Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 472719, 7 pages
doi:10.1155/2008/472719

Research Article

Unsupervised Learning in Detection of Gene Transfer

L. Hamel,[1] N. Nahar,[1] M. S. Poptsova,[2] O. Zhaxybayeva,[3] and J. P. Gogarten[2]

[1] *Department of Computer Science and Statistics, University of Rhode Island, Kingston, RI 02881, USA*
[2] *Department of Molecular and Cell Biology, College of Liberal Arts and Sciences, University of Connecticut, CT 06269, USA*
[3] *Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS, Canada B3H 1X5*

Correspondence should be addressed to L. Hamel, hamel@cs.uri.edu

Received 13 September 2007; Accepted 14 December 2007

Recommended by Daniel Howard

The tree representation as a model for organismal evolution has been in use since before Darwin. However, with the recent unprecedented access to biomolecular data, it has been discovered that, especially in the microbial world, individual genes making up the genome of an organism give rise to different and sometimes conflicting evolutionary tree topologies. This discovery calls into question the notion of a single evolutionary tree for an organism and gives rise to the notion of an evolutionary consensus tree based on the evolutionary patterns of the majority of genes in a genome embedded in a network of gene histories. Here, we discuss an approach to the analysis of genomic data of multiple genomes using bipartition spectral analysis and unsupervised learning. An interesting observation is that genes within genomes that have evolutionary tree topologies, which are in substantial conflict with the evolutionary consensus tree of an organism, point to possible horizontal gene transfer events which often delineate significant evolutionary events.

1. INTRODUCTION

Evolutionary history of species is now inferred from the evolutionary histories of their genomes. Genomes can be viewed as a collection of genes and whole genome evolution is concluded from the evolution of individual genes. If the majority of genes followed the same evolutionary history, supertree approaches can be used to calculate a majority consensus tree. However, evolutionary trees of individual genes can differ from the majority [1], and in this case, the consensus tree is embedded in a network represented by the histories of the different genes. Evolutionary tree topologies of genes that conflict with the consensus tree are strong indicators of horizontal gene transfer events. Given this, it is clear that organismal evolution cannot be inferred from studying the evolution of just a few genes but must be inferred from studying as many (orthologous) genes as possible.

To construct and evaluate an evolutionary consensus tree based on multiple genes for a set of genomes, it is advisable to construct all possible evolutionary tree topologies for these genomes and measure the support of each topology by the (orthologous) genes within the genomes. Unfortunately, evaluating all possible tree topologies is computationally intractable for any but a very small set of genomes, since the number of possible tree topologies grows factorially with the number of participating genomes. An approach based on the spectral analysis of genomic data using bipartitions [2, 3] allows the inference of consensus trees from smaller quanta of phylogenetic information, side stepping some of the difficult computational issues. Table 1 shows the number of possible trees versus the number of possible bipartitions given a fixed set of genomes. With n taxa there are $(2n-5)!/[2^{(n-3)}(n-3)!]$ different unrooted tree topologies. The number of possible nontrivial bipartitions for n taxa is given by the formula $2^{(n-1)} - n - 1$, and it grows much slower with an increasing number of species than the number of different trees. We refer to the approach based on bipartitions as *spectral genome analysis*.

It is worth noting that when a single tree is calculated from the combination of all genes, including genes that were horizontally transferred, the topology of the resulting tree might not represent the plurality of gene histories. Therefore, a detailed analysis of the evolutionary histories of the participating genes is of interest. The techniques outlined here support this kind of analysis.

Table 1: Number of possible trees and bipartitions given a fixed set of genomes.

Number of genomes	Number of unrooted trees	Number of nontrivial bipartitions
4	3	3
5	15	10
6	105	25
7	945	56
8	10,395	119
9	135,135	246
10	2,075,025	501
20	2.22E + 20	5.24E + 05
50	2.84E + 74	5.63E + 14
n	$(2n-5)!/[2^{(n-3)}(n-3)!]$	$2^{(n-1)}-n-1$

In spectral genome analysis, each set of orthologous genes (a gene family) is associated with a particular set of bipartitions (its *spectrum*) that define its evolutionary tree. Thus, we can envision a gene family as a point in the space spanned by all possible bipartitions of a set of genomes. Here, we apply unsupervised learning in the form of self-organizing maps [4] to this space and obtain a visual representation of clusters of gene families with similar spectra. The spectra of the gene families within a particular cluster allow us to infer the consensus tree for that cluster. It is now possible to investigate whether the consensus tree topologies of the clusters are compatible or conflicting with the overall consensus tree. If a cluster of gene families is discovered that conflicts with the consensus tree topology, then this is a strong indication for a horizontal gene transfer event. The advantage of this approach is that we not only see a distinction between consensus and conflicting trees, but that we can detect trends of agreement between the conflicting genes. This additional insight might provide biological clues as to the nature of the origin of these genes.

Unsupervised learning has been used in genomic analyses before (e.g., [5]). However, our approach seems to be novel in that we do not apply unsupervised learning directly to DNA sequence data but instead analyze the much more abstract representation of the genomic data in the form of bipartitions. We have constructed a web service called Gene Phylogeny eXplorer (GPX, http://bioinformatics.cs.uri.edu/gpx) that supports spectral genome analysis [6].

2. MATERIALS AND METHODS

2.1. Spectral analysis of evolutionary trees

Given n entities, there are $2^{n-1}-1$ different ways to assign the entities to two different nonempty sets. That is, there are $2^{n-1}-1$ different *bipartitions* of n entities including trivial bipartitions. An (unrooted) tree can be viewed as a model of the evolutionary relationships between n entities or taxa such as species, genes, molecules, and so forth. Trees and bipartitions are related as follows. Each edge in a tree can be seen as dividing the tree into a bipartition: the leaf nodes that can be reached from one end of the edge form one set of taxa and the leaf nodes that can be reached from the other end of the edge form the other set of taxa. A binary tree with n leaf nodes has exactly $2n-3$ edges. Thus, an evolutionary tree relating n taxa gives rise to $2n-3$ bipartitions. It is easy to see that $2n-3 < 2^{n-1}-1$, that is, the number of bipartitions defined by an evolutionary binary tree of n taxa is much smaller than the number of possible bipartitions of n entities.

Trivial bipartitions, which is bipartitions where one of the partitions is a singleton set, do not contain any phylogenetic information. Thus, given n entities, there are $2^{(n-1)}-n-1$ different nontrivial bipartitions. However, in an unrooted binary tree with n leaf nodes there are $n-3$ interior edges and therefore $n-3$ nontrivial bipartitions. An interior edge is an edge that is not incident to a leaf node of a tree. It is evident that $n-3 < 2^{n-1}-n-1$, that is, the number of nontrivial bipartitions generated by a tree is much smaller than the number of possible nontrivial bipartitions.

Let t_n be an evolutionary tree over n taxa, then we define the bipartitions of t_n as the *spectrum* of t_n, denoted as $S(t_n)$. It is convenient to adopt a vector notation for the spectrum $S(t_n) = (b_1, \ldots, b^n_{2-1}) = (0, 1, 1, 0, \ldots, 0, 0)$, where b_k denotes bipartition k with $1 < k < 2^{n-1}-1$. Here, $b_k = 1$ if the spectrum of the tree includes bipartition b_k, and $b_k = 0$ otherwise. Note that the vector notation is a representation over all possible bipartitions. Given this, we can now refer to a *bipartition space* and we can readily see that a spectrum of a particular evolutionary tree t_n represents the coordinates of a point in that space. In our case, where the tree represents the evolutionary relationship between orthologous genes in n genomes, we often refer to the spectrum as the gene family spectrum and therefore a gene family is denoted by a point in bipartition space.

Figure 1(a) is an unrooted tree relating five taxa A through E. The arrows indicate branches defining the nontrivial bipartitions in this tree. Figure 1(b) represents a bipartition corresponding to the left arrow in Figures 1(a) and 1(c) represents a bipartition corresponding to the right arrow in Figure 1(a), respectively. Observe that the sub-tree topologies in the bipartitions are unresolved.

By further generalizing and interpreting the values in the spectrum vectors as arbitrary real numbers, as we will do in what follows when we assign confidence values to bipartitions, a bipartition space can be viewed as a $2^{n-1}-1$ dimensional real vector space. An interesting consequence of this is

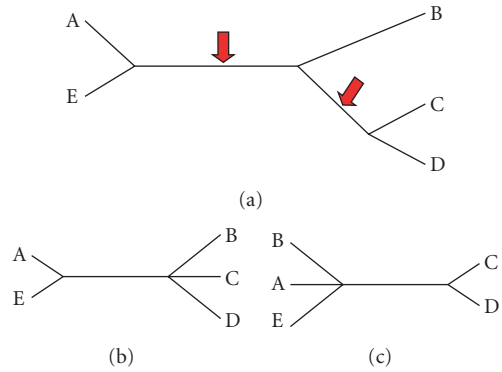

Figure 1: (a) An unrooted tree with 5 taxa, (b) the bipartition corresponding to the left arrow above, (c) the bipartition corresponding to the right arrow above.

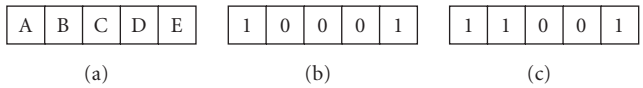

Figure 2: (a) A binary vector indexed by taxa names, (b) a binary representation of the bipartition in Figure 1(b), (c) a binary representation of the bipartition in Figure 1(c).

that we can now measure the difference between spectra as the Euclidean distance between the two corresponding spectrum points in a bipartition space. Let t_1, t_2, and t_3 be three different evolutionary trees of n taxa and let $S(t_1)$, $S(t_2)$, and $S(t_3)$ be the respective spectra, then we say that $S(t_2)$ is more similar to $S(t_1)$ than $S(t_3)$ if $\|S(t_1) - S(t_2)\| < \|S(t_1) - S(t_3)\|$, here the operator $\|\cdot\|$ denotes the Euclidean distance between two points in bipartition space.

2.2. Representation of bipartitions

Let A be a set of n elements, and b is a bipartition defined on a set A. Each bipartition b splits a set A into two subsets m and its complement m^C, such that $A = m \cup m^C$.

We say that two bipartitions are *compatible* if there exists a tree whose spectrum includes both bipartitions. We say that two bipartitions are *conflicting* if they cannot appear in the same spectrum. In set notation, two bipartitions are compatible if a set (either m or m^C) of one bipartition is a subset of one of the sets of the second bipartition; or, in other words, bipartitions b_1 and b_2 are compatible if and only if one of four possible conditions is satisfied:

$$(m_1 \subset m_2), (m_1 \subset m_2^C), (m_1^C \subset m_2), \text{ or } (m_1^C \subset m_2^C). \quad (1)$$

To handle bipartitions computationally in an efficient way, we can represent them effectively as binary masks. Figure 2(a) shows a binary vector indexed by the taxa in Figure 1(a). Figure 2(b) shows the binary representation of the bipartition in Figure 1(b) arbitrarily assigning 1 and 0 to the left and right bipartition, respectively. Figure 2(c) shows the binary representation of the bipartition in Figure 1(c).

Given our binary representation of bipartitions, there is a simple computation to test for compatibility between

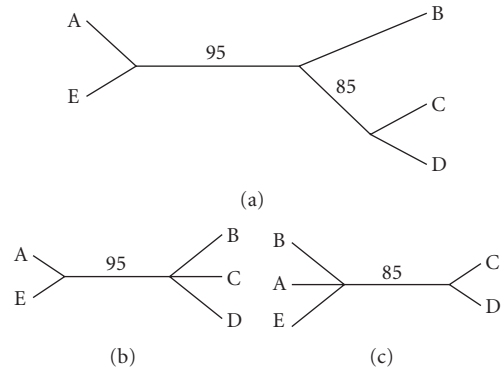

Figure 3: (a) Bootstrapped consensus tree with 5 taxa, (b) a bipartition with a 95% bootstrapped confidence value, (c) a bipartition with an 85% bootstrapped confidence value.

bipartitions. We say that two bipartitions are compatible if the following returns true:

$$\begin{aligned} &((b_1 \mid b_2) == b_1) \parallel, \\ &((b_1 \mid b_2) == b_2) \parallel, \\ &((b_1 \mid \sim b_2) == b_1) \parallel, \\ &((b_1 \mid \sim b_2) == \sim b_2) \parallel, \end{aligned} \quad (2)$$

where b_1 and b_2 denote bipartitions. Here the "|" operator represents the bitwise OR operation, the "~" operator represents the bitwise negation, the "||" operator represents the logical OR operation, and "==" represents the bitwise equality operator. Given the two masks from Figures 1(b) and 1(c), it is easy to see that they are compatible:

$$10001 \mid 11001 == 11001. \quad (3)$$

On the other hand, the bipartitions 11001 and 10011 are conflicting.

2.3. Consensus trees

It is customary to compute confidence values for the edges in an evolutionary tree via bootstrapping [7]. The computed tree represents a consensus tree over the bootstrap samples. The confidence values are typically chosen between 0 and 100. With this, a bipartition derived from a particular edge in the bootstrap consensus tree inherits the confidence value of that edge. This allows us to refine our spectrum vector notation, for example, $S(t_n) = (0, 67, 85, 0, \ldots, 15, 0)$, where t_n is now a bootstrapped consensus tree and the values in the vector represent the confidence values for the individual bipartitions.

Figure 3(a) shows a bootstrapped consensus tree with five taxa. The values on the edges represent the bootstrapped confidence values. Figures 3(b) and 3(c) show nontrivial bipartitions of the tree. Notice that the bipartitions inherit the confidence value of the edge that corresponds to the bipartition.

By computing a consensus tree on the bootstrap samples, it is possible to introduce biases due to the fact that phylogenies that do not agree with the plurality are suppressed. This

is particularly critical in our case where the biases of this kind of computation might compound during an analysis. A different approach that avoids computing a consensus tree too early in an analysis is by taking advantage of the spectra of the bootstrap samples. Before we can describe this construction, we need to define what we mean by an *average spectrum*. Given m spectra, $S_1, \ldots, S_m$, in a bipartition space of n taxa, we define the average spectrum S_a as

$$S_a = \frac{1}{m} \sum_{k=1}^{m} S_k. \tag{4}$$

The summation of spectra is well defined as vector additions in bipartition space and the multiplication of a scalar and a vector simply scales the components of the vector.

The bootstrap approach can be summarized as follows.

(1) For the phylogenetic tree of each bootstrap sample, compute the corresponding spectrum.
(2) Compute the average spectrum S_a over the bootstrap spectra.
(3) The values that appear in the vector for the average spectrum can now be interpreted as *confidence values*.

In step 3, we could multiply the average spectrum by 100 to make it compatible with the traditional bootstrap confidence values. A consequence of this approach is that the average spectrum is no longer guaranteed to represent a phylogenetic tree due to possible bipartition conflicts and this represents an extension of our definition of spectrum above that did not admit any conflicts. However, even in this extended definition of a spectrum we can retrieve a consensus tree from the average spectrum S_a as follows:

(1) Sort the bipartitions in S_a according to their confidence values.
(2) Delete all bipartitions in S_a that conflict with more strongly supported bipartitions in S_a.
(3) Incrementally construct a consensus tree from the remaining bipartitions in S_a, starting with the bipartition with the strongest support to the bipartition with the weakest support.

Observe that computing the consensus tree for the average spectrum is a lossy operation (step 2) as before. However, the advantage of this approach is that we can defer this lossy operation as long as necessary. Note that we need only $n - 3$ top nonconflicting bipartitions. If conflicts are singular or minor events, they will not appear in the top $n - 3$ bipartitions because their confidence values will be low. If the conflicting bipartitions are among top $n - 3$, then the case deserves special attention. If the confidence values for bipartitions are rather small and randomly distributed over the data, this can serve as an indication that the data do not have a clean phylogenetic signal.

An interesting application of this is the construction of a consensus tree of multiple spectra in a bipartition space. If we interpret the spectra $S_1, \ldots, S_m$ as a cluster in bipartition space, then the average spectrum can be viewed as the *centroid* of that cluster.

The following constructs a centroid consensus tree of m given spectra, $S_1, \ldots, S_m$.

(1) Compute S_a for $S_1, \ldots, S_m$
(2) Sort the bipartitions in S_a according to their confidence values.
(3) Delete all bipartitions in S_a that conflict with more strongly supported bipartitions in S_a.
(4) Incrementally construct a consensus tree from the remaining bipartitions in S_a, starting with the bipartition with the strongest support to the bipartition with the weakest support.

Note that this is essentially the same algorithm as above with the exception that the spectra, $S_1, \ldots, S_m$ are not bootstrapped samples but arbitrary points in some bipartition space.

2.4. Unsupervised learning in bipartition space

Self-organizing maps [4] were introduced by Kohonen in 1982 and can be viewed as tools to visualize structure in high-dimensional data. Self-organizing maps are considered members of the class of unsupervised machine learning algorithms, since they do not require a predefined concept but will learn the structure of a target domain without supervision.

Typically, a self-organizing map consists of a rectangular grid of processing units. Multidimensional observations are represented as vectors. Each processing unit in the self-organizing map also consists of a vector called a reference vector or reference model. In our case, the multidimensional observations are spectra, where the number of possible bipartitions given n taxa governs the dimensions of the spectra. The dimensions of processing elements of the map match the dimensionality of the observations.

The goal of the map is to assign values to the reference models on the map in such a way that all observations can be represented on the map with the smallest possible error. However, the map is constructed under constraints in the sense that the reference models cannot take on arbitrary values but are subject to a smoothing function called the neighborhood function. During training the values of the reference models on the map become ordered so that similar reference models are close to each other on the map and dissimilar ones are further apart from each other. This implies that similar observations will be mapped to similar regions on the map. Often reference models are referred to as centroids since they typically describe regions of observations with large similarities.

The training of the map is carried out by a sequential process, where $t = 1, 2, \ldots$ is the step index. For each observation $\mathbf{x}(t)$ at time t, we first identify the index c of some reference model which represents the best match in terms of Euclidean distance by the condition

$$c = \arg\min_i \|x(t) - m_i(t)\| \quad \forall i. \tag{5}$$

Here, the index i ranges over all reference models on the map. The quantity $\mathbf{m}_i(t)$ refers to the reference model at po-

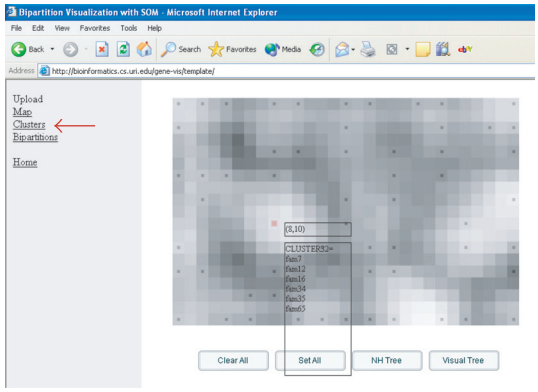

FIGURE 4: A typical visualization computed by GPX.

sition i on the map at time step t. Next, all reference models on the map are updated with the following regression rule where model index c is the reference model index as computed above:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}[\mathbf{x}(t) - \mathbf{m}_i(t)] \quad \forall i. \tag{6}$$

Here, h_{ci} is the neighborhood function that is defined as follows:

$$h_{ci} = \begin{cases} 0 & \text{if } |c-i| > \beta, \\ \eta & \text{if } |c-i| \leq \beta, \end{cases} \tag{7}$$

where $|c-i|$ represents the distance between the best matching reference model at position c and some other reference model at position i on the map, β is the neighborhood distance and η is the learning rate. It is customary to express η and β also as functions of time. This computation is usually repeated over the available observations many times during the training phase of the map. Each iteration is called a training epoch.

An advantage of self-organizing maps is that they have an appealing visual representation. That is, the structure of the input domain is graphically represented as a 2-dimensional map. Figure 4 shows a typical map computed in GPX (here the map reconstructed from bipartition matrix of 14 Archaeal species).

Each square in the map represents a reference model. The shading of the map represents the level of quantization or mapping error for the map. Light shading represents a small quantization error; that is, the reference models in those areas match the observations very closely. Dark shading represents a large quantization error; that is, there is a poor match between reference models and observations. Contiguous areas of low quantization error represent clusters of similar entities. Figure 4 shows an interactive cluster layout of the GPX tool. Each cluster contains a set of orthologous families that we put together by the SOM algorithm. By moving a mouse pointer over the map, a user is able to highlight and select clusters of interest and reconstruct phylogenetic trees for the selection.

Here, we make use of this ability of self-organizing maps to visualize high-dimensional spaces in order to visualize similarities and dissimilarities of high-dimensional tree spectra. We would expect points in bipartition space that represent similar spectra to map close together on the visualization and vice versa. Once we have identified clusters of spectra, we can proceed to compute consensus trees for those clusters. Furthermore, we can now compare the trees calculated from individual clusters to the overall consensus tree, and we can investigate whether there exists substantial conflict between the bipartitions of various clusters. Furthermore, the clusters that result from this unsupervised learning allow the biologist to detect trends in the evolutionary histories of the participating genes which might provide insight into events such as horizontal transfers of individual genes or whole metabolic pathways. The fact that the spectra of individual gene families can be visualized as consensus trees and that it is possible to compute the average of several selected spectra and the corresponding majority consensus tree on the fly distinguishes our approach from other spectral approaches (e.g., [3, 8]).

2.5. The construction of gene families

One of the insights of recent evolutionary biology is that it is not sufficient to use one or a few genes to infer phylogenetic relationships among species. Therefore, we propose to use as many genes as possible in our analysis based on the notion of a *gene family*. A gene family is a collection of genes from different genomes that are related to each other and share a common ancestor. In general, a gene family may include both orthologs and paralogs [9]. Here, we consider only sets of putatively orthologous genes where each species contributes only one gene into a family. The evolutionary history of an individual gene family is a phylogenetic tree.

We select common gene families based on reciprocal best BLAST [10] hit criteria [11] with relaxation (see below). The reciprocal best BLAST hit method requires strong conservative relationships among the orthologs so that if a gene from species 1 selects a gene from species 2 as the best hit when performing a BLAST search with genome 1 against genome 2, then the gene 2 must in turn select gene 1 as the best hit when genome 2 is searched against genome 1. The requirement of reciprocity is very strict and often fails in the presence of paralogs. To select more orthologous sets, we relax the criteria of strict reciprocity by allowing a fixed number of broken connections.

The gene families are aligned with Clustalw version 1.83 using default parameters [12]. For each family, 100 bootstrapped replicates are generated and evaluated with the Phyml program [13] using the JTT model, four relative substitution rate categories, and an estimated shape parameter for the gamma distribution describing among site rate variation.

All 100 generated trees are split into their corresponding bipartition spectra and corresponding bootstrap support values are assigned to each bipartition by calculating how many times each bipartition is present in a family (the bootstrap procedure discussed in detail above). The result of these calculations is a spectrum for each gene family. Observe that trees calculated from individual bootstrap samples contain edges that are not part of a majority consensus tree, that

is, the spectrum for a gene family can contain bipartitions that conflict with other bipartitions in the spectrum. For our purposes, this is important since it prevents information loss and avoids bias during our analyses.

We can now use the machinery developed above to investigate the consensus tree of the collection of gene families and whether there exist spectra that have a significant conflict with the overall consensus tree.

3. APPLICATION OF GPX

GPX, a tool based on the techniques developed above supports an active, investigation-style analysis where the user can interact with the visualization. The user is able to select centroids on the map and investigate consensus trees and conflicting bipartitions in the respective spectra. A detailed description of an experiment using GPX appears in [6]. In a first experiment, we analyzed 123 gene families of 14 archaea species. We found that sets of gene families exhibited substantial conflict with the overall organismal consensus tree corroborating findings of frequent gene transfers between organisms sharing the same or similar ecological niches [14, 15]. In the consensus over all 123 gene families, the representative of the Methanosarcinales (*Methanosarcina acetivorans*) grouped with the Haloarchaea (*Haloarcula marismortui and Halobacterium salinarum*) as expected from the analysis of ribosomal RNAs and enzymes involved in transcription and translation [16, 17]. Two clusters of gene families were recognized that strongly supported a conflicting bipartiton that places the homolog from *Methanosarcina acetivorans* with *Archaeoglobus fulgidus*. For one of these clusters, the relationships among the other archaea remained otherwise compatible with the consensus, suggesting gene transfer events between the ancestors of *Methanosarcina* and *Archaeoglobus*. However, in case of the second cluster formed by a single gene family, prolyl tRNA synthetases (prolylRS), the Haloarchaea grouped at the base of the euryarchaeota. This placement suggests that the ancestor of the Haloarchaea might have acquired this enzyme from outside the archaeal domain, a finding that was corrabborated through more detailed phylogenetic analysis (Gogarten, unpublished). While the haloarchaeal prolylRS are more similar to bacterial than to archaeal homologs, database searches did not identify any sequence from an extant organism that is specifically related to the haloarchaeal prolyl tRNA synthetases. The donor of the haloarchaeal prolylRS is not a member of any of the bacterial or archaeal phyla that have prolylRS sequences in the current nonredundant or environmental databases; possibly the lineage that donated this enzyme has gone extinct as a distinct lineage, and only those genes that were donated to other lineages in the past survived into the presence [18]. These results were obtained by means of an originally developed interactive tool [6], which combines computationally expensive analysis of complex data with convenient visual representation of phylogenetic information.

4. CONCLUSIONS

We developed a comparative genomic analysis technique based on bipartition spectra and unsupervised learning. We have incorporated the techniques developed here into a web-based tool and have used this tool successfully in a set of analyses. The tool allows the user to reconstruct the evolutionary history shared by the plurality of gene family histories present within a collection of genomes; gene families with histories that are in conflict with the plurality are detected, and families which share conflicting histories can be recognized, thereby facilitating the discovery of major "highways of gene sharing" [15].

Bipartition spectrum analysis is not restricted to the SOM algorithm, other clustering algorithms, such as principal component analysis (PCA) [19] and local linear embedding (LLE) [20], can be applied to the analysis of large data sets. A new algorithm, generative topographic mapping (GTM) [21], displays maps similar to SOM but uses an expectation maximization (EM) algorithm instead of relying on neural network convergence. An alternative-to-traditional PCA is kernel PCA [22]. This algorithm is based on support vector machines, which allows it to easily deal with very wide datasets. ISOMAP [23] is an algorithm similar to LLE but distinguishes itself from LLE in that there is no need to solve a set of linear equations. To make comparative genomic studies a reality, we need to be able to include large numbers of genomes. This implies that we need to be able to handle large amounts of data. Future efforts will revolve around scaling up methodologies to include as many species as possible and testing different clustering algorithms for extraction of important phylogenetic information.

ACKNOWLEDGMENTS

This work was in part supported through the NASA Applied Information System Research Program (NNG04GP90G). O. Zhaxybayeva was supported through a Postdoctoral Fellowship from the Canadian Institute of Health Research.

REFERENCES

[1] J. P. Gogarten, W. F. Doolittle, and J. G. Lawrence, "Prokaryotic evolution in light of gene transfer," *Molecular Biology and Evolution*, vol. 19, no. 12, pp. 2226–2238, 2002.

[2] M. D. Hendy and D. Penny, "Spectral analysis of phylogenetic data," *Journal of Classification*, vol. 10, pp. 5–24, 1993.

[3] O. Zhaxybayeva, P. Lapierre, and J. P. Gogarten, "Genome mosaicism and organismal lineages," *Trends in Genetics*, vol. 20, no. 5, pp. 254–260, 2004.

[4] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, Germany, 3rd edition, 2001.

[5] T. Abe, H. Sugawara, M. Kinouchi, S. Kanaya, and T. Ikemura, "Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples," *DNA Research*, vol. 12, no. 5, pp. 281–290, 2005.

[6] N. Nahar, M. S. Poptsova, L. Hamel, and J. P. Gogarten, "GPX: a tool for the exploration and visualization of genome evolution," in *Proceedings of the 7th IEEE International Symposium*

on Bioinformatics & Bioengineering (BIBE '07), pp. 1338–1342, Boston, Mass, USA, October 2007.

[7] J. Felsenstein, "Confidence limits on phylogenies: an approach using the bootstrap," *Evolution*, vol. 39, no. 4, pp. 783–791, 1985.

[8] G. M. Lento, R. E. Hickson, G. K. Chambers, and D. Penny, "Use of spectral analysis to test hypotheses on the origin of pinnipeds," *Molecular Biology and Evolution*, vol. 12, no. 1, pp. 28–52, 1995.

[9] W. M. Fitch, "Homology: a personal view on some of the problems," *Trends in Genetics*, vol. 16, no. 5, pp. 227–231, 2000.

[10] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, pp. 403–410, 1990.

[11] O. Zhaxybayeva and J. P. Gogarten, "Boostrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses," *BMC Genomics*, vol. 3, p. 4, 2002.

[12] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[13] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.

[14] R. Jain, M. C. Rivera, J. E. Moore, and J. A. Lake, "Horizontal gene transfer accelerates genome innovation and evolution," *Molecular Biology and Evolution*, vol. 20, no. 10, pp. 1598–1602, 2003.

[15] R. G. Beiko, T. J. Harlow, and M. A. Ragan, "Highways of gene sharing in prokaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 40, pp. 14332–14337, 2005.

[16] C. Brochier, P. Forterre, and S. Gribaldo, "An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences," *BMC Evolutionary Biology*, vol. 5, p. 36, 2005.

[17] C. R. Woese, "Bacterial evolution," *Microbiology and Molecular Biology Reviews*, vol. 51, pp. 221–271, 1987.

[18] O. Zhaxybayeva and J. P. Gogarten, "Cladogenesis, coalescence and the evolution of the three domains of life," *Trends in Genetics*, vol. 20, no. 4, pp. 182–187, 2004.

[19] I. Jolliffe, *Principal Component Analysis*, Springer, Berlin, Germany, 2002.

[20] L. K. Saul and S. T. Roweis, "Think globally, fit locally: unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, no. 2, pp. 119–155, 2004.

[21] C. M. Bishop, M. Svensen, and C. K. I. Williams, "GTM: the generative topographic mapping," *Neural Computation*, vol. 10, no. 1, pp. 215–234, 1998.

[22] B. Scholkopf, A. Smola, and K. R. Muller, "Kernel principal component analysis," in *Advances in Kernel Methods-Support Vector Learning*, pp. 327–352, MIT press, Cambridge, Mass, USA, 1999.

[23] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.