Submitted to: RECOMB Workshop Dec 2-4, 2005, San Diego
Topic: modeling and recognition of regulatory elements

An Interactive Map of Regulatory Networks of Pseudomonas aeruginosa Genome

Weihui Wu[1], Yongling Song[2], Shouguang Jin[1], Su-Shing Chen[2]
[1]Department of Molecular Genetics and Microbiology
[2]Department of Computer Information Science and Engineering
University of Florida
Gainesville, FL 32611

Abstract

For studying gene regulatory and protein signaling networks, we have developed an interactive map for the *Pseudomonas aeruginosa* genome. We first represent genes, proteins and their regulatory networks in a relational database. Known regulatory networks of the genome in the PubMed literatures are extracted by a manual and later a semi-automated text-mining method. Then a graphical interface displays these networks upon the query of specific genes, proteins or subsystems (i.e., groups of genes or proteins) on these networks. The interactive map has another capability of browsing those networks. The method can be extended to any other genome. Our objective is to develop this interactive map for the *Pseudomonas aeruginosa* community so that new research results may be ingested into the database, while annotations may be developed incrementally on the existing regulatory elements. Eventually some standards might be necessary for a long-term modeling and compilation of regulatory networks.

Bibliography

Weihui Wu is a graduate student of MGM and Yongling Song is a graduate student of CISE. Shouguang Jin is a professor of MGM with PhD from the University of Washington, and Su-Shing Chen is a professor of CISE with PhD from the University of Maryland. They are also members of the Systems Biology Lab at the UF Genetics Institute.

Introduction

With the exponential growth of available genomic sequences, various bioinformatics tools became available for comparative genomic analysis to annotate gene functions and build metabolic pathways (4, 7). However, establishing a functional regulatory network still relies largely on experimental approaches. Understanding the relationships among various biological "subsystems" enable us to understand the real biology at organism level, making bioengineering as well as drug discovery easier. *Pseudomonas aeruginosa* is an environmental bacterium, which causes serious human infections, especially those with reduced immunity, patients with Cystic Fibrosis or severe burns (2, 9). A large number of virulence genes and regulatory genes encoded by this organism make this bacterium one of the most successful pathogen on the earth. A complicated regulatory network coordinates the expression of various virulence genes as well as different functional groups of genes for an efficient host infection and survival in hostile host environments (5, 13). Prolonged treatments with antibiotics often result in multi-drug resistant isolates, which eventually cause death in the infected individuals (11). Therefore, there is an urgent need to develop new antimicrobial strategies for an effective control of this deadly bacterium.

Through decades of active research, tremendous amount of experimental data are available on the gene function and their regulation in *P. aeruginosa* (3, 8). However, these information are embedded in tens of thousands published literatures, thus difficult for individual researchers to extract the information for a comprehensive view of the field. Also, as more and more new research data become available, it is difficult for individual scientists to keep up with all the new information. There is a need to develop an interactive database which not only compiles experimental evidences but also logically integrate the knowledge related to gene function and regulation in *P. aeruginosa*.

The whole genome sequence of this microorganism has been completed several years ago and freely available to the public (12)[1]. The complete sequence of the genome was the largest bacterial genome sequenced to date when published, with 6.3-Mbp in size encoding 5570 predicted genes (12). Functions of only 480 of those encoded proteins have been demonstrated experimentally while the rest, including 1059 where functions of strongly homologous genes have been demonstrated experimentally in other organisms, 1524 whose functions were proposed based on the presence of conserved amino acid motif, structural feature or limited homology, and 2507 which are homologs of previously reported genes of unknown function, or no homology to any previously reported sequences. Most interestingly, as high as 8% of the genome encodes transcriptional regulators, which is consistent with the observed bacterial adaptability to various growth environments through alteration of gene expression pattern (6, 12).

The Knowledge Base of *P. aeruginosa*

[1] http://www.pseudomonas.com/AnnotationByPAU.asp?PA=PA1077

In the current project in progress, we intend to collect phenotypes of *P. aeruginosa* mutants and construct an interactive map of regulatory networks based on published literatures. A regulatory network will show relationships among genes, operons, regulons and stimulons. This regulatory network map will help researchers have a global view on the function of one gene and the relationship among several regulatory elements, facilitating the acquisition of relative information and design of future experiments. The database will be searchable by gene names or PA numbers, which have links to provide the following information:

(1) Mutant phenotypes, which include genotypes and phenotypes of the mutants and the parent strains as well as published references[2].

(2) Gene regulation at each subsystem level.

(3) Relationship among subsystems.

A regulatory network will include the genes if they belong to certain regulons/stimulons and related signals for activation/repression. Regulation at the transcriptional level (activation or repression), posttranscriptional (mRNA stability), translational or post-translational (protein-protein interaction/modification) will be indicated. As a reference for data reliability, the evidence for regulation – whether it was based on genetic evidence, biochemical tests or sequence-based prediction - will also be included. Using functional subsystem as a unit, the relative position of each gene in the regulatory cascade will be placed in the map. Also, placing regulatory genes in center, their regulatory role on various subsystems will also be marked.

In order to achieve the objectives, we have extracted information from 150 published papers in a systematic manner, which can be automated or semi-automated by natural language processing techniques. Phenotypes of all the mutants and references have been recorded. A database of gene regulatory networks has been developed, which is the basis for the computer-generated interactive maps. In this work, we have collected the virulence genes relevant to human infections, including the type IV pili subsystem involved in adhesion and twitching motility (10) and flagellar subsystem for the bacterial motility (1). Each subsystem involves tens of genes and is tightly and coordinately regulated. In existing literatures, some gene regulatory networks have been drawn manually, such as Flagella (1). However our database provides not only detailed gene relationships within subsystems, but also relationships between different subsystems so that users will have a global view of gene regulatory networks, and find specific potential connections between subsystems.

Upon completion, this interactive database will be released to the public. The interactive system may be further improved and updated by the research community. Our long-term goals are to build a database on the gene regulation for a comprehensive view of the regulatory networks at the genomic level; to automate the data extraction from published literatures; and to automate regulatory network building tools for various organisms based on this modeling methodology.

---

[2] PubMed and the University of North Carolina Microarray Data Bank

Modeling of Regulatory Elements

The modeling consists of a database and a graphical tool. The graphical tool will be interoperable with the web services so that users can search and visualize during any session. The database stores all necessary knowledge about regulatory networks. The database schema has 4 tables. The System Table presents subsystems and their relationships. The Gene Table includes all basic information about genes and proteins, which will be nodes in the regulatory networks, while subsystems may be either nodes or regulatory networks. The Edge Table describes the edge information, including the attributes: nodes, directions, types, relationships and subsystems. Directions are Active or Repressive - represented as positive and negative – and Be Activated and be Repressed. Types indicate DNA, RNA, and Protein Binding, Signal Molecule Production, Signal Sensing, and Signal/Molecule Binding. The Gene Information Table includes all other information about genes, such as references, genotypes, strains, phenotypes, and comments. The following figure illustrates a very small portion of the graph. Various annotations and symbols are introduced to model the regulatory networks. The other two figures – Figures 3 and 4 – can be zoomed out to above 400% to visualize more details of regulatory networks.

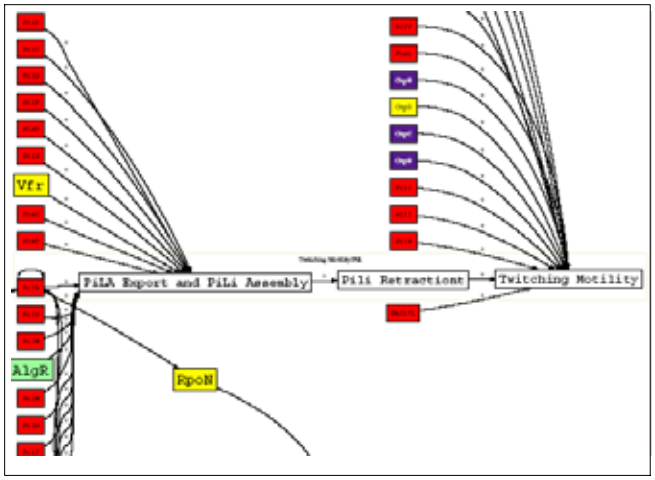

Figure 1. A Glimpse of the Pili subsystem

The system architecture includes three parts: Data Collection and Management System (DCM), Dynamic Graph Generation System (DGG) and Web-based User Navigation System (WUN). The DCM system will collect the regulatory network knowledge from the PubMed database, extract and represent regular elements into a spreadsheet table. Then by using a parser program, we insert the data in the spreadsheet table into our Pseudomonas aeruginosa Database **(**PADB) server. When users access the WUN system, they can query some special requests, for example: one subsystem graph, the whole system graph, or other graphs based on previous search results. The WUN system sends users' requests to the WWW server. The WWW server sends users' requests to the DGG system. DGG will do three things: first, it will analyze the users' requests and generate some SQL commands based on users' requests. Then it will pass the SQL to the PADB

database server, the database server will execute the SQL commands and return the resulting data to the DGG system. DGG will analyze the data and generate text file with the DOT language format. Then DGG will call the Graphviz software, which will generate the graph visualization based on the DOT file and output the graph into a PDF file. Then DGG will pass the PDF file into WUN. Now a user can view the requested graph of a regulatory network. From any displayed regulatory network, a user can further click on specific genes or other elements to search for other regulatory networks. The system architecture is depicted in the following figure:

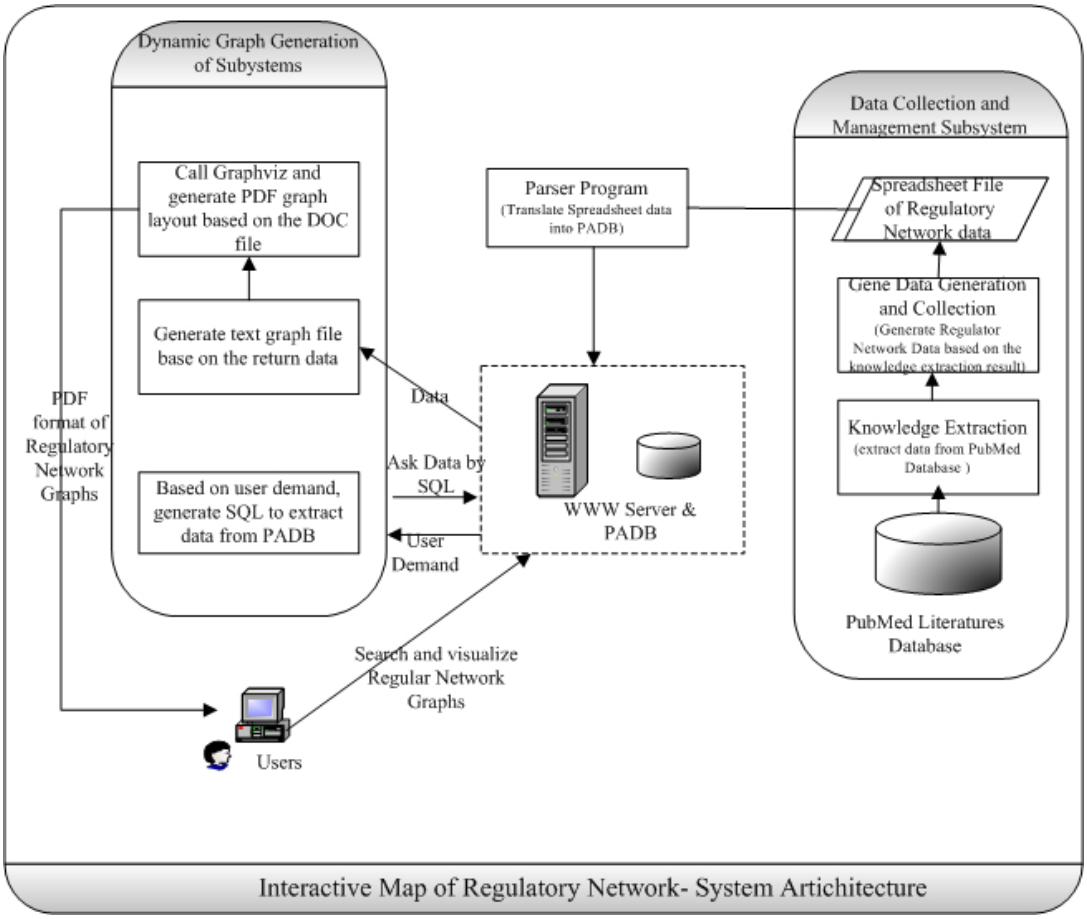

Figure 2. System Architecture of Interactive Map of Regulatory Networks

Automated Natural Language Processing for Information Extraction

We have developed natural language methods for text mining of MEDLINE or PubMed databases (16). For this work, we are developing automated natural language processing techniques to extract two kinds of information: phenotypes of mutants and relationships between genes from literatures in the PubMed database. The basic techniques are noun phrases extraction and listing of noun phrases of special headings. First, we need to find out how many mutants have been studied in a paper on Pseudomonas aeruginosa. Usually, there is a table in the paper showing all the strains used in the study and their genotypes. If such a table is not provided, we look for information about phenotypes of mutants in the papers, such as

"a mutation of A gene is constructed in one strain",
"a gene is knocked out in one stain",
"a gene mutant shows…".
"the X phenotype of A gene mutant is…".
"compared to wild type strain",
"we assayed the X phenotype in …".
After these phrases, we will find the genotypes usually. To extract information about relationships between genes or one gene to one subsystem, we look for information, such as
"*A gene actives the expression of B gene"* [3]
"A gene or protein+ description of relationship+ gene or subsystem"
The key in information extraction is that whenever one gene or protein is mentioned in the text, the description of its function or relationship with other gene/subsystem may be somewhere close.

Pseudomonas aeruginosa Subsystems and their Significance

So far, the interactive map of gene regulation networks of *Pseudomonas aeruginosa* contains eight subsystems: flagellum, pili, Type III ssecretion, Iron acquisition, quorum sensing, biofilm, alginate synthesis and multi-drug efflux. These subsystems compose the overall regulatory networks of *P. aeruginos.* Among them, there is interaction between genes in different subsystems. In this abstract, we only demonstrate the first two subsystems. As we are constructing the database, more subsystems will be extracted and integrated. The overview of these two subsystems are in the Figure3. Figure 4 gives the search result page when the user search for the gene "RpoN". The interactive map consists of several layers, the top view is the global view of all subsystems, while the lower layers display zoom-in and zoom-out a subsystem map.

Flagellum serves as a motive organelle on the surface of bacterium. The flagellum consists of basal body, hook, flagellar filament and motor. The basal body anchors the flagellum on the surface of the bacterium; the hook functions as a joint, connecting the filament to the basal body. The filament functions as a propeller and the motor generates the rotation of the flagellum. By rotating flagellum, bacteria can move in the surrounding environment. Two types of movement depend on flagella, swimming and swarming. Swimming is a movement of bacteria in the surround liquid and swarming is a surface translocation by groups of bacteria. Besides flagellum, *P. aeruginosa* produces another motive organelle named type IV pilus. Pilus is a polar filament structure, mediating attachment to host epithelial cells and one type of surface translocation named twitching motility. The pilus is composed of a small subunit (pilin). Pilin is synthesized in the cytoplasm as pre-pilin and translocated through inner membrane, cell wall and outer membrane to the surface of bacterium. During translocation, pre-pilin is cleaved to pilin.

[3] In the regulatory network, this means an arrow from gene A toward gene B.

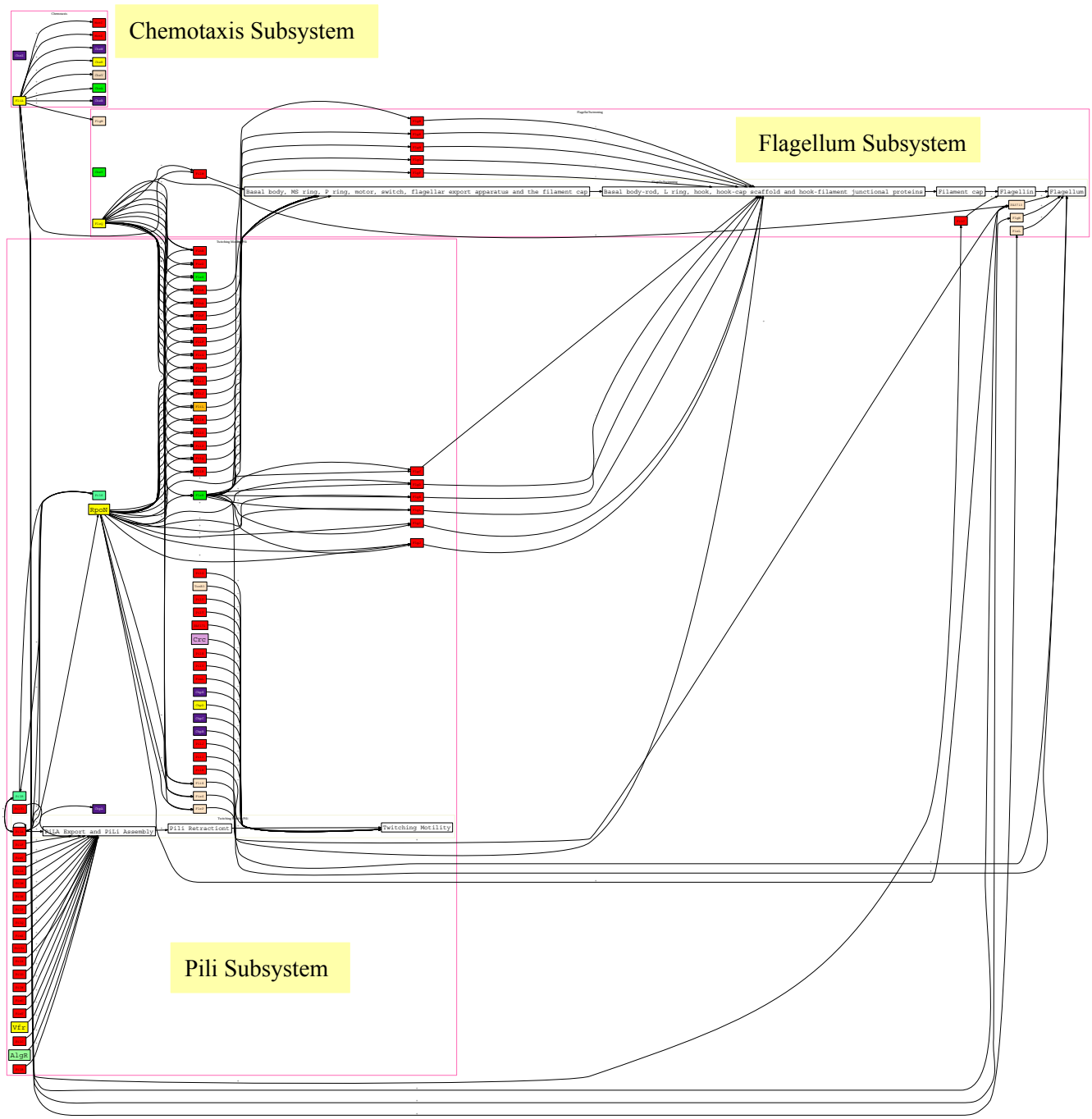

Fig 3 Three Subsystems in Regulatory Networks of Pseudomonas Aeruginosa Genome

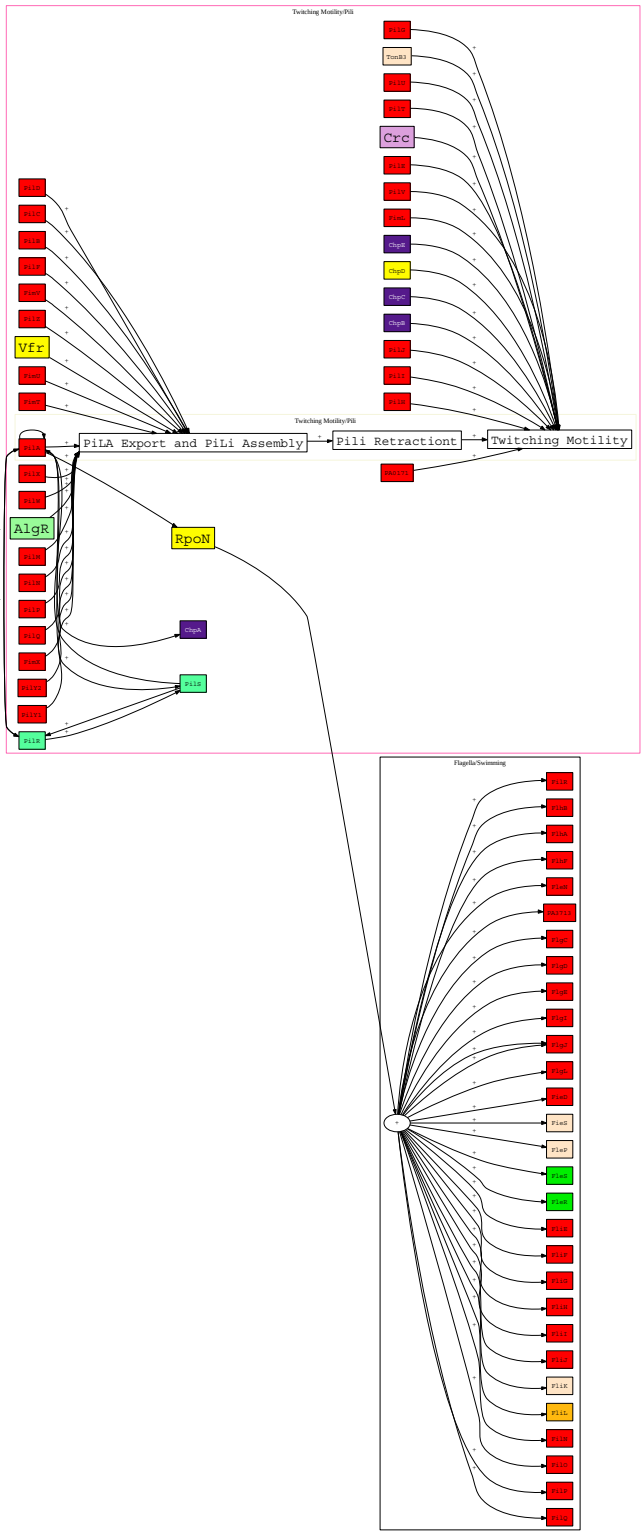

Figure 4 The search result of Gene "Rpon" in the Regulator Networks

The pilus is able to extend and retract, resulting in the surface translocation (twitching motility). The Type III secretion system (TTSS) is a potent virulence factor possessed by *P. aeruginosa*. The TTSS contains a syringe like apparatus, which can directly inject effector proteins from bacterium cytoplasm into host cell cytosol, causing cell death. Four effector proteins have been found in *P. aeruginosa*, ExoS, ExoT, ExoY, ExoU. Expression and secretion of the TTSS regulon can be stimulated by a direct contact with the host cell or by growth under low Ca^{2+} environment. Iron is essential for the metabolism and survival of *P. aeruginosa*. To acquire iron from surrounding environment *P. aeruginosa* produces and secrets iron-chelating compound, named siderophore. Two types of siderophores, pyoverdine and pyochelin, are produced by *P. aeruginosa*. The pyoverdine and pyochelin synthesis genes and receptors are under the negative control of a regulator, Fur and under iron deplete environments, the expression of these genes are derepressed. *P. aeruginosa* involves a signaling system for cell-cell communication, named quorum sensing. *P. aeruginosa* possesses three quorum sensing systems, known as las, rhl and PQS (pseudomonas quinolone signal). Each system contains a small molecule involved in signal communication. The las and rhl systems use acyl-homoserine lactones, C4-HSL and 3OC12-HSL as signal molecules, respectively. The signal molecule of PQS system is quinolone. The signal molecules are secreted into the surrounding environment and when their concentrations reach a threshold, they can interact with their own receptors and change the gene expression in other cells. The three quorum sensing systems can interact with each other. In *P. aeruginosa* many genes, including virulence genes are under the control of quorum sensing systems.

During chronic infection of cystic fibrosis (CF) patient's airway, *P. aeruginosa* forms biofilm. The biofilm is composed of bacteria microcolonies surrounded by exopolysaccharide. The formation of biofilm requires flagella, pili, exopolysaccharide and quorum sensing systems. In biofilm, *P. aeruginosa* is highly resistant to host immune system and antibiotics. Most *P. aeruginosa* clinical isolates from CF patients display a mucoid phenotype. The mocoidy is caused by over production of an exopolysaccharide, alginate. The production of alginate is repressed by an inner membrane protein: *MucA*. A high proportion of clinical isolates from CF patients has been found to contain mutation in *mucA* gene, resulted in the over production of alginate. *P. aeruginosa* is highly resistant to multiple antimicrobial agents. One mechanism of its intrinsic resistance is chromosomally encoded multi-drug efflux system. The multi-drug efflux system contains three components, an inner membrane drug transporter, a channel-forming outer membrane protein and a periplasm protein connecting the other two. The multi-drug efflux system can pump out antibiotics from cytoplasm and periplasm. *P. aeruginosa* possesses several multi-drug efflux systems, with different substrate specificity.

Conclusion

This work in progress builds an interactive map of regulatory networks for the Pseudomonas aeruginosa Genome. We have developed a semi-automated technique to

extract information about regulatory networks from the PubMed database. We are still working on fully automated methods developed in our earlier work (16).

REFERENCES

1. **Dasgupta, N., M. C. Wolfgang, A. L. Goodman, S. K. Arora, J. Jyot, S. Lory, and R. Ramphal.** 2003. A four-tiered transcriptional regulatory circuit controls flagellar biogenesis in *Pseudomonas aeruginosa*. Mol Microbiol **50:**809-24.
2. **Donaldson, S. H., and R. C. Boucher.** 2003. Update on pathogenesis of cystic fibrosis lung disease. Curr Opin Pulm Med **9:**486-91.
3. **Erwin, A. L., and D. R. VanDevanter.** 2002. The *Pseudomonas aeruginosa* genome: how do we use it to develop strategies for the treatment of patients with cystic fibrosis and *Pseudomonas infections*? Curr Opin Pulm Med **8:**547-51.
4. **Gibson, G., and E. Honeycutt.** 2002. The evolution of developmental regulatory pathways. Curr Opin Genet Dev **12:**695-700.
5. **Goodman, A. L., and S. Lory.** 2004. Analysis of regulatory networks in *Pseudomonas aeruginosa* by genomewide transcriptional profiling. Curr Opin Microbiol **7:**39-44.
6. **Greenberg, E. P.** 2000. Bacterial genomics. Pump up the versatility. Nature **406:**947-8.
7. **Lange, B. M., and M. Ghassemian.** 2005. Comprehensive post-genomic data analysis approaches integrating biochemical pathway maps. Phytochemistry **66:**413-51.
8. **Larbig, K., C. Kiewitz, and B. Tummler.** 2002. Pathogenicity islands and PAI-like structures in *Pseudomonas* species. Curr Top Microbiol Immunol **264:**201-11.
9. **Lyczak, J. B., C. L. Cannon, and G. B. Pier.** 2000. Establishment of *Pseudomonas aeruginosa* infection: lessons from a versatile opportunist. Microbes and Infection **2:**1051-1060.
10. **Mattick, J. S.** 2002. Type IV pili and twitching motility. Annu Rev Microbiol **56:**289-314.
11. **Rossolini, G. M., and E. Mantengoli.** 2005. Treatment and control of severe infections caused by multiresistant *Pseudomonas aeruginosa*. Clin Microbiol Infect **11 Suppl 4:**17-32.
12. **Stover, C. K., X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrener, M. J. Hickey, F. S. Brinkman, W. O. Hufnagle, D. J. Kowalik, M. Lagrou, R. L. Garber, L. Goltry, E.**
13. **Tolentino, S. Westbrock-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbig, R. Lim, K. Smith, D. Spencer, G. K. Wong, Z. Wu, I. T. Paulsen, J. Reizer, M. H. Saier, R. E. Hancock, S. Lory, and M. V. Olson.** 2000. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. Nature **406:**959-64.
14. **Woods, D. E.** 2004. Comparative genomic analysis of *Pseudomonas aeruginosa* virulence. Trends Microbiol **12:**437-9.
15. **Ellson, J., E.R Gansner, L. Koutsofios, S.C. North, and G. Woodhull,** 2003. Graphviz and Dynagraph – Static and Dynamic Graph Drawing Tools, chapter in Graph Drawing Software, M. Jugner and P. Mutzels, eds., Springer-Verlag.
16. **Kim, H. and Chen, S. ,** 2005. Ontology search and text mining of MEDLINE database, DATA MINING IN BIOMEDICINE, edited by P.M. Pardalos et al, Springer, to appear.