2nd Progress Report for NASA Applied Information System Research Program (AISRP)
"Exploration of Novel Methods to Visualize Genome Evolution"
period: 9/1/05-8/31/06
Grant No: NNG04GP90G
Proposal Number: AISP03-0037-0008

PI: J. Peter Gogarten, University of Connecticut
Co PI: Lutz Hamel, University of Rhode Island

Current personnel:
Maria Poptsova, Ph.D., University of Connecticut
Neha Nahar, University of Rhode Island

Former personnel supported in part under this proposal:
Olga Zhaxybayeva, Ph.D., University of Connecticut, currently at Dalhousie University, Halifax, NS, Canada
Jinling Huang, Ph.D., University of Connecticut and NASA Astrobiology Institute at the Marine Biological Laboratories, Woods Hole, MA, currently at East Carolina University

Publications and presentations that resulted form the sponsored research:

Publications:

Huang, Jinling and J. Peter Gogarten (2006):
"Ancient horizontal gene transfer can benefit phylogenetic reconstruction",
Trends in Genetics, in press
This manuscripts reports on the usefulness of transferred genes in reconstructing organismal evolutionary history. Traditionally gene transfer is considered a complication in reconstructing a genomes history: if different parts of a genome have different histories, lumping the different genes together in a single analysis might result in phylogenies that represent neither the individual genes nor the genomes [1]. However, transferred genes that are retained in the recipient lineage provide a shared derived character for the recipient lineage that characterizes this lineage as a natural group.

Zhaxybayeva, Olga, and J. Peter Gogarten,
Robert L. Charlebois, W. Ford Doolittle and R. Thane Papke:
"Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events."
Genome Research, accepted for publication with minor revisions.
Introduces quartet decomposition analysis as tool to detect horizontally transferred genes. The work builds on the analysis of embedded quartets [2], an approach that solves the taxon-sampling problem usually encountered in quartet-based analyses. The reliability of the quartet-based approach is assessed using simulated genome evolution. The quartet decomposition approach is applied to the evolution of cyanobacteria, we also discuss implications of the microbial species concept.

Zhaxybayeva, O. and Gogarten, JP. (in revision):
"Horizontal gene transfer, gene histories and the root of the tree of life"
Proceedings of the McMaster Astrobiology conference", Paul Higgs (ed.),
Cambridge University Press.

We discuss horizontal gene transfer, coalescence of gene lineages, and the placement of the root of the tree of life based on ancient gene duplications. In discussing genome evolution, we extend population genetic concepts to early microbial evolution. While the most recent common ancestors (MRCA) of different genes existed in different organisms and at different times, most ancient duplicated genes place the MRCA between the archaeal and bacterial domains.

Maria Poptsova and J. Peter Gogarten:
"The power of phylogenetic approaches to detect horizontally transferred genes"
BMC Evolutionary Biology, submitted.
We use in silico gene transfer to test different approaches to detect horizontally transferred genes. Using the gamma proteobacteria as a test case we show that most methods, while accurately estimating the rate of false positives, have an unacceptably high rate of false negatives, i.e., most methods greatly underestimate the amount of gene transfer events. We also show that for some datasets bipartition spectra based approaches [3] are more efficient (fewer false negatives) than the popular AU test [4].

Gogarten, J. P. and Townsend J. P. (2005):
Horizontal gene transfer, genome innovation, and evolution
Nature Reviews in Microbiology 3(9) 679-687 (online version, pdf)
Also part of the brochure "*Focus on Horizontal Gene Transfer*", published with support from DOE, NSF, and NASA.
http://www.nature.com/nrmicro/focus/genetransfer/index.html
We review the role of gene transfer in microbial evolution, assess examples where frequent gene transfer has generated a signal that gave rise to the wrong organismal phylogenetic reconstruction. We introduce a nearly neutral theory for horizontal gene transfer, and promote the application of population genetic concepts to the evolution of higher taxonomic categories.

From year 1:

Hamel L., Zhaxybayeva O., Gogarten J.P. (2005):
PentaPlot: A software tool for the illustration of genome mosaicism.
BMC Bioinformatics 2005, 6:139
This manuscript introduces simple maps of gene families into tree space (an extension of ml mapping to 5 genomes) to identify genes that have different evolutionary histories.

Invited lectures (year two only)**:**

J. Peter Gogarten:
"Is there a "Tree of Life"?"
Invited lecture at the ISSI ASTROBIOLOGY WORKSHOP "STRATEGIES OF LIFE DETECTION", Bern, Switzerland, April 24-28, 2006

J. Peter Gogarten:
"Oxygen Producing Photosynthesis and the Molecular Record"
Invited lecture at the Agouron Instute's Oxygen meeting, Santa Fe, April 6-10, 2006

J. Peter Gogarten:
"Horizontal gene transfer and microbial evolution: Is the Tree-of-Life a Tree?"
Invited seminar at the Astrobiology Forum, Harvard University, December 13th, 2005

J. Peter Gogarten:
"Horizontal gene transfer and microbial evolution: Is the Tree-of-Life a Tree?"

Invited seminar at the EEB Department at Yale University, November 30th, 2005

J. Peter Gogarten:
"Horizontal gene transfer and microbial evolution: Is the 'Tree of Life' a Tree?"
Presentation at the Astrobiology/Exobiology PI meeting, NASA Ames, August 16, 2005

J. Peter Gogarten:
"Prokaryotic Evolution: Is the 'Tree-of-Life' a Tree?' "
Invited lecture at the World Summit on Evolution in the Galapagos, June 9-12, 2005

Maria Poptsova:
"Horizontal Gene Transfer and Microbial Evolution: Testing Methods of Detection",
Invited lecture at the University of Rhode Island, February 2006

Olga Zhaxybayeva:
"Spectral Analyses of Cyanobacterial Genomes",
Lecture at the Annual Meeting of CIAR Program in Evolutionary Biology, Vancouver Island, BC, September 15-19, 2005

Other presentations at meetings

Maria Poptsova, Timothy J. Harlow and J. Peter Gogarten: "*The Power of Phylogenetic Approaches to Detect Horizontally transferred Genes*", presented at the Phylogenomics Conference, Saint-Adèle, Québec, Canada, March 15-19 2006.

Olga Zhaxybayeva, J. Peter Gogarten, Robert L. Charlebois, W. Ford Doolittle and R. Thane Papke: "*Phylogenetic Analyses of Cyanobacterial Genomes: Quantification of Horizontal gene Transfer Events"*, Phylogenomics Conference, Sainte-Adele, Quebec, Canada, March 15-19, 2006.

Pascal Lapierre (grad student, advisor J. Peter Gogarten): "*Size of the Prokaryotic Protein Universe*" selected for an oral presentation at the Astrobiology Science Conference, Washington DC, March 26-30, 2006

Greg Fournier (grad student, advisor J. Peter Gogarten) is invited to give an oral presentation on "*Evolution of Methanogenesis: An Ancient Transfer Event?"* at the Origin of Life Gordon Research Conference at Bates College in Lewiston, ME July 23-28, 2006

Kaiyuan Shi (grad student, Gogarten lab) was selected to give oral presentation on "*Parametric Bootstrap Analyses of Bacterial 16S rRNA Mosaicism*" (collaboration with Olga Zhaxybayeva, and Sushma Samala) at the National Academy of Sciences organized conference on the "Tapestry of Life: Lateral Transfers of Heritable Elements", which was held at Beckman Center, Irvine, California USA, on December 12-13, 2005.

Jinling Huang (postdoc, Gogarten lab, salary provided through a NASA Astrobiology Institute postdoctoral fellowship) presented a poster on "*Ancient gene transfer benefits phylogenetic reconstruction*" (Jinling Huang, Peter Gogarten) at the NAS Sackler Tapestry of Life Colloquium in Dec 2005.

Olga Zhaxybayeva (participation supported through the NSF) presented a poster on "*Spectral Analyses of Cyanobacterial Genomes: Quantification of Horizontally Transferred Genes"* (Olga Zhaxybayeva, R. Thane Papke, W. Ford Doolittle and J. Peter Gogarten) at the World Summit on Evolution, Galapagos Islands, Ecuador, June 9-12, 2005.

ONGOING RESEARCH

Development of a new algorithm to select gene families.

We developed the new phylogenetic algorithm, BranchClust, for the automated assembly of orthologous gene families. The reciprocal best blast hit method [5, 6] is very conservative, but often fails in the presence of paralogs, i.e., orthologous genes that in one lineage were recently duplicated (so-called inparalogs[7]) are excluded from the analyses. An improved reciprocity method with broken connections, that we developed and implemented last year, improved the selection quality by approximately 20% but still did not assemble many anciently duplicated gene families. In addition, both methods have a restriction on the number of taxa that can be analyzed due to restrictions of MySQL query engine. We developed an algorithm for the automated selection of orthologous families that recognizes orthologous genes from different species in a phylogenetic tree for any number of taxa. The algorithm is capable of distinguishing complete (containing all taxa) and incomplete (not containing all taxa) families and recognizes in- and out-paralogs[7]. The BranchClust algorithm is implemented in Perl with the use of the BioPerl module for parsing trees and will be made freely available. An unexpected outcome from the application of this algorithm is that the set of genes with orthologs in the two prokaryotic domains is about twice as big as previously assumed. A manuscript on the branchclust algorithm is in preparation.

Software tool for the visualization of genome mosaicism

We built the web-based software tool for the interactive analysis of cluster diagrams. The current implementation uses maps generated by a Self Organizing Map algorithm from bipartition tables (giving bootstrap support for all possible bipartitions for all gene families). The current version is available at http://bioinformatics.cs.uri.edu/gene-vis/template/, it demonstrates how SOM maps can be used for the comparative analysis of genomes using 14 archaeal genomes as example. The tool requires a browser that supports Java 1.4.2.

The site is divided into three sections: Map, Clusters and Bipartitions.
The Map contains the SOM map reconstructed from the bipartition matrix. The latter needs to be calculated independently and contains for each gene family a list of bootstrap support values for all possible bipartitions.

The bipartition section provides for each bipartition a a 3D function giving support values over the SOM. This allows the users to find regions in the SOM that contain gene families supporting a phylogenetic pattern.

The Clusters section allows an interactive analysis of individual clusters, or selected groups of clusters. Each point on the map corresponds to a group of gene families. When the user moves a mouse over the SOM, a pop-up window shows the coordinates of the cluster on the map and the list of families that constitute it. For any combination of clusters, a majority consensus phylogeny can be reconstructed on the fly using the data from the original bipartition matrix. The tree is provided using bootstrap support values for branch lengths, the tree can be further manipulated using the ATV module [8].

Work planned for the next funding period

We will improve the interface to interactively study the generated maps. Currently we use the ATV module to depict the phylogenies for the selected gene families. The use of ATV limits the interactivity and feedback provided by the program. We aim for an implementation that on demand provides annotation lines of selected gene families, allows to interactively display the individual gene phylogenies (and the support they provide for individual bipartitions), and the majority consensus trees for selected clusters of gene families. We also will add an interface for users to upload their own data matrices for analyses.

We will repeat analyses with larger datasets. Currently we only analyze gene families identified by the very restrictive reciprocal best-hit criterion. Utilizing BRANCHCLUST (see above) many more families are assembled and can be included in the analyses.

We also will apply the clustering algorithm to support values for embedded quartets. This will allow including gene families that are absent in some of the genomes. Together with the previous task, this will allow to include families catalyzing important metabolic pathways, a step crucial to begin correlating the molecular, fossil and geological records of early life.

We will explore Principal Component Analyses (PCA) and Local Linear Embedding (LLE) algorithms as alternatives to the Self Organizing Maps (SOM) to cluster gene families based on their bipartition data.

We will apply the analyses to different selection of genomes, with emphasis to those groups that might allow correlation to the fossil and geological records (cyanobacteria and the evolution of photosynthesis, archaea and the evolution of methanogenesis).

References:

1. Gogarten, J.P., and Townsend, J.P. (2005) Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* 3, 679-687
2. Zhaxybayeva, O., and Gogarten, J.P. (2003) An improved probability mapping approach to assess genome mosaicism. *BMC Genomics* 4, 37
3. Zhaxybayeva, O*., et al.* (2004) Genome mosaicism and organismal lineages. *Trends Genet* 20, 254-260
4. Shimodaira, H. (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51, 492-508
5. Zhaxybayeva, O., and Gogarten, J. (2002) Bootstrap, Bayesian probability and maximum likelihood mapping: Exploring new tools for comparative genome analyses. *BMC Genomics* 3, 4
6. Montague, M.G., and Hutchison, C.A., 3rd (2000) Gene content phylogeny of herpesviruses. *Proc Natl Acad Sci U S A* 97, 5334-5339.
7. Sonnhammer, E.L., and Koonin, E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* 18, 619-620
8. Zmasek, C.M., and Eddy, S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics* 17, 383-384