GPX: A Tool for the Exploration and Visualization of Genome Evolution

(This is a longer version of the paper published in the proceedings)

Neha Nahar
Lutz Hamel
Department of Computer Science and Statistics, University of Rhode Island,
9 Greenhouse Road,
Kingston, RI 02881.
nnahar@cs.uri.edu,
hamel@cs.uri.edu

Maria S. Popstova
J. Peter Gogarten
Department of Molecular and Cell Biology,
University of Connecticut,
91 North Eagleville Road,
Storrs, CT 06269-3125.
maria.poptsova@uconn.edu,
gogarten@uconn.edu

Abstract

Early life on Earth has left many traces that can be utilized to reconstruct the history of life. This information is present in the form of fossils, geological records and also in information retained in living organisms. Gene sequences are now recognized as an invaluable document of life's history on Earth. Ever since Darwin the Tree of Life has provided a framework to study the evolution of organisms. However, comparative genome analyses have shown that genomes are mosaics where different parts have different histories. One of the reasons for this is the exchange of genes between species. Due to this horizontal gene transfer the Tree of Life concept is transforming to a Web of Life where different parts of a genome possess different evolutionary histories compared to the accepted evolutionary history of the corresponding species. Clustering gene families based on the phylogenetic information they retain allows extracting a majority consensus for the genomes' evolution, and the determination of genes that have a conflicting phylogeny. The latter is of interest in the context of comparative genomics of prokaryotes because these conflicts point towards possible horizontal transfers of genes and metabolic pathways between divergent organisms.

We have created a web-based tool Gene Phylogeny eXplorer (GPX) that facilitates comparative genome analysis of different species. GPX displays results as an interactive map that allows users to explore and interpret genomic data representing gene evolution. It allows the visualization of consensus and conflicting evolutionary histories of genes. The novel aspect of our approach is that we do not try to analyze DNA sequences directly but instead use self-organizing maps to find structure (clusters) in a space spanned by all possible evolutionary relationships between the genomes in questions. Since the number of possible evolutionary trees grows factorially with the number of genomes we use smaller quanta of phylogenetic information, in particular we use bipartitions, to represent the evolutionary relationships between genomes. The number of possible bipartitions grows exponentially with the number of genomes and therefore grows much slower than the number of evolutionary trees making it amenable for a computational approach. The structure of the resulting clusters and in particular the patterns of bipartition support within these clusters provide important information on the origin of individual genes. If a strongly supported bipartition for a gene conflicts with the consensus tree then it is most probably due to a horizontal gene transfer event.

1. Introduction

The *Tree Of Life* has provided a framework to study the evolution of organisms [1]. Phylogenetic trees are used to depict the evolution of organisms or of molecules. However, comparative genome analyses have shown that genomes are mosaics where different parts have different histories [2-5]. These findings questioned the validity of the tree concept, especially for prokaryotic species [6, 7]. Individual genes may travel from one species to another, a core of infrequently transferred genes might represent a tree-like organismal history, genomes that had independent evolutionary histories might have fused to form a new line of descent, and highways of gene sharing [8] might overwhelm the signal retained in non transferred genes [9]. The *Tree of Life* concept needs to be amended by fusing lines of descent and by connecting threads, representing gene transfer events that embed the organismal lines of descent into a *Web of Life* [10]. Without selection of gene families that were refractory towards transfer, phylogenetic trees calculated using super-tree [11, 12] or super-matrix [13] approaches might neither reflect the history of the organism nor the history of the genes [9]. Thus the task of a comparative genomics is to identify the genes that share a common history, genes whose evolution is different from the majority consensus, and to identify groups of genes that might have been transferred together. The last are of special interest because they point towards crucial events in evolutionary history.

Our web-based tool Genome Phylogeny eXplorer (GPX) facilitates comparative genome analysis by allowing visual and interactive exploration and inspection of either individual gene histories or groups of closely related families identified as those through clustering based on a self-organizing map [14] approach. Our tool also allows for locating gene families whose histories are in significant disagreement with a majority consensus, the type of phylogenetic conflict that is in most cases attributed to horizontal gene transfer.

Table 1. Number of trees and bipartitions required to represent phylogenetic data.

Number of genomes	Number of trees	Number of bipartitions
4	3	3
6	105	25
8	10,395	119
10	2,075,025	501
13	1.37E + 10	4,082
20	2.22E + 20	5.24E + 05
50	2.84E + 74	5.63E + 14

Evolutionary relationships between organisms are usually represented as a phylogenetic tree. For n different taxa the number of possible trees grows very fast. With n taxa there are $(2n-5)!\ /[2(n-3)(n-3)!]$ different unrooted tree topologies (see Table 1), and it is an impossible computational task to iterate through all possible trees (1.37E+10) even for only 13 taxa. Alternatively, a phylogenetic tree can be divided into quanta of phylogenetic information such as a bipartition. A bipartition as shown in Figure 1 is the division of a tree into two parts that are connected by a single branch. The number of possible 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 (see Table 1). Other advantages of bipartition analysis are that different bipartitions can either be compatible (they can reside in one tree) or conflicting (they cannot co-exist in one tree) [15, 16], and that the statistical support for bipartitions can be assessed readily through bootstrap analyses [17]. By identifying compatible bipartitions that are supported by the majority of gene families one can find a plurality consensus phylogenetic signal. Bipartitions that are in significant conflict with the plurality consensus bipartitions are most likely related to a horizontal gene transfer or to systematic artifacts of phylogenetic reconstruction. Although self-organizing maps have been used in comparative genomics before (e.g. [18]), the novel aspect of our approach is that we do not try to analyze the evolutionary relationships between DNA sequences directly but instead use self-organizing maps to find structure (clusters) in a space spanned by all possible evolutionary relationships between the genomes in question represented as bipartitions. Only this more abstract representation makes this problem computationally feasible. Also novel is our use of *emergent* self-organizing maps, which allows for more precise elucidation of inter- and intra- cluster relationships [19].

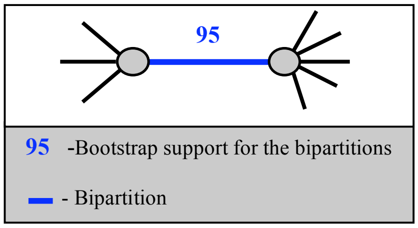

Figure 1. Bipartition of a phylogenetic tree.

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 (genes that arose due to a speciation event) and paralogs (genes that arose due to duplication) [20]. 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 that can be represented as a set of bipartitions. GPX uses a bipartition matrix (see Results section for the explanation of how the bipartition matrix is calculated) as input to represent the phylogenetic information contained in gene families. The format of a bipartition matrix is shown in Table 2. It is a matrix where rows represent gene families and columns give the bootstrap support values for the particular bipartitions calculated for each gene family.

In bipartition analysis, each gene family can be treated as a point in a high dimensional space of all possible bipartitions. The coordinates of such a point are the bootstrap support values of the individual bipartitions. Clustering using self-organizing maps is able to project the high dimensional bipartition space of gene family vectors onto a two-dimensional plane and gene families with similar phylogenetic signals are grouped together. Conflicting and non-conflicting families are discovered through a visual and interactive exploration of the map.

As tested in [15] bipartition analysis under some conditions outperforms the AU test [21] for detecting conflicting signals in phylogenetic data. Our visually oriented tool has an advantage over Lento plot [16, 22] analysis by allowing interactive and visual inspection of the areas on the map that comprise families that generated the conflicts. The consensus phylogeny for these families can be investigated by one click. The group of conflicting families can also be visually divided into clusters according to self-organizing maps. This type of analysis is not provided by Lento plot based approaches and provides interesting information regarding groups of horizontally transferred genes with similar phylogenetic information content.

Section 2 provides an overview of GPX. In Section 3 we work through a simple example that demonstrates the capabilities of GPX. We follow up with conclusions in Section 4 and finally close the paper with some remarks on future work in Section 5.

2. Methods

GPX is an online application that performs bipartition analysis using emergent self-organizing maps. The input to GPX is a bipartition matrix composed from common gene families of a given set of species. Using this input, GPX generates clusters of gene families with similar phylogenetic signals. It allows an interactive reconstruction of phylogenetic trees for any combination of the resulting clusters of gene families. Finally, the tool also reports strongly supported and conflicting bipartitions.

2.1. Emergent Self-Organizing Maps (ESOM)

ESOM models [23] are not substantially different from the classical notion of self-organizing maps (SOMs) introduced by Kohonen [14] with the exception that we consider much larger maps typically with thousands of neural elements. These substantially larger maps allow for emergent phenomena not possible to observe in the standard statistical application of SOMs [24]. In order to accommodate training for these larger maps Ultsch et al. have introduced new block-based training algorithms [25]. ESOM models are two-dimensional projections of high-dimensional data that preserve the topology of clusters as much as possible. The canonical way to interpret ESOM models is through the unified distance matrix (u-matrix) [26]. In these visualization areas of small quantization error, that is, areas that form tight clusters show up as light areas. Conversely, areas of large quantization error, that is, areas that do not form clusters or only very sparse clusters, show up as dark areas. Often these dark areas form separation boundaries between tight clusters.

2.2. GPX Framework.

GPX is composed of different components (as shown in Figure 2) such as web server, interface, analysis programs, and data storage space. The web server component provides the web interface. The interface component provides the user management capabilities. Analysis programs are scripts that perform the analyses and generate the visual output. These results are stored in the data storage for the user to access for a given period of time. Once the server performs the analyses and generates the required results, a hyperlink to a results page is displayed to the user. The results are accessible to the user for a given period of time before they expire.

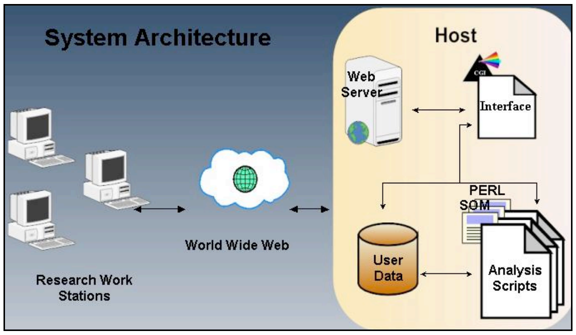

Figure 2. GPX Framework.

2.3. GPX Interface.

GPX allows for the analysis of different genomes with different bootstrap cutoff values. A bootstrap cutoff value acts as a filter and only allows those bipartitions to participate in an analysis whose support is greater or equal to the bootstrap cutoff value. Once the visualization map has been built and other static analyses are completed using user-set parameters, the user is issued a link to a dynamic analysis interface. This interface is organized around a hyperlinked menu.

By selecting the “Map’’ link the user can view the gene family map that displays the gene family clusters.

The next menu point “Clusters” provides an interactive way for users to select specific clusters on the map and generate majority consensus trees using those clusters. This consensus tree is generated on the fly from the combined set of supported bipartitions of the families in the selected clusters. The tree is calculated according to the majority consensus rule algorithm [27] as implemented in CONSENSE [20]. Only the first (n-3) highly supported non-conflicting bipartitions supported by the selected gene families are taken into consideration. Bootstrap support values reported for individual branches in the consensus tree reflect the average support for this branch in the individual gene families constituting the selected cluster (see 3.2.2. for a more detailed description). The consensus tree is viewed and can be manipulated using the ATV tree viewer applet [28]. The user also can view the map co-ordinates of individual clusters and the families in those clusters by hovering the mouse over the appropriate cluster visualization.

The final menu point “Bipartitions” displays a list of bipartitions that are supported by at least some gene families at or above the bootstrap cutoff value followed by a list of conflicting bipartitions. One way to view these plots is as slices from the ESOM generated map where each slice corresponds to a particular bipartition. Each such slice shows exactly where on the map a particular bipartition is or is not supported and by which gene families. For easier comprehensibility these slices are also represented as 3D plots. These representations can be further investigated to identify horizontal gene transfer events (see Results for the screenshots of the tool for a special case of the analysis of 14 archaeal genomes).

3. Results

3.1. Assembling gene families and building the bipartitions matrix.

14 complete genomes of archaea containing all genes as amino acid sequences were downloaded from the NCBI ftp-site on July 2005. The list of 14 archaeal taxa is shown in Table 2.

Table 2: List of 14 archaeal taxa

Aeropyrum	Ethanothermobacter
Archaeoglobus	Nanoarcheum
Haloarcula	Pyrobaculum
Halobacterium	Pyrococcus
Methanococcus	Sulfolobus
Methanopyrus	Thermococcus
Methanosarcina	Thermoplasma

Common gene families were selected based on reciprocal best BLAST [29] hit criteria [30, 31] with relaxation. 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 broken connections. Applying this approach to 14 archaea, 123 gene families were selected with up to 3 broken connections (109 families were selected by strict reciprocal BLAST hits method and 14 families with up to 3 broken connections). Gene families were aligned with Clustalw version 1.83 using default parameters [32]. For each family a maximum likelihood tree was calculated by Phyml [33] 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 were generated and evaluated with the phyml program. All 100 generated trees were split into a set of bipartitions and corresponding bootstrap support values were assigned to each bipartition by calculating how many times each bipartition is present in a family. The bipartition matrix was composed from bootstrap values for bipartitions for each gene families with rows corresponding to gene families and columns to all possible bipartitions for a given set of taxa. For 14 species there are total 8177 different bipartitions. In theory, the bipartition matrix should contain 123 x 8191 elements. In practice, we did not include those bipartitions that are not supported by a single family, so columns with all zeros were removed. As a result we had matrix with 123 x 1646 elements, filled with bootstrap support values for remaining bipartitions for each family.

3.2. Static analysis

The SOM algorithm [19] is applied to bipartition matrix using the following SOM parameters: x-dimension = 15, y-dimension = 10, rectangular topology, with radius equal to the x-dimension and 100,000 training iterations [34]. The experiment we report on uses a bootstrap cutoff value of 70% (Figure 4). In Figure 3 we show the maps produced by bootstrap cutoff values of 0% and 90%.

Clusters become more pronounced with the increase of the cutoff value because smaller numbers of bipartitions remain for the analysis after the cutoff filter is applied, and families are grouped together only according to highly supported splits. For 90% cutoff (only bootstrap support values higher than 90% are considered) there exits one big cluster that includes majority of the families with a similar phylogenetic signal. For our tested case of 14 archaea these splits are

Haloarcula-Halobacterium and *Thermococcus-Pyrococcus*. Families that fall into one big cluster in the top right corner of the map for 90% cutoff value are the families that share high support for these two bipartitions.

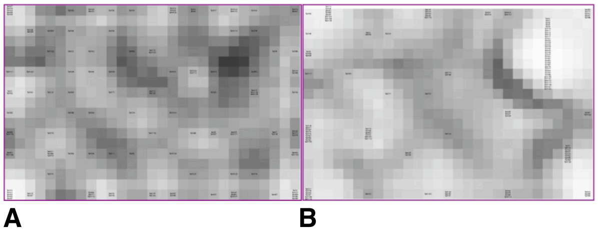

Figure 3. SOM maps from bipartition matrix generated from 14 archaea species: A – 0 cutoff, B – 90% cutoff.

A bootstrap support of 70% is considered to be still reliable in analyzing phylogenies and it allows more bipartitions to be included in the analysis. Here we demonstrate how GPX works on the example of 70% cutoff value.

3.3. Dynamic analysis

This section describes the interaction of a user with GPX. We provide the screenshots of the interactive analysis for a test case of bipartition matrix for 14 archaea.

3.3.1. Map. Here we show the analysis for 14 archaea with 70% cutoff value after the static analysis completes. As a first step we show the interface to the generated map (Figure 4).

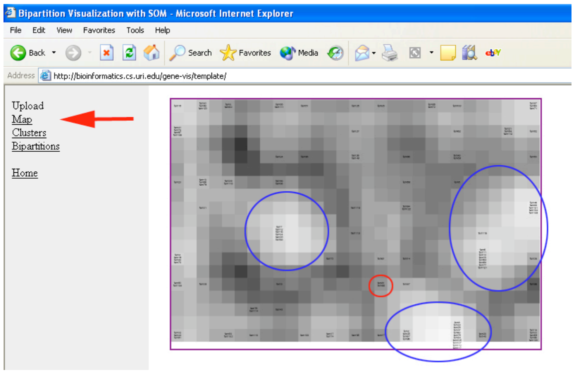
Bipartition Visualization with SOM - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address http://bioinformatics.cs.uri.edu/gene-vis/template/
Upload
Map
Clusters
Bipartitions
Home

Figure 4. SOM generated map from bipartition matrix of 123 families with a cutoff of 70%.

The map displays light colored areas (shown with blue circles), grey areas, and dark areas. Light colored areas show clusters that have small quantization errors and are dense. Grey areas display clusters that are sparse, and dark areas represent regions with very little similarity information. In the above map white areas indicate groups of gene families whose phylogenetic signal is distinctly different from the surrounding genes. Genes that are in conflict (shown in red) with plurality bipartitions are grouped separately from the clusters containing genes conforming to the plurality phylogeny. A clearer picture of this emerges when one studies the support of individual bipartitions

3.3.2. Clusters. The Clusters link will direct a user to an interactive map with clusters of gene families (see Figure 5).

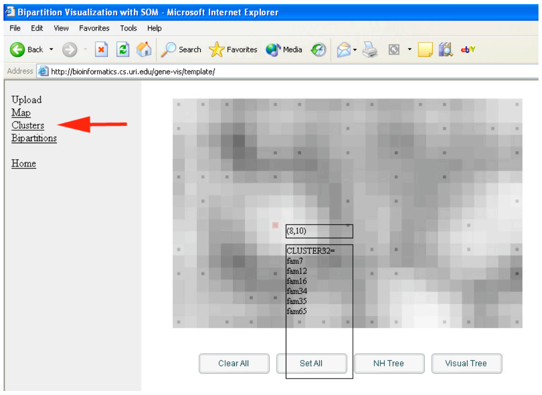
Bipartition Visualization with SOM - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Media
Address http://bioinformatics.cs.uri.edu/gene-vis/template/
Upload
Map
Clusters
Bipartitions
Home
(8,10)
CLUSTER32=
fam7
fam12
fam16
fam34
fam35
fam65
Clear All
Set All
NH Tree
Visual Tree

Figure 5. Clusters of gene families created by SOM.

When the user moves the mouse over a neuron that contains gene families, a pop-up window displays the coordinates of the neuron on the SOM map and the gene families it contains. By clicking on the map, the user can select any set of neurons that have gene families mapped to them and then visualize a consensus phylogenetic tree. Selected neurons are highlighted as red squares on the map. Figure 6 depicts consensus tree reconstructed from all clusters (red squares on the SOM map). The ATV tree viewer applet [28] is used to visualize the tree. ATV has many options that allow the user to modify the view of the tree. For example, one can re-root a tree with particular outgroup.

3.3.3. Bipartitions. The Bipartitions link directs the user to the list of bipartitions with their visual representations. This page begins with a list of bipartitions that are supported by at least some gene families at or above the bootstrap cutoff value (see Figure 7). Support values for a given bipartition are given in brackets (see blue arrow on Figure 7).

A three-dimensional bipartition support representation is used to depict areas on the map that highly support a given bipartition. The same information is given by a 2-D representation with the gene family cluster information added in where white areas correspond to the regions that highly support this bipartition while black represent regions of conflicts (see black-and-white map on Figure 7).

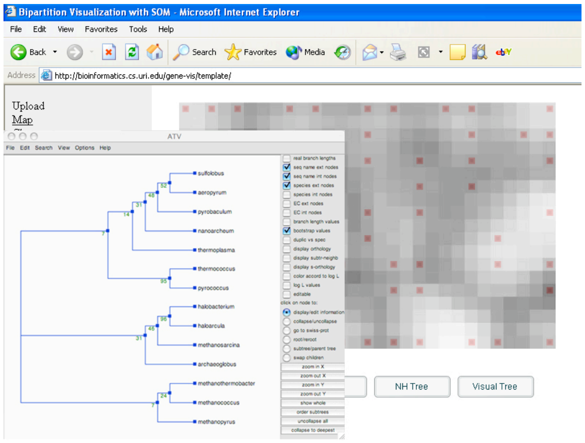

Figure 6. Consensus tree generated from all clusters. Consensus tree is displayed in ATV tree viewer applet.

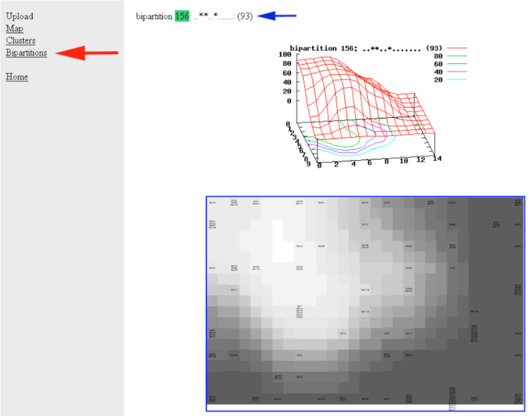

Figure 7. Bipartition 3D function over SOM map together with a slice of the SOM map that corresponds to the bipartition.

Here we give an example of how one can explore in detail the phylogeny of the families that support a given bipartition. The bipartition 156, whose support is depicted in Figure 7, groups *Halobacterium, Haloarcula* and *Methanosarcina* together. This bipartition is in agreement with the small subunit ribosomal RNA phylogeny [35] and the consensus calculated from the transcription and translation machinery [36]. Using 2D black-and-white map for the bipartition 156 (see Figure 7) as a guide, one can select all clusters from the white area on the SOM map (selected neurons inside blue circle on Figure 8). A phylogenetic tree, reconstructed from the clusters in the white area, would represent a history of the families that are in agreement with the selected bipartition, here with the bipartition 156 (see Figure 8).

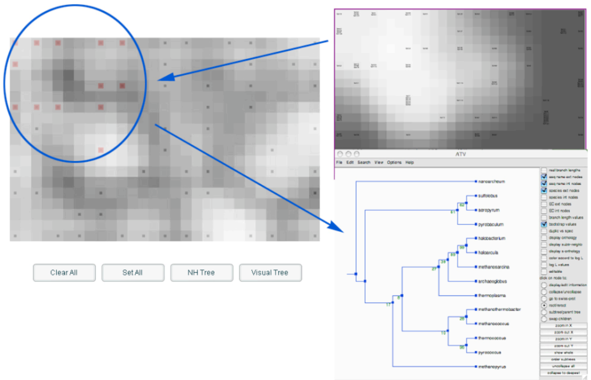

Figure 8. Tree reconstructed from the selected clusters (red dots on the left map) that fell into white areas on the bipartition superposition map (on the right).

To find families whose phylogenetic histories are in conflict with bipartition 156, one can scroll down in the list of bipartitions and find bipartitions that show conflicts with bipartition 156. The same phylogenetic analysis described above for the bipartition 156, can be also done for the conflicting bipartitions.

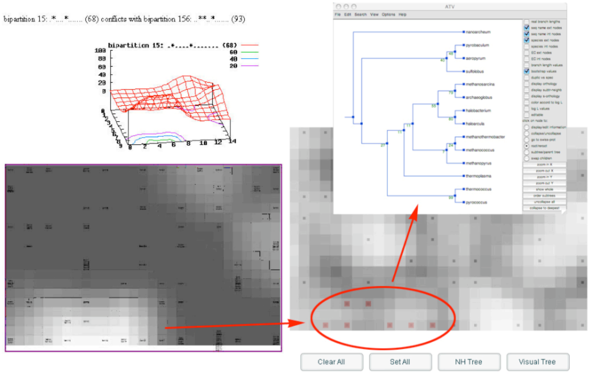
bipartition 15: .*....*....... (68) conflicts with bipartition 156: ..**..*....... (93)
bipartition 15: .*....*....... (68)
100
80
60
40
20
0
60
40
20
0
1
2
3
4
5
6
7
8
9
0
2
4
6
8
10
12
14
ATV
File Edit Search View Options Help
Clear All
Set All
NH Tree
Visual Tree

Figure 9: Analysis of the conflicting bipartition (see text for explanation).

Figure 9 shows the analysis of conflicting bipartition 15, which corresponds to a split where Archaeoglobus groups together with Methonosarcina. This is a bipartition that is in conflict with the consensus phylogeny of conserved genes [36]. White areas on a bipartition support map show clusters that support this conflicting bipartition (selected neurons inside red circle on Figure 9). The phylogenetic tree, reconstructed from the selected families, confirms that Archaeoglobus - Methanosarcina branch has a high bootstrap support value in the gene families of these clusters. This finding suggests a highway of gene sharing [8] between the Methanosarcina and Archaeoglobus lineages. This kind of exploratory phylogenetic analysis is unique to our tool and is not provided by any other system we are aware of.

4. Conclusion

GPX provides an exploratory interface for biologists to perform analysis and knowledge discovery on genomic data. The user can interactively conduct a phylogenetic analysis of any gene family, or clusters of gene families with the ease of a single click. The web-service allows the researcher to view the genomic information contained within a bipartition matrix from a number of difference perspectives:

(a) Gene family clusters based on similarity of bipartition support.
(b) Gene family consensus tree.
(c) Bipartition support together with conflicting bipartitions.

An important advantage of GPX is an interactive and visual identification of horizontally transferred genes. This function is implemented in a detection of conflicting signals in bipartition matrix. This kind of functionality is not available in other techniques or tools. Our abstract representation of evolutionary relationships makes this approach computationally possible and enables the interactive, exploratory nature of this tool.

5. Future work

Bipartition analysis requires gene families that contain representatives from all species of interest, thus only a relatively small number of families can be included in an analysis. A quartet-based approach allows including incomplete gene sets where only four or more species are present [37, 38]. We developed a phylogenetic algorithm called BranchClust [39] that performs reliable and effective selection of both complete and incomplete gene families. As a result the total number of gene families selected for the analysis is considerably increased. For example for our test case of 14 archaea this number is increased from 123 complete families to 1800 both complete and incomplete gene families with a minimum of four species being included. The idea of a quartet analysis is essentially the same as for bipartitions [16], the only difference that only four species are considered at a time. To avoid taxon sampling problems support values for the quartets can be calculated using embedded quartets [40]. The tree is decomposed into a set of quartets and phylogenetic conflicts are searched between individual quartets. The quartet matrix will be similar to bipartition matrix with rows representing families and columns corresponding to the different possible quartets. Each quartet can have three possible topologies. The number of possible quartets is given by the formula: n!/(n-4)!4! and is equal to 1001 for 14 species. The dimension of quartet matrix will be 1800x3003 given the three different topologies for each quartet, and its analysis will be greatly simplified by applying our visualizing technique using SOM approach. To avoid the border effects in the current map generated by SOM, we plan to use boundless maps (such as toroid maps) [41].

6. Acknowledgement

This work was supported through the NASA Applied Information System Research Program (NNG04GP90G).

7. References

[1] C. R. Darwin, *The Origin of Species by Means of Natural Selection. Or the Preservation of Favoured Races in the Struggle for Life.* Adamant Media Corporation,

[2] E. Hilario and J. P. Gogarten, "Horizontal transfer of ATPase genes--the tree of life becomes a net of life," *BioSystems,* vol. 31, pp. 111-119, 1993.

[3] E. V. Koonin, A. R. Mushegian, M. Y. Galperin and D. R. Walker, "Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea," *Mol. Microbiol.,* vol. 25, pp. 619-637, 1997.

[4] W. F. Doolittle, "Phylogenetic classification and the universal tree," *Science,* vol. 284, pp. 2124-2129, Jun 25. 1999.

[5] 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.

[6] W. Doolittle, "Phylogenetic classification and the universal tree." *Science,* vol. 284, pp. 2124-2129, 1999.

[7] 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.

[8] 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.

[9] J. P. Gogarten and J. P. Townsend, "Horizontal gene transfer, genome innovation and evolution," *Nat. Rev. Microbiol.,* vol. 3, pp. 679-687, Sep. 2005.

[10] W. Martin, "Mosaic bacterial chromosomes: a challenge en route to a tree of genomes," *Bioessays,* vol. 21, pp. 99-104, 1999.

[11] O. R. P. Bininda-Emonds, "The evolution of supertrees," *Trends in Ecology and Evolution,* vol. 19, pp. 315-322, 2004.

[12] C. J. Creevey, "Clann: investigating phylogenetic information through supertree analyses," *Bioinformatics,* vol. 21, pp. 390-392, 2005.

[13] F. Delsuc, H. Brinkmann and H. Philippe, "Phylogenomics and the reconstruction of the tree of life," *Nature Reviews Genetics,* vol. 6, pp. 361-375, 2005.

[14] T. Kohonen, *Self-Organizing Maps.* Springer, 2001,

[15] M. S. Poptsova and J. P. Gogarten, "The power of phylogenetic approaches to detect horizontally transferred genes," *BMC Evol. Biol.,* vol. 7, pp. 45, Mar 21. 2007.

[16] G. M. Lento, R. E. Hickson, G. K. Chambers and D. Penny, "Use of spectral analysis to test hypotheses on the origin of pinnipeds," *Mol. Biol. Evol.,* vol. 12, pp. 28-52, Jan. 1995.

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

[18] 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.

[19] A. Ultsch and F. Morchen, "ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM," *Data Bionics Research Group, University of Marburg,* vol. 17, 2005.

[20] W. M. Fitch, "Homology," *TRENDS IN GENETICS,* vol. 16, pp. 227-231, 2000.

[21] H. Shimodaira, "An Approximately Unbiased Test of Phylogenetic Tree Selection," *Syst. Biol.,* vol. 51, pp. 492-508, 2002.

[22] O. Zhaxybayeva, P. Lapierre and J. P. Gogarten, "Genome mosaicism and organismal lineages," *Trends Genet.,* vol. 20, pp. 254-260, May. 2004.

[23] A. Ultsch, "Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series," *Kohonen Maps,* pp. 33–46, 1999.

[24] W. Venables and B. D. Ripley, *Modern Applied Statistics with S.* Springer, 2002,

[25] M. Noecker, F. Moerchen and A. Ultsch, "Fast and reliable esom learning." *Proceedings 14th European Symposium on Artificial Neural Networks ,* 2006.

[26] A. Ultsch, "Self-organizing neural networks for visualization and classification." *Information and Classification,* pp. 307–313, 1993.

[27] T. Margush and F. R. McMorris, "Consensusn-trees," *Bull. Math. Biol.,* vol. 43, pp. 239-244, 1981.

[28] C. M. Zmasek and S. R. Eddy, "ATV: display and manipulation of annotated phylogenetic trees," *Bioinformatics,* vol. 17, pp. 383-384, 2001.

[29] 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.

[30] O. Zhaxybayeva and J. P. Gogarten, "Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses," *Feedback,* 2004.

[31] M. G. Montague and C. A. Hutchison 3rd, "Gene content phylogeny of herpesviruses," *Proc. Natl. Acad. Sci. U. S. A.,* vol. 97, pp. 5334-5339, May 9. 2000.

[32] 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.

[33] 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.

[34] T. Kohonen, J. Hynninen, J. Kangas and J. Laaksonen, "SOM_pak: the selforganizing map program package, release 3.1," *Laboratory of Computer and Information Science, Helsinki University of Technology, Finland,* 1998.

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

[36] C. Brochier, P. Forterre, S. Gribaldo, Y. Zivanovic and F. Confalonieri, "An emerging phylogenetic core of Archaea: phylogenies of transcription and translation machineries converge following addition of new genome sequences," *Feedback,* 2006.

[37] O. Zhaxybayeva and J. P. Gogarten, "An improved probability mapping approach to assess genome mosaicism," *Feedback,* 2004.

[38] 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, 2006.

[39] M. S. Poptsova and J. P. Gogarten, "BranchClust: a phylogenetic algorithm for selecting gene families," *BMC Bioinformatics,* vol. 8, pp. 120, Apr 10. 2007.

[40] O. Zhaxybayeva and J. P. Gogarten, "An improved probability mapping approach to assess genome mosaicism," *Feedback,* 2004.

[41] A. Ultsch, "Maps for the Visualization of high-dimensional Data Spaces," *Proc.Workshop on Self Organizing Maps,* pp. 225-230, 2003.