UCRL-TR-228471

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

De Novo Identification of Regulatory Regions in Intergenic Spaces of Prokaryotic Genomes

Patrick Chain, Emilio Garcia, Kevin Mcloughlin ,
Ivan Ovcharenko

February 28, 2007

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

FY06 LDRD Final Report

De Novo Identification of Regulatory Regions in Intergenic Spaces of Prokaryotic Genomes

LDRD Project Tracking Code: 04-ERD-103
Patrick Chain, Principal Investigator

Co-Investigators: Emilio Garcia (BIO), Kevin Mcloughlin (EEBI/BIO), Ivan Ovcharenko (EEBI/BIO)

Abstract

This project was begun to implement, test, and experimentally validate the results of a novel algorithm for genome-wide identification of candidate transcription-factor binding sites in prokaryotes. Most techniques used to identify regulatory regions rely on conservation between different genomes or have a predetermined sequence motif(s) to perform a genome-wide search. Therefore, such techniques cannot be used with new genome sequences, where information regarding such motifs has not yet been discovered. This project aimed to apply a *de novo* search algorithm to identify candidate binding-site motifs in intergenic regions of prokaryotic organisms, initially testing the available genomes of the *Yersinia* genus. We retrofitted existing nucleotide pattern-matching algorithms, analyzed the candidate sites identified by these algorithms as well as their target genes to screen for meaningful patterns.

Using properly annotated prokaryotic genomes, this project aimed to develop a set of procedures to identify candidate intergenic sites important for gene regulation. We planned to demonstrate this in *Yersinia pestis*, a model biodefense, Category A Select Agent pathogen, and then follow up with experimental evidence that these regions are indeed involved in regulation. The ability to quickly characterize transcription-factor binding sites will help lead to a better understanding of how known virulence pathways are modulated in biodefense-related organisms, and will help our understanding and exploration of regulons - gene regulatory networks - and novel pathways for metabolic processes in environmental microbes.

Motivation: Current techniques for identifying transcription factor binding sites in prokaryotic promoters are not suitable for whole genome discovery and are not accessible for use on available genomes. The proposed effort focused on implementing, testing and experimentally validating the results of integration and execution of procedures and computational algorithms for the genome-wide identification of candidate transcription factor binding sites in prokaryotic organisms. This would provide an important tool for genome, transcriptome, and systems biology. It would allow scientists to explore gene regulation to better understand regulatory networks in prokaryotic organisms, and to apply this pathway information to those microbes relevant to the Laboratory's environmental (microbe-based clean-up of metal-contaminated sites) and biodefense (understanding pathogen virulence mechanisms) missions.

Introduction/Background

The most prevalent technique used for identification of putative functional regulatory motifs has been comparative genomics, where collinear regions of two or more closely related genomes are aligned at the nucleotide level, and conserved regions are identified (7, 8, 16). The underlying theory is that conserved sequences in intergenic regions (usually proximal and upstream of genes) likely have a biological functional basis and may be involved in gene regulation. This method has been successfully applied to numerous eukaryotic systems, including fungi, plants, primates and other animals (4, 11). Most other techniques rely on prior knowledge (tested experimentally) of a given transcription factor's binding site motif, to then search for other copies of the motif in the entire genome (9, 10).

This project had as a goal to improve upon and combine a number of approaches such that they can be robustly applied to identify candidate, biologically active, transcription factor binding site motifs using a genome-wide approach. The main search algorithms are based on local alignment approaches that allow the preferential identification of sequences in promoter regions of genes that are under functional selective pressure. By studying the statistical distribution of evolutionary-conserved sequence motifs in the promoters of genes in comparison to the background (in the non-coding sequence of the genome), one can detect short motifs that are specific to a co-regulated set of genes (12, 13). Transcription factor binding site recognition correlated with detected motifs are an extremely powerful approach towards modeling gene regulatory networks.

We used as our target organism, *Yersinia pestis*, the causative agent of bubonic and pneumonic plague, a most devastating disease. It has been suggested that *Y. pestis* is a very recently evolved clone of *Y. pseudotuberculosis*, which causes a much less severe gastroenteritis (1, 2). Like *Escherichia coli*, *Yersinia* are gram negative Gammaproteobacteria, which has already facilitated the use of genetics to target genes of interest for mutagenesis experiments (Lao et al., manuscript in preparation).

Our original plan was to rigorously evaluate the success of the binding site motif predictions in identifying functional regulatory elements by creating defined motif mutants and looking at differences in expression of reporter gene constructs. Two types of knockout mutants were to be designed, those harboring deletions of the conserved motif, and those with nucleotide substitutions at the most conserved nucleotides found within the motif. Understanding the determinants of gene regulation would have a great impact in understanding how microbes develop, survive, adapt and interact with their environments (and other organisms). For example in *Y. pestis* and its relatively benign near-neighbor *Y. pseudotuberculosis*, differences in gene regulation may provide clues into *Y. pestis* pathogenicity, its life cycle and the evolution of plague. Our ultimate goal was to gain a better understanding of microbial genome regulons and regulatory networks by applying these methods to other prokaryotic systems, including those of interest to the DOE GTL-Genomics program, and to NIH biodefense targets. Unfortunately, due to funding constraints, the last half of this project was not undertaken. Here we report our progress in developing methods to identify candidate, conserved transcription factor binding sites, as applied to a Select Agent, *Yersinia pestis*. A *de novo* search algorithm for such sites would provide an important tool for genome, transcriptome, and systems biology studies.

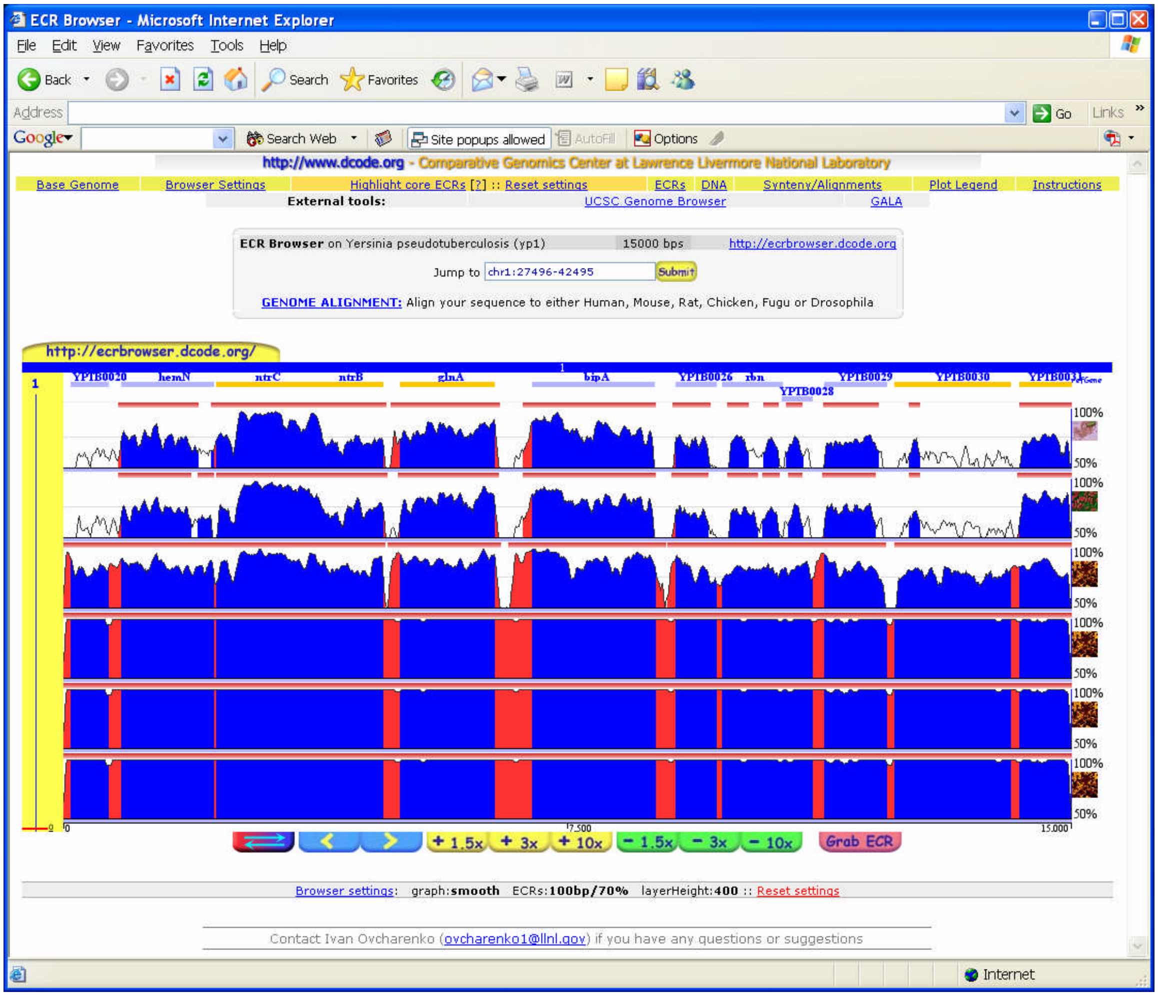
Research Activities and Results

As part of our project outline, we designed and tested a genomics/bioinformatics approach for genome-wide prediction of putative transcription factor binding sites in prokaryotic systems. We utilized *Yersinia pestis* as a model organism to study the application of transcription factor binding site-detecting alignment algorithms. Available to us at the time, due to our sequencing initiative, were seven complete genomes, four of which were published [the published *Y. pseudotuberculosis* strain IP32953 and *Y. pestis* strains CO92, KIM and 91001 (5, 6, 15, 17) as well as three other unpublished genomes at the time, *Y. pestis* strains Antiqua, Nepal516 and PestoidesF], their gene sets and expression microarray data (14). Combined, this allowed us to carefully map specific regulons that function in *Yersinia spp*., which could then be applied to a wide range of bacteria.

First, we collated all known genomic data for the Yersinia. We added the complete sequences of the recently sequenced *Yersinia pestis* strains Pestoides F, Antiqua and Nepal516 to perform annotation and alignments. We generated comparison profiles of the previously sequenced complete genomes of *Yersinia* using a version of the Basic Local Alignment Search Tool (3, 16), then integrated these into local data structures to create a web browser displaying both coding and non-coding sequence conservation (see Figure 1) using the reference genome of *Y. pseudotuberculosis* IP32953.

Multiple alignments of all available genomes for *Yersinia* confirmed that intergenic regions were conserved only among members of a species, with divergence seen between orthologous genes in different species of *Yersinia* (here, *Y. pseudotuberculosis* is treated as the same species as *Y. pestis* despite their different clinical etiologies; in fact, the two share greater than 95% sequence identity and genome-genome hybridization experiments both suggest that the two should be reclassified as the same species). The genomes of *E. coli* and *Salmonella* are very divergent from *Y. pestis/Y. pseudotuberculosis* and *Y. enterocolitica* varies specifically in the intergenic regions, suggesting changes in regulation of these genes or divergence of transcription factors as well as their binding sites.

Next, we analyzed *Y. pestis* microarray gene expression patterns as a function of temperature (26° vs 37°C) to establish sets of strongly up- and down-regulated genes and identified members of co-regulated gene sets in *Yersinia*. It is of utmost importance to be able to test and train the search algorithm on a set of genes known to be co-regulated upon exposure to certain conditions. A *Yersinia* DNA chip was constructed by LLNL and differential gene expression was measured from cells grown at 26 and 37 degree Celsius (14). Of the predicted ~4,500 genes in the *Yersinia* genome, 235 genes were upregulated and 274 were downregulated upon shift from 26 to 37°C (see Figure 2).



http://www.dcode.org - Comparative Genomics Center at Lawrence Livermore National Laboratory

ECR Browser on Yersinia pseudotuberculosis (yp1) 15000 bps http://ecrbrowser.dcode.org

Jump to chr1:27496-42495

GENOME ALIGNMENT: Align your sequence to either Human, Mouse, Rat, Chicken, Fugu or Drosophila

Browser settings: graph:**smooth** ECRs:**100bp/70%** layerHeight:**400** :: Reset settings

Contact Ivan Ovcharenko (ovcharenko1@llnl.gov) if you have any questions or suggestions

Figure 1. Screenshot of similarity across a multiple genome alignment. The reference genome of *Y. pseudotuberculosis* IP32953 (genes are below the blue horizontal bar, colored either yellow or light blue – representing top and bottom coding strand) was used for aligning the genomes of several bacterial genomes and all the *Yersinia* genomes. Displayed are *Salmonella typhi*, *Escherichia coli*, *Y. enterocolitica*, and *Y. pestis* strains 91001, CO92, and KIM10+ (from top to bottom). The red horizontal bars above each graph represent those sequences that are conserved with the reference genome. The X axis represents a 15 kb stretch of the reference genome; the Y axis represents the level of similarity (from 50-100%), broken down into the six genomes being compared. Red represents conserved intergenic regions, blue represents conserved coding regions and white represents less- or non-conserved sequences.

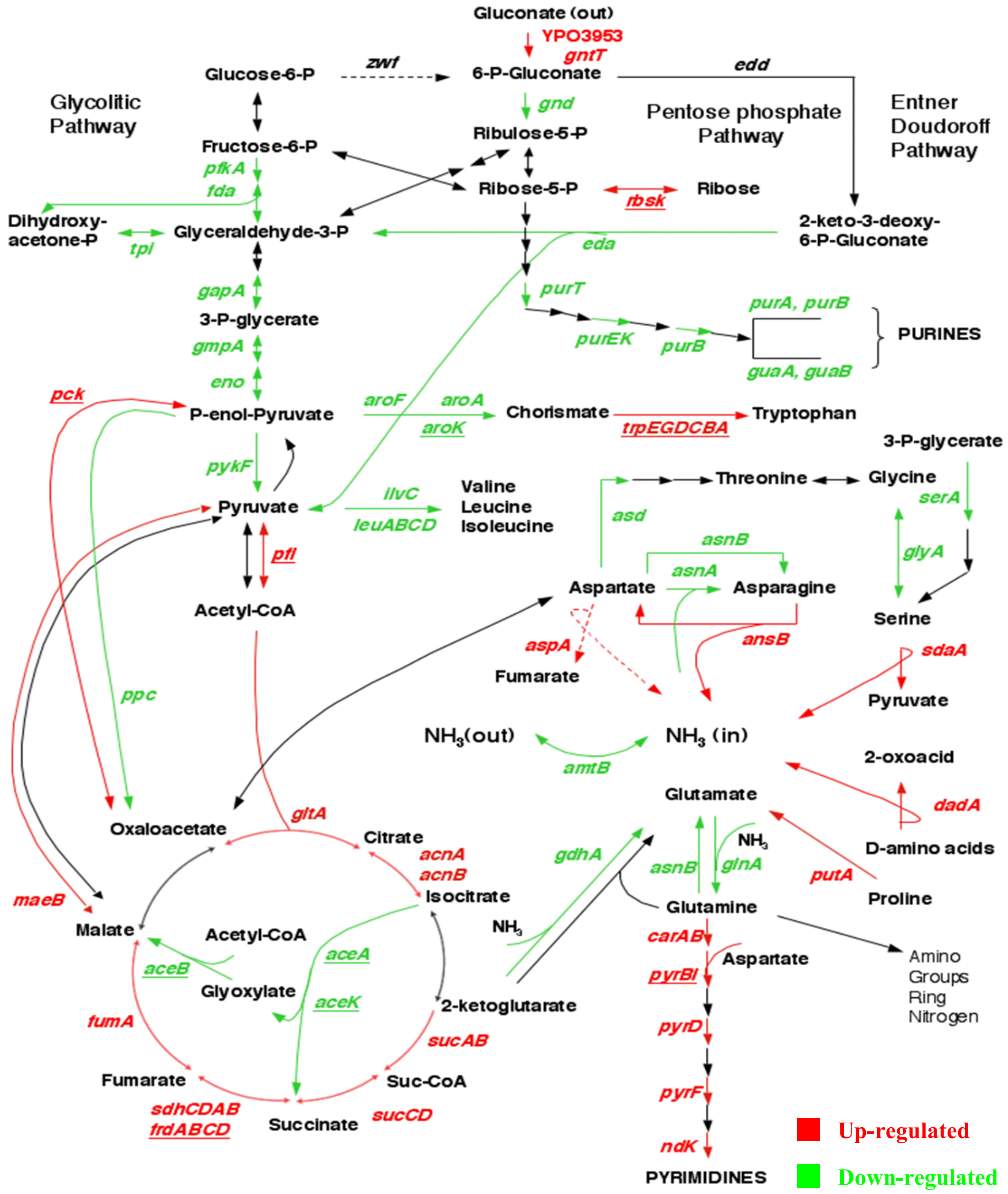
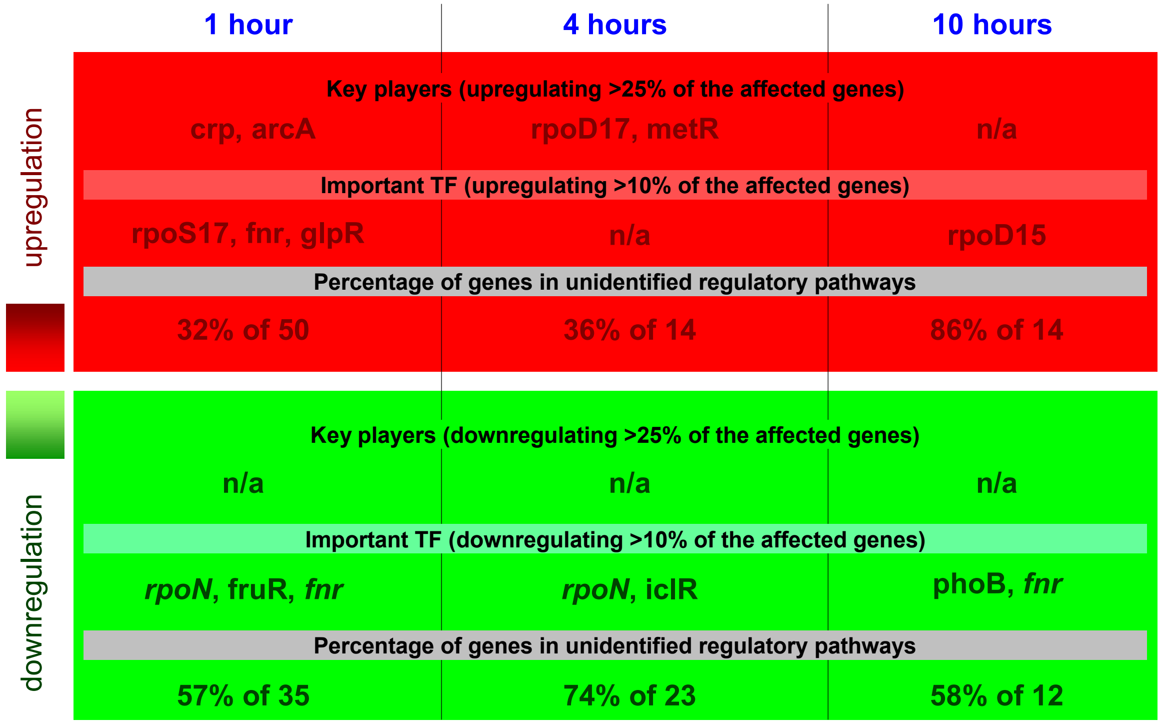


Figure 2. *Y. pestis* primary metabolic pathways affected by shift of growth temperature. The colors show the expression changes upon shift in growth temperature from 26 to 37°C. The red arrows and the green arrows represent the genes induced and repressed at 37°C, respectively. The gene names are colored in the same way as the arrows with the late-regulated gene names underlined. The dashed lines indicate enzymes known to be inactive in *Y. pestis* (i.e. genes *zwf* and *aspA*). The gene names are designated according to the annotated *Y. pestis* CO92 genome.

Utilizing curated gene annotations, as well as gene ontology (GO) and molecular interaction (KEGG) databases, we clustered these thermally-regulated gene datasets into categories of similar function and began to define conserved elements that corresponded to transcription factor binding sites. Through multiple iterations of clustering co-regulated genes and identifying putative conserved motifs, we refined our model to allow identification of average bacterial promoter architecture. We also created position weight matrices for putative transcription factor binding sites and used these to search for additional, previously missed motifs elsewhere in the genome to identify other members of the regulon, corroborating this with expression data. Modifications and multiple iterations of its implementation are expected to be necessary. We identified a number of statistically overrepresented sequence motifs in the regions upstream of thermally co-regulated *Y. pestis* chromosomal genes and have shown a strong correlation between known transcription factor binding sites and observed up- and down-regulated metabolic pathways in *Y. pestis*.

To complement these results, we began to explore whether binding site motifs and gene networks predicted in *E. coli*, another member of the Gammaproteobacteria Class, can be extrapolated to *Yersinia*. These putative and known functional regulatory sequence elements were used to search the *Yersinia* genomes for similar conserved motifs. We then evaluated the gene products of possibly co-regulated genes and their involvement in known metabolic pathways as well as correlation with the microarray data (see Figure 3). We identified a number of motifs in the promoters of *Yersinia* genomes that corresponded to the transcription factor binding sites of several global regulators in *E. coli*, such as Crp and ArcA. The genes downstream of these binding sites responded as one would expect if they were regulated by Crp and ArcA *Yersinia* homologs in the microarray experiments involving exposing *Y. pestis* strains to different temperatures (26 vs 37 degrees Celcius), which essentially mimicks a lifestyle change from a relatively aerobic environment for one that is microaerophilic or anoxic. Other regulators identified include RpoN (NtrA), Fnr, IclR and several others (see Figure 3). Encouragingly, both methods, the *de novo* identification and using known motifs from a relatively close relative genome, resulted in a largely overlapping set of putative transcription factor binding sites (data not shown) that correlate well with known expression profiles.

For future experimental validation of both the putative transcription factor binding sites as well as the regulators that are predicted to bind these sites and affect change, a set of vectors suitable for use in *Yersinia* were constructed, tested. We developed a specific set of vectors that were already used for generating targeted mutations in both *Y. pestis* and *Y. pseudotuberculosis* (Lao et al., manuscript in preparation). These constructs now have restricted applications due to regulations on the use of antibiotic resistance genes and cassettes in Select Agent pathogens, thus the development of a novel system would be required to carry out experimental analyses.



	1 hour	4 hours	10 hours
upregulation	Key players (upregulating >25% of the affected genes)		
	crp, arcA	rpoD17, metR	n/a
	Important TF (upregulating >10% of the affected genes)		
	rpoS17, fnr, glpR	n/a	rpoD15
	Percentage of genes in unidentified regulatory pathways		
🟥	32% of 50	36% of 14	86% of 14
🟩	Key players (downregulating >25% of the affected genes)		
downregulation	n/a	n/a	n/a
	Important TF (downregulating >10% of the affected genes)		
	rpoN, fruR, *fnr*	*rpoN*, iclR	phoB, *fnr*
	Percentage of genes in unidentified regulatory pathways		
	57% of 35	74% of 23	58% of 12

Figure 3. Computationally identified binding sites for global regulators in *Y. pestis* correlated with gene expression patterns over time in response to temperature shift.

Future Directions

There now exist involved methods to determine statistically overrepresented sequences given a set of global expression data, and there are procedures in place to identify sequences that are similar to known transcription factor binding sites. Future directions within this field should focus on expanding the set or database of known transcription factor binding sites to include putative binding sites derived from studying global expression, including those under the control of unknown regulators. This will be particularly problematic for pathogens with only distant relatives that have been sequenced and studied, such as *Francisella tularensis*. One of the future goals for work on transcription factor binding sites in pathogens is to identify possible virulence targets based on shared binding site motifs with known virulence factors. The confirmation of the regulatory function of a small number of predicted transcription factors and their identified binding sites in *in vivo* studies will be crucial to better understand the transcriptional landscape. The implementation of algorithms for *de novo* identification of putative transcription factor binding sites can, in principle, be applied to any sequenced microbial genome, thus opening up a very large sphere of research in both applied environmental and biodefense applications.

Summary

This project examined a *de novo* whole genome search technique to identify candidate binding site motifs in microbial systems, specifically *Yersinia pestis*. The ability to identify these binding sites is an important tool for genome, transcriptome, and systems biology. We have developed a number of computational programs to parse data generated by DNA microarrays, and to cluster genes into groups that have similar expression patterns. We have also combined alignment and motif searching tools to identify DNA regions that are statistically overrepresented among thermoregulated genes and thus may be involved in adapting to a new environment. We identified that motifs corresponding to the *E. coli* Crp and ArcA regulator binding sites, were present in promoters of many genes strongly upregulated 1 hour after *Y. pestis* is grown at human body temperature, while Fnr and RpoN appear to be involved in downregulating genes during this environmental switch, providing new insights and targets for future studies. This fits particularly well with the Crp, Fnr, and ArcA pathways, involving catabolite repression and aerobic respiration, since *Yersinia* is required to switch to a different carbon source and must adapt quickly to a microaerobic environment. Overall, the identification of regulons under transcription factor control allows scientists to explore gene regulation and better understand regulatory networks in microbial organism(s). This impacts the field of genomics, particularly in exploring and understanding global regulation and response to environmental cues, and the difficult translation from genome sequence to genome function.

References

1. **Achtman, M., G. Morelli, P. Zhu, T. Wirth, I. Diehl, B. Kusecek, A. J. Vogler, D. M. Wagner, C. J. Allender, W. R. Easterday, V. Chenal-Francisque, P. Worsham, N. R. Thomson, J. Parkhill, L. E. Lindler, E. Carniel, and P. Keim.** 2004. Microevolution and history of the plague bacillus, Yersinia pestis. Proc Natl Acad Sci U S A **101:**17837-42.
2. **Achtman, M., K. Zurth, G. Morelli, G. Torrea, A. Guiyoule, and E. Carniel.** 1999. Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. Proc Natl Acad Sci U S A **96:**14043-8.
3. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J Mol Biol **215:**403-10.
4. **Berman, B. P., Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen.** 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A **99:**757-62.
5. **Chain, P. S., E. Carniel, F. W. Larimer, J. Lamerdin, P. O. Stoutland, W. M. Regala, A. M. Georgescu, L. M. Vergez, M. L. Land, V. L. Motin, R. R. Brubaker, J. Fowler, J. Hinnebusch, M. Marceau, C. Medigue, M. Simonet, V. Chenal-Francisque, B. Souza, D. Dacheux, J. M. Elliott, A. Derbise, L. J. Hauser, and E. Garcia.** 2004. Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. Proc Natl Acad Sci U S A **101:**13826-31.
6. **Deng, W., V. Burland, G. Plunkett, 3rd, A. Boutin, G. F. Mayhew, P. Liss, N. T. Perna, D. J. Rose, B. Mau, S. Zhou, D. C. Schwartz, J. D. Fetherston, L. E. Lindler, R. R. Brubaker, G. V. Plano, S. C. Straley, K. A. McDonough, M. L. Nilles, J. S. Matson, F. R. Blattner, and R. D. Perry.** 2002. Genome sequence of Yersinia pestis KIM. J Bacteriol **184:**4601-11.
7. **Down, T. A., and T. J. Hubbard.** 2002. Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res **12:**458-61.
8. **Down, T. A., and T. J. Hubbard.** 2004. What can we learn from noncoding regions of similarity between genomes? BMC Bioinformatics **5:**131.
9. **Hughes, J. D., P. W. Estep, S. Tavazoie, and G. M. Church.** 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. J Mol Biol **296:**1205-14.
10. **Kreiman, G.** 2004. Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. Nucleic Acids Res **32:**2889-900.
11. **Lee, T. I., N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young.** 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science **298:**799-804.
12. **Loots, G. G., and I. Ovcharenko.** 2004. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. Nucleic Acids Res **32:**W217-21.

13. **Loots, G. G., I. Ovcharenko, L. Pachter, I. Dubchak, and E. M. Rubin.** 2002. rVista for comparative sequence-based discovery of functional transcription factor binding sites. Genome Res **12:**832-9.

14. **Motin, V. L., A. M. Georgescu, J. P. Fitch, P. P. Gu, D. O. Nelson, S. L. Mabery, J. B. Garnham, B. A. Sokhansanj, L. L. Ott, M. A. Coleman, J. M. Elliott, L. M. Kegelmeyer, A. J. Wyrobek, T. R. Slezak, R. R. Brubaker, and E. Garcia.** 2004. Temporal global changes in gene expression during temperature transition in Yersinia pestis. J Bacteriol **186:**6298-305.

15. **Parkhill, J., B. W. Wren, N. R. Thomson, R. W. Titball, M. T. Holden, M. B. Prentice, M. Sebaihia, K. D. James, C. Churcher, K. L. Mungall, S. Baker, D. Basham, S. D. Bentley, K. Brooks, A. M. Cerdeno-Tarraga, T. Chillingworth, A. Cronin, R. M. Davies, P. Davis, G. Dougan, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, A. V. Karlyshev, S. Leather, S. Moule, P. C. Oyston, M. Quail, K. Rutherford, M. Simmonds, J. Skelton, K. Stevens, S. Whitehead, and B. G. Barrell.** 2001. Genome sequence of Yersinia pestis, the causative agent of plague. Nature **413:**523-7.

16. **Schwartz, S., W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller.** 2003. Human-mouse alignments with BLASTZ. Genome Res **13:**103-7.

17. **Song, Y., Z. Tong, J. Wang, L. Wang, Z. Guo, Y. Han, J. Zhang, D. Pei, D. Zhou, H. Qin, X. Pang, J. Zhai, M. Li, B. Cui, Z. Qi, L. Jin, R. Dai, F. Chen, S. Li, C. Ye, Z. Du, W. Lin, J. Yu, H. Yang, P. Huang, and R. Yang.** 2004. Complete genome sequence of Yersinia pestis strain 91001, an isolate avirulent to humans. DNA Res **11:**179-97.