

Projects	Award Information - ER63201-1017853-0007174
Action Due Now	ID: ER63201-1017853-0007174
All My Projects	Principal Investigator: Gary J. Olsen 217-244-0616
Project Detail	Co-PIs:
Proposal	Institution: University of Illinois
Award	Title: Bioinformatics for Genome Analysis
Progress	SC Division: SC-23.1
Lead PI Info	Program Manager: John C. Houghton 301-903-8288
Msg to Proj Mgr	Research Areas:
Other Items	Project Progress
FAQs	Most recent report of results to date:
My Preferences	Final Report, July 1, 2001 - June 30, 2005
Change Password	Effect of sequence alignment order on alignments and inferred trees (this project partially overlaps gene transfer research supported by <u>DE-FG02-01ER63146</u>)
Logout	Nesbo, Boucher and Doolittle (2001) used phylogenetic trees of four taxa to assess whether euryarchaeal genes share a common history. They have suggested that of the 521 genes examined, each of the three possible tree topologies relating the four taxa was supported essentially equal numbers of times. They suggest that this might be the result of numerous horizontal gene transfer events, essentially randomizing the relationships between gene histories (as inferred in the 521 gene trees) and organismal relationships (which would be a single underlying tree).
Help	Motivated by the fact that the order in which sequences are added to a multiple sequence alignment influences the alignment, and ultimately inferred tree, we were interested in the extent to which the variations among inferred trees might be due to variations in the alignment order. This bears directly on our efforts to evaluate and improve upon methods of multiple sequence alignment. We set out to analyze the influence of alignment order on the tree inferred for 43 genes shared among these same 4 taxa. Because alignments produced by CLUSTALW are directed by a rooted guide tree (the dendrogram), there are 15 possible alignment orders of 4 taxa. For each gene we tested all 15 alignment orders, and as a 16th option, allowed CLUSTALW to generate its own guide tree. If we supply all 15 possible rooted guide trees, we expected that at least one of them should be as good as CLUSTAL's own guide tree, but most of the time they differed (sometimes being better than CLUSTAL's default tree and sometimes being worse. The difference seems to be that the user-supplied tree is not given meaningful branch lengths, which effect the assumed probability of amino acid changes. We examined the practicality of modifying CLUSTALW to improve its treatment of user-supplied guide trees. This work became ever increasing bogged down in finding and repairing minor bugs in the CLUSTALW code. This effort was put on hold as we feel that our other proposed approaches will ultimately be better.
	Analyzing gene expression
	We have been completing work initiated under a previous DoE grant to study gene expression and regulation. Archaeal (<i>Methanococcus jannaschii</i>) promoter sequences are experimentally identified by in vitro binding of transcription initiation proteins. Fragments of genomic DNA that include promoters bind the proteins and are retarded in a gel mobility shift assay. Shifted DNA is isolated, amplified, cloned and sequenced. We found that the shifted DNA is dominated by upstream sequences of tRNA genes (among the most highly express genes in the cell).
	In this manner we assembled the largest collection of experimentally identified archaeal promoters. Most of these promoters include a canonical TATA consensus sequence and a "B-responsive element" (even though most genes in the genome do not have recognizable forms of these elements). Information analyses of the sequences has extended the corresponding consensus regions beyond their previously defined boundaries. We have an unpublished manuscript describing these results.
	Work on other aspects of archaeal gene expression and regulation (including publication of earlier DoE-funded proteomics work) has been supported, but is being moved to other funding (NASA and University of Illinois).
	In analyzing the archaeal promoter sequences compiled, we have created a minor modification of the sequence LOGOs program (in Perl, not the original PASCAL) that adjusts the analysis for G+C content of the DNA. The intergenic regions of the DNA being analyzed is extremely low in G+C, so this substantially changes the profile generated. One counter intuitive result that we have encountered that when the bias is great enough (say 10% G, 10% C, 40% A, 40% T), the most abundant observed nucleotide at an alignment position (say 35% A) can have a negative score in the profile because it is occurring at below

random expectation. Although there is no new computational issue here, there is an presentation issue of having the largest letter in the LOGO actually be a disfavored residue type. We are considering an appropriate representation to help those viewing the profile.

Analyzing codon usage versus gene expression level

It has been commonly asserted that (in some organisms) high-expression genes are detectable by their characteristic codon usage. As part of his thesis research in open reading frame detection, Jonathan Badger computationally identified a set of genes in *M. jannaschii* that appear to constitute a high-expression class. We are currently mining protein abundance data from two-dimensional gel analyses previously performed by Carol Giometti (Argonne National Laboratory), Claudia Reich (here at UIUC) and John Yates, III (Scripps Research Institute) to provide expression-level estimates of specific *M. jannaschii* genes and correlate them with the codon usage data. This work is part of an effort to automate interpretation of codon usage data. (Most of the applications of these methods will be funded by a grant for gene transfer research.)

John Sabo found that due to mixtures of proteins being present in many spots, and many proteins being spread over multiple spots by varying levels of modification (and perhaps degradation in a few cases), that there quantitative statements become very complex (a give protein is present in not more than some amount, or at least one of the two proteins present must be particularly abundant). Regardless, there does seem to be a correlation of protein level and codon bias in several of the organisms examined.

More strikingly, has been a graphical representation of the codon bias of orthologous genes between genome. To do this he first assess the codon bias of each gene in one organism by their positions along the first principle axis in a correspondence analysis (the form of Principle Component Analysis used by the CodonW program) of all the genes in that genome (this differs from all of the other measures of codon bias). This codon bias value is then used to color code the presumptive orthologous genes from another organism. These color coded genes are then plotted according the the first two principle axes of correspondence analysis of all the genes from the second genome. (For the correspondence analyses, all of the genes are used, but the color coding is only applied to those with presumptive orthologs.) In organisms where Jonathan Badger observed shifts in codon usage for ribosomal proteins, the correlations are striking. In at least one case with no obvious bias for the ribosomal proteins, the present analysis has revealed that there is a clear systematic correlation, but it is not the primary explanation of variations in codon usage for the particular organism (thus it did not align well with the plane defined by the first two principle components and was not readily visible). Overall, these results suggest that defining the set of proteins with similar expression levels can be automated by seeking orthologs of more than just the ribosomal proteins and the translation elongation factors.

The graphical display led to several observations:

1. There only a few (0-2) genes with "high-expression" codon bias that do not have an annotated function in each genome analyzed.
2. Most "high-expression" genes have identifiable orthologs in the other genomes analyzed.
3. A lower percentage of the "moderate-expression" genes have identifiable orthologs in the other genomes analyzed.
4. A still lower percentage of the "low-expression" genes have identifiable orthologs in the other genomes analyzed.
5. Essentially none of the "alien" genes (identified solely by unusual codon usage) have identifiable orthologs in the other genomes analyzed.

These observations were initially made in for archaeal genomes. They have now been repeated and extended with a diverse array of bacterial genomes, examining separations ranging from strains-to-strain differences, to different bacterial "Phyla".

In support of the analysis of new genomes, we have made tools for: (1) identifying orthologs between the new genome and the original "reference" genome from which expression level is inferred; (2) projecting the expression level from the reference genome to the orthologs in the new genome; (3) converting the expression level to a color scheme and applying it to the gene codon usage coordinates (from CodonW); and (4) the display of the resulting figure in 2-D or 3-D graphics. The move to 3-D graphics has led to a number of surprising, and, as yet unexplained, categories of genes in several of the published genomes.

We are we have prototyped some software tools that exploit these observations to simplify the detection of codon usage features in new genomes.

Automated orthologue identification

The most common method of identifying orthologous genes is the use of "bidirectional best hits". That is, find pairs of proteins in two genomes (A from genome 1 and B from genome 2) such that if B is the genome 2 protein most similar to protein A, and A is the genome 1 protein most similar to B. To facilitate a more comprehensive analysis of the role of gene transfer in thermal adaptation of an organism, we have extended this to 3 and 4 genome comparisons. (The ultimate use of the protein sets was for phylogenetic analysis, which is a still more reliable method for assessing orthology of genes.)

Phylogenetic inference

We continued our work on improving phylogenetic analysis software, primarily based upon fastDNAm1. This includes collaborations with Craig Stewart and coworkers at Indiana University. They are continuing to work on the parallel code and are interested in incorporating features suggested by other research groups.

Finding an individual capable of making progress on this project was much more difficult than anticipated. In collaboration with Bruce Parella (Argonne National Laboratory), fastDNAm1 was converted to a more objected-oriented C++ program, creating the program fastSEQm1. The program structure was generalized to allow the input of both protein and nucleotide sequences (as well as even more general probability vectors). The computing was restructured to allow integration of arbitrary models of change, and even different models at different sites. The code has been tested and compared with the original to verify it results.

Genome sequence analysis

We continued assisting in the analysis of several genomes, and used these as opportunities to develop and test new tools. Some comparative analyses of genome structure in Salmonella serovars have been published.

Katie Karberg has written scripts that automatically categorize the results of direct genome comparisons in terms of the relative status of homologous genes (and paralogs). In particular, she has automatically identified 210 pseudogenes in Salmonella typhi that correspond to apparently functional genes in Salmonella typhimurium. This is 6 more than reported by the original authors; we are not sure where our criteria differ from theirs, though hand checking of differences supports our conclusions.

The programs were also designed to handle the comparison of a complete genome with a partial genome, so they are being applied to the data available from other serovars. Before we published, similar software was published by Ochman's laboratory.

The SEED project (<http://theseed.uchicago.edu/FIG/index.cgi>)

Ross Overbeek, a long term collaborator, started a project to produce the SEED, an open source comparative genome analysis platform. We established close contact with this effort, so as to make tools that will complement the project. We have contributed Perl libraries for some of the core functions. These PERL packages include:

1. general sequence reading, writing and manipulation;
2. one of the most complete libraries for phylogenetic tree manipulation and analysis (a Perl port of a Prolog library that we wrote 10 years ago for the Ribosomal Database Project);
3. the ability to add sequences to an alignment by aligning on the set of most related sequences (an approximation of aligning on the related subtree);
4. the very flexible program for producing representative sets of sequences (similar to clustblast, but faster by virtue of not computing all pairwise similarities); and
5. a very flexible BLAST output parser that permits the calling program to request data all at once, one query at a time, one subject sequence at a time, or one HSP at a time.

Among the novel tools introduced are the ability to sort similarities by bit-score-per-position or percent sequence identity, allowing helping to prevent high similarity matches to short sequences from being lost at the bottom of the similarities list. In addition, we have introduced a unique coloring system to the user interface that adjusts the color scheme of the displayed similarity to reflect the extent to which the similarity covers the the query and subject sequences. The shorter the region of similarity, the more intense the color, with the hue reflecting the position in the molecule (red near beginning, green near middle and blue near the end). This visual feedback helps to rapidly recognize a fused or partial gene.

Another feature added to the display of similarities is the ability to collect all matches to the same genome together, an invaluable tool in working with paralogues. Other unique tools for genome analysis include phylogenetic analysis of the related genes, again helping to sort through paralogues.

Other software created and integrated into the SEED include a graphical display of blast matches against genes or proteins, with genome context. In the case of tblastn searches, this is integrated with the ability to add the region of similarity as a new protein-coding gene. The giving the user the ability to modify the assumed genetic code, this feature has been used to curate most of the known selenocysteine-containing protein genes in the SEED.

Small subunit ribosomal RNA gene analysis

An area that the PI has been involved in since its inception is the use of rRNA genes to analysis microbial communities. Periodically experts in the field including S. J. Giovannoni, E. DeLong, and N. R. Pace have been asked about whether they have recently reevaluated the rDNA primers used in the light of over 200,000 sequences and the new metagenomic data. Surprisingly the answer has always been "no". Given current volume of information, manual analysis was out of the question. In addition, it requires finding an unknown sequence to observe its variation. We have created programs and algorithms for analyzing rRNA

sequences, genomes and metagenomes for the sequences at any desired location (we are interested in primer-binding sites, but any location will work) based on its context. It is sufficiently efficient to analyze large data sets. We extracted the data for >200,000 rRNA sequences in the RDP, for the complete Sargasso Sea metagenomic data set and for all of the genomes in the SEED. This has led to our proposing modifications to the design of some of the primers (as part of other work).

Most recent products delivered:

This work contributed to the following publications:

McCloskey, J. A., Graham, D. E., Zhou, S., Crain, P. F., Ibba, M., Konisky, J., Soll, D., and Olsen, G. J. 2001. Posttranslational modifications of archaeal tRNAs: Identities and phylogenetic relations of nucleotides from mesophilic and hyperthermophilic Methanococcales. *Nucleic Acids Res.* 29: 4699-4706.

Edwards, R. A., Olsen, G. J., and Maloy, S. R. 2002. Comparative genomics of closely related *Salmonellae*. *Trends Microbiol.* 10: 94-99.

Edwards, R. A., Olsen, G. J., and Maloy, S. R. 2002. The importance of complete genome sequences. *Trends Microbiol.* 10: 220-220.

Kurland, C., and Olsen, G. 2002. Genomics. *Curr. Opin. Microbiol.* 5: 497-498.

Giometti, C. S., Reich, C., Tollaksen, S., Babnigg, G., Lim, H., Zhu, W., Yates, J., III, and Olsen, G. 2002. Global analysis of a "simple" proteome: *Methanococcus jannaschii*. *J. Chromatog. B.* 782: 227-243.

Lim, H., Eng, J., Yates, J. R., III, Tollaksen, S. L., Giometti, C. S., Holden, J. F., Adams, M. W. W., Reich, C. I., Olsen, G. J., and Hays, L. G. 2003. Identification of 2D-gel proteins: A comparison of MALDI/TOF peptide mass mapping to microLC-ESI tandem mass spectrometry. *J. Am. Soc. Mass Spectrom.* 14: 957-970.

Zhu, W., Reich, C. I., Olsen, G. J., Giometti, C. S., and Yates, J., III. 2004. Shotgun proteomics of *Methanococcus jannaschii* and insights into methanogenesis. *J. Proteome Res.* 3: 538-548.

Best, A. A., Morrison, H. G., McArthur, A. G., Sogin, M. L., and Olsen, G. J. 2004. Evolution of eukaryotic transcription: Insights from the genome of *Giardia lamblia*. *Genome Res.* 14: 1537-1547.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes, S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A. C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G. D., Rodionov, D. A., Rickert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., and Vonstein, V. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1,000 genomes. *Nucleic Acids Res.* 33: 5691-5702.

Qiao, B., Goldberg, T. L., Olsen, G. J., and Weigel, R. M. 2006. A computer simulation analysis of the accuracy of partial genome sequencing and restriction fragment analysis in the reconstruction of phylogenetic relationships. 6: 323-330.

Elements of this work have been presented in the following talks at scientific meetings:

Olsen, G. J. 2001. The persistence of an essence: Archaea as defined by their genomes. Gordon Research Conference on Archaea: Ecology, Metabolism and Molecular Biology, Proctor Academy, Andover, NH, Aug. 5-10, 2001.

Olsen, G. J. 2001. You gain some, you lose some: The comings and goings of genes, and the history of life. Keynote Lecture, American Society for Rickettsiology / Bartonella Joint Conference, Big Sky, MT, Aug. 17-22, 2001.

Olsen, G. J. 2001. Microbial Genomics. USDA Comparative Insect Genomics Workshop, Arlington, VA, Oct. 28-30, 2001.

Olsen, G. J. 2001. Environments and Genomes: The mix and match of genes in adaptation and innovation. Geological Society of America Annual Meeting, A Geo-Odyssey, Boston, MA, Nov. 5-8, 2001.

Olsen, G. J. 2002. Temperature and evolution of archaeal genomes. Gordon Research Conference on Molecular Evolution, Sheraton Harbortown, Ventura, CA, Jan. 13-18, 2002.

Olsen, G. J. 2002. Historical Records in Genes and Genomes: Insights from faded, shredded documents. A Discussion by Experimentalists of the Origin of Life, 10th Annual Suddath Symposium, Georgia Technical University, Atlanta, GA, Apr. 19-20, 2002.

Olsen, G. J. 2002. Evolutionary Rhythm and Beat: Lessons from comparative genomics. American Society for Microbiology 102th General Meeting, Salt Lake City, UT, May 19-23, 2002.

Olsen, G. J. 2002. You Gain Some, You Lose Some: The comings and goings of genes and the history of life. American Society for Microbiology 102th General Meeting, Salt Lake City, UT, May 19-23, 2002.

Olsen, G. J. 2002. Prokaryotic Genome Evolution: Continuity in the presence of flux. Molecular Evolution: Evolution, Genomics, Bioinformatics. International Society of Molecular Evolution and Society for Molecular

Biology and Evolution, Sorrento, Italy, June 13-16, 2002.

Olsen, G. J. 2002. Historical records in genes and genomes: Insights from faded shredded documents. Extremophiles 2002: The 4th International Congress. Complesso Universitario Monte Sant'Angelo of the University "Federico II" of Naples, Naples, Italy, Sept. 22-26, 2002.

Olsen, G. J. 2003. Evolution of cells, with a special look at parasites and pathogens. Joint Annual Meeting 2003 Swiss Society for Microbiology Swiss Society for Infectious Diseases Swiss Society of Tropical Medicine and Parasitology. Basel, Switzerland, Mar. 6-7, 2003.

Olsen, G. J. 2003. The limits of molecular phylogeny. American Society for Microbiology 103th General Meeting, Washington DC, May 18-22, 2003.

Olsen, G. J. 2003. Ribosomal RNA, gene transfer and the meaning of a tree of life. American Society for Microbiology 103th General Meeting, Washington DC, May 18-22, 2003.

Olsen, G. J. 2003. A tree in the jungle of gene histories: or, don't trip on the vines. Gordon Research Conference on The Origins of Life. Bates College, Lewiston, ME, July 13-18, 2003.

Olsen, G. J. 2003. Horizontal gene transfer and the triumph of Darwin. 2003 Crafoord Days Symposium. Lund, Sweden, Sep. 22, 2003 and Stockholm, Sweden, Sep. 23, 2003.

Olsen, G. J. 2003. Horizontal gene transfer and the global gene pool. European Conference on Prokaryotic Genomes. Gottingen, Germany, Oct. 5-8, 2003.

Olsen, G. J. 2004. Comparative analysis of protein thermal adaptation. American Chemical Society Annual Meeting. Anaheim, CA, Mar. 28-31, 2004.

Olsen, G. J. 2004. Comparative genomics. EMBO Conference on Molecular Microbiology: Exploring Prokaryotic Diversity. EMBL, Heidelberg, Germany, Apr. 22-26, 2004.

Olsen, G. J. 2004. Sharing genes and sharing environments: Two versions of cooperation. Okazaki Biology Conference 2: Terra Microbiology. Mielpearl, Ise-Shima, Japan, Sep. 26-30, 2004.

Olsen, G. J. 2004. Genomic biology as a pathway to the future of sustainable technologies. American Oil Chemists' Society meeting "Industrial Applications of Renewable Resources." Chicago, IL, Oct. 12-14, 2004.

Olsen, G. J. 2005. Thermal adaptation in Archaea: Life at many temperatures. Archaea in the Environment. University of Southern California Wrigley Institute for Environmental Studies. USC Wrigley Marine Science Center, Catalina Island, CA, June 25, 2005.

Most recent notes concerning the project:

Publications containing elements of this work:

Stewart, C. A., Hart, D., Berry, D. K., Olsen, G. J., Wernert, E., and Fischer, W. 2001. Parallel implementation and performance of fastDNAMl -- a program for maximum likelihood phylogenetic inference. Proceedings of SC2001, Denver, CO, November 2001.

Edwards, R. A., Olsen, G. J., and Maloy, S. R. 2002. Comparative genomics of closely related Salmonellae. Trends Microbiol. 10: 94-99.

Giometti, C. S., Reich, C., Tollaksen, S., Babnigg, G., Lim, H., Zhu, W., Yates, J., III, and Olsen, G. 2002. Global analysis of a "simple" proteome: Methanococcus jannaschii. J. Chromatog., in press.

Elements of this work have been presented in 8 talks at scientific meetings:

Olsen, G. J. 2001. The persistence of an essence: Archaea as defined by their genomes. Gordon Research Conference on Archaea: Ecology, Metabolism and Molecular Biology, Proctor Academy, Andover, NH, Aug. 5-10, 2001.

Olsen, G. J. 2001. You gain some, you lose some: The comings and goings of genes, and the history of life. Keynote Lecture, American Society for Rickettsiology / Bartonella Joint Conference, Big Sky, MT, Aug. 17-22, 2001.

Olsen, G. J. 2001. Microbial genomics. USDA Comparative Insect Genomics Workshop, Arlington, VA, Oct. 28-30, 2001.

Olsen, G. J. 2001. Environments and genomes: The mix and match of genes in adaptation and innovation. Geological Society of America Annual Meeting, A Geo-Odyssey, Boston, MA, Nov. 5- 8.

Olsen, G. J. 2002. Historical records in genes and genomes: Insights from faded, shredded documents. A Discussion by Experimentalists of the Origin of Life, 10th Annual Suddath Symposium, Georgia Technical University, Atlanta, GA, April 19-20, 2002.

Olsen, G. J. 2002. Evolutionary rhythm and beat: Lessons from comparative genomics. American Society for Microbiology 102th General Meeting, Salt Lake City, UT, May 19-23, 2002.

Olsen, G. J. 2002. You gain some, you lose some: The comings and goings of genes and the history of life. American Society for Microbiology 102th General Meeting, Salt Lake City, UT, May 19-23, 2002.

Olsen, G. J. 2002. Prokaryotic genome evolution: Continuity in the presence of flux. Molecular Evolution: Evolution, Genomics, Bioinformatics. International Society of Molecular Evolution and Society for Molecular Biology and Evolution, Sorrento, Italy, June 13-16, 2002.

Elements of this work have been presented in a seminar and a workshop:

"It's not quite chaos, but ...: Nature's piecemeal approach to genomes." NEC Research Institute, Princeton, NJ, Oct. 5, 2001.

"Protein sequence analysis." University of Illinois Biotechnology Center Bioinformatics Conference/Workshop, Urbana, IL, May 3, 2002.

Elements of this work have been presented in 3 posters:

Hendrickson, E. L., Olson, M., Olsen, G., and Leigh, J. A. 2002. Genome sequence of *Methanococcus maripaludis*, a genetically tractable methanogen. DOE Genome Contractor-Grantee Workshop IX, Oakland, CA, Jan. 27-31, 2002.

Li, E., Best, A. A., Colon, G. M., Reich, C. I., and Olsen, G. J. 2002. A genome-wide search for archaeal promoter elements. DOE Genome Contractor-Grantee Workshop IX, Oakland, CA, Jan. 27-31, 2002.

Li, E., Reich, C., and Olsen, G. 2002. Whole genome approach to identify promoter sequences in *Methanococcus jannaschii*. American Society for Microbiology 102th General Meeting, Salt Lake City, UT, May 19-23, 2002.

Other Project Information Sources:

Project URL:

None

Related URL at institution:

None

Contact: RIMSAdmin@science.doe.gov

[Search OBER Abstracts Database](#) Note: There is a delay in posting abstracts to allow for Program Manager review.