Unsupervised Learning in Spectral Genome Analysis

Lutz Hamel[1], Neha Nahar[1], Maria S. Poptsova[2], Olga Zhaxybayeva[3], J. Peter Gogarten[2]
[1] Department of Computer Sciences and Statistics, University of Rhode Island, USA
[2] Department of Molecular and Cell Biology, University of Connecticut, USA
[3] Department of Biochemistry and Molecular Biology, Dalhousie University, Canada
hamel@cs.uri.edu, nnahar@cs.uri.edu, maria.poptsova@uconn.edu, olgazh@dal.ca, gogarten@uconn.edu

Abstract

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 that are in significant conflict with the evolutionary consensus tree of an organism point to possible horizontal gene transfer events which often delineate significant evolutionary events.

1. Introduction

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 [1]. This discovery calls into question the notion of a single evolutionary tree describing organismal evolution and gives rise to the notion of an evolutionary consensus tree that is based on the evolutionary patterns of the majority of genes in a genome. This 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 anything 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] allows the inference of consensus trees from smaller quanta of phylogenetic information, side stepping some of the difficult computational issues. We refer to this approach 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.

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 [3] 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 of 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 (e.g. [4]). However, our approach seems to be novel in that we do not apply unsupervised learning directly to DNA data but instead analyze the much more abstract representation of the genomic data in the form of bipartitions. We have constructed a webservice called GPX (Genome Phylogenetic explorer)[1] that supports this kind of analysis [5].

2. Spectral analysis of evolutionary trees

Given n entities, there are 2^n-1 different ways to assign the entities to two different sets. That is, there are 2^n-1 different *bipartitions* of n entities. A (unrooted) tree can be viewed as a model of the evolutionary relationships between n entities or taxa such as species, genes, molecules, *etc*. 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 $2n-3 < 2^n-1$, that is, the number of bipartitions defined by an evolutionary tree of n taxa is much smaller than the number of possible bipartitions of n entities.

Let t_n be an evolutionary tree over n taxa, we define the bipartitions defined by t_n as the *spectrum* of t_n, denoted as $S(t_n)$. It is convenient to adopt a vector notation for the spectrum $S(t_n) = (b_1,\ldots, b_{2^n-1}) = (0,1,1,0,\ldots,0,0)$, where b_k denotes bipartion k with $1 < k < 2^n-1$. Here, $b_k = 1$ if the spectrum of the tree includes bipartition b_k, otherwise $b_k = 0$. 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 tree spectrum as the gene family spectrum and therefore a gene family is denoted by a point in bipartition space.

It is customary to compute confidence values for the edges in an evolutionary tree via bootstrapping [6]. 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, $S(t_n) = (0,67,85,0,\ldots,15,0)$, where t_n is now a bootstrapped consensus tree and the values in the vector represent the confidence values for the individual bipartitions.

Figure 1a shows a bootstrapped consensus tree with five taxa. The values on the edges represent the bootstrapped confidence values. Figures 1b and 1c show two possible bipartitions of the tree. Notice that the bipartitions inherit the confidence value of the edge that corresponds to the bipartition. Also note that the subtrees on either end of the bipartition edge do not preserve the topologies of the original subtrees in the evolutionary tree.

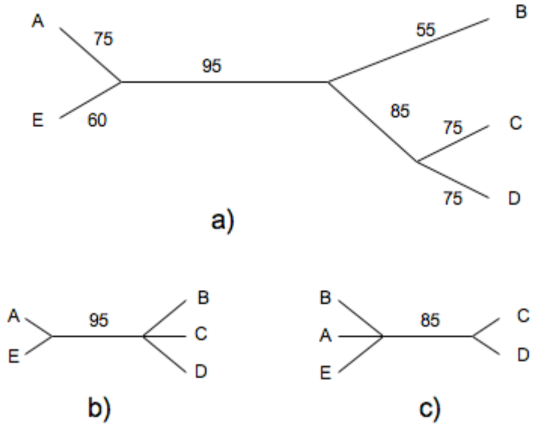


Fig. 1: a) Bootstrapped consensus tree with 5 taxa. b) A bipartition with a 95% bootstrapped confidence value. c) A bipartition with an 85% bootstrapped confidence value.

An interesting consequence of our notion of bipartition space is that we can now measure the difference between spectra as the Euclidean distance between the two corresponding spectrum points in 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.

3. Bipartitions and consensus trees

To handle bipartitions computationally in an efficient way we can represent them effectively as binary masks. Figure 2a shows a binary vector indexed

[1] http://bioinformatics.cs.uri.edu/gpx

by the taxa names of the tree in Figure 1a. Figure 2b shows the binary representation of the bipartition in Figure 1b and Figure 2c shows the binary representation of the bipartition in Figure 1c.

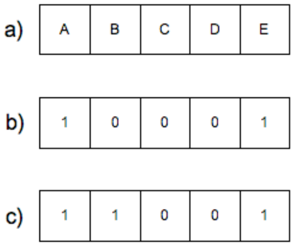
We say that two bipartitions are *compatible* is there exists a tree whose spectrum includes both bipartitions. We say that two bipartitions are *conflicting* if they cannot appear in the same spectrum. Given our binary representation of bipartitions, there is a simple computation to test for compatibility between bipartitions. We say that two bipartitions are compatible if the following returns true:

$$((b_1 \mid b_2) == b_1) \;\|$$
$$((b_1 \mid b_2) == b_2) \;\|$$
$$((b_1 \mid \sim b_2) == b_1) \;\|$$
$$((b_1 \mid \sim b_2) == \sim b_2),$$

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 '==' the bitwise equality operator. Given the two masks from Figures 1b and 1c, it is easy to see that they are compatible:

$$10001 \mid 11001 == 11001$$

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



a)	A	B	C	D	E
b)	1	0	0	0	1
c)	1	1	0	0	1

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

An interesting application of this is the construction of a consensus tree of multiple spectra in a bipartition space. 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,

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

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. 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 algorithm constructs a consensus tree given m 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 bootstrap support 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 our definition of average spectrum implies that it can contain conflicting bipartitions making step 3 necessary in order to construct a tree.

4. Unsupervised learning in bipartition space

Self-organizing maps [3] 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 regression process, where t = 1, 2,... 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,

$$c = \underset{i}{\operatorname{argmin}} \left\| \mathbf{x}(t) - \mathbf{m}_i(t) \right\|, \ \forall i$$

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

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

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

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

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 3 shows a typical map computed in GPX.

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.

In this paper 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.

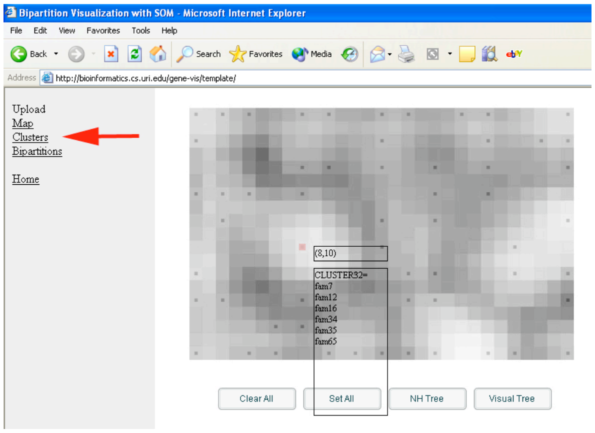


Fig. 3: A typical visualization computed by GPX.

5. The construction of gene families

One of the insights of recent molecular biology is that it is not enough 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 [7]. 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 [8] hit criteria [9] 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 a 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 [10]. For each family a maximum likelihood tree is calculated by Phyml [11] using the JTT model, four relative substitution rate categories, and an estimated shape parameter for the gamma distribution describing among site rate variation.

For each gene family tree, 100 bootstrapped replicates are generated and evaluated with the Phyml program. 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 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 that 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.

6. GPX

We have developed a tool based on the techniques developed above. Furthermore, the tool 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 [5]. In this 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 [12, 13].

7. Conclusions

Here we described 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. A big drawback of the techniques given here is the reciprocity requirement in the gene families severely limiting the number of gene families we can use for our analyses. A new approach based on embedded quartet spectra [14] promises to lift this restriction.

10. References

[1] J. P. Gogarten, W. F. Doolittle and J. G. Lawrence, "Prokaryotic Evolution in Light of Gene Transfer," *Mol. Biol. Evol.,* vol. 19, 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] T. Kohonen, *Self-Organizing Maps.* ,3rd ed., vol. 30, Berlin ; New York: Springer, 2001, pp. 501.

[4] 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 Res.,* vol. 12, pp. 281-290, 2005.

[5] 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 IEEE 7th International Symposium on Bioinformatics & Bioengineering (BIBE07),* to appear,

[6] J. Felsenstein, "Confidence Limits on Phylogenies: An Approach Using the Bootstrap," *Evolution,* vol. 39, pp. 783-791, 1985.

[7] W. M. Fitch, "Homology: a personal view on some of the problems," *TRENDS IN GENETICS,* vol. 16, pp. 227-231, 2000.

[8] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.,* vol. 215, pp. 403-410, 1990.

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

[10] 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 Res.,* vol. 22, pp. 4673-4680, 1994.

[11] S. Guindon and O. Gascuel, "A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood," *Syst. Biol.,* vol. 52, pp. 696-704, 2003.

[12] R. Jain, M. C. Rivera, J. E. Moore and J. A. Lake, "Horizontal gene transfer accelerates genome innovation and evolution," *Mol. Biol. Evol.,* vol. 20, pp. 1598-1602, Oct. 2003.

[13] R. G. Beiko, T. J. Harlow and M. A. Ragan, "Highways of gene sharing in prokaryotes," *Proceedings of the National Academy of Sciences,* vol. 102, pp. 14332-14337, 2005.

[14] O. Zhaxybayeva, J. P. Gogarten, R. L. Charlebois, W. F. Doolittle and R. T. Papke, "Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events," *Genome Res.,* vol. 16, pp. 1099-1108, 2006.