right-hand cluster), eukaryotes (represented by just one species on the left, y), and archaea (the remaining six species in the left-hand cluster on Figure 1). Here, this tree is assumed to be true, so that the evolutionary histories of COGs are analysed against it.

Let us consider COG1889 (Fibrillarin-like rRNA methylase), which includes genes from seven species, y,o,z,m,k,a,p, which constitute the archaeo-eukaryotic cluster (the left-hand cluster in the species tree in Figure 1. Two evolutionary scenarios can be proposed for this COG. First, the gene could have been present in LUCA (in the root of the species tree) and then lost in the ancestor of the bacterial cluster but retained in archaea and eucaryotes. Second, the gene might not have been present in LUCA and emerged in the common ancestor of archaea and eukaryotes. In spite of the major difference between these scenarios in the biological sense, they assume the same number of evolutionary events explaining the

phyletic pattern of this COG, namely only one event, either loss or emergence during the time covered by the tree.

Note that we attach a special significance to the emergence event if it occurred within the evolutionary time under consideration; the presence or emergence of a COG at the tree root is not considered an event.

Let us turn to COG1102 (Cytidylate kinase), which includes seven species, a,o,m,p,k,z,t, six of which are archaea and one is a bacterium. A parsimonious scenario for this pattern may be this: the gene emerged in the last common ancestor of the archaea and then was horizontally transferred to the spirochetes, altogether two evolutionary events.

A somewhat more complex case of COG1490 (D-Tyr-tRNAtyr deacylase), which includes twelve species, y,b,l,d,r,n,f,g,h,e,q,v, can be explained by either seven losses or the same number of horizontal transfer events. The losses correspond to those clusters of species in which the given COG is not represented, e.g. the archaea. In contrast, the horizontal transfers correspond to those clusters in which the given COG is represented, assuming that it originally evolved in any one of these clusters. Of course, mixed scenarios including both losses and horizontal transfers are also possible and, in some cases, are likely to be the most realistic ones. For instance, in this particular case, a transfer from bacteria to eukaryotes is a distinct possibility. However, in the absence of specific additional data, when there are several equally parsimonious scenarios, we tend to accept the one that involves the most ancient derivation of the COGs and, accordingly, gives preference to losses over horizontal transfer. In general, this is biologically justified because a loss is a more common event than horizontal transfer. Thus, for COG1490, the first scenario, under which the COG was present in LUCA and was subsequently lost on seven independent occasions should be accepted.

The loss and horizontal transfer events occurring in different parts of the tree may both be accepted if they lead to the most parsimonious scenario. Consider, for instance, COG0420 (DNA Repair exonuclease), which includes species y,o,z,m,k,a,p,b,l,d,c,r,f,g,e,t,q,v. Obviously, this COG covers the seven-element cluster on the left in tree Figure 1, but the scenario in the bacterial part of the tree could not be determined unequivocally. It is reasonable to assume that the gene was present in the last common ancestor of cluster b,l,w,d,c,r (with one follow-up loss, at w) but not in the last common ancestor of its sibling cluster n,s,f,g,h,e,x,j,u (assuming it was acquired horizontally by the last common ancestor of cluster f,g,h,e and then lost in h). A scenario with this COG evolving in LUCA adds one more loss, in i, and is the most parsimonious scenario for this COG, with the five evolutionary events as shown in Figure 2.

4.2. Algorithm for determining the inconsistency. The examples above show that the phylogenetic pattern of any COG can be explained in terms of emergence, loss and horizontal transfer events. Since, at the given level of resolution, we cannot distinguish between the events of emergence and horizontal transfer of a COG, we will refer to either of them as a gain. The total number of loss and gain events in a scenario shows the extent of consistency between the evolutionary tree and the given COG according to a particular scenario. Among all possible scenarios, at the given level of resolution, we select the most parsimonious one(s), i.e. the one(s) with the minimum total number of loss/gain events. This minimum number of evolutionary events according to the most parsimonious scenario will be referred to as the COG inconsistency index.

It appears that the problem of building a parsimonious evolutionary scenario for a COG's phylogenetic profile according to a binary evolutionary tree, such as that in Figure 2, can be formalised in terms of an iterative process running through the tree bottom-up

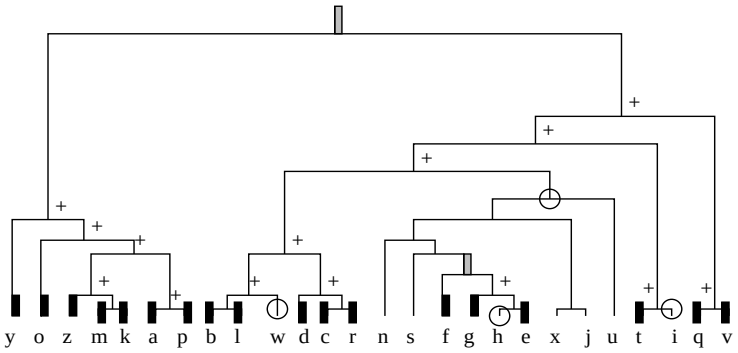

FIGURE 2. A parsimonious evolutionary scenario for COG0420 DNA Repair exonuclease: black rectangles pertain to species represented in the COG; grey rectangles correspond to the COG emergence events; circles show loss events; and + denotes passing the COG from parent to offspring.

by building a parsimonious scenario for a parent, given parsimonious scenarios for its children [**11**]. However, this requires maintaining, at each node of the tree, loss and gain events under both the assumption that the COG was inherited at the given node and the assumption that it was not. To make it clearer, consider any parent-children triple. Two pairs of loss/gain events are assigned to each node in the triple. Sets G_i and L_i assigned to a node consist of those nodes in the subtree descending from the node under consideration which have been, respectively, gained or lost at the node under the assumption that the COG has been inherited by this node. Sets G_n and L_n have similar meaning but under the assumption that the COG has not been inherited by the node. Let us denote the total number of events by $e_i = |G_i| + |L_i|$, under the inheritance assumption, and by $e_n = |G_n| + |L_n|$, under the non-inheritance assumption.

If we assume that the COG has been inherited by the parent, then sets G_i and L_i can be derived under either of the following two alternative scenarios: (i) the COG has been lost in the parent and (ii) the COG has not been lost in the parent. In the first case, the lost COG could not be inherited by the children and, thus, sets L_{n1} and L_{n2} are to be considered loss events and sets G_{n1} and G_{n2} are to be considered gain events. Then the sets at the parent are determined by combining corresponding sets at the children:

$$G_i = G_{n1} \cup G_{n2},\ L_i = L_{n1} \cup L_{n2} \cup \{parent\} \tag{13}$$

The parent is added in the latter equation according to the assumed loss event. In the second case, the COG has been inherited; thus, the loss/gain event sets will be determined by the other pairs of children sets:

$$G_i = G_{i1} \cup G_{i2},\ L_i = L_{i1} \cup L_{i2} \tag{14}$$

Of the two alternatives, the parsimony principle suggests that the scenario with the smaller number of events is accepted. According to the equations in scenario (i), the total number of events will be $e_i = e_{n1} + e_{n2} + 1$, and under scenario (ii), the total will be $e_i = e_{i1} + e_{i2}$. Thus the parsimony principle suggests that we select the scenario with the minimal total number of events:

$$e_i = \min(e_{n1} + e_{n2} + 1, e_{i1} + e_{i2}) \tag{15}$$

If we assume that the COG has not been inherited by the parent, there also can be two alternative scenarios for sets L_n and G_n: (i) the COG has been gained by the parent and (ii) the COG has not been gained by the parent. In the first case, obviously,

$$G_n = G_{i1} \cup G_{i2} \cup \{parent\},\ L_n = L_{i1} \cup L_{i2} \tag{16}$$

and, in the second case,

$$G_n = G_{n1} \cup G_{n2},\ L_n = L_{n1} \cup L_{n2} \tag{17}$$

The scenario with minimal total inconsistency is to be accepted:

$$e_n = \min(e_{i1} + e_{i2} + 1,\ e_{n1} + e_{n2}) \tag{18}$$

These equations show how to combine all the needed events by iteratively proceeding from leaves to the root to find a parsimonious scenario and the corresponding inconsistency index. The paper [**11**] contains this and other developments related to modelling losses and horizontal transfers in a species tree.

5. Applying LBH to phyletic profiles of COGs

5.1. LBH tree with all COGs. Figure 3 shows the genome tree that was built on the basis of the complete matrix of representations of 26 genomes in 3166 COGs by using the approach outlined above.

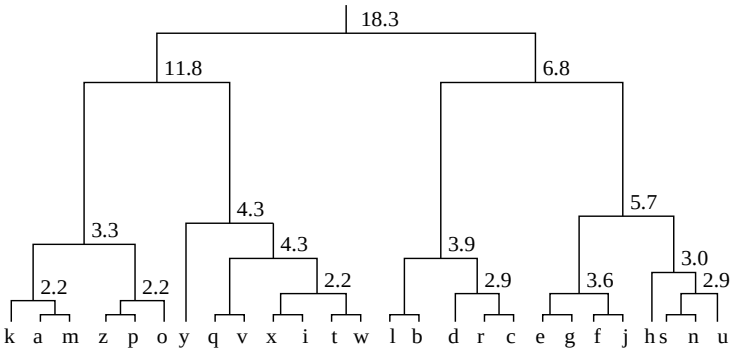

FIGURE 3. The LBH inter-genome tree. The percentage values at the splits show the proportion of the data scatter accounted for by each split.

Clearly, the deep branchings in the tree shown in Figure 3 do not reflect phylogenetic relationships betwen species. Indeed, the first split, which accounts for almost 20% of the data scatter separates archaea, eukaryotes and the parasitic bacteria with very small genomes (chlamydia, spirochetes and rickettsia) from the rest of the bacteria. While the grouping of archaea and eukaryotes is phylogenetically valid according to the standard model of early evolution [**18**], the inclusion of the bacterial parasites in the same cluster cannot be explained within this paradigm. However, examination of the lists of COGs that make the greatest contribution to the split (ftp://ncbi.nlm.nih.gov/pub/koonin/mirkin) clarifies the basis for this bifurcation. Specifically, the composition of the two largest clusters seems to be a combined product of a negative contribution of the loss of many characteristic bacterial genes in chlamydia, spirochetes and rickettsia and a positive contribution of several COGs that are indeed shared by this subset of parasitic bacteria and

archaea and/or eukaryotes, to the exclusion of the rest of bacteria. This small list consists primarily of the subunits of vacuolar/archaeal proton ATPase and also includes the class I lysyl-tRNA synthetase. The most likely explanation of the presence of these genes in the small bacterial genomes involves horizontal transfer from archaea [**5**]). Similar explanations based on a negative contribution corresponding to lineage-specific gene loss and positive contributions that are formed both by common heritage and by horizontal transfer can be offered for other deep bifurcations of the tree. In contrast, some of the terminal clusters appear to carry predominantly phylogenetic information. In addition to the obvious unifications such as, for example, the three lineages of euryarchaea (*A. fulgidus, M. jannaschii/M. thermoautotrophicum* and the two *Pyrococci*) or *B. subtilis* with *C. acetobutilicum*, some of these assemblages might reflect still undetected evolutionary relationships between major bacterial lineages. Such putative high-level evolutionary clustering, which seems to be compatible with the results produced by other methods of whole-genome evolutionary analysis [**19**] includes the chlamydia-spirochete group, and the group formed by *Deinococcus, Mycobacterium* and *Synechocystis* (Figure 3).

5.2. LBH tree built with relevant COGs. The poor performance of the LBH method on the COG data as assessed against the species tree, described in the previous section, may be thought of as resulting from the inconsistency of many COGs with the species tree, which is supported by the fact that other methods of tree-building, such as Neighbour-Joining, using the same data also led to inconsistent trees [**19**]. Thus, it seemes reasonable to remove those COGs, in which the evolutionary signal has been washed out by multiple loss and horizontal transfer events, so that they have become irrelevant to the evolutionary tree building. To do so, we use an index that takes into account the inconsistency index relative to the COG size. Thus, we define the irrelevance index for a COG c, which includes n species, as the ratio $i(c) = I(c)/n$. For a COG c that requires five loss/gain events in a parsimonious scenario as shown in Figure 2, the irrelevance index is $i(c) = 0.5$ if the COG includes 10 species and $i(c) = 0.25$ if the COG includes 20 species.

Removing COGs with a large irrelevance index may drastically reduce the number of COGs available for analysis. For instance, there are only 1073 COGs (of 3166) with $i(c) \leq 0.4$. Figure 4 shows the LBH tree built with this set of COGs.

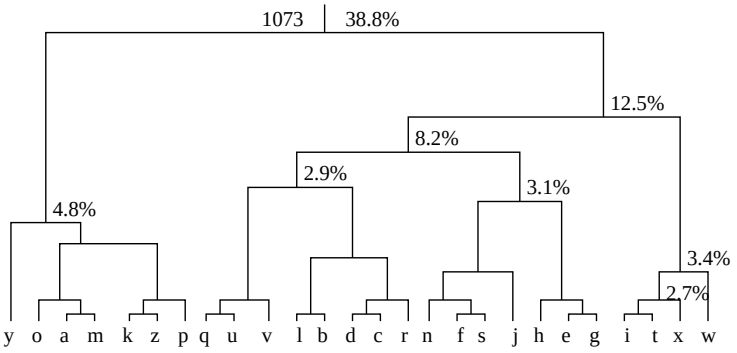

FIGURE 4. The inter-genome tree based on COGs with the irrelevance index less than 0.4.

Obviously, this tree is much closer to the hypothetical species tree in Figure 1 than the tree in Figure 3. In particular, the three kingdoms: eucaryotes, archaea and bacteria, are separated correctly.

5.3. Removing "worked-out" COGs. To further improve the resulting tree, we are going to exploit the structure of our top-down approach as an evolutionary model. Indeed, each evolutionary divergence of species is accompanied by divergence of some characters. The diverged characters can be realistically assumed to never diverge again. Thus, in our process of building divergences one by one we may reasonably assume that those COGs which went to only one of the branches after a divergence can be removed from further divergences as they represent characters that already have been "worked-out" by evolution.

Unfortunately, because of the lack of evolutionary consistency among COGs, it is a rare event that a COG belongs in only one part of a split of the genome set. Thus, we introduce a measure of the extent of COG's divergence according to the linear theory described above. For a COG k and split $\{S_{w1}, S_{w2}\}$ of cluster S_w, the divergence index is defined as $d(k) = (\frac{p_{w1k}}{p_{w1}} - \frac{p_{w2k}}{p_{w2}})^2$ in (10), the squared difference of probability of k in different split parts. For instance, if COG k is present in 80% of genomes in one part, and 10% of genomes in the other part of a split, then $d(k) = (0.8 - 0.1)^2 = 0.49$. The greater the difference the larger $d(k)$ so that $d(k) = 1$ if and only if k completely covers one part of a split and does not belong in the other part at all.

The algorithm for building an evolutionary tree is the same as the sequential splitting algorithm based on criterion (5) or, equivalently, (8), except for the similarity matrix (a_{ij}), which is built anew after each split, at the split cluster, by using only COGs k with $d(k) \leq s$ where s is a pre-specified threshold. In Figure 5, a tree built by this method with the divergence threshold of $s = 0.5$ is shown.

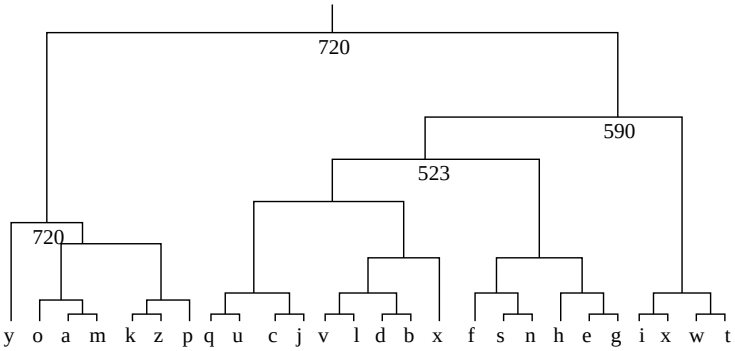

FIGURE 5. The LBH inter-genome tree built with 1073 relevant COGs and consecutively removing the worked-out COGs from further splits. The numbers under splits show the numbers of relevant cogs left after all meaningful (that is, taking into account more than 2.5% of the data scatter) splits.

The tree in Figure 5 does not much differ from the tree in Figure 4, although some changes in the terminal branches are seen. For instance, j has been moved out of the cluster n, s, f, g, h, e, which now coincides with the corresponding cluster in the species tree in Figure 1. It should be noted, however, that most of the terminal bifurcations in the tree appear to be unreliable because they account for only a small proportion of the data scatter.

6. Conclusion

In this paper, we describe a new method for building and analysing evolutionary trees. Unlike other tree building algorithms, this method builds a tree by sequentially splitting the set of species top-down, thus attempting to mimic the actual sequence of events during evolution. This feature allows us to additionally utilise the following features:

(1) stopping the splitting process when it becomes unreliable as determined on the basis of the concept, following from the LBH theory, of the proportion of the data scatter accounted for by a split;

(2) estimating the extent of involvement of a COG in a splitting event by calculating the divergence index $d(k)$ and removing worked-out COGs from further splits.

On the mathematical side, the linear binary hierarchy approach allows one to explicitly establish a mathematical relationship between the presence/absence character-to-genome data and the corresponding genome-to-genome similarity data. Moreover, the genome-to-genome similarity index derived using this theory is advantageous to the well-known Jaccard index: (1) in contrast to the Jaccard index, our measure does not underestimate the similarity betwene genomes of different size, and (2) it involves not only comparison of species according to the presence/absence patterns, but also assigns diferent characters with different weights according to their frequencies.

On the biological side, it appears that, in the majorty of the COGs, about two-third, the phylogenetic signal is confounded by multiple gene loss and horizontal transfer events and these COGs should be removed before applying a tree building method to the COG presence/absence data. The trees built with relevant COGs agree with the species tree on the level of deep branching. The recent branchings appear to be unreliable because they account for only a small part of the data scatter.

We believe that LBH is a potentially useful approach for in-depth analysis of the relationships between genomes. The principal application of this methodology, in our view, involves constructing most parsimonious evolutionary scenarios for all COGs and combining them in order to reconstruct the gene sets of ancestral life forms.

References

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth International Group, Belmont, Ca., 1984.

[2] W. F. Doolittle, Phylogenetic classification and universal tree, *Science* **284(5423)** (1999), 2124–9.

[3] J. Felsenstein, *PHYLIP 3.6: Phylogeny Inference Package*, http://evolution.genetics. washington.edu/phylip/, 2001.

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

[5] E. V. Koonin, L. Aravind, and A. S. Kondrashov, The impact of comparative genomics on our understanding of evolution, *Cell* **101** (2000), 573–576.

[6] J. G. Lawrence, Gene transfer, speciation, and the evolution of bacterial genomes, *Curr. Opin. Microbiol.* **2** (1999), 519–523.

[7] W. Li and D. Graur, *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, MA, 1991.

[8] K. S. Makarova, L. Aravind, M. Y. Galperin, N. V. Grishin, R. L. Tatusov, Y. I. Wolf, and E. V. Koonin, Comparative genomics of the Archea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell, *Genome Research* **9** (1999), 608–628.

[9] B. Mirkin, Linear embedding of binary hierarchies and its applications, in (B. Mirkin, F. McMorris, F. Roberts, and A. Rzhetsky, eds.), *Mathematical Hierarchies and Biology*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science **37**, American Mathematical Society: Providence, 1997, 331–356.

[10] B. Mirkin, Least-squares structuring, clustering, and related computational issues, *The Computer Journal* **41** (1998), 518–536.

[11] B. Mirkin, T. Fenner, and E. Koonin, Modelling loss and horizontal transfer of clusters of orthologous groups in evolutionary trees (2002), submitted.
[12] R. D. M. Page and E. C. Holmes, *Molecular Evolution: A Phylogenetic Approach*, Blackwell Scientific, Oxford, 1998.
[13] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy*, W. H. Freeman, San Francsco, 1973.
[14] B. Snel, P. Bork, and M. A. Huynen, Genome phylogeny based on gene content, *Nat. Genet.* **21** (1999), 108–10.
[15] D. C. Swofford *PAUP* Version 4: Tools for Inferring and Interpreting Phylogenetic Trees*, Sinauer Associates, Sunderland, MA, 2002.
[16] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, The COG database: a tool for genome-scale analysis of protein function and evolution, *Nucleic Acids Research* **28** (2000), 33–36.
[17] F. Tekaia, A. Lazcano and B. Dujon, The genomic tree as revealed from whole proteome comparisons, *Genome Research* **9** (1999), 550–557.
[18] C. R. Woese, O. Kandler, and M. L. Wheelis, Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87** (1990), 4576–4579.
[19] 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 Evolution Biology* **1** (2001), 8–15.

SCHOOL OF COMPUTER SCIENCE AND INFORMATION SYSTEMS, BIRKBECK COLLEGE, MALET STREET, LONDON, WC1E 7HX, UK
E-mail address: mirkin@dcs.bbk.ac.uk

NCBI, NIH, BLDG. 38A, RM. 5N503, 8600 ROCKVILLE PIKE, BETHESDA, MD 20894, USA
E-mail address: koonin@ncbi.nlm.nih.gov