

Granlibakken HPC workshop 2009

Gary Wesenberg
University of Wisconsin-Madison
BACTER Institute





Genomics:GTL Mission and Goal

- Provide sufficient scientific understanding of plants and microbes to develop robust new strategies to produce biofuels, clean up waste, or sequester carbon.
- This includes research that supports the development of computational models to direct the use and design of improved organisms carrying out these processes.
- The ultimate scientific goal is to achieve a predictive, systems-level understanding of plants, microbes, and biological communities.



uwbacter.org

Bringing Advanced Computational Techniques to Energy Research

BACTER is a research training initiative funded by the US Department of Energy to promote the development of computational biology and bioinformatics techniques that are of direct value to energy science.



BACTER Training Model and Resources

- **Training Goals**

- Teach students how to **collaborate** and **communicate** outside their discipline
- Allow the pursuit of a PhD within a **traditional program** while fostering **interdisciplinary skills**
- Guidance in **writing** prelim proposals, first papers and dissertation outlines, language skills
- Guidance and support for turning prototype codes into professional **quality software** and web-based tools

- **Training Model**

- “Village” mentoring
- Weekly Journal Club/Seminar
- Short courses
- Brainstorming sessions
- Funding and participation generally spans **entire graduate career**

- **Training Resources**

- Computing cluster
- Travel funds
- Science **writer**
- Scientific **programmer**

Interdisciplinary and Collaborative

- Departments/Programs
 - Mathematics
 - Computer Sciences
 - Biostatistics
 - Chemical and Biological Engineering
 - Biophysics
 - Chemistry
 - Biochemistry
 - Genetics
 - Microbiology
- Research Areas
 - Electron/Proton Transfer
 - Allosteric Mechanisms
 - DNA Recognition
 - Pathway Reconstruction
 - Chemical Kinetics
 - Chemotaxis
 - Uranium Reduction
 - Photosynthesis Machinery
 - Gene Regulatory Networks
 - Microarray Design

Languages: Java, Perl, Matlab/Octave, Python, C/C++, FORTRAN

BACTER Projects With Possible HPC Needs

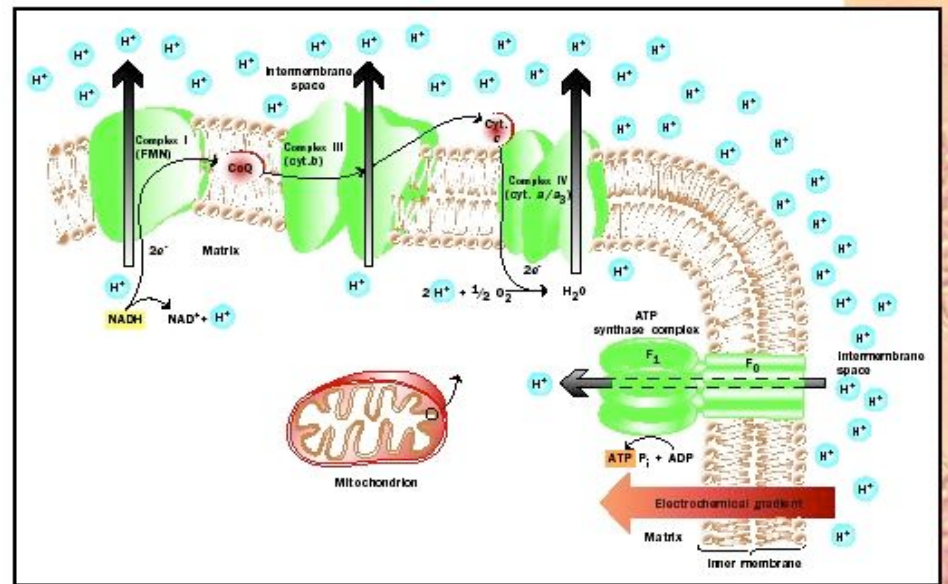
- A** • Simulating proton transfer through biological membranes with QM/MM methods
- M** • Modeling the enzymatic action of cellulase
- C** • Modeling metabolic pathways of bacterial cells to improve ethanol production
- E** • Designing gene chips used in gene regulation experiments

QM/MM simulations

Proton pumps: play a critical role in the bioenergetics of cells - inside the mitochondrion, for example, a proton gradient used by cells to synthesize the energy-storing molecule ATP.

Fundamental questions: How does pump “decide” when to send protons to participate in the reduction of oxygen to water, and when to transport them across the membrane? How is the backflow of protons prevented while pumping them against their concentration gradient?

Approach: QM/MM simulation, which combines the accuracy of quantum mechanical (QM) calculations with the speed of molecular mechanical (MM) techniques.



Goal: Identify the amino acid residue (or residues) that perform “gating.” Calculate the energetics of proton translocation.

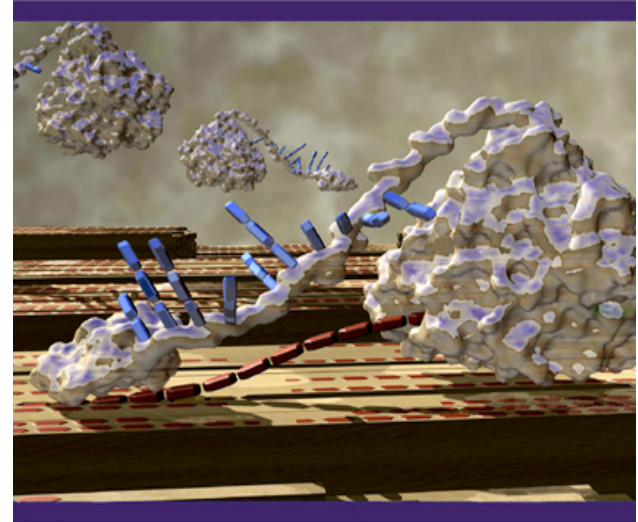
Shuo Yang, BACTER Predoctoral Fellow
Biophysics Program
Faculty Advisor: Qiang Cui

Modeling cellulase and cellulosomes

Cellulose degradation: a multi-scale problem of such complexity that neither chemical kinetics models, at the scale of cells, nor molecular dynamics models, at the scale of atoms and molecules, can adequately describe it.

Approach: use **Lattice Monte Carlo methods for simulating the dynamics** of free enzyme systems and cellulosomes, LARGE extracellular enzyme complexes that degrade cellulose.

Goal: to understand the synergism of the different modules within a single, free enzyme in conjunction with the dynamics of the cellulosome.



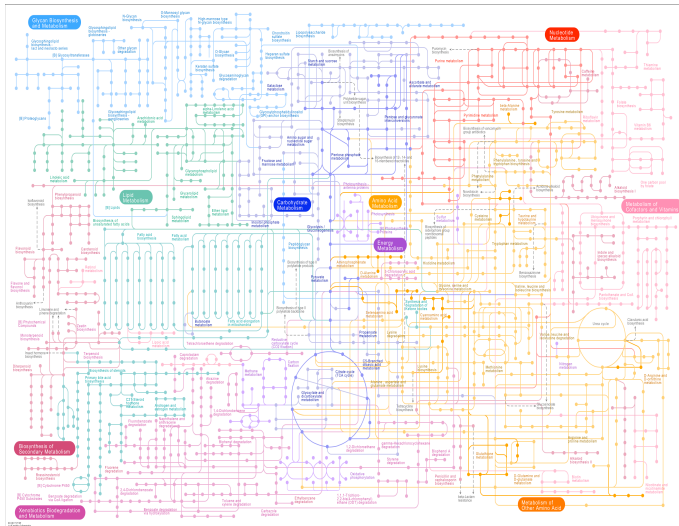
He Zhang ("Crane"), BACTER Predoctoral Fellow
Biophysics Program
Faculty Advisor: George Phillips

Modeling metabolic pathways of bacterial cells

CONSTRAINED FLUX ANALYSIS OF METABOLISM WITH GENE REGULATION

Constructing and validating computational models that represent the reconstructed metabolic network of E. coli K12. Extend genome annotation and comparative genomics to generate models of bacteria to aid in the bioengineering of an ideal ethanologen.

David Baumler, BACTER Postdoctoral Fellow
Genome Center of Wisconsin
Faculty Advisor: Nicole Perna



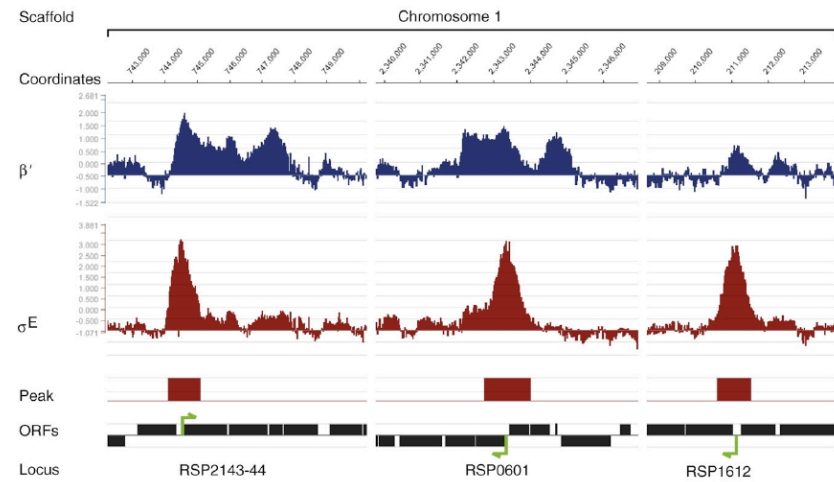
IMPROVED GENE CHIP DATA ANALYSIS

Modeling the networks that are involved in controlling how E.coli regulates its response to the carbon source(s) available to it. Sequence-based models provide predictive accuracy that is better than similar models without sequence-based parameters.

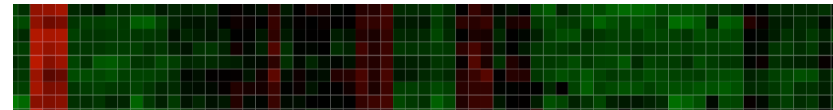
Yue Pan, BACTER Predoctoral Fellow
Department of Computer Sciences
Faculty Advisor: Mark Craven

ChipD Microarray Design

- High density CHIP-Chip arrays are used to analyze the interactions between proteins and DNA
 - One can **search for binding sites** of gene regulatory proteins across the entire genome
 - Algorithm first scores the quality of the sequence (repeats, stretches of the same base, etc..) then tries to pick the best probes.
 - Varies lengths to keep the chip isothermal.
 - The algorithm leaves no gaps on the coverage even in difficult regions of the sequence.
 - Every other probe is is reversed so both strands of the DNA is represented.
 - This method has been shown to be superior to commercial methods.



chipd.uwbacter.org



Yann Dufour, BACTER Predoctoral Fellow
Microbiology Program
Faculty Advisor: Tim Donohue

My Workshop Goals

Gain an understanding of how Leadership Class machines can be used in the various BACTER projects.

Be able to provide guidance to BACTER trainees in using HPC machines:

- Identifying code which will benefit by HPC
- Porting and optimizing code (tools)
- Accessing HPC machines

chipD code as sandbox:

Basic Questions on Porting code

- Avoid two step process?
 - 1) java → c++ (garden variety)
 - 2) c++ (garden variety) → c++ (big machine)
- Mapping java generics to c++ templates?
- Whether or not to use or avoid specific STL classes or algorithms?
- Best data types to use for genomic sequence data?

Thank you to all who made this workshop possible!