Research Statement

1. Research accomplishments

In the past four years, my research interests have been focused on computational inference and modeling of metabolic and regulatory networks in prokaryotic cells, and computational modeling and simulation of structures of protein/lipid complexes. My PhD thesis work was mainly on experimental studies of store-operated ion channels in mammalian cells using molecular biology and biophysical techniques.

1.1. Computational inference and modeling of regulatory networks in prokaryotes

In the past three years, I have been appointed as a group leader of a few postdocs with diverse background in Dr. Ying Xu's lab then at Oak Ridge National Laboratories and now at the Institute of Bioinformatics, University of Georgia. In collaboration with my colleagues, I have developed a computational framework (1) and a suite of software (2-7) for inference of regulatory networks in prokaryotes through mining various sources of high-throughout data including protein-protein interaction data, microarray gene expression data, and genome sequences. Using these tools, my colleagues and I have been able to build rather sophisticated network models for phosphorus assimilation (1), carbon fixation (8) and nitrogen assimilation networks (9) in cyanobacterium *Synechococcus* sp. WH8102. The major components of this computational framework are as follows:

A. Prediction of operons: Since genes involved in a pathway in prokaryotes are often encoded in an operon, the structure of operons/transcriptional units in a genome is therefore a valuable piece of information for pathway inference. In collaboration with my colleagues, we have developed a comparative genomics based algorithm/software for operon predictions called JPOP (Joint Prediction of Operons) (2, 3), which can be applied to any prokaryote genome. This work received the Best Paper Award at the Fifth International Conference on Genome Informatics (GIW) in 2004 held in Japan (2).

B. Mapping known pathways in well-studied organisms onto a target genome: The first step of the protocol to construct a pathway in a target genome (which is usually a newly sequenced and less well-studied genome) is to map onto it the corresponding pathways in some well-studied genomes to construct what we call an initial pathway model (1). Although pathway mapping is to some extent equivalent to orthology mapping, however, when the sequence similarity between two genes in two genomes is low, or there are multiple paralogs in both genomes, identification of orthologous relationships among these genes becomes non-trivial. To solve this problem, we have subsequently developed two algorithms (4, 5) that not only considers sequence similarity between genes, but also other information about gene functions as well, such as the predicted operon structures and shared *cis*-regulatory binding sites (for details see below). Using such a constrained mapping method, we are able to establish true orthologous relationships between genes when methods based only on sequence similarity fail (4).

C. Prediction of functional modules: In collaboration with my colleagues, we have developed a Bayesian network based algorithm/software for prediction of functional modules using various sources of information that suggests possible functional links between two genes in a genome, including GO (gene ontology) terms, phylogenetic profiles, and gene neighborhoods on the chromosome and microarray data if available (7).

D. Prediction of protein-protein interaction map: I have used orthologous mapping to known interaction pairs in well studied genomes and protein fusion event detections across the whole sequence space to predict pair-wise protein-protein interactions in a target genome (1, 9).

F. Prediction of *cis*-regulatory binding sites and regulons: Genes regulated by the same transcriptional factor (these genes are collectively called a regulon) are usually involved the same biological process, thus prediction of *cis*-regulatory binding sites and regulons is very important for pathway inference. I have recently developed a novel algorithm for prediction of *cis*-regulatory elements at a genome scale (6). This algorithm uses a comparative genomics approach, and identifies the binding site using a statistical

classifier. A very high prediction accuracy can be achieved with a false-positive rate being 40 orders of magnitude lower than conventional methods without suffering the loss of sensitivity (6). Using this algorithm, we have predicted (6) and verified (9) the molecular basis of the coupling between photosynthesis and nitrogen assimilation in a cyanobacterial cell.

G. Construction of regulatory pathways/networks through information integration: The final pathway/network model is constructed by recruiting additional proteins into the mapped initial model based on the aforementioned predictions of functional modules, protein-protein interactions and *cis*-regulatory binding sites/regulons (1, 8, 9). The models constructed by the protocol are in excellent agreement with whole genome microrray data (9).

1.2. Structure modeling and molecular dynamics simulation of human high density lipid (HDL) proteins----done with Dr. Stephen Harvey

As the major protein component of high density lipid protein (HDL) particles, apo A-I plays an important role in reducing cholesterol levels in the bloodstream. However the structure of the lipid binding form of apo A-I is largely unknown. Based on recent data that support a belt model for discoidal HDL particles, I have constructed a planar circular model of the lipid-binding domain of apo A-I (amino acid residues 41-234) (10), and conducted 10 ns molecular dynamics (MD) simulations on three distinct rotamers of the belt model to compare the roles of interhelical salt bridges *vs.* proline kinks in determining the registration of the apo A-I dimer. I found that both interhelical salt bridges and proline kinks play critical roles in stabilizing the protein/lipid complex. Since proline kinks occurr at a much faster time scale (<100 ps) than do relative rotations of the monomers, they exert strong constraints on the rotations of apo A-I after binding to the lipids. This suggests that the registration of the apo A-I dimer would occur well before or at least when apo A-I binds to the lipids. In addition, proline kinks reduce the area available to the lipids within the apo A-I ring that defines the perimeter of the HDL particle, therefore, proline kinkings also play a role in the conformational changes necessary for accommodating different lipid:protein ratios in HDL particles (Su et al, paper in preparation).

1.3. Experimental investigations of store-operated ion channels---PhD thesis related works

Capacitative Ca^{2+} entry (CCE) is an important biological phenomenon observed in practically all non-excitable cells and probably many types of excitable cells (11). It is widely believed that CCE is mediated by so-called store-operated ion channels (SOCs) that are activated upon depletion of internal Ca^{2+} stores (11). Experimental data has suggested that SOCs comprise an extended family whose members are distinguished by degree of Ca^{2+} selectivity, ranging from non-selective to among the most Ca^{2+} selective channel known. While some members in the TRP (transient receptor potential) subfamily have been shown to be store-operated, none of them has been unequivocally linked to a native SOC (12). Thus, the molecular composition of SOCs as well as the mechanism that links store depletion to channel activation is hitherto an intensive research area (12). My research on SOCs has been on the activation, regulation and molecular characterization of SOCs in immune cells.

1.3.1. Activation mechanism of CRAC channels. Ca^{2+} release activated Ca^{2+} (CRAC) channels that were first described in human T lymphocytes and rat mast cells are the most extensively studied SOC from a biophysical standpoint and serve as a prototypic SOC. The CRAC channel is distinguished from other SOCs by its extremely high Ca^{2+} selectivity and extremely small unitary conductance. I have previously demonstrated that a diffusible compound extracted from a mutant yeast with its endoplasmic reticulum Ca^{2+} ATPase pump genetically disrupted, could accelerate the activation of CRAC channels by removal of an inhibitory component homologous to the INAD protein of *Drosophila* (13, 14).

1.3.2. The mechanism of fast inactivation of Na^+ current through CRAC channels. In the absence of any divalent cations in the extracellular space, CRAC channels become permeable to monovalent cations such as Na^+ and K^+, with a larger conductance. However, the Na^+ current inactivates rapidly. This phenomenon is not seen in its equally Ca^{2+} selective counterpart, i.e, the voltage-gated Ca^{2+} channels (CaV). I have shown that Ca^{2+} disassociation from a site on the outer vestibule of CRAC channels is responsible for the inactivation of the Na^+ current, Ca^{2+} binding to this site otherwise potentiates the channel activities (15) by increasing its open probability (Su et. al, unpublished data).

1.3.3. Characterization of a native nonselective SOC in human T lymphocytes. Depletion of Ca^{2+} stores not only triggers Ca^{2+} influx but also Na^{+} influx in lymphocytes. I have characterized a non-selective cation channel in human lymphocytes using a combination of ion fluorescence imaging, patch clamp recording and RNA inactivation methods (16, 17). Interestingly, the channel can be activated by a few cellular manipulations including Ca^{2+}-store depletion. Store depletion induced Na^{+} influx through this channel was greatly attenuated by an anti-sense oligo RNA targeted to the LTRPC2 channel, suggesting that it might be a component of this SOC (17).

2. Research plans

In the future, I plan to use an integrated approach of computational and experimental methods to investigate the following biological problems through collaborations with experimentalists or setting up a small wet laboratory if it is possible. I believe that my rigorous training in computer science, computational biology/bioinformatics and experimental biology (cell biology, molecular biology, biophysics and physiology), has well prepared me to adapt such an interdisciplinary approach to tackle very challenging computational and biological problems.

2.1. **Development of novel algorithms/tools for genome-wide predictions of operons, regulons and metabolic and regulatory pathways**

As before, development of algorithms and tools for solving biological problems will be an important part of my research. In the near future, I plan to further improve the computational framework and tools that my colleagues and I have developed previously (Please see **Research accomplishments** section of this statement) by developing novel algorithms/tools for more accurate genome-wide predictions of operons, regulons and networks/pathways in any sequenced prokaryotic genomes using various sources of high throughput data.

2.1.1. Prediction of operons. Though a few algorithms for operon prediction have been published (18-22) including our own (2, 3), however they are either likely to be biased on the training data set (2, 3, 20) or dependent on third party data sources (2, 3, 18, 22), thus their performances are not guaranteed when applying to a remotely related genome, or they are not user-friendly. They also often suffer from the missing data problem, since the information required by the algorithms is not always available for each gene in the target genome. I plan to develop a user-friendly software for operon predictions using a Bayesian network based approach. Under Bayesian framework, missing data problem can be consistently dealt with, and different sources of information can be easily integrated, and thus high prediction accuracy are likely to be achieved.

2.1.2. Genome-wide prediction of *cis*-regulatory binding sites and regulons. So far, no general purpose algorithm/tool has been published for genome-wide predictions of *cis*-regulatory binding sites and regulons due to the high false positive rates of current motif prediction programs. I have recently developed a novel algorithm for prediction of *cis*-regulatory elements with very high prediction accuracy as mentioned before (6). This tool will be modified to predict the *cis*-regulatory binding sites at genome scales under the phylogenetic footprinting schema (23). Specifically, orthologues of the genes of a target genome will be identified in a group of closely related sequenced genomes using existing tools or a tool to be developed. Putative promoter regions are extracted for each group of orthologues genes based on the operon predictions by the tools developed in 2.1.1. Multiple putative motifs will be identified in each of these pooled sequences. These putative sites will be filtered by the modified algorithm to discriminate the true binding sites from the spurious ones. The final set of the predicted binding sites will be clustered using an algorithm yet to be developed, and each statistically significant cluster will constitute a regulon.

2.1.3 Prediction of metabolic and regulatory pathway. We have previously developed an algorithm P-MAP (4) for mapping well characterized metabolic pathways from one organism(template) to another(target). I plan to generalize this program so that it can handle more general cases of pathway mapping. For instance, the current program assumes that the proteins having the same functions in two genomes are orthologues, thus when the functions of a protein in the template is substituted by a protein of different evolution origin (a phenomenon called non-orthologous gene displacement), then the program

fails to find the "orthologue" in the target genome, and a hole is created in the mapped pathway. This problem will be fixed in the new program by considering the phylogenetic profiles (24) of genes, since non-orthologous gene displacement events can be detected by their inverted phylogenetic profiles of the relevant genes. In addition, although we have automated the most parts of the original protocol of pathway inference (1), the information integration step of the protocol still involves heavy manual works. I plan to develop tools to automate it. In the long run, I plan to extend the current computational pathway inference framework to the eukaryotic cells through mining various types of high throughput experimental data available in the public domain.

2.2. Rules that govern the evolution and adaptation of genomic structures in some specific groups of environmentally or medically important bacteria

One of my long term goals is to derive the rules that govern the evolution and adaptation of operons, regulons and pathways in a specific group of environmentally (e.g., cyanobacteria) or medically(e.g., mycobacteria or bacillus) important bacteria. To achieve this goal, the tools developed in 2.1 will be used to predict operons, regulons and networks/pathways in such a selected group of bacteria, of which a larger number of genome sequences of species, strains and ecotypes with diverse habitats are currently available or to be available soon. The prediction accuracies will be validated through collaborations with experimentalists as I have done previously (9) or by my own wet lab if setting up such a lab is possible. Comparative analyses of these functional units will relate the differences among them to their hosts' respective living environments [as I have done before for the NtcA regulons in cyanobacteria (6) and for the phosphonate assimilation pathways in a diverse group of bacteria (25)] or to specific parasite-host interactions. These differences, which may be associated with specific subgroups, could reflect features associated with fitness under different environmental conditions or parasite-host interactions and may also define points of evolutionary divergence. Such analyses should offer me clues about how evolutionary processes have altered genome organization; how well operon, regulon and pathway structures have been conserved; how operon, regulon and pathway structures have been modified among the organisms with different life style or habitats; which genes have been introduced into operons in specific subsets of the groups and at what point in evolutionary time; and how operon, regulon and pathway structures and the introduction of additional genes might reflect evolutionary changes in operon, regulon and pathway functions and the tailoring of the metabolism of the organism to a specific environmental niche or parasite-host interactions. One of the specific questions that I will address concerns the identification of niche-specific or parasite-specific 'survival' genes, operons, regulons, metabolic pathways and pathway modifications. In addition, selective use of operon/regulon and pathway information in other eubacteria may provide supporting data for questions and hypotheses generated from comparative analyses of a specific of group of genomes (e.g. are specific changes in operon structure/content associated with thermophily or parasitic life?)

2.3. Construction and modeling of transcriptional regulatory networks in prokaryotic and eukaryotic cells

Deciphering transcriptional regulatory networks encoded in either prokaryotic or eukaryotic genomes remains a very challenging task in both computational and experimental biology in the post genome era. Another goal I want to pursue in the near future is to develop a combined approach of computational predictions and experimental investigations (1) to map each *cis*-elements/regulons identified in 2.1.2 and 2.2 to its counterpart transcriptional factor, and (2) to construct and model transcriptional regulatory networks in a prokaryotic cell or a specific eukaryotic cell type. I plan to approach these aims by the following steps, which will be first tested on a prokaryote, eg. *Synechococcus* WH8102 or *E. coli* for their simpler transcriptional regulatory systems (e.g. only ~33 transcription factors are annotated in WH8102; and 314 in *E. coli*). After acquiring some experience, the techniques will be modified and moved to eukaryotic systems, e.g, *S. cerevisiae, D. melanogaster, C. elagans* or human cells.

2.3.1. Prediction of regulatory DNA binding proteins. Knowing the number and different types of regulatory DNA binding proteins in a genome will serve as a guide and constrain on transcriptional regulatory networks construction. All regulatory DNA binding proteins in a target genome (e.g., *synechococcus* WH8102) will be predicted using a novel algorithm that I will develop. Very likely, in addition to annotated transcription factors, a few new ones will be predicted by the new method because I plan to develop a more sensitive algorithm.

2.3.2. Prediction of *cis*-regulatory elements and regulons at genome scale. *cis*-regulatory binding sites and regulons in the target genome will be predicted using the tools developed in 2.1.2. The goal of this step is to uncover as many as possible true binding sites while keeping a low false positive rate of predictions. Each predicted binding site will be represented by a probability profile.

2.3.3. Mapping each set of predicted *cis*-regulatory elements/regulon to its counterpart DNA binding protein. I will develop a novel algorithm to map each predicted *cis*-regulatory element/regulons to its counterpart regulatory protein based on the observation that a regulator and the genes in its regulon are often coevolved, form clusters on the chromosome or have relatively short distances (26), and have either similar (activator) or inverted(repressor) expression profiles. All or some of the predictions will be verified by tandem oligonucleotide affinity column chromatography(27) followed by mass spectrometry identification of retarded protein (for prokaryotic cells) or by Chip-chip experiments (28) (for eukaryotic cells). An oligonucleotide sequence for an affinity column will be designed to have the minimal free binding energy according to its profile predicted by the 2.3.2, such that it will have the strongest binding affinity to its putative counterpart regulator.

2.3.4. Construction and refinement of genome-wide transcriptional regulatory networks. An initial wiring diagram of a gene regulatory network will be constructed based on the predicted *cis*-regulatory elements, and the predicted and experimentally validated pairs of *cis*-regulatory elements and their counterpart transcription factors. However, this diagram is likely to be incomplete, and spurious connections probably also exist. Thus, the model will be refined by time series microarray data or the other high throughput data using Bayesian network analysis (29). Missed connections will be added to the model whereas spurious ones will be removed during the analysis. Since the most parts of the initial model of the regulatory networks are expected to be correct, the search space of the Bayesian learning is greatly reduced. Hopefully, the prediction accuracy will be greatly improved. Therefore, this approach is likely to be superior to the *de novo* network inference that is solely based on microarray data.

Reference List

1. Su,Z., Dam,P., Chen,X., Olman,V., Jiang,T., Palenik,B. and Xu,Y. (2003) Computational inference of regulatory pathways in microbes: an application to phosphorus assimilation pathways in Synechococcus sp. WH8102. *Genome Inform Ser Workshop Genome Inform*, **14**, 3-13.
2. Chen,X., Su,Z., Xu,Y. and Jiang,T. (2004) Computational Prediction of Operons in Synechococcus sp. WH8102. *Genome Inform Ser Workshop Genome Inform*, **15**, 211-222.
3. Chen,X., Su,Z., Dam,P., Palenik,B., Xu,Y. and Jiang,T. (2004) Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome. *Nucleic Acids Res*, **32**, 2147-2157.
4. Mao,F., Su,Z., Olman Victor, Dam,P., Liu,Z. and Xu,Y. (2005) Mapping of Orthologous Genes in the Context of Biological Pathways: an Application of Integer Programming. *Proc Natl Acad Sci U S A*, **in press**.
5. Olman,V., Peng,H., Su,Z. and Xu Ying (2004) Mapping of Microbial Pathways through Constrained Mapping of Orthologous Genes. *Proceedings of IEEE Computational Systems Biology Conference*, 363-370.
6. Su,Z., Olman,V., Mao,F. and Xu,Y. (2005) Comparative genomics analysis of ntcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis. *Nucleic Acid Res*, **33**, 5156-5171.
7. Wu,H., Su,Z., Mao,F., Olman,V. and Xu,Y. (2005) Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Res*, **33**, 2822-2837.
8. Dam,P., Su,Z., .Olman,V. and Xu,Y. (2005) In silico reconstruction of the carbon fixation pathway in *Synechococcus sp.* WH8102. *J Biol Sys*, **12**, 97-125.
9. Su,Z., Mao,F., Dam,P., Palenik,B. and Xu,Y. (2005) Computational inference and experimental validation of nitrogen assimilation regulatory networks in cyanobacterium *Synechococcus sp.* WH8102. *Submitted.*

10. Catte,A., Patterson,J., Weinstein,G., Su,Z., Li,L., Chen,J., Jones,M., , Aliste, arcela *et al.* (2005) Novel Minimal Surface Conformations of Nascent High Density Lipoproteins through Molecular Dynamics. *Biophysical Journal*, **submitted**.
11. Putney,J.W., Jr. (2003) Capacitative calcium entry in the nervous system. *Cell Calcium*, **34**, 339-344.
12. Prakriya,M. and Lewis,R.S. (2003) CRAC channels: activation, permeation, and the search for a molecular identity. *Cell Calcium*, **33**, 311-321.
13. Csutora,P., Su,Z., Kim,H.Y., Bugrim,A., Cunningham,K.W., Nuccitelli,R., Keizer,J.E., Hanley,M.R., Blalock,J.E. and Marchase,R.B. (1999) Calcium influx factor is synthesized by yeast and mammalian cells depleted of organellar calcium stores. *Proc Natl Acad Sci U S A*, **96**, 121-126.
14. Su,Z., Barker,D.S., Csutora,P., Chang,T., Shoemaker,R.L., Marchase,R.B. and Blalock,J.E. (2002) Regulation of Ca^{2+} Release Activated Ca^{2+} Channels by INAD and Calcium Influx Factor. *Am J Physiol Cell Physiol*, **284**, C497-C505.
15. Su,Z., Shoemaker,R.L., Marchase,R.B. and Blalock,J.E. (2004) Ca2+ modulation of Ca2+ release-activated Ca2+ channels is responsible for the inactivation of its monovalent cation current. *Biophys J*, **86**, 805-814.
16. Su,Z., Csutora,P., Hunton,D., Shoemaker,R.L., Marchase,R.B. and Blalock,J.E. (2001) A store-operated nonselective cation channel in lymphocytes is activated directly by Ca^{2+} influx factor and diacylglycerol. *Am J Physiol Cell Physiol*, **280**, C1284-C1292.
17. Su,Z., Guo,X., Barker,D.S., Shoemaker,R.L., Marchase,R.B. and Blalock,J.E. (2005) A store-operated nonselective cation channel in human lymphocytes. *Cell Mol Neurobiol*, **25**, 625-647.
18. Jacob,E., Sasikumar,R. and Nair,K.N. (2005) A fuzzy guided genetic algorithm for operon prediction. *Bioinformatics*, **21**, 1403-1407.
19. Westover,B.P., Buhler,J.D., Sonnenburg,J.L. and Gordon,J.I. (2005) Operon prediction without a training set. *Bioinformatics*, **21**, 880-888.
20. De Hoon,M.J., Imoto,S., Kobayashi,K., Ogasawara,N. and Miyano,S. (2004) Predicting the operon structure of Bacillus subtilis using operon length, intergene distance, and gene expression information. *Pac Symp Biocomput*, 276-287.
21. Zheng,Y., Szustakowski,J.D., Fortnow,L., Roberts,R.J. and Kasif,S. (2002) Computational identification of operons in microbial genomes. *Genome Res*, **12**, 1221-1230.
22. Price,M.N., Huang,K.H., Alm,E.J. and Arkin,A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res*, **33**, 880-892.
23. McGuire,A.M., Hughes,J.D. and Church,G.M. (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res*, **10**, 744-757.
24. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, **96**, 4285-4288.
25. Huang,J., Su,Z. and Xu,Y. (2005) The evolution of Microbial phosphonate Degradation Pathways. *J Mol Evol*, **61**, 1-11.
26. Tan,K., McCue,L.A. and Stormo,G.D. (2005) Making connections between novel transcription factors and their DNA motifs. *Genome Res*, **15**, 312-320.
27. Kadonaga,J.T. and Tjian,R. (1986) Affinity purification of sequence-specific DNA binding proteins. *Proc Natl Acad Sci U S A*, **83**, 5889-5893.
28. Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306-2309.
29. Pe'er,D. (2005) Bayesian network analysis of signaling networks: primer. *Sci STKE*, pl4.

Teaching experiences and interests

I have extensive teaching experience in biological sciences (Biochemistry and Animal Physiology) at the undergraduate level before moving to United States. Here at the University of Georgia in the last two years, I have co-taught twice a graduate level course: Computational Methods in Bioinformatics (BCMB8210). In the future, I am interested in teaching both undergraduate and graduate level courses in both biological sciences and computational science. Specifically, I am interested in and capable of teaching the following courses:

1. Introductory or advanced course in bioinformatics emphasizing algorithms and applications;
2. Algorithms in computational biology emphasizing theory and algorithms;
3. Computational comparative genomics emphasizing genomic structures and genome evolution in prokaryotes or eukaryotes;
4. Introductory course in algorithm and data structure;
5. Advanced course in systems biology emphasizing technology development and regulatory network analysis and modeling;
6. Introductory or advanced course in molecular evolution;
7. Introductory or advanced course in microbiology;
8. Introductory or advanced course in cell biology or cellular physiology, or medical physiology.