1st Progress Report for NASA Applied Information System Research Program (AISRP)
"Exploration of Novel Methods to Visualize Genome Evolution"
period: 9/1/04-8/31/05
Grant No: NNG04GP90G
Proposal Number: AISP03-0037-0008
PI: J.Peter Gogarten

Publications and presentations that resulted form the sponsored research:

Publications:
Hamel L., Zhaxybayeva O., Gogarten J.P. (2005):
PentaPlot: A software tool for the illustration of genome mosaicism.
BMC Bioinformatics 2005, 6:139

Gogarten, J. P., Townsend, J. P. (2005):
Horizontal gene transfer, genome innovation, and evolution
Nature Reviews in Microbiology, accepted for publication

Software release:
http://pentaplot.sourceforge.net
Pentaplot allows visualizing the phylogentic information content of 5 genomes.

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

"Is the "Tree of Life" a Tree?" J. Peter Gogarten, Seminar at the Origins Institute conference/workshop on: "Astrobiology and the Origins of Life", May 24-28, 2005

"Horizontal Gene Transfer and Microbial Evolution: Is the "Tree of Life" a Tree?", J. Peter Gogarten, Invited Seminar at the Bioinformatics Institute at the University of Georgia, Wednesday, May 4, 2005

"Horizontal gene transfer and the early evolution of life", J. P. Gogarten,
Invited Departmental Seminar at the Department of Biology and Biochemistry at the University of Houston, Texas, January 26, 2005

"New Tools for Visualizing Genome Evolution", Lutz Hamel* and J. Peter Gogarten,
Presentation at the NASA AISR Principle Investigator Meeting, Mountain View, April 2005

Bioinformatics at the University of Rhode Island, Lutz Hamel, Bioinformatics Workshop at Framingham State College, Mass., February 2005.

Poster presentations at meetings
Olga Zhaxybayeva, R. Thane Papke, Robert Charlesbois, W. Ford Doolittle and J. Peter Gogarten: "*Spectral analyses of completely sequenced genomes using Bipartitions and*

embedded Quartets", presented at the Origins Institute conference/workshop on: "Astrobiology and the Origins of Life", May 24-28, 2005

Olga Zhaxybayeva, R. Thane Papke, W. Ford Doolittle and J. Peter Gogarten "*Spectral Analyses of Cyanobacterial Genomes: Quantification of Horizontally Transferred Genes*" presented at the International Conference on Microbial Genomes April 13 - 16, 2005; Halifax, Nova Scotia, Canada

Workshops attended
NCBI PowerScripting, Maria Poptsova, Workshop at NCBI, NLM, NIH, Bethesda, MD, April 22-25, 2005.

ONGOING RESEARCH

Initially, Drs. Zhaxybayeva, Hamel and Gogarten worked on the project. Maria Poptsova was hired as postdoctoral fellow effective February 3, 2005. Dr. Poptsova has training in computational applications in biophysics and in software development, which already has benefited the ongoing research. Dr. Zhaxybayeva has accepted a position at Dalhousie University in Halifax. She continues to collaborate on the project; however, her salary will be provided through a fellowship from the Canadian Institutes for Health Research.

Software tool for the visualization of genome mosaicism

We have implemented a software tool for the visualization of genome mosaicism based on self-organizing maps. We successfully validated this implementation against results obtained with alternative methodologies. We feel that the visualization with self-organizing maps is particularly attractive when compared to dekapentagonal maps [1] and Lento plots [2] due to the explicit treatment of inter-cluster relationships on the maps. We currently are investigating constructing self-organizing maps on challenging genome bipartition data. We are very encouraged by these results will begin investigating using manifold learning based on locally linear embedding to construct even more effective maps. We also will investigate bipartition data with missing values.

Improvement of algorithms to select gene families

The effective assembly of genes from different genomes into gene families continues to be a controversial subject in comparative genomics. The use of a strict reciprocal best hit relationships is an effective, but very conservative approach [2-4]. Especially, in case of deep phylogenetic relationships a large number of gene families are not recognized under this criterion (e.g., the ATP synthase catalytic and non-catalytic subunits [5]). We explored alternative measures to define the "best hit", such as the percent identity and bitscore per alignments length, and we explored the impact of gradually relaxing the requirement that every member of a gene family has every other member of the gene family as top scoring hit in a search of the other genome. We find that allowing for a small number of broken connections (violations of the strict reciprocity) greatly improves the detection gene families without increasing the number of false positives.

We have automated the process to assemble gene families, so that the strict reciprocal hit criterion (with and without broken connections) can be used on a large number of

genomes -- previously, the number of genomes was limited due limitations of the database (MySQL) query interface. A manuscript reporting on these findings is in preparation. The developed software will be made available under the GNU license.

Bootstrap values from maximum likelihood analyses

Recent progress in phylogenetic analyses makes it possible to calculate more reliable confidence measures for bipartitions. In the past, we used TREE-PUZZLE [6] to calculate distance matrices from pseudo-samples generated through bootstrapping, and we analyzed these distance matrices through neighbor joining, a fast algorithmic approach to building phylogenetic trees [7]. PHYML is a new software package that allows the fast and accurate calculation of maximum likelihood phylogenies from large datasets [8]. We have developed scripts that utilize PHYML on a local cluster to calculate bootstrap support values. First results indicate that the support measures calculated through PHYML are qualitatively similar to the ones obtained through neighbor joining; however, the use of an optimality criterion will lead to a wider acceptance of the obtained data, and for divergent gene families the increase in reliability and resolution might be crucial to draw an accurate picture of early evolutionary history.

During the **next funding period** we will add gene families to the analyses that are absent in some of the genomes; we will implement strategies to combine bipartition and quartet data, and we will explore the extent to which certain types of proteins (enzymes, housekeeping genes, metabolic pathways) share phylogenetic information content.

References

1. Zhaxybayeva O, Hamel L, Raymond J, Gogarten JP: Visualization of the phylogenetic content of five genomes using dekapentagonal maps. *Genome Biol* 2004, 5:R20.
2. Zhaxybayeva O, Lapierre P, Gogarten JP: Genome mosaicism and organismal lineages. *Trends Genet* 2004, 20:254-260.
3. Zhaxybayeva O, Gogarten JP: Bootstrap, Bayesian probability and maximum likelihood mapping: Exploring new tools for comparative genome analyses. *BMC Genomics* 2002, 3:4.
4. Montague MG, Hutchison CA, 3rd: Gene content phylogeny of herpesviruses. *Proc Natl Acad Sci U S A* 2000, 97:5334-5339.
5. Zhaxybayeva O, Lapierre P, Gogarten JP: Ancient Gene Duplications and the Root(s) of the Tree of Life. *Protoplasma* 2005, in press.
6. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 2002, 18:502-504.
7. Zhaxybayeva O, Gogarten JP: An improved probability mapping approach to assess genome mosaicism. *BMC Genomics* 2003, 4:37.
8. Guindon S, Gascuel O: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 2003, 52:696-704.