Building genome classification trees with linear binary hierarchies

Boris Mirkin*† 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. The resulting classification tree has the potential of revealing previously unnoticed but potentially biologically relevant relationships between the corresponding organisms.

Introduction

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 (Snel, Bork, Huynen, 1999; Fitz-Gibbon, House, 1999; Tekaia, Dujon, 1999). The critical issue with these attempts on 'whole-genome' tree building is, to what extent the topology of the resulting trees reflects phylogeny and to what extent it is shaped by other evolutionary processes such as lineage-specific gene loss and horizontal gene transfer.

Here we use the theory of linear binary hierarchies (Mirkin 1997, 1998) that combines features of both character-based and distance-based analyses to a more detailed examination of the representation of genomes in orthologous families. The latter were derived from the system of Clusters of Orthologous Groups of proteins (COGs) (Tatusov et al., 1997, 2000; http://www.ncbi.nlm.nih.gov/COG/). The resulting classification tree primarily reflects similarities between the lifestyles of the respective organisms although the terminal branches appear to carry a phylogenetic signal also.

*School of Computer Science and Information Systems, Birkbeck College, Malet Street, London, WC1E 7HX, UK; e-mail: mirkin@dcs.bbk.ac.uk

†Communicating co-author.

‡NCBI, NIH, Bldg. 38A, Rm. 5N503, 8600 Rockville Pike, Bethesda, MD 20894, USA; e-mail: koonin@ncbi.nlm.nih.gov

1 Linear theory for binary hierarchies

1.1 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 (Mirkin, 1997). 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 restrict our attention to binary trees in which every interior node has two children, which is the normal situation with evolutionary trees. The values a_w and b_w are (essentially) uniquely defined by the condition that the vector ϕ_w be centered and normalized. The former 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. It is not difficult to prove that the vectors ϕ_w are 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.

1.2 Building an LBH for a data matrix

This construction 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 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 have been 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 $C^* = \Phi^T Y$ and

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

where $\mu_w^2 = \frac{N_{w1} N_{w2}}{N_w} d^2(y_{w1}, y_{w2})$, 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 $d^2(y_1, y_2) = \sum_k (y_{1k} - y_{2k})^2$, the square of the 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 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 (Mirkin, 1998), that is, by doing splits iteratively, one split at a time.

1.3 Deriving an LBH for a similarity matrix

A square genome-to-genome similarity matrix $a = (a_{ij})$ can be similarly decomposed into a linear combination of the matrices P_w, whose elements are $\phi_{iw}\phi_{jw}$, and residuals to be minimized.

$$a_{ij} = \sum_{w=1}^{N-1} \lambda_w \phi_{iw} \phi_{jw} + e_{ij} \tag{4}$$

Matrix $a = (a_{ij})$ must be preliminarily double-centered by subtraction of within-column averages from the columns and within-row averages from the rows with consecutive adding the original grand mean to all the entries.

Due to orthonormality of the set P_w, $w \in W$, for any binary hierarchy S_W, the squared data scatter also can be decomposed in the explained and unexplained parts:

$$\sum_{i,j \in I} a_{ij}^2 = \sum_{w=1}^{N-1} \lambda_w^2 + \sum_{i,j \in I} e_{ij}^2 \tag{5}$$

whenever there is orthogonality of the residual matrix $E = (e_{ij})$ to the rest, as it is when $a = (a_{ij})$ has been projected in the subspace generated by matrices $\{P_w : w \in S_W\}$ in the parallel or sequential manner. Moreover, it can be found from (4) that the least-squares optimal coefficients are the scalar products (a, P_w) satisfying the equations $\lambda_w = \sum_{i,j \in I} a_{ij} \phi_{iw} \phi_{jw}$. With little arithmetic this can be transformed into:

$$\lambda_w = \frac{N_{w_2} N_{w_1}}{N_w} (a(w_1, w_1) + a(w_2, w_2) - 2a(w_1, w_2)) \tag{6}$$

where

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

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

When an element $i \in S_{w1}$ is moved into S_{w2}, the change $\Delta(i)$ of the criterion λ_w can be easily expressed through elements of matrix a and the average similarities.

The algorithm suggested for maximization of λ_w 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_{w1} and S_{w2} along with corresponding λ_w. 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 algorithm above, all members of a cluster are more attracted to their own cluster than 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.

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

The following features of our approach will be emphasized here:

- reformulation of the criterion with regard to 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.

2.1 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:

$$y_{wk} = p_{wk}/p_w - p_k. \tag{7}$$

which implies

$$c^2_{wk} = \frac{n_{w1}n_{w2}}{n_w}\left(\frac{p_{w1k}}{p_{w1}} - \frac{p_{w2k}}{p_{w2}}\right)^2 \tag{8}$$

The first factor here "takes care" of making the split closer to halving (which corresponds to the information concepts of the search theory), while the second factor maximizes the difference between the frequencies of 1's in the subclusters. It should be noted that this measure closely resembles the so-called "twoing rule" used in the CART techniques for deriving decision trees; see Breiman et al. (1984).

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

$$\mu^2_w = \frac{n_{w1}n_{w2}}{n_w} d^2(p(w1), p(w2)) \tag{9}$$

where $p(w)$ is the vector of (conditional) probabilities of categories k in cluster S_w.

2.2 Contribution of a split and stopping the splitting process

The quantity μ^2_w in (9) 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.

2.3 Most contributing COGs

A feature's contribution to the data scatter is equal to the value of c^2_{wk} in (8). 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 can be exploited for interpretation of the split.

2.4 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 (see, for instance, Snel, Bork and Huynen (1999)). 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, while 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 intuition on the evolutionary relationships between genomes i and j because the smaller of them 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, referred to as the Maryland Bridge coefficient, has all the advantages of the Jaccard coefficient (such as 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 one more similarity measure derived by multiplication of the equation (1) by its transposed version. Under this measure, the elements of the similarity matrix $A = YY^T$ are equal to

$$a_{ij} = |ij| - |i| - |j| + h(i) + h(j) + g \qquad (10)$$

where $h(i)$ (or $h(j)$) is the total weight of COGs represented in i (or, in j). The weight of a COG 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 formula (10) is a linear analogue to the Maryland bridge coefficient. However, it is adjusted according to the information content of the COGs in i and j.

3 Results

Figure 1 shows the genome tree that was built on the basis of the complete matrix of representation of 26 genomes in 2575 COGs by using the approach outlined above.

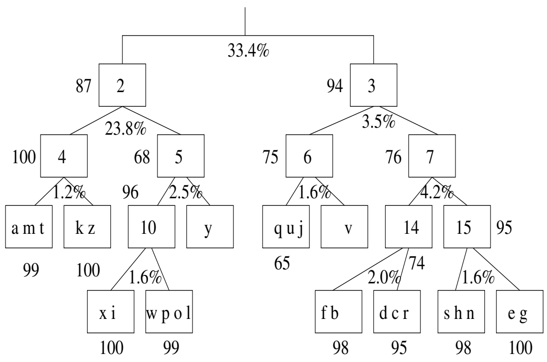

Figure 1: The inter-genome tree; the genomes are: a - *Archaeoglobus fulgidus*, m - *Methanococcus jannaschii*, t - *Methanobacterium thermoautotrophicum*, k - *Pyrococcus horikoshii* (and *Pyrococcus abyssi*), z - *Aeropyrum pernix*, y - *Saccharomyces cerevisiae*, q - *Aquifex aeolicus*, v - *Thermotoga maritima*, d - *Deinococcus radiodurans*, r - *Mycobacterium tuberculosis*, b - *Bacillus subtilis*, c - *Synechocystis*, e - *Escherichia coli*, f - *Clostridium acetobutilicum*, g - *Vibrio cholera*, h - *Haemophilus influenzae*, s - *Xylella fastidiosa*, n - *Neisseria meningitidis*, u - *Helicobacter pylori*, j - *Campylobacter jejuni*, x - *Rickettsia prowazekii*, o - *Borrelia burgdorferi*, l - *Treponema pallidum*, i - *Chlamydia trachomatis* (and *Chlamydia pneumoniae*), w - *Ureaplasma urealyticum*, p - *Mycoplasma genitalium* (and *Mycoplasma pneumoniae*).The numbers next to the clusters show how many times a cluster occurred in 100 bootstrap trials (Felsenstein, 1996); the percentage values show the proportion of the data scatter accounted for by each split. The splitting process was terminated at 1%.

Clearly, the deep branchings in the tree shown in Figure 1 do not reflect phylogenetic relationships. Indeed, the first split which accounts for 1/3 of the data scatter separates archaea, eukaryotes and the parasitic bacteria with very small genomes (chlamydia, spirochetes and rickettsia) from the rest of bacteria. While the grouping of archaea and eukaryotes is phylogenetically valid according to the "standard model" of early evolution (Woese, Kandler, Wheelis, 1990), 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 (Koonin, Aravind, Kondrashov, 2000). Similar explanations based on a 'negative' combination 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 euryarchaea (*A. fulgidus, M. jannaschii and M. thermoautotrophicum*) 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 (Grishin, Rogozin, Wolf, Tatusov, Koonin, unpublished observations) includes the chlamydia-spirochete group, and the group formed by *Deinococcus, Mycobacterium* and *Synechocystis* (Figure 1).

We believe that LBH is a promising approach for in-depth analysis of the relationship between genomes. Further developments should include attempts to explicitly incorporate the horizontal gene transfer and differential gene loss phenomena into the classification procedure.

References

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

[2] J. Felsenstein (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods, *Methods Enzymol.*, 266, 418-427.

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

[4] E.V. Koonin, L. Aravind, A. S. Kondrashov (2000) The impact of comparative genomics on our understanding of evolution, *Cell*, 101, n. 6, 573-576.

[5] B. Mirkin (1997) 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, V. 37, American Mathematical Society: Providence, 331-356.

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

[7] B. Snel, P. Bork, and M.A. Huynen (1999) Genome phylogeny based on gene content, *Nat. Genet.*, 21, no. 1, 108-10.

[8] R.L. Tatusov, M.Y. Galperin, D.A. Natale, and E.V. Koonin (2000) The COG database: a tool for genome-scale analysis of protein function and evolution, *Nucleic Acids Research*, 28, no.1, 33-36.

[9] R.L. Tatusov, E.V. Koonin, D. J. Lipman (1997) A genomic perspective on protein families, *Science*, 278(5338), 631-637.

[10] F. Tekaia and B. Dujon (1999) Pervasiveness of gene conservation and persistence of duplicates in cellular genomes, *J. Mol. Evol.*, 49, 591-600.

[11] C. R. Woese, O. Kandler, and M. L. Wheelis (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya, *Proc Natl Acad Sci U S A*, 87, 4576-4579.