Final Technical Report DE-FG02-04ER63942

Abstract

The transcription regulatory network is arguably the most important foundation of cellular function, since it exerts the most fundamental control over the abundance of virtually all of a cell’s functional macromolecules. The two major components of a prokaryotic cell’s transcription regulation network are the transcription factors (TFs) and the transcription factor binding sites (TFBS); these components are connected by the binding of TFs to their cognate TFBS under appropriate environmental conditions. Comparative genomics has proven to be a powerful bioinformatics method with which to study transcription regulation on a genome-wide level. We have further extended comparative genomics technologies that we introduced over the last several years. Specifically, we developed and applied statistical approaches to analysis of correlated sequence data (i.e., sequences from closely related species). We also combined these technologies with functional genomic, proteomic and sequence data from multiple species, and developed computational technologies that provide inferences on the regulatory network connections, identifying the cognate transcription factor for predicted regulatory sites. Arguably the most important contribution of this work emerged in the course of the project. Specifically, the development of novel procedures of estimation and prediction in discrete high-D settings has broad implications for biology, genomics and well beyond. We showed that these procedures enjoy advantages over existing technologies in the identification of TBFS. These efforts are aimed toward identifying a cell’s complete transcription regulatory network and underlying molecular mechanisms.

Summary

Accomplishments under this award fall into two general categories: bioinformatics technology development and bioinformatics applications to transcription regulation studies of bacterial species of environmental interest. In addition, we describe efforts toward resource sharing – providing open-source software as well as web services to the scientific community. Scientific publications resulting from this award are provided at the end of this report.

I. Bioinformatics technology development

A. Regulatory networks

We developed three regulatory network technologies. (i) We extended our previous Gibbs sampling models to incorporate phylogenetic relationships. We employed the Metropolis-Hasting sampling algorithm to draw samples from a full phylogenetic model and used centroid estimators for the cis-regulatory predictions\(^1\). We demonstrated that predictions using this model had improved sensitivity and positive predictive value over established methods, and showed that centroid estimators out-perform MAP estimators. (ii) We developed a technology (PhyloScan) for the identification of statistically significant matches to position weight matrices for related clades of species\(^2\). (iii) In collaboration with Gary Stormo, we developed a method to predict cognate transcription factors (TFs) for identified regulons, and thus extend regulatory network predictions to cis-trans connections\(^3\).

B. Comparative genomics

We developed a sequence weighting procedure that minimizes the variances of parameter estimates\(^4\). We showed that, even with optimal weights, estimates of base frequency parameters
are inefficient for a clade of bacterial species, as well as a clade of mammalian species. We developed a method for estimating the effective sample size of sequences from clades of phylogenetically related species. We developed tools to collect percent identity data for global alignments of orthologous sequences and generate summary statistics to assist the selection of appropriate species for motif finding studies.

C. Statistical inference in discrete high-D spaces

Arguably, the most important products of this research were the theoretical and methodological development of procedures for statistical inferences in discrete high-dimensional (high-D) spaces. First, we focused on the general problem of parameter estimation in discrete high-D spaces. We applied statistical decision theory to show that there is no principled reason to expect the popular highest-scoring (HS) estimators, including MLEs, MAPs, maximum similarity, and minimum free energy, to be representative of the data’s implication, and we found that the probability of an HS solution was often very small. To address these limitations, we developed alternative “centroid” estimators. We showed that centroids enjoy theoretical advantages over all of the HS estimators. Specifically, we showed that for binary and nominal variables, centroid estimators minimize expected pth power loss functions. In an important class of problems, centroids correspond to consensus estimators, and under squared-error loss the centroid is the feasible solution that is nearest to the mean. Accordingly, the centroid garners information from the entire ensemble of solutions to find an estimator that is representative of the entire posterior space. Secondly, we developed a general procedure for Bayesian confidence limits, a.k.a. credibility limits, of point estimates in discrete spaces, and illustrated its application to sequence alignment. These confidence limits report on the global reliability of an alignment. Using promoters sequences from Shewanella species, we found that the reliability of alignments for orthologous sequence pairs vary widely gene-to-gene and species-pair to species-pair. There is now clear evidence of the advantages of centroid estimators to predict ground truth standards in three important applications: the prediction of RNA secondary structures, protein structure prediction by homology, and motif finding. Since in each case the probabilistic model of the centroid and the HS estimator are identical, these improvements in the prediction of ground truth reference sets stem entirely from differences in the estimation procedure.

II. Bioinformatics applications

A. Rhodopseudomonas palustris

We completed a genome-scale phylogenetic footprinting study focused on R. palustris. This alpha-proteobacterial species carries out three of the chemical reactions that support life on this planet: the conversion of sunlight to chemical-potential energy, the conversion of carbon dioxide to cellular material, and the fixation of atmospheric nitrogen into ammonia. Our objective was to elucidate regulons in this bacterium using comparative data from 7 other alpha-proteobacterial species. Motifs were predicted upstream of 2,044 R. palustris genes and operons and clustered using the Bayesian Motif Clusterer (BMC). Analysis of the resulting 101 motif clusters produced a number of significant findings, including: (i) the PpsR regulon, which controls the expression of many genes of the photosynthetic apparatus; (ii) the FlbD regulon, which controls flagellar synthesis; (iii) four nitrogen regulons (FixK2, NnrR, NtrC, Sigma54), representing an important first step in understanding nitrogen fixation in this species; (iv) a predicted cobalamine riboswitch; and (v) a organic hydroperoxide resistance regulon. Complete results are available at http://bayesweb.wadsworth.org/prokreg.html. We have also collaborated with the Center for Molecular and Cellular Systems (http://mippi.ornl.gov/) to identify motifs
upstream of genes encoding *R. palustris* proteins that interact to form complexes *in vivo*, thereby delineating the transcriptionally co-regulated genes (unpublished, with investigators of the CMCS). The PpsR motif was identified upstream of several photosynthesis and oxidation/reduction proteins that interact, and a novel motif was identified upstream of several interacting chemotaxis-related proteins.

**B. Shewanella oneidensis**

In collaboration with the Shewanella Federation, we examined the genome of *S. oneidensis* MR-1 for repetitive elements, transposons and pseudogenes, the delineation of which has improved the gene calls for hundreds of genes\(^1\), and which will facilitate regulatory predictions by removing these repetitive sequences from the intergenic regions. These efforts, and the availability of several more *Shewanella* genomes, provide an exceptional data set for regulatory motif and regulon prediction. We have generated the orthologous promoter data sets of 17 *Shewanella* species and, using the Bayesian credibility limits\(^8\), each data set is being analyzed to determine the appropriate clades of alignable sequences for analysis by the phylogenetic Gibbs sampler\(^1\). This work is on-going. Also, in collaboration with the Shewanella Federation, we have analyzed microarray expression data sets. Identification of the regulatory sites in co-expression data delineates the directly co-regulated genes (regulons) from those genes subject to secondary regulatory effects. In an analysis of 712 genes with altered expression profiles in a iron regulatory protein (Fur) mutant, we identified 56 genes that are likely directly regulated by Fur, by identifying the Fur regulatory motif in their upstream intergenic regions\(^1\). In addition, we identified 73 genes likely co-regulated by the regulatory protein EtrA, among 610 genes with altered expression in an EtrA mutant (unpublished, with J. Tiedje’s lab).

**III. Resource sharing and web site development supported by DE-FG02-04ER63942**

All software developed and analyses of genomic data performed under this grant are available through web interfaces at [http://ccmbweb.ccv.brown.edu/](http://ccmbweb.ccv.brown.edu/) and [http://www.wadsworth.org/resnres/bioinfo/](http://www.wadsworth.org/resnres/bioinfo/). Software tools include: Gibbs Sampler\(^1\), \(^1\)\(^6\), PhylScan\(^2\), and software for microbial species comparisons\(^6\). The websites allow the user to analyze sequence data, and provide access to user manuals, as well as online tutorials for phylogenetic footprinting and the analysis of prokaryotic co-expression data. Usage results are available for the Gibbs Sampler, which has processed over 4700 data sets. The open source Gibbs Sampler software is also distributed under GNU General Public License version 2 ([http://www.gnu.org/copyleft/gpl.html](http://www.gnu.org/copyleft/gpl.html)). Since June 1, 2007, 1130 copies of the standalone versions of the Gibbs Sampler have been downloaded. This software has been used extensively in the identification of cis-regulatory signals\(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\), \(^1\).

**IV. Students and Post-docs trained in part under funding from DE-FG02-04ER63942**

**Sean Conlan**, Post-doc, Wadsworth Center. Dr. Conlan was a post-doctoral fellow at the Wadsworth Center with Drs. Lawrence and McCue prior to their moves (Dr. Lawrence to Brown University in mid-2004 and Dr. McCue to Pacific Northwest National Laboratory in late-2005). Dr. Conlan remained at the Wadsworth Center through early 2006, continuing to work remotely with Drs. Lawrence and McCue. Dr. Conlan was primarily responsible for our work with *Rhodopsseudomonas palustris*\(^4\), was instrumental to our work developing the phylogenetic Gibbs centroid sampler\(^1\), and contributed significantly to several publications aimed at describing the use of our software tools\(^6\), \(^1\), \(^1\).
Luis Carvalho, Graduate Student, Applied Math, Brown University. Mr. Carvalho is a Ph.D. student at Brown University, currently working with Dr. Lawrence on the properties of centroid estimators in discrete high-dimensional spaces.

Thomas Smith, Graduate Student, Computer Science, Rensselaer Polytechnic Institute. Dr. Smith was a graduate student at RPI (graduated spring 2005), where Dr. Lawrence had a joint appointment during his time at the Wadsworth Center. Dr. Smith contributed to the development of the phylogenetic Gibbs centroid sampler.

V. References


VI. Publications supported wholly or in part by DE-FG02-04ER63942

A. Research papers, technology development (in chronological order):


This paper analyzes the added statistical power that is obtainable from multi-species sequence data sets. It quantifies the extent to which additional genomes will be useful when attempting to locate and characterize transcription factor binding sites.


This paper describes the development of computational methods to connect transcription factors and DNA motifs, using *Escherichia coli* as a model system. Our method uses three types of mutually independent information that are combined to calculate the probability of a given transcription-factor-DNA-motif pair being a true pair.


This paper evaluates the use of sequence weights in inferences drawn from multiple evolutionarily related sequences.

This paper describes, and provides a web address for easy download of, the tools that we use for calculating percent identities of promoter regions between species. These tools aid in the identification of species appropriate and useful for cross-species promoter analysis.


This paper describes the scanning algorithm, PhyloScan, that searches a genome-scale database with a position weight matrix by combining evidence from matching sites found in orthologous data from several related species with evidence from multiple sites within an intergenic region to increase the statistical power of regulon prediction.


http://dx.doi.org/10.1093/bioinformatics/btm241.

This paper describes a version of the Gibbs recursive sampler that incorporates the phylogeny of the input sequences through the use of an evolutionary model and calculates an ensemble centroid motif solution. Using simulated data, we show that false positive predictions, caused by correlation among the sequences, are dramatically reduced by these added features.


http://dx.doi.org/10.1073/pnas.0712329105.

This paper reports on a novel procedure, centroid estimation, to obtain point estimates in discrete high-D spaces. Properties of these estimators are identified in four theorems, and evidence of improvements in the prediction of ground truth standards using these estimators compared to very popular highest scoring procedures is summarized.


http://dx.doi.org/10.1371/journal.pcbi.1000077.

This paper describes the development of Bayesian credibility limits to describe the uncertainty associated with high-dimensional inference problems, with a specific application to pairwise sequence alignment. We showed that credibility limits of the alignments of promoter sequences of 125 orthologous sequence pairs from six *Shewanella* species vary widely, and that centroid alignments dependably have tighter credibility limits than traditional maximum similarity alignments.

### B. Research papers & book chapters, resource sharing (in chronological order):


http://dx.doi.org/10.1002/0471250953.bi0208s10.
This book chapter describes the basic operation of the web interface to Gibbs and advanced examples of its use for locating transcription factor binding sites in unaligned DNA sequences.


http://dx.doi.org/10.1093/nar/gkm265. This paper, in the Web Server issue of *Nucleic Acids Research*, describes the advanced features of the web interface to Gibbs that allow identification of the centroid solution.

**C. Research papers, biology applications (in chronological order):**

http://dx.doi.org/10.1128/JB.186.24.8385-8400.2004. This paper is the result of collaboration with members of the Shewanella Federation to examine regulation by the transcription factor Fur in *Shewanella oneidensis* MR-1 by integrating DNA microarrays, proteomics, and promoter sequence analysis.

http://dx.doi.org/10.1073/pnas.0409111102. This paper was a joint effort by many members of the Shewanella Federation. The focus of this report was to annotate hypothetical ORFs in the *S. oneidensis* MR-1 genome.

http://dx.doi.org/10.1128/AEM.71.11.7442-7452.2005. This paper describes a phylogenetic footprinting study focused on the identification of regulons in *R. palustris*. A total of 4,963 regulatory motifs were predicted and clustered into 101 putative regulons.
This paper is the result of collaboration with members of the Shewanella Federation to examine the *Shewanella oneidensis* MR-1 genome for repetitive elements, transposons, and pseudogenes.