Final report on the project DE-FG02-01ER63205 (DOE), 2004:

**New Markov Model Approaches to Deciphering Microbial Genome Function and Evolution: Comparative Genomics of Laterally Transferred Genes**

The project started in September 2001.

In the project we have made—considerable progress, specifically in the following areas:

- exploring opportunity of developing more sensitive Markov models for gene finding, such as variable order and interpolated Markov models;
- developing a specific tool for gene prediction in draft genomes;
- building a bioinformatic analysis pipeline to rigorously test each of the gene candidates within an explicit phylogenetic framework.

**ABSTRACT**

Algorithmic methods for gene prediction have been developed and successfully applied to many different prokaryotic genome sequences. As the set of genes in a particular genome is not homogeneous with respect to DNA sequence composition features, the GeneMark.hmm program utilizes two Markov models representing distinct classes of protein coding genes denoted “typical” and “atypical”. Atypical genes are those whose DNA features deviate significantly from those classified as typical and they represent approximately 10% of any given genome. In addition to the inherent interest of more accurately predicting genes, the atypical status of these genes may also reflect their separate evolutionary ancestry from other genes in that genome. We hypothesize that atypical genes are largely comprised of those genes that have been relatively recently acquired through lateral gene transfer (LGT). If so, what fraction of atypical genes are such *bona fide* LGTs? We have made atypical gene predictions for all fully completed prokaryotic genomes; we have been able to compare these results to other "surrogate" methods of LGT prediction.

**Specific Aim 1. Improving accuracy of atypical gene prediction.**

In the paper published in *Bioinformatics* entitled “Effects of choice of DNA sequence model structure on gene identification accuracy” (by Rajeev K. Azad and Mark Borodovsky, full citation: *Bioinformatics*, 2004, Vol. 20, No. 7, pp. 993-1005) we have explored the issue of improving the gene finding accuracy in prokaryotic genomes. Gene finding methods based on Markov models are known to perform consistently well for prokaryotic genomes. GeneMark, a frequently used gene finding algorithm, uses fixed order homogeneous and inhomogeneous Markov models to locate genes in genomic sequences. While for a limited training data set one can select the best performing order of the Markov model, it has been suggested that interpolation methods which combine models of different orders can produce even better accuracy. It is thus desirable to analyze the performance of such methods in the context of gene finding. In this work, we
implement several variants of interpolation technique in the Markov model training and present a comparative analysis of results produced by the GeneMark algorithms using interpolated Markov models and fixed order Markov models. For genomes with a mid-range GC content, the model built by deleted interpolation slightly outperformed other models under some conditions. For genomes with high or low GC content, we observed that, in several important cases, a fixed order model performs better in detecting the coding potential of a sequence segment as compared to other types of models. In many cases, similar results were observed when accuracy was measured by ability of an algorithm to identify genes in real genomes as compared with annotated genes.

In another paper, with the title “GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses” (by John Besemer and Mark Borodovsky), we describe several versions of GeneMark software, including versions is to derive models and analyze nearly sequenced (but not complete) genomes. Many such “draft” genome sequences are available at the DOE Joint Genome Institute (http://www.jgi.doe.gov/JGI_microbial/html). Our initial runs have shown a reasonable match between our predictions and those made by other programs currently in use at JGI. Most of the differences are seen in the annotation of translation starts and short genes and also in “atypical genes” which could be laterally transferred genes.

Still the previous version of GeneMarkS was successfully used for annotation of many DOE genomes including the Shewanella genome (Daraselia, N., Dernovoy, D., Tian, Y., Borodovsky, M., Tatusov, R., and Tatusova, T. “Re-annotation of Shewanella oneidensis genome”; paper submitted to OMICS A Journal of Integrative Biology), though we were not able yet to confidently identify the fraction of predicted atypical genes which are LTGs. This could be done after completion of the second part of the project (specific aim 2)

**Specific aim 2. Predicting the functions and evolutionary relationships of atypical genes**

In order to validate the use of atypical genes for LGT detection, we have been building a bioinformatic analysis pipeline to rigorously test each of the gene candidates within an explicit phylogenetic framework. This process starts with gene predictions and ends with a phylogenetic reconstruction of each candidate. From the set of bona fide LGTs that we have identified, we are determining the LGT parameters to which our gene finding programs are most sensitive (i.e. time scale of transfers, phylogenetic distance from transfer source, etc.). We are utilizing this pipeline to estimate the extent and pattern of LGT in a selection of genomes, both complete and nearly complete, with the long term goal of analyzing all such sequences.

The first full-scale test of the completed pipeline was near completion for Thermotoga maritima—a genome with a high putative fraction of LGTs which serves as a clear test-case for our analysis pipeline. An iterative BLASTP analysis was performed on all 52 GeneMark.hmm atypical predicted genes from T. maritima. The resulting phylogenetic distributions revealed 14 cases (27%) capable of supporting an LGT hypothesis and,
therefore, worthy of further study. While the iterative BLAST analyses are computationally intensive, they are important in order to ensure our study does not suffer from inadequate taxon sampling which has plagued many attempts at LGT detection in the past. Our iterative BLAST procedure has been found to be robust with respect to E-value, producing nearly congruent result sets across differences in E-value as high as ten orders of magnitude (e^{-10} to e^{-20}). These genes are currently being processed by the phylogenetic portion of our pipeline: multiple alignment via T-COFFEE; alignment cleaning via GBlocks; distance calculation via TREE-PUZZLE; constructing a Neighbor-Joining tree via Neighbor†; bootstrap calculation via Seqboot†, Protdist†, Neighbor† and Consense†. We are also exploring other phylogenetic algorithms for inclusion in our pipeline, including Bayesian methods. Still we have experienced difficulties with getting accurate identification of the source of LGT, this difficulty brought us to conclusion that the task is too difficult and just the presence of LGT should be a result of the analysis rather than more elaborate prediction of where the gene came from.

Simultaneously, we are testing a set of putative LGTs from the T. maritima genome identified by Garcia-Valive et al [1] based on aberrant G+C content, codon or amino acid usage, etc. Interestingly, of the 194 total putative LGTs for T. maritima, only 22 were also predicted as atypical by the GeneMark.hmm program. The vast majority of those remaining (164) were predicted as typical by GeneMark.hmm. Eight genes were not predicted at all and may simply be false positive genes, as the Garcia-Valive data is based on the GenBank annotation. Of these 164 genes, iterative BLAST analysis produced only 15 (9%) phylogenetic distributions which may be indicative of potential LGT.

The pipeline test results on several LGT candidates identified as atypical genes were presented at the “Genomes to Life” Contractor-Grantee Workshop I in Arlington, VA in February 2003 and at the “New Horizons in Genomics” DOE Joint Genome Institute meeting in Santa Fe, NM in March-April 2003.

Specific aim 3. Creating a new resource for comparative microbial genomics. This aim was eliminated upon DOE correction of the budget and specific aims.

† Neighbor, seqboot, protdist and consense are all part of the PHYLIP [2] package.
