Bringing Advanced Computational Techniques to Energy Research (BACTER)

Final Technical Report (DE-FG02-04ER25627)
PI: Julie C Mitchell, University of Wisconsin – Madison

This report briefly summarizes the achievements of the BACTER training program in computational biology, sponsored by the US Department of Energy's SciDAC and Genomics:GTL programs. We thank them for their support and generosity over the course of this interdisciplinary training initiative. We began in 2004 with a small group of dedicated graduate students and postdocs. Our goal was to enhance training of quantitative scientists in biological areas of relevance to the DOE mission, as well as to integrate biological scientists and cross-train them in areas of quantitative research.

Our early students have now matriculated and gone on to postdoctoral positions at Yale, UC Berkeley, UC Irvine, and Imperial College, while others have taken positions in local biotech firms here in Madison based on their positive experiences with the research climate here. Two former postdocs went into faculty jobs (University of California Merced and Northern Illnois University) while others went on for additional postdoctoral training (Purdue, Duke) and still others went on into industry.

Over the course of time, we have experimented with new training activities, including embedded science writers and scientific programmers. Our final training experiment was to train two computational and experimental students/postdocs to work in unison on new projects, to see whether this would accelerate transfer of knowledge and scientific discovery. As evidenced by the high-quality publications and career trajectories of our trainees, and the rapid progress of cross-training projects initiated during our final year, our efforts were very successful in understanding how to better advance the role of quantitative science in bioenergy research and to train the next generation of researchers to help make biofuels a viable energy source for to help meet the needs of the US population.

Trainee Activities

One of the innovative features of our program was the introduction of a science writer and programmer. The students found these resources invaluable to their training. The improvement in student writing as the result of working with our science writer was marked. Similarly, having a scientific programmer made high-level computation accessible to a range of students and helped move student prototype code to production, including a number of web servers that are actively used to design gene arrays, analyze protein-protein interfaces, and simulate metabolic networks. These activities not only strengthened the training of our students but also enhanced and accelerated dissemination of their research.

Throughout the course of BACTER, we have run a seminar and journal club. Newer students present papers on a topic of interest to them, while advanced students present their own research. A number of faculty have actively participated in these activities, lending feedback to students over the course of their research. In 2006, we also hosted a genomics workshop with a number of outside speakers. An agenda for this meeting is included with our report. In subsequent years, we hosted many visitors to the university to give special lectures.

Trainee Diversity

Over the course of the program, we have had a diverse pool of students and postdocs. Female students have been well-represented within our program, whereas their representation in highly

quantitative fields is a known challenge. Our recruiting pool for underrepresented and underprivileged students and postdocs remains small, but we are fortunate to see truly outstanding outcomes from these trainees. Saheed Imam is a student in Microbiology, expanding his already excellent quantitative skills through his affiliation with BACTER. He is making rapid progress toward his PhD work on genome-scale reconstruction of the regulatory network of *Rhodobacter*. Roummel Marcia was a BACTER postdoc working in the area of applied mathematics and optimization and was recently promoted to Associate Professor of Applied Mathematics at the University of California at Merced.

Trainee Research

The most important aspect of the BACTER training program has been the research it has facilitated. It is difficult to summarize the vast range of research that has been accomplished by our trainees over the course of 8 years. In the interest of being both succinct and comprehensive, and because many details have been provided in our annual reports, we've divided our summary into two sections. For trainee research that has already been completed and published, and described in prior year annual reports, we've included a complete list of publications.

In addition to this past work, we give detailed descriptions that are ongoing at the time of completion of this award. This includes the work of our current long-standing trainees as well as recent awardees during the no-cost extension period. Last year, we requested a one-year no-cost extension to our grant in order to spend down some surplus funds and engage in one final training experiment based on our observations of the course of the previous 7 years and the objectives of our training program in relation to the DOE mission.

The BACTER program always funded primarily computational students, some of whom occasionally engaged in a small amount of experimental work as a means to understand their data and its limitations. These constraints are often necessitated by funding streams at the agency level, and at the student level, it is difficult to cross-train people rigorously in both computation and experiment. At the same time, there is great power in being able to create an active feedback loop between computation and experiment. For this reason, we requested for our no-cost extension to be able to fund a number of one-year "startup" projects between experimental and computational groups, with varied goals that reflect the needs of energy science across many scales. This has been amazingly successful, inspiring a number of new campus projects that will persist beyond the life of the BACTER award.

To select projects, we advertised across the campus looking to inspire new collaborations between the biological and quantitative scientists. Some projects are at the genomic scale, looking at how to reengineer transcriptional regulation in bioenergy crops and microorganisms. Other projects are at a macroscale, applying the same types of network models and industrial engineering techniques toward optimizing energy production at the regional level. We've included brief project reports as of the end of grant term (August 31, 2012) for ongoing work to give a flavor of the range of activities and accomplishments, along with publications in progress or expected.

Trainee Publications

(Note: 15+ publications are in preparation as noted in project writeups, with added publications to result from projects completed as trainees matriculate in the coming few years.)

1. Dufour YS, Imam S, Koo B-M, Green HA, Donohue TJ (2012) Convergence of the transcriptional responses to heat shock and singlet oxygen stresses. PLoS genetics 8: e1002929. doi:10.1371/journal.pgen.1002929.

2. Demerdash ONA, Mitchell JC (2012) Density-cluster NMA: A new protein decomposition technique for coarse-grained normal mode analysis. Proteins 80: 1766–1779. doi:10.1002/prot.24072.

3. Sen SM, Binder JB, Raines RT, Maravelias CT (2012) Conversion of biomass to sugars via ionic liquid hydrolysis: process synthesis and economic evaluation. Biofuels, Bioprod Bioref 6: 444–452. doi:10.1002/bbb.1336.

4. Liang K, Keles S (2012) Detecting differential binding of transcription factors with ChIP-seq. Bioinformatics 28: 121–122. doi:10.1093/bioinformatics/btr605.

5. Correll M, Albers D, Franconeri S, Gleicher M (2012) Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12 New York, New York, USA: ACM Press. pp. 1095–1104. doi:10.1145/2207676.2208556.

6. Alfeld S, Barford C, Barford P (2012) Proceedings of the 3rd International Conference on Future Energy Systems Where Energy, Computing and Communication Meet - e-Energy '12 New York, New York, USA: ACM Press. pp. 1–4. doi:10.1145/2208828.2208837.

7. Liang K, Keles S (2012) Normalization of ChIP-seq data with control. BMC Bioinformatics 13: 199. doi:10.1186/1471-2105-13-199.

8. Albers D, Dewey C, Gleicher M (2011) Sequence Surveyor: leveraging overview for scalable genomic alignment visualization. IEEE Trans Vis Comput Graph 17: 2392–2401. doi:10.1109/TVCG.2011.232.

9. Fleishman SJ, Whitehead TA, Strauch E-M, Corn JE, Qin S, et al. (2011) Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol 414: 289–302. doi:10.1016/j.jmb.2011.09.031.

10. Gleicher M, Albers D, Walker R, Jusufi I, Hansen CD, et al. (2011) Visual comparison for information visualization. Information Visualization 10: 289–309. doi:10.1177/1473871611416549.

11. Yang S, Cui Q (2011) Glu-286 rotation and water wire reorientation are unlikely the gating elements for proton pumping in cytochrome C oxidase. Biophys J 101: 61–69. doi:10.1016/j.bpj.2011.05.004.

12. Simons JE, Milewski PA (2011) The volcano effect in bacterial chemotaxis. Mathematical and Computer Modelling: An International Journal 53.

13. Imam S, Yilmaz S, Sohmen U, Gorzalski AS, Reed JL, et al. (2011) iRsp1095: a genome-scale reconstruction of the Rhodobacter sphaeroides metabolic network. BMC Syst Biol 5: 116. doi:10.1186/1752-0509-5-116.

14. Kontur WS, Ziegelhoffer EC, Spero MA, Imam S, Noguera DR, et al. (2011) Pathways involved in reductant distribution during photobiological H2 production by Rhodobacter sphaeroides. Applied and Environmental Microbiology 77: 7425–7429.

15. Zhu X, Mitchell JC (2011) KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. Proteins 79: 1097–0134. doi:10.1002/prot.23094.

16. Baumler DJ, Peplinski RG, Reed JL, Glasner JD, Perna NT (2011) The evolution of metabolic networks of E. coli. BMC Syst Biol 5: 182. doi:10.1186/1752-0509-5-182.

17. Demerdash ONA, Buyan A, Mitchell JC (2010) ReplicOpter: a replicate optimizer for flexible docking. Proteins 78: 3156–3165. doi:10.1002/prot.22811.

18. Dufour YS, Kiley PJ, Donohue TJ (2010) Reconstruction of the core and extended regulons of global transcription factors. PLoS genetics 6: e1001027. doi:10.1371/journal.pgen.1001027.

19. Albers D, Gleicher M (2010) Perceptual principles for scalable sequence alignment visualization ACM. doi:10.1145/1836248.1836286.

20. Cipriano G, Wesenberg G, Grim T, Phillips GN, Gleicher M (2010) GRAPE: GRaphical Abstracted Protein Explorer. Nucl Acids Res 38: W595–W601. doi:10.1093/nar/gkq398.

21. Dufour YS, Wesenberg GE, Tritt AJ, Glasner JD, Perna NT, et al. (2010) chipD: a web tool to design oligonucleotide probes for high-density tiling arrays. Nucl Acids Res 38: W321–W325. doi:10.1093/nar/gkq517.

22. Wang Y, Dufour YS, Carlson HK, Donohue TJ, Marletta MA, et al. (2010) H-NOX-mediated nitric oxide sensing modulates symbiotic colonization by Vibrio fischeri. Proc Nat Acad Sci USA 107: 8375–8380. doi:10.1073/pnas.1003571107.

23. Zhu X, Koenig P, Hoffmann M, Yethiraj A, Cui Q (2010) Establishing effective simulation protocols for β- and α/β-peptides. III. Molecular mechanical model for acyclic β-amino acids. J Comp Chem: NA–NA. doi:10.1002/jcc.21493.

24. Demerdash ONA, Daily MD, Mitchell JC (2009) Structure-based predictive models for allosteric hot spots. PLoS Comput Biol 5: e1000531. doi:10.1371/journal.pcbi.1000531.

25. Dynerman D, Butzlaff E, Mitchell JC (2009) CUSA and CUDE: GPU-accelerated methods for estimating solvent accessible surface area and desolvation. J Comput Biol 16: 523–537. doi:10.1089/cmb.2008.0157.

26. Baumler DJ, Hung K-F, Jeong KC, Kaspar CW (2008) Molybdate treatment and sulfate starvation decrease ATP and DNA levels in Ferroplasma acidarmanu. Archaea 2: 205–209.

27. Dufour YS, Landick R, Donohue TJ (2008) Organization and evolution of the biological response to singlet oxygen stress. J Mol Biol 383: 713–730. doi:10.1016/j.jmb.2008.08.017.

28. Bannen RM, Suresh V, Phillips GN, Wright SJ, Mitchell JC (2008) Optimal design of thermally stable proteins. Bioinformatics 24: 2339–2343. doi:10.1093/bioinformatics/btn450.

29. Bae E, Bannen RM, Phillips GN (2008) Bioinformatic method for protein thermal stabilization by structural entropy optimization. Proc Nat Acad Sci USA 105: 9594–9597. doi:10.1073/pnas.0800938105.

30. Zhu X, Koenig P, Gellman SH, Yethiraj A, Cui Q (2008) Establishing Effective Simulation Protocols for β- and α/β-Peptides. II. Molecular Mechanical (MM) Model for a Cyclic β-Residue. J Phys Chem B 112: 5439–5448. doi:10.1021/jp077601y.

31. Riccardi D, König P, Guo H, Cui Q (2008) Proton transfer in carbonic anhydrase is controlled by electrostatics rather than the orientation of the acceptor. Biochemistry 47: 2369–2378. doi:10.1021/bi701950j.

32. Bitto E, Bingman CA, Kondrashov DA, McCoy JG, Bannen RM, et al. (2008) Structure and dynamics of gamma-SNAP: insight into flexibility of proteins from the SNAP family. Proteins 70: 93–104. doi:10.1002/prot.21468.

33. Jeong K, Hung K, Baumler DJ, Byrd JJ, Kaspar CW (2008) Acid stress damage of DNA is prevented by Dps binding in Escherichia coli O157:H7. BMC Microbiology 8: 181. doi:10.1186/1471-2180-8-181.

34. Mincheva M, Craciun G (2008) Multigraph Conditions for Multistability, Oscillations and Pattern Formation in Biochemical Reaction Networks. Proc IEEE 96: 1281–1291. doi:10.1109/JPROC.2008.925474.

35. Bannen RM, Bingman CA, Phillips GN (2007) Effect of low-complexity regions on protein structure determination. J Struct Funct Genomics 8: 217–226. doi:10.1007/s10969-008-9039-6.

36. Baumler DJ, Hung K-F, Jeong KC, Kaspar CW (2007) Production of methanethiol and volatile sulfur compounds by the archaeon "Ferroplasma acidarmanus". Extremophiles 11: 841–851. doi:10.1007/s00792-007-0108-8.

37. Lall R, Mitchell J (2007) Metal reduction kinetics in Shewanella. Bioinformatics 23: 2754–2759. doi:10.1093/bioinformatics/btm400.

38. Marcia RF, Mitchell JC, Wright SJ (2007) Global optimization in protein docking using clustering, underestimation and semidefinite programming. Optimization Methods and Software 22: 803–811. doi:10.1080/00207170701203756.

39. Craciun G, Pantea C (2007) Identifiability of chemical reaction networks. J Math Chem 44: 244–259. doi:10.1007/s10910-007-9307-x.

40. Marcia RF, Mitchell JC, Rosen JB (2007) Multi-funnel optimization using Gaussian underestimation. Journal of Global Optimization 39.

41. Pan Y, Durfee T, Bockhorst J, Craven M (2007) Connecting quantitative regulatory-network models to the genome. Bioinformatics 23: i367–i376. doi:10.1093/bioinformatics/btm228.

42. McCoy JG, Bitto E, Bingman CA, Wesenberg GE, Bannen RM, et al. (2007) Structure and dynamics of UDP-glucose pyrophosphorylase from Arabidopsis thaliana with bound UDP-glucose and UTP. J Mol Biol 366: 830–841. doi:10.1016/j.jmb.2006.11.059.

43. Kondrashov DA, Van Wynsberghe AW, Bannen RM, Cui Q, Phillips GN (2007) Protein structural variation in computational models and crystallographic data. Structure 15: 169–177. doi:10.1016/j.str.2006.12.006.

44. Davis J, Ong I, Struyf J, Burnside E, Page D, et al. (2007) Change of representation for statistical relational learning In: Proceedings of the 20th International Joint Conference on Artificial Intelligence (M. M. Veloso, ed.), pp. 2719-2726.

45. Riccardi D, König P, Prat-Resina X, Yu H, Elstner M, et al. (2006) “Proton holes” in long-range proton transfer reactions in solution and enzymes: A theoretical analysis. J Am Chem Soc 128: 16302–16311. doi:10.1021/ja065451j.

46. Baumler DJ, Hung KF, Bose JL, Vykhodets BM, Cheng CM, et al. (2006) Enhancement of acid tolerance in Zymomonas mobilis by a proton-buffering peptide. Appl Biochem Biotechnol 134: 15–26.

47. Han BW, Bingman CA, Mahnke DK, Bannen RM, Bednarek SY, et al. (2006) Membrane association, mechanism of action, and structure of Arabidopsis embryonic factor 1 (FAC1). J Biol Chem 281: 14939–14947. doi:10.1074/jbc.M513009200.

48. Riccardi D, Schaefer P, Yang Y, Yu H, Ghosh N, et al. (2006) Development of effective quantum mechanical/molecular mechanical (QM/MM) methods for complex biological processes. J Phys Chem B 110: 6458–6469. doi:10.1021/jp056361o.

49. Glasner JD, Rusch M, Liss P, Plunkett G, Cabot EL, et al. (2006) ASAP: a resource for annotating, curating, comparing, and disseminating genomic data. Nucl Acids Res 34: D41–D45. doi:10.1093/nar/gkj164.

50. Anderson BD, Gilson MC, Scott AA, Biehl BS, Glasner JD, et al. (2006) CGHScan: finding variable regions using high-density microarray comparative genomic hybridization data. BMC Genomics 7: 91. doi:10.1186/1471-2164-7-91.

Summary of Future/Continuing Projects Initiated Under DE-FG02-04ER25627

Transcriptional Mapping of Rhodobacter Sphaeroides

Saheed Imam and Timothy Donohue

We propose studying the metabolic and regulatory networks of the model photosynthetic bacterium, *Rhodobacter sphaeroides*. To achieve this, I proposed using a combination of experimental and computational approaches, which I grouped into the following 3 specific aims: (1) reconstruct, validate and refine a genome-scale metabolic network for *R. sphaeroides*; (2) infer a *R. sphaeroides* condition-dependent transcriptional regulatory network; and (3) construct an integrated *R. sphaeroides* regulatory-metabolic network.

(1) Reconstruct, validate and refine a genome-scale metabolic network for *R. sphaeroides*. Having previously constructed and validated a genome-scale metabolic model for *R. sphaeroides*, namely iRsp1095, utilizing then available genomic and experimental data, over the last year I have carried out an extensive refinement and extension of this model utilizing a combination of high-throughput phenotypic experiments, mutational analysis and computational approaches. These new data, which represent a significant enhancement of our knowledge of the metabolic activities of this organism, were used as input to update iRsp1095 leading to the identification and addition of 267 reactions, 74 metabolites and 31 genes to enable modeling of the metabolism of these substrates. This refined model, currently called iRsp1125, was validated against previously generated chemostat data.

(2) Infer a *R. sphaeroides* condition-dependent transcriptional regulatory network. I previously generated 2 transcriptional networks using the information present in gene expression data only. To improve on these predicted networks, I am currently utilizing a phylogenetic footprinting approach to identify conserved regulatory motifs in individual operons. Using 5 relatively closely related α-proteobacteria, I have been able to identify 807 conserved motifs in the promoter regions of *R. sphaeroides* genes. This data, which is currently being refined by the addition of more genomes and utilization of alternative motif finding approaches, will serve as input for motif clustering and subsequent transcriptional network inference, which will utilize both gene expression and sequence information. Once I complete this computationally inferred network, it will need to be experimentally validated. To achieve this, I proposed using approaches such as ChIP-seq. To verify that ChIP-seq would be a viable approach in *R. sphaeroides*, as well as to gain added insight about the function of FnrL – the master regulator of photosynthetic growth in *R. sphaeroides*, I performed a paired-end ChIP-seq experiment for FnrL in the WT and FnrL deletion strains. With this analysis I verified that ChIP-seq will be a viable approach for assessing transcription factor (TF) binding in *R. sphaeroides* and identified a total of 69 FnrL binding sites including 45 previously unidentified sites. In addition to work on FnrL, I also did work on the 2 paralogs of the heat shock σ factor in *R. sphaeroides*, $RpoH_I$ and $RpoH_{II}$, verifying ChIP-Chip identified sites by β-galactosidase assays and creating a promoter library for assessing bases important for activity and specificity of $RpoH_I$ and $RpoH_{II}$.

Publications

Dufour YS, **Imam S**, Koo BM, Green HA, Donohue TJ: **Convergence of the transcriptional responses to heat shock and singlet oxygen stresses**. *Plos Genetics*. In press

Imam S, Yilmaz S, Sohmen U, Gorzalski AS, Reed JL, Noguera DR, Donohue TJ: **iRsp1095: a genome-scale reconstruction of the *Rhodobacter sphaeroides* metabolic network**. *BMC Syst Biol* 2011, **5**:116.

Kontur WS, Ziegelhoffer EC, Spero MA, **Imam S**, Noguera DR, Donohue TJ. **Pathways involved in reductant distribution during photobiological H_2 production by *Rhodobacter sphaeroides*.** Appl Environ Microbiol. 77(20):7425-9. (2011).

Imam S, Donohue TJ: **A systematic analysis of the metabolism of a photosynthetic microbe.** In preparation. (in preparation)

Improving ChIP-seq data analysis for mapping regulatory networks relevant to biofuel producing E. coli

Tricia Kiley, Sunduz Keles and Bob Landick

We have developed tools to 1) Accurately identify closely spaced transcription factor binding sites (dPeak - Dongjun Chung) and 2) Identify and quantify differential binding across multiple conditions (DBChIP - Kun Liang) using ChIP-seq data from the Kiley lab. As an initial test case, we generated paired-end and single-end ChIP-seq data for the σ^{70} subunit of RNA polymerase from cells grown under both aerobic and anaerobic conditions. NCIS analysis of the σ^{70} data allowed us to determine genome wide oxygen dependent transcriptional regulation. This data has been incorporated into a manuscript, which describes global transcriptional regulation mediated by the anaerobic transcription factor, FNR [in progress]. Further analysis of the σ^{70} data with dPeak revealed many novel transcription start sites and demonstrated that paired-end sequencing data is better able to resolve closely spaced transcription factor binding sites compared with single-end ChIP-seq data. Several start site predictions were validated using primer extension, a biochemical technique used to map transcription start sites (Dan Park). A manuscript describing the dPeak algorithm and σ^{70} ChIP-seq analysis has been written and will be submitted shortly. dPeak was also used to analyze paired-end ChIP-seq data for the transcription factor ArcA (Dan Park), which is known to recognize multiple closely spaced binding sites within regulatory regions. The role of two such binding sites in the transcriptional regulation of *cyoA* was confirmed by systematically mutating the binding sites and assaying transcription using galactosidase assays. DNase 1 footprinting assays will be performed to verify ArcA binding at additional predicted sites and a manuscript will be prepared in the next year.

Studies of Protein Mutagenesis Effects

Xiaolei Zhu, Omar Demerdash, Spencer Ericksen and Julie C Mitchell

We develop data-driven models for the study of protein mutagenesis effects upon protein-protein and protein-DNA binding. Our models derive new features based on protein electrostatics and interface plasticity to examine the effects of amino acid substitutions on binding affinity and specificity. Our KFC2 hot spot model for alanine substitution is a gold standard in the field. More recently, we have developed models for general sidechain substitution (19 subs/residue), as well as for identifying *de novo* designed protein-protein interfaces that bind when assessed experimentally. We recently developed a highly accurate model for identifying DNA-binding sites on the surface of proteins that is currently being adapted to identify sites able to kink and bind DNA selectively, with the ultimate goal of identifying the cognate sequence from surface features alone.

These studies lead toward understanding the basis of macromolecular recognition and specificity. Models able to predict such phenomena are a challenge, as classical force fields and statistical potentials return poor results for predictions of changes to binding affinity resulting from a mutation. Yet, the implications across science to understanding how to predict these changes are vast, from understanding the basis of functional evolution, to the ability to design transcriptional regulation networks. The study of mutagenesis effects also a critical link between structure and systems biology. The functional effect of a mutation is often realized as a geometric or chemical change in the surface of a protein. As the genomics era continues to advance, structural biology has not kept the same pace. It is only through integration of data into predictive models that we can readily merge the molecular and systems scales at the data level, helping to create models at the structural level that also help to understand the evolution of regulatory and metabolic networks in terms of loss of function mutations at the structural level.

Publications

Demerdash ONA, Daily MD, Mitchell JC (2009) Structure-based predictive models for allosteric hot spots. PLoS Comput Biol 5: e1000531. doi:10.1371/journal.pcbi.1000531.

Zhu X, Mitchell JC (2011) KFC2: a knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features. Proteins 79: 1097–0134.

Zhu X, Ericksen SE and Mitchell JC (2012) DBSI: DNA Binding Site Identifier (in review)

Demerdash ONA and Mitchell JC (2012) Distinguishing Native Binding Interface from *de novo* Designed Interfaces That Do Not Bind (in preparation)

Lattice Monte Carlo Simulation of Cellulose Degradation

Crane Zhang and George Phillips

We developed a Lattice Monte Carlo model of cellulose degradation system. The processive reaction of cellobiohydrolases (CBHs) on crystalline cellulose was reproduced. We also simulated the reaction-diffusion process in cellulose degradation by CBHs. The simulation results indicated that the initial burst and fast inhibition is caused by the heterogeneity of cellulose hydrolysis.

To further investigate the bioprocessing of cellulose by different types of cellulases, we are going to apply this Lattice Monte Carlo model to the study of "endo-exo" synergy and "exo-exo" synergy. The previous results suggest that endo-glucanases may help smooth the cellulose surface and release trapped CBHs. For "exo-exo" synergy, details of the interactions between two types of CBHs (CBH I and II) will be tested.

Publications

He Zhang, Xiaolin Cheng and George N. Phillips, Jr. "A Lattice Monte Carlo model of cellulose degradation by processive cellobiohydrolases" (in preparation)

Computational Mapping of ChIP-seq data

Kun Liang and Sunduz Keles

We have worked on various projects that use high-throughput sequencing to detect transcriptional events. The main tool used is ChIP-seq (Chromatin immunoprecipitation followed by sequencing), which has become an important tool for identifying genome-wide protein-DNA interactions. In ChIP-seq experiments, ChIP samples are usually coupled with their matching control samples. Proper normalization between the ChIP and control samples is an essential aspect of ChIP-seq data analysis. We have developed a novel normalization method for ChIP-seq that can improve power and achieve proper error rate control. In our second project, we focused on in-depth analysis of transcription factor binding with ChIP-seq. Although there are large numbers of ChIP-seq programs that detect transcription factor binding in a single experiment condition, no program exists to detect differential binding across multiple conditions. To this end, we have developed a powerful and flexible program, called DBChIP, to detect differentially bound sharp binding sites across multiple conditions, with or without matching control samples. By assigning uncertainty measure to the putative differential binding sites, DBChIP facilitates downstream analysis. Finally, we applied above methods in the study of E. coli transcriptional network. In particular, we studied genome-wide binding of FNR under anaerobic condition and evaluated the differential binding of sigma-70 between aerobic and anaerobic conditions. Our study leads to insight on how FNR facilitates the rapid response to environmental signals

Publications

Liang, K. and Keles, S. (2012), Detecting differential binding of transcription factors with ChIP-seq, Bioinformatics, 28, 121-122

Liang, K. and Keles, S. (2012), Normalization of ChIP-seq data with control, BMC Bioinformatics, to appear

Kevin Myers, Huihuang Yan, Irene Ong, Dongjun Chung, Kun Liang, Sündüz Keles, Robert Landick, Patricia Kiley, Integrating genome-wide analyses reveals a flexible transcriptional network of the global anaerobic transcription factor FNR in Escherichia coli, in preparation

Engineering *E. coli* for Production of Branched Chain Amino Acids

Xiaolin Zhang and Jennifer Reed

Two separate approaches were used to predict and engineer *E. coli* strains able to increase production of branched chain amino acids (BCAAs), including leucine. The first approach involved use of a co-culture of *E. coli* auxotrophs to improve production, while the second approach involved using a computational tool (MOMA) to identify gene deletions that would improve production. Using an *E.coli* co-culture model, Xiaolin predicted ~40 auxotrophic mutants could support growth of a leucine auxotroph in co-culture. After testing different pairs of auxotrophs experimentally, she selected *a* Δ*lysA* strain and a Δ*leuA* strain to adaptively evolve in co-culture for up to a month. Xiaolin has characterized the Δ*lysA*+Δ*leuA* co-culture evolution by analyzing the growth rates, death rates, population compositions, and nutrient requirements. She also studied the phenotypes of evolved isolates in monoculture and co-culture to determine how individual isolates in the co-culture contribute to the fitness of the population, and how individual isolates evolve in the co-culture. Unfortunately, none of the isolates examined appeared to produce detectable levels of leucine when grown in monoculture. A computational model of the co-culture was used to study the effect of increasing uptake rates and production rates on the population distributions and growth rates, and predictions were compared to experimental data. The modeling and experimental results suggest that one co-culture improved growth by a better exchange of leucine (or its precursors) while the other co-culture used a different strategy. Xiaolin also used a second approach to improve BCAAs that did not involve co-cultures. Xiaolin used an existing computational approach (MOMA) to predict how gene deletions would affect BCAA production. Based on these results she identified a large number of gene deletion strategies for improving leucine production. Xiaolin has begun implementing these strains experimentally and evaluating their ability to produce leucine. She also plans to use a more recent approach (RELATCH) developed in the Reed Lab to identify additional genetic strategies as well. A paper describing the characterization of the un-evolved and evolved co-cultures is in preparation and will be submitted in September 2012. Xiaolin also presented her results as a talk at the 5th International *E. coli* Alliance Conference in December 2011 and is planning to present her work at the ASM Conference Beneficial Microbes meeting in October 2012.

Publications

Xiaolin Zhang and Jennifer L Reed. “Adaptive Evolution of a Mutualistic Microbial Community”. In preparation.

Xiaolin Zhang and Jennifer L Reed. “Engineering E. coli for Branched Chain Amino Acids Production”. Expected Publication.

Protein Synthesis and Turnover in *Rhodobacter Sphaeroides*

Timothy Donohue, Daniel Noguera, and Yury Bukhman.

Collaborators at PPNL: Gordon Anderson, Mary Lipton, and Gordon Slysz.

Cellular protein levels reflect a dynamic balance between protein synthesis and turnover. We have used stable isotope labeling (15N) and loss of non-labeled (14N) proteins across the entire proteome to estimate synthesis and turnover rates in *Rhodobacter sphaeroides*. Mass spectral measurements were acquired at different time points for cells growing under four different conditions (with and without light and/or oxygen). Time-dependent changes in incorporation of 15N were used to estimate protein synthesis rates while the time-dependent loss of 14N-labeled proteins provided the means to measure protein turnover rates. Extracting this information from mass spectra required application of advanced statistical methods and bioinformatics. Using these estimates of synthesis and turnover rates we identified several proteins that we believe to be affected by the presence of radical oxidative species during photo-oxidative stress. Additional experiments will be undertaken in the future to confirm these protein identifications.

COMPUTATIONAL AND EXPERIMENTAL APPROACHES FOR TUNING GENE EXPRESSION

Jennifer Reed and Brian Pfleger

One of the goals of our project was to develop a novel computational tool to predict flux distributions in response to changes in gene expression. Previous approaches could predict flux distributions in response to metabolic gene deletions. Using similar principles we have developed a framework that is capable of accounting for changes in gene expression on a continuous level. This new framework is based on volumetric constraints due to molecular crowding within a cell. One important parameter needed is a term that relates reaction fluxes to the volume taken up by the enzymes. An initial attempt to acquire these parameters was not very successful and used a kinetic model of *E. coli* from a recent project. We have since tried a different approach that takes advantage of available transcriptomic, proteomic, and fluxomic data. Using this information, we have estimated the required parameters that enable conversion of fluxes to enzyme volumes. We are in the process of evaluating the method and comparing predictions to experimental results for gene deletions, for which data is available. We then plan on using the approach to evaluate data produced by the Pfleger Lab. We anticipate generating a manuscript detailing the methods and results for the new approach within the next 6 – 12 months.

The second goal of our project was to develop a set of characterized promoters that could control gene expression at various levels in cyanobacteria. Members of this set would be used to alter the expression of key metabolic and/or regulatory genes and test predictions made by metabolic models developed by the Reed lab. We constructed a plasmid expression system for altering the expression of two reporter proteins in the model cyanobacterium *Synechococcus* sp. PCC7002. The reporters included a yellow fluorescent protein (YFP) and a thermostable green fluorescent protein (GFP). We cloned a mutagenized promoter library in front of these reporter proteins and obtained a wide distribution of fluorescence phenotypes in *E. coli*. When individual members of the YFP library were cloned into PCC7002, cultures generated highly variable fluorescence values (even among replicates). This motivated our work with the thermostable GFP. Upon further examination, we discovered that our constructs were not fully segregated meaning that some cells had more copies of the YFP cassette than the wild-type target locus, whereas others had the opposite. We have since developed a more stringent set of selection conditions and developed a counter-selection procedure that facilitates full segregation. A manuscript detailing the mode of toxicity used in the counter selection is currently under review. A second manuscript detailing the method of counter selection will be written by the end of 2012. We are currently reconstructing the library in PCC7002 using the counter-selection method and will be screening the fluorescence libraries thereafter. We plan to use the library to test predictions made by the Reed lab using their metabolic model of PCC7002 in 2013.

Publications

Wai Kit Ong and Jennifer Reed. “Predicting Metabolic Responses to Gene Expression Changes”. Expected Publication.

Begemann MB, Schmitt EF, Pfleger BF. “Increasing tolerance of the cyanobacterium Synechococcus sp. PCC 7002 to 3-hydroxypropionic acid and acrylic acid”. (Submitted)

Chemical Hydrolysis of Lignocellulosic Biomass into Fermentable Sugars

Christos Maravelias and Ronald T Raines

The goal of this study is to investigate the technoeconomic feasibility of a large-scale process for fermentable sugar production from ionic liquid (IL) pre-treated biomass. Toward this aim, we synthesized an integrated strategy for the production of sugars from corn stover. Our strategy employs four types of systems: (i) IL-based hydrolysis, (ii) liquid-solid separations, (iii) power generation, and (iv) simulated-moving-bed (SMB) separations. The modeling of the first is based on an experimental study for a series of two IL hydrolyses to produce fermentable sugars from crude biomass (Binder and Raines, 2010). The modeling of the second and third, which are established industrial unit operations, is based on data from a previous National Renewable Energy Laboratory (NREL) report (Aden et al., 2002). The modeling of the fourth system, the SMB-based separation of IL from fermentable sugars and water, is based on a literature search for a similar system (Xie et al., 2005). At the basic design, the minimum selling price (MSP) of the fermentable sugar product was found to be $6.72/kg, which is significantly higher than the current market price of sugars. To identify the major cost drivers and technology gaps, we carried out sensitivity analysis studies with respect to a series of technical (e.g. IL recovery rate and IL:biomass ratio) and economic (e.g. IL price, tax rate, return on investment discount rate) parameters. As a result of the sensitivity analyses, the lowest MSP was found to be at $1.14/kg. The cost of ILs appears to be the major barrier for the implementation of the process. Therefore, our analysis suggests that process alternatives with lower IL consumption and/or separation strategies that would allow higher recycle of ILs should be studied.

References

J. B. Binder and R. T. Raines, PNAS, 2010, 107, 4516–4521.

A. Aden, M. Ruth, K. Ibsen, J. Jechura, K. Neeves, J. Sheehan, B. Wallace, L. Montague, A. Slayton and J. Lukas, National Renewable Energy Laboratory (NREL) Report, NREL/TP-510-32438, 2002.

Y. Xie, C. Y. Chin, D. S. C. Phelps, C. H. Lee, K. B. Lee, S. Mun and N. H. L. Wang, Industrial & Engineering Chemistry Research, 2005, 44, 9904–9920.

Publications

S. M. Sen, J. B. Binder, R. T. Raines and C. T. Maravelias, Biofpr, 2012, 6, 444–452.

Evolution of substrate specificity in CBM33 and GH61 polysaccharide monooxygenases families

Cameron R. Currie, George N. Phillips and Brian G. Fox

The fungal GH61 and bacterial CBM33 families have recently been determined to be copper-containing polysaccharide monooxygenases (PMOs) capable of cleaving crystalline cellulose or chitin. Although these two families are structurally conserved, there is little sequence similarity between the two families. Given the large number of PMO sequences available, we analyzed the phylogeny of 254 sequences from the GH61 family and 373 sequences from the CBM33 family. The analyses suggest that convergent evolution may have produced the biochemical activity of these families from unrelated ancestral proteins. By annotating crystal structures of PMO proteins with the breadth of functional activities now reported for these proteins, we propose an approach for classifying the substrate specificity in each enzyme family. The classification has been applied to representatives of the bacterial CBM33 family with newly determined transcriptional and functional properties. A manuscript covering this work is in preparation.

Adaptive Optimization of Energy Production in the US Power Grid

Carol Barford and Paul Barford

Our project has focused on the U.S. electric power grid. We have used statistical machine learning and data analysis to model and understand the macro-level behavior of the electric grid, with the ultimate goal of improving efficiency and determining where and how to best incorporate green energy sources such as biomass.

The large majority of U.S. electricity comes from fossil fuels, and considerable efforts have been made in both industry and academia to develop more environmentally friendly alternatives. Green energy is more dependent on time than its fossil fuel counterparts, as it is dramatically affected by seasonal and daily cycles. For example, coal can be added o a furnace whenever it is needed, but solar power works only when the sun is shining.

To maximize the efficiency of additional green energy, as well the profits of the green power plant owners, one must understand the behavior of the electric power grid over time. In May, S.A. presented Toward an Analytic Framework for the Electric Power Grid at the e-Energy 2012 conference in Madrid, Spain. In this work, we created a framework for analyzing the markets in which electricity is bought in sold, and discovered underlying behavior in the markets useful for adopting green energy technologies.

Our more recent work has focused on load forecasting. Companies that direct the flow of electricity must also predict how much is going to be used and where. A small-scale example is predicting how much electricity a house is going to use during each minute of a day, while a large-scale example is predicting Madison's total usage for each hour of the next week. We are focusing on improving large-scale forecasts, and have improved the accuracy of MISO's (the Midwest Independent System Operator) Midterm Load Forecasts (MTLF, paper in progress).

In addition to load forecasting, we have begun work with a more theoretical focus. The price and usage of electricity are functions of time, as are the influential factors (e.g., weather, seasons, holidays). Some of the methods we've developed in our analysis are general and applicable to more than only electricity. We are exploring how our techniques and analysis might advance the field of Time Series Analysis. Specifically, we are working on parallel data streams and forecasting signals with seasonality.

Over the past year, we have developed collaborations with a number of industry partners. Genscape, a company that measures and reports real-time energy use fundamentals, has provided us with a year's worth of data and we are in on-going communications with them. In addition, Madison Gas and Electric's green energy initiative relates directly to our work. We are currently negotiating with MG&E to obtain load profiles (records of electricity use by different parties over time) for the Madison area. Our strongest partnership is with MISO. MISO controls the market and flow of electricity for the Midwest U.S. and Canada, and we have been using their publicly available data over the last year. We are currently planning a visit to their headquarters for next month, and discussing future NDA's with their legal department.

Publications

Toward an Analytic Framework for the Electric Power Grid. Scott Alfeld, Carol Barford, and Paul Barford. Published in the proceedings of e-Energy 2012.

Improving load forecasts in the US (in progress)

PROJECT SUMMARY: A TRANSCRIPTIONAL ANALYSIS OF CELLULOLYTIC RUMINAL BACTERIAL CO-CULTURES

Christina Kendziorski and Garret Suen

The goal of this study was to gain insight into the genetic underpinnings of cellulose degradation by rumen bacteria and to identify synergistic microbial interactions in bacterial co-cultures grown on next-generation cellulosic feedstocks. To begin, we examined the genetic underpinnings of cellulose degradation in a single ruminal bacterial species, *Ruminococcus albus 7*. Importantly, *R. albus* 7 has the ability to ferment cellulose directly into ethanol. We tested the ability of this bacterium to degrade cellulose and ferment ethanol when grown on either cellulose or cellobiose. A transcriptomics approach using Illumina-based RNA-seq was employed. Specifically, we aligned the reads generated from Illumnia sequencing using BWA (Li *et al.*, 2009b), an approach that allows for small gaps (indels) and mismatches, producing very good (>90%) alignment with our reference genome in almost all samples. These aligned reads were then enumerated using SAMtools (Li *et al.*, 2009a) and investigated for differential expression using a technique developed in the Kendziorski lab, called EBseq (Leng *et al.*, 2012).

We identified several novel attributes of cellulose degradation, including the surprising finding that this organism lacks a traditional cellulosome, an enzyme complex that is typically used for cellulose recently, we have begun investigating co-cultures of *R. albus* 7 with the ruminal bacterium degradation by other species of this genus. Moreover, we found that tryptophan synthesis was upregulated when *R. albus* 7 was grown on cellulose, and that this organism has an unusually high enrichment for tryptophan residues in cellulose-active enzymes. We also found that when grown on crystalline cellulose, this organism initiated a broad transcriptional response that included a range of enzymes important for plant cell wall deconstruction. This broad transcriptional response included upregulation of several hemicellulases and the downregulation of several enzymes involved in the synthesis of nitrogenous metabolites. These findings have been incorporated into a manuscript that will be submitted shortly. More *Prevotella ruminicola 28* singly and in co-culture. A major challenge in working with these mixed cultures has been the in determining their relative abundance. We have approached this by using qPCR and have formulated the appropriate innoculation ratios and conditions to pursue a normalized RNA-Seq analysis for this co-culture combination on either cellulose or cellobiose. We anticipate obtaining these data relatively soon. The collaboration between the Kendziorski and Suen labs has been profitable in that we have established a pipeline for analyzing RNA-seq using tools developed in the Kendziorski lab. These tools, primarily employed for Eukaryotic systems, have now been demonstrated to be effective for bacteria. The resulting initial publication will not only underscore the importance of this pipeline in analyzing bacterial RNA-seq data, but will also provide insights into how *R. albus* 7 leverages its genetic potential to ferment ethanol from cellulose. Importantly, this work will fuel future collaborations between the Suen and Kendziorski labs.

References

Leng N., Dawson, J. A., Stewart, R. M., Ruotti, V., Rissman, A., Smits, B., Haag, J., Gould, M. N., Thomson, J. A. and Kendziorski, C. EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. Technical Report 226, University of Wisconsin- Madison, Department of Biostatistics and Medical Informatics, 2012.

Li, H. and Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-1760, 2009.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R, and and the 1000 Genome Project Data Processing Subgroup. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25:2078-2079, 2009.

Publications

Christopherson, M.R., Dawson, J.A., Bramhacharya, S., Stevenson, D.M., Weimer, P.J., Kendziorski, C. & G. Suen. A Global Analysis of Gene Expression in the Genome Sequence of *Ruminococcus albus* 7 Reveals an Atypical Cellulose Fermenting Ethanologen. In preparation.

Studying curvature-mediated mechanisms of intracytoplasmic protein sorting in the model photosynthetic bacterium *Rhodobacter sphaeroides*

Qiang Cui and Doug Weibel

The general goal of our project is to use a combined experimental/computational approach to understand the role of cardiolipin (CL) in recruiting specific proteins to specific regions of the cellular/organelle membrane. The project was based on our current study of the interaction between *E. coli* DNA repair protein RecA and CL-rich membrane (*ms submitted*), and a long-range goal of this research is to understand how *R. sphaeroides* recruits photosynthetic proteins to chromatophores.

From an experimental angle, we took a two pronged-approach to lay the groundwork for our localization studies in *R. sphaeroides*. 1) To present a strong positive control for our experiments, we characterized the CL-binding sequences of *E. coli* RecA and human β-2-glycoprotein 1 (β-2GP1; *ms in preparation*). Our approach has been to identify peptide sequences that bind to CL domains tightly and specifically. We envision that these peptides will enable us to create fluorophore-labeled peptides that we can engineer for visualization of CL in *R. sphaeroides* cells. We have made excellent progress on this aim in the past year and have identified candidate β-2GP1 CL-binding sequences that we are currently characterizing in vivo and in vitro. 2) To regulate and control CL levels in *R. sphaeroides* in vivo, we are engineering strains that have controllable CL expression. We found that the cardiolipin synthase (CLS) knockout in *R. sphaeroides* reduces but does not eliminate CL formation. We assume that this phenotype arises from CLS homologs and redundant CL biosynthesis. Using the sequence for CLS, we identified a candidate homology in *R. sphaeroides* and are in the process of knocking it out. We have also engineered CLS expression systems in *E. coli* and have demonstrated in a complete CL null that we are able to precisely control CL levels in vivo. We are currently moving this expression system into the *R. sphaeroides* CL null and will characterize how CL concentration influences photosynthesis as a prelude to working on the molecular interactions of CL microdomains in the membrane with photosynthetic proteins.

On the computational side, going beyond our recent work on RecA-membrane interaction using the simple IMM1 model, we have been pushing forward a multi-stage approach for efficiently predicting the binding conformation and energetics for proteins at the water/membrane interface. The simplest model uses a more sophisticated dielectric model for membranes (GBSW). Moreover, to describe the negatively charged CL, we have developed a "pseudo-quadruple" layer model that overcomes the limitations of popular Gouy-Chapman model. We are in the process of testing the model based on the partitioning of single amino acids, small cationic peptides and RecA at the water/membrane interface (*ms in preparation*). These calculations have laid the ground work for more elaborate computational studies using explicit lipids, which are required to elucidate the role of chemical interactions between CL and protein motifs in determining the specificity of CL-driven protein localization.

Publications

M. Rajendram, S. Yang, L. Zhang, Q. Cui, D.B. Weibel, "Identification and characterization of the cardiolipin-binding site of *Escherichia coli* RecA." (in preparation)

Z. Wu, L. Zhang, A. Yethiraj and Q. Cui, "An analysis of the Gouy-Chapman model for studying protein binding to anionic membranes." (in preparation)

Electron Transfer Pathways

Shuo Yang and Qiang Cui

(1) Gating element(s) in the protein Cytochrome c Oxidase (CcO)[20]

Cytochrome c Oxidase is the terminal component in the electron transfer pathway in the cellular respiration. Inside the mitochondrion, for example, it reduces oxygen to water and uses the energy liberated by this reaction to pump protons across the mitochondrial membrane. This action, in turn, helps generate the proton concentration gradient which can be used to synthesize the energy-storing molecule ATP. The working mechanism of CcO is still elusive after decades of study. One of the key unsolved issues concerns the identity of the gating element(s) that prevents the backflow of protons. Many researchers have focused on the conformational preference of a key side-chain (E286) and water molecules in the key hydrophobic cavity.

Those previous studies, however, employed rather crude computational models; for example, only part of the protein was included and the calculations were done without any lipid membrane or solvent molecules. These simplifications may lead to qualitative differences in the behavior of a charge side chain (E286) in the protein interior. Parameters in the molecular model for CcO are carefully calibrated. Specifically, the protonation states of titratable groups are determined using results from multiconformation continuum electrostatics (MCCE) calculations; the dielectric contributions from the membrane and solvent are included in the model by using the GSBP protocol; the number of water molecules in the protein interior is tested with the grand canonical Monte Carlo (GCMC) simulations. Using this model, we carry out free energy simulation with umbrella sampling to evaluate the relative stability of different orientations of E286. Two-dimensional free energy simulations using metadynamics, which we implemented into CHARMM, are used to further validate the result. All the results consistently show that neither E286 nor water molecules is the gating element.

(2) Structural and dielectric properties of CcO from simulations with explicit membrane and solvent (ms in preparation)

In most of previous computational studies of CcO, due to its large size, a large part of the protein was frozen and/or no explicit lipid membrane and solvent environment was included in the model. Even though rigid models with constrains on the system are sufficient for some purposes, we believe it is necessary to have a more realistic model in which CcO is embedded without any artificial restrain in the membrane and bulk solvent environment. This is potentially important in correctly describing the dielectric properties of key regions.

In this work (ms in progress), we setup a computational model of CcO by putting it in the membrane and the bulk solvent following the Periodic Boundary Condition. We then study three intermediate states in the functional cycle of CcO and show that the protonation/oxidation states of some key co-factors significantly affect the solvation of the hydrophobic cavity near the active site, which in turn changes the effective dielectric constant of this functionally important region of the protein.

We note that in a previous study by Stuchebrokhov and coworkers, E286 has a very high pK_a value if both the protein and the hydrophobic cavity have a dielectric constant of 4, regardless of the oxidation states of the active site. In the current work, we carry out microscopic pK_a calculations of E286 in two states, with and without a proton at the D-propionate of the $Heme_{a3}$. In both cases, the pK_a values are high, and the difference between them corresponds to an effective dielectric constant of about 4 for the hydrophobic cavity, in agreement with our analysis of dielectric properties of CcO based on the Kirkwood-Fro □hlich model. Based on this and our other analyses, we propose a new proton pumping mechanism in which E286 is never negatively charged in the functional cycle. Rather, it can accept a proton and become doubly protonated transiently (separate work by P. Goyal and Q. Cui, in progress).

Another new feature emerging from our unrestrained simulation is the significant movement of a nearby loop, which seems to have a major impact on the structural and solvation properties of the hydrophobic cavity. We are carrying out two additional simulations for several mutants of this loop to test its functional importance.

(3) Understanding the metal ion binding specificity of the regulator protein CueR (work in progress)

Metal ions are essential co-factors to many biomolecules, especially metalloenzymes that catalyze complex chemical transformations. For example, a recent survey indicated that, among 1,371 different enzymes for which three-dimensional structures are available, ~ 47% contain metal ions with 41% hosting metals at the catalytic site. For the proper engineering of metalloproteins of novel functions (e.g., new transcription factors), a fundamental

challenge is to design the active site such that specific metal ion(s) is bound with high affinity and selectivity. Our general goal is to develop novel computational methods to help accomplish this.

We are developing the methodology using two proteins of the MerR family of transcription factors, CueR and CupR, as examples. These two proteins have similar metal binding sites but very different relative affinity towards Cu^+ and Au^+. Previous experimental studies by Soncini and coworkers showed that the response of CueR to Cu^+ is mainly from the metal-binding loop, while the binding of Au^+ is related to other factors. Further studies narrowed the focus down to a single residue on this loop, the mutation of which made CueR completely lose its response to Cu^+ (Soncini, et. al, personal communications), although the underlying mechanism is far from clear. In the current work (S. Yang and Q. Cui, in progress), we are trying to understand how this mutation impacts the structural and solvation properties of the metal site and therefore its Cu^+/Au^+ binding affinity.

To quantitatively evaluate the binding affinity of transition metal ions to proteins, ab initio QM/MM free energy simulations are required. Since such calculations are very expensive, we have developed a set of classical models for the Cu^+ and Au^+ binding sites, so that we can use classical simulations to generate trajectories for QM/MM free energy perturbation calculations. This combined classical and QM/MM approach has been found valuable in previous metalloenzme studies. We have also carried out short QM/MM MD simulations to validate the classical model. Free energy perturbation calculations are currently in progress.

Publications

Yang S, Cui Q (2011) Glu-286 rotation and water wire reorientation are unlikely the gating elements for proton pumping in cytochrome C oxidase. Biophys J 101: 61–69.

Several additional publications will be forthcoming, as noted in the text.