Bioinformatics, Modeling, and Computation

48

SimPheny™: A Computational Infrastructure for Systems Biology

Christophe H. Schilling* (cschilling@genomatica.com), Sean Kane, Martin Roth, Jin Ruan, Kurt Stadsklev, Rajendra Thakar, Evelyn Travnik, Steve van Dien, and Sharon Wiback

Genomatica, Inc., San Diego, CA

The Genomics:GTL (GTL) program has clearly stated a number of overall goals that will only be achieved if we develop "a computational infrastructure for systems biology that enables the development of computational models for complex biological systems that can predict the behavior of these complex systems and their responses to the environment." At Genomatica we have developed the SimPheny™ (for Simulating Phenotypes) platform as the computational infrastructure to support a model-driven systems biology research paradigm. SimPheny enables the efficient development of genome-scale metabolic models of microbial organisms and their simulation using a constraint-based modeling approach.

With SBIR grants from the DOE Genomics:GTL program we are in the process of enhancing the system-level functionality within SimPheny based on a collection of well qualified input from our pilot users. We have successfully deployed SimPheny into a series of academic laboratories and addressed key issues related to deployment strategies, collaboration requirements, training/support services, and experimental data integration needs, as well as modeling and simulation requirements. The findings that have resulted from these efforts have led to the identification of a series of high priority system-level requirements that need to be incorporated into the SimPheny platform in order to facilitate the use of such technologies to establish a model-driven research paradigm that can be broadly utilized by both expert and non-expert users.

We are now focusing on the software research and development activities necessary to achieve these system requirements. They include a major rework of the SimPheny underlying architecture and the related addition of functionality to improve remote connectivity for collaboration. Furthermore we are developing functionality to enable model and simulation data to be exported and imported within SimPheny. We are also developing capabilities to enable model comparison and the publishing of model content.

An integrated Fermentation Module has recently been added to SimPheny. This module allows users to manage, analyze, visualize, and integrate fermentation data within the existing modeling infrastructure. The module has the capability to check for the consistency of fermentation data by conducting elemental material balances and to convert raw concentration measurements into flux data that can be used to constrain and further validate simulation studies within SimPheny.

We are also in the process of integrating flux measurements within the context of genome-scale metabolic models to further our ability to analyze, interpret, and predict the behavior of biological systems. Designs are currently under way to incorporate data from 13C-labeling experiments into SimPheny and interpret the resulting flux measurements in the context of the predictive genome-scale models. Such functionality in SimPheny will provide experimental verification for our constraint-based

predictions of metabolic flux distributions as well as lead to a streamlined method for isotope label tracing experiment analysis that can be implemented by the non-expert user.

These enhancements and overall system improvements will create an extremely stable foundation that will enable SimPheny to be deployed to a broader range of users with more targeted functionality. These efforts will lead to the creation of a comprehensive package of software, model content, and training/support for the delivery of cellular modeling technologies to the metabolic research community.

49

Hybrid Bacterial Cell Models: Linking Genomics to Physiological Response

Jordan C. Atlas[1]* (jca33@cornell.edu), Mariajose Castellanos[1], Anjali Dhiman[1,2], Bruce Church[2], and Michael L. Shuler[1]

[1]Cornell University, Ithaca, NY and [2]Gene Network Sciences, Ithaca, NY

A major challenge in the biological sciences is to relate cell physiology to genomic structure. Models that explicitly link genomic and proteomic data to physiology are necessary to take full advantage of bioinformatics. We have developed a whole cell hybrid model that captures the dynamics of a single celled chemoheterotrophic prokaryote. By hybrid model we refer to inserting a genomically/molecularly detailed sub-model into a coarse-grained model which is embedded in a representation of the cell's environment. The initial step is to construct a coarse-grained model with lumped "pseudochemical species" (lumped components of similar chemical species). All subsystems of the coarse-grained model can be "de-lumped" into genomically complete, chemically distinct subsystems with corresponding genes and gene products. Using this coarse-grained host model structure we expect to quickly build a complete coarse-grained model of any given chemoheterotrophic bacteria using data from chemostat or other growth experiments. By combining molecularly detailed modules within the coarse grained host model, we capture not only the internal details of the dynamics of the molecular subsystem, but also can evaluate that mechanism within the context of a whole cell and its environment. The whole cell modeling approach presented here is being augmented by statistical mechanics methods for parameter estimation that allow us to rapidly develop parameter sets for new modules as they are added.

This framework has been applied to create Cornell's Minimal Cell Model (MCM). The MCM is a theoretical construction that attempts to develop our understanding of the relationship between cellular function and genetics using a "bottom up" approach; the necessary model functions are selected by rationally deciding what machinery a cell needs to live and reproduce. A "minimal cell" is a hypothetical free living cell possessing the functions required for sustained growth and reproduction in a maximally supportive culture environment. The MCM simulates the growth of a minimal cell. Ultimately, we aim to model the complete functionality of a minimal cell. The Shuler group has demonstrated the "modularity" of hybrid models by constructing a genomically and chemically detailed model of nucleotide metabolism within the MCM (PNAS v. 101(17), pp. 6681-6686). The current work focuses on incorporating amino acid supplementation into the coarse grained model. Another system of interest is *Shewanella oneidensis*, which has the potential to help remove metal pollutants from the environment. We believe that these techniques will ultimately allow us to build a model for *Shewanella* that creates a connection from the organism's genomics, to its molecular functions, to the whole cell, and to the environment.

* Presenting author

50

Identification of the Most Probable Biological Network Using Model Discrimination Analysis

Andrea L. Knorr and Ranjan Srivastava* (srivasta@engr.uconn.edu)

University of Connecticut, Storrs, CT

In seeking to understand the behavior of biological systems, whether at the molecular, cellular, or higher level, it is possible to develop multiple hypotheses of how the system of interest functions. These hypotheses may often be formulated into different network descriptions of the system. Using a Bayesian-based model discrimination technique, it is possible to determine which network, and as a consequence, which hypothesis is most probable. It is important to note that this method of network determination is not a data-mining approach, but rather is hypothesis-driven.

As an illustration of model discrimination, our work evaluating and identifying the most probable model of HIV-1 viral dynamics will be presented. Four different models of viral dynamics accounting for uninfected cells, infected cells, viral level, and/or the immune response were either taken from the literature or developed by our group. Parameters for the models were estimated from a cohort of 338 patients monitored for up to 2,484 days. Model discrimination was applied to determine which of the models, based on how they best captured overall viral dynamics, was most probable. The model determined as most likely was overwhelmingly favored relative to the remaining three models. It accounted for uninfected cells, infected cells, and cytotoxic T lymphocyte dynamics. Interestingly it was the only model that did not explicitly account for viral load, suggesting that none of the models to date have captured the appropriate network connectivity relating to viral load.

The technique of model discrimination is generic enough that it may easily be used to analyze biochemical kinetic pathways or identify the most likely genetic regulatory network in a given system of interest. To make such analysis readily available to a larger user base, we are in the process of developing a software package for carrying out model discrimination. Specifically, the package will allow the user to enter their models using SBML, as well as the appropriate data. The package will then determine the most probable model, presenting statistical analysis and comparative graphical output of actual and predicted network behavior.

51

Rhodopseudomonas palustris Regulons Detected by a Cross-Species Analysis of the α-Proteobacteria

Sean Conlan[1,*] (sconlan@wadsworth.org), Charles E. Lawrence[1,2], and Lee Ann McCue[1]

[1]The Wadsworth Center, Albany, NY and [2]Brown University, Providence, RI

The objective of this study is to elucidate transcription regulatory mechanisms of the environmentally significant bacterium, *Rhodopseudomonas palustris*. This α-proteobacterial species carries out three of the chemical reactions that support life on this planet: the conversion of sunlight to chemical-potential energy, the absorption of carbon dioxide which it converts to cellular material, and the fixation of atmospheric nitrogen into ammonia. We predicted regulatory signals genome-wide by applying a Gibbs sampling algorithm[1,2,3] to orthologous intergenic regions; specifically, those upstream of 2,044 genes/operons from *R. palustris* and seven other α-proteobacterial species (*Bradyrhizobium japonicum*, *Brucella suis*, *Caulobacter crescentus*, *Rhodobacter sphaeroides*, *Rhodospirillum rubrum*, *Hyphomonas neptunium*, and *Novosphingobium aromaticivorans*). A Bayesian motif clustering algorithm[4] was then used to cluster the cross-species motifs to identify genes that are likely regulated by the same transcription factor (i.e., a regulon). Of the 101 putative regulons detected, several were of particular interest: an organic hydroperoxide resistance regulon, a flagellar regulon, a photosynthetic regulon, the LexA regulon, and four regulons involved in nitrogen metabolism ($FixK_2$, NnrR, NtrC, σ^{54}). In addition, a highly conserved repeat sequence was detected downstream of over 100 genes.

Cognate transcription factor identification: Organic hydroperoxide resistance

At the core of the transcription regulatory network is the interaction between a transcription factor and its cognate binding site. Currently, there is no reliable way to infer these *cis-trans* connections from sequence data alone. It has been estimated, however, that ~50% of transcription factors in *E. coli* are auto-regulatory. Given that, auto-regulatory site(s) for a transcription factor should cluster with sites for additional genes that are regulated. We investigated eight *R. palustris* clusters (mentioned above) with biochemical and genetic data in the literature and found that four of those clusters contain motifs upstream of the likely cognate transcription factor. In particular, a motif upstream of *rpa0828*, a MarR family transcription factor of unknown specificity, clustered with motifs upstream of two genes involved in resistance to organic hydroperoxides (*ohr*, *rpa4101*). RPA0828, and its orthologs, all contain a highly-conserved cysteine residue required for the activity of the organic hydroperoxide resistance regulator, OhrR, from *Xanthamonas campestris* (Fig. 1). Therefore, it seems likely that RPA0828 is a regulator of hydroperoxide resistance in *R. palustris*. An additional OhrR homolog, RPA4102, was found in the *rpa4101-4103* operon and may have a similar or redundant function.

Figure 1. Alignment of the oxidation sensitive regions of several OhrR orthologs.

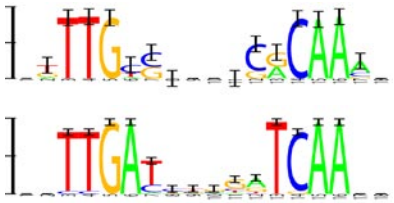
Organism	Protein	Alignment
X. campestris	OhrR	LDNQL**C**FALYS
R. palustris	RPA0828	LETQL**C**FALYS
R. palustris	RPA4102	LDRQV**C**FLLYA
B. japonicum	BLR0736	LDNQI**C**FAVYS
B. suis	BRA0886	LADML**C**FAVYS
H. neptunium	n.d.	LDHAL**C**FAIYS

* Presenting author

Discriminating between members of a transcription factor family: $FixK_2$ and NnrR

A difficulty with any clustering procedure is determining how many clusters are present in the data set, and many clustering approaches require this knowledge *a priori*. The Bayesian motif clustering algorithm (BMC), used in our work, determines the number of clusters based on sequence evidence and a tunable parameter (q), which influences whether a motif forms a cluster by itself or joins an existing cluster. This ability of the BMC algorithm to infer the number of clusters, while detecting subtle differences between motif types, is demonstrated by the $FixK_2$ and NnrR clusters. $FixK_2$ and NnrR regulate genes involved in respiration and nitric oxide metabolism, respectively, and both belong to the Fnr/Crp transcription factor family. Genes involved in these two pathways (respiration and nitric oxide metabolism) formed distinct clusters with the logos shown in Figure 2. Despite the high similarity between the binding consensus sequences, and 55% identity between the helix-turn-helix regions of $FixK_2$ and NnrR, BMC correctly separated the motifs of these regulons.

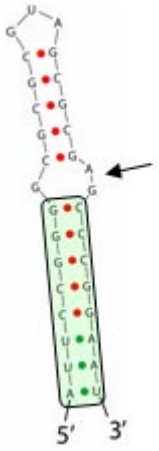
Figure 2. Sequence logos of the $FixK_2$ (top) and NnrR (bottom) clusters.



Detection of a novel inverted repeat

A benefit of the genome-wide approach used in this study is that conserved regulatory signals, regardless of their mechanism, can be detected. By not limiting the search to a particular set of genes (*e.g.*, a set of genes identified by a microarray experiment), or to a particular transcription factor, we were able to find a highly conserved inverted repeat downstream of over 100 genes. This repeat was found almost exclusively in intergenic regions at the 3' ends of genes. It was very rarely found between divergently transcribed genes or in coding regions. It is composed of a variable region, flanked by invariant inverted repeats. The variable region, which is 10-52 bp in length, is palindromic in 89% of the cases. Analysis of the repeats using *S*fold[5], demonstrated that many likely fold into a structure composed of an invariant helix, followed by a bulge and a variable-length hairpin (Fig. 3). While the repeat has features reminiscent of a mobile DNA element or transcriptional terminator, neither of these elements provide satisfactory explanations for the non-random distribution and perfectly conserved ends.

Figure 3. Putative structure of a 32 bp inverted repeat. The invariant helix is boxed and the bulge is indicated with an arrow.



References

1. Thompson, W, Rouchka, EC and Lawrence, CE. *NAR* **31**:566-84 (2003).
2. McCue, LA, Thompson, W, Carmack, CS, Ryan, JS, Liu, JS, Derbyshire, V and Lawrence, CE. *NAR* **29**: 774-782 (2001).
3. McCue, LA, Thompson, W, Carmack, CS and Lawrence, CE. *GenRes* **12**: 1523-32 (2002)
4. Qin, ZS, McCue, LA, Thompson, W, Mayerhofer, L, Lawrence, CE and Liu, JS. *NatBiotech* **21**: 435-39 (2003).
5. Ding, Y, Chan, CY and Lawrence, CE. *NAR* **32**: W135-41 (2004).

52

Exploring Evolutionary Space

Timothy G. Lilburn[1]* (tlilburn@atcc.org), Yun Bai[2], Yuan Zhang[2], James R. Cole[2], and George M. Garrity[2]

[1]American Type Culture Collection, Manassas, VA and [2]Michigan State University, East Lansing, MI

The use of principal components analysis (PCA) to visualize the evolutionary relationships among thousands of sequences was developed by us as a tool for aiding in the definition of higher-level prokaryotic taxonomy. It not only helped define higher taxa by revealing naturally occurring clusters within the data, but also proved invaluable in highlighting errors in classification and annotation of sequences. Such PCA projections can be viewed as maps of evolutionary space for the sequences and, by extension, the organisms (and genomes) from which the sequences are obtained. Maps based on SSU rRNA sequences show large gaps between some phylogenetic groups. Presumably, this white space is due to the constraints on the evolution of these molecules that arise from their functional requirements. Sequences that might appear there either simply cannot occur in nature or belong to extinct species. Although PCA and other projection techniques can provide a reasonable approximation of the topology hidden within a dataset, some distortion is inevitable and can be attributed to methodological biases and biases that may exist within the data. Previously, we had demonstrated that we could improve the accuracy of projections for a test case having a known topology and coordinate system by using a set of uniformly distributed external benchmarks. However, neither the true topology nor the coordinate system of the prokaryotic evolutionary space has been defined. Therefore, to understand the distortion, we would need to first define the limits of this space. In this study, we examine the use of a limited number (n=179) of internal reference points (benchmarks) on the transformation of the evolutionary distance data into the new coordinate system defined by PCA. We look at ways of making our maps of evolutionary space more accurate and explore why the white space exists. Methods explored include the generation of synthetic polychimeras, *in silico* random mutation, and complementation of a set of 179 proposed benchmark sequences. Our results are presented as a set of PCA plots that are evaluated in terms of their resolution and their concordance with the current taxonomy and with each other.

53

PhyloScan: a New Tool for Identifying Statistically Significant Transcription Factor Binding Sites by Combining Cross-Species Evidence

Lee A. Newberg[1,2]*, C. Steven Carmack[1], Lee Ann McCue[1] (mccue@wadsworth.org), and Charles E. Lawrence[3]

[1]Wadsworth Center, Albany, NY; [2]Rensselaer Polytechnic Institute, Troy, NY; and [3]Brown University, Providence, RI

If there are known transcription factor binding sites (TFBSs) for a particular transcription factor (TF), then it is possible to construct a motif model or position weight matrix with which to scan a

* Presenting author

sequence database for additional sites, thereby predicting a regulon. However, scanning a genome for additional TFBSs typically results in finding few statistically significant sites. Specifically, the statistical significance of a sequence match (p-value) to a motif can be assessed by comparison with the probability of observing a match with a score as good or better in a randomly generated search space of identical size and nucleotide composition -- the smaller the p-value the greater the evidence that the match is not due to chance alone. Staden [1] presented an efficient method that exactly calculates this probability, and Neuwald *et al.* [2] described an implementation of this method. In practice, when scanning a genome or the promoter regions of a genome, it is frequently difficult to identify (below a chosen level of statistical significance) even the known TFBSs that were used in the construction of the motif, to say nothing of additional novel sites for that TF. Essentially, given the statistical nature of this approach, only a relatively small number of TFBSs will be identified that could possibly be considered significant (low sensitivity, high specificity).

With the goal of increasing the statistical power of scanning a genome sequence database with a regulatory motif, we have developed a scanning algorithm, PhyloScan, that combines evidence from matching sites found in orthologous data from several related species. Specifically, we have extended Staden's method [1] to allow scanning of orthologous sequence data that is either multiply-aligned, unaligned or a combination thereof (aligned and unaligned). PhyloScan statistically accounts for the phylogenetic dependence of the species contributing aligned data and returns a p-value for the sequence match; importantly, the statistical significance is calculated directly, without employing training sets.

To evaluate this method we chose the *Escherichia coli* Crp and PurR motifs and gathered genome sequence data for several gamma-proteobacteria. Among the species chosen for this study (*E. coli*, *Salmonella enterica* Typhi, *Yersinia pestis*, *Haemophilus influenzae*, *Vibrio cholerae*, *Shewanella oneidensis*, and *Pseudomonas aeruginosa*), only *E. coli* and *S. typhi* exhibit extensive homology in the promoter regions [3]. Thus we aligned orthologous intergenic regions for these two species, and combine statistical evidence from scanning the aligned *E. coli* and *S. typhi* data with statistical evidence from scanning unaligned orthologous intergenic regions from the remaining five more distantly related species. This method enhances the identification of TFBSs in *E. coli* by several-fold over scanning the set of *E. coli* intergenic regions alone.

1. Staden, R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci* **5:**89-96.
2. Neuwald, A. F., Liu, J. S. and Lawrence, C. E. (1995) Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Sci* **4:**1618-32.
3. McCue, L. A., Thompson, W., Carmack, C. S. and Lawrence, C. E. (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res* **12:**1523-32.

54

Predicting Protein Interactions via Docking Mesh Evaluator

Roummel F. Marcia[1], Susan D. Lindsey[2], Erick A. Butzlaff[1], and Julie C. Mitchell[1]* (mitchell@math.wisc.edu)

[1]University of Wisconsin, Madison, WI and [2]University of California, San Diego, CA

Introduction

The Docking Mesh Evaluator (DoME) is a software package for protein docking and energy evaluation. It uses fast energy evaluation methods and optimization algorithms to predict the docked configurations of proteins with DNA, ligands, and other macromolecules. A fully parallelized package, DoME can run on supercomputers, clusters, and linked independent workstations.

Description

Previously, DoME's energy model was based on solvent effects defined implicitly using adaptive mesh solutions to the Poisson-Boltzmann equation and van der Waals energy terms. While this earlier version achieved moderate success [1], more accurate results were obtained by incorporating hydrogen bond and solvation energy terms. The hydrogen bond term uses the spatial relationship between a potential acceptor atom and a donor-hydrogen pair both to determine the existence of a hydrogen bond and to compute the bond's potential energy. The solvation term uses a modification to the method of Zhang et al. [2] to estimate the effective atomic contact energy of a complex in water. In addition, the energy function employs weighting and switching parameters to measure the individual contribution of each energy term to the overall interaction.

Global optimization methods were developed to determine the lowest function value of this energy model. In particular, the General Convex Quadratic Approximation [3] constructs a sequence of convex underestimators to a collection of local minima to predict possible areas of low energies. Yukawa potentials are used as analytic solutions to the linearized Poisson-Boltzmann equations so that gradient-based local optimization can be performed. Initial scanning of the energy landscape for favorable configurations as initial seeds for underestimation improves algorithm performance. This coupled use of scanning and optimizing is more effective in determining points of low energy values than scanning or optimizing alone [4].

We present results from the standard benchmarking test set of Chen et al. [5] for testing protein-protein docking algorithms. We consider both bound and unbound crystalline structures for determining proper docked configurations. We also highlight which energy terms are significant in each test case. (Of the 59 the benchmarking test set contains, 22 are enzyme-inhibitor complexes, 19 are antibody-antigen complexes, 11 are various complexes, and 7 are difficult test cases whose solutions have considerable structural changes.)

References

1. R.F. Marcia, J.C. Mitchell, and J.B. Rosen, Iterative convex quadratic approximation for global optimization in protein docking, *Comput. Optim. Appl.*, Accepted for publication.
2. C. Zhang, G. Vasmatzis, J.L. Cornette, C. DeLisi, Determination of atomic desolvation energies from the structures of crystallized proteins, *J. Mol. Biol*, 1997 Apr. 4; 267(3); 707-26.

* Presenting author

3. J.B. Rosen and R.F. Marcia, Convex quadratic approximation, *Comput. Optim. Appl.*, 28, pp.173-184, 2004.

4. J.C. Mitchell, J.B. Rosen, A.T. Phillips, and L.F. Ten Eyck, Coupled optimization in protein docking, in *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, ACM Press, 1999, pp. 280-284.

5. R. Chen, J. Mintseris, J. Janin, and Z. Weng, A protein-protein docking benchmark, *Prot. Struct. Fun. Gen.*, 52, pp. 88-91, 2003.

55

UC Merced Center for Computational Biology

Michael Colvin[1]* (mcolvin@ucmerced.edu), Arnold Kim[1], and Felice Lightstone[2]

[1]University of California, Merced, CA and [2]Lawrence Livermore National Laboratory, Livermore, CA

We are establishing a Center for Computational Biology (CCB) at the newest campus of the University of California. The CCB will sponsor multidisciplinary scientific projects in which biological understanding is guided by computational modeling. The center will also facilitate the development and dissemination of undergraduate and graduate course materials based on the latest research in computational biology. The Center is starting a number of activities that aim to recast biology as an information science:

1. Host multidisciplinary research projects in computational and mathematical biology that will provide a rich environment for graduate and undergraduate research.
2. Develop new mathematical and computational methods that are widely applicable to predictive modeling in the life sciences.
3. Develop and disseminate computational biology course materials that translate new research results into educational resources.
4. Extend the successes in achieving these objectives to other universities and "university-feeder" institutions such as community colleges.

This project is a multi-institutional collaboration including the new University of California campus at Merced, Rice University, Rensselaer Polytechnic Institute, and Lawrence Livermore National Laboratory, as well as individual collaborators at other sites. The CCB will foster a number of research projects that emphasize the role of predictive simulations in guiding biological understanding by funding graduate students and post-doctoral fellows with backgrounds in the mathematical and computational sciences to work on collaborative biology projects. Additionally, the center will work to translate this research into educational materials. UC Merced, as the first U.S. research university of the 21st century, offers many advantages for this new center, including an ideal venue for developing and implementing new courses and degree programs, a highly multidisciplinary faculty and organizational structure, and a strong commitment to educational outreach to diverse and underrepresented groups. New computational biology courses will be used and assessed in the new UC Merced Biological Sciences major that will be accepting freshman and junior transfers in Fall 2005. Graduate courses will be implemented in the Quantitative Systems Biology Graduate Group that began accepting graduate students in Fall 2004. All course materials will be released under an open

* Presenting author

public license using the Connexions courseware system developed at Rice University. We anticipate that this new biology curriculum will be effective in attracting students to biology who have an interest and aptitude in mathematics and computational sciences, as well as broaden the horizons of students expecting a traditional biology program. The electronic, modular course materials produced by the center will facilitate linkages to feeder schools at the state university, community college, and high school levels.

The long-term impact of the CCB will be to help train a new generation of biologists who bridge the gap between the computational and life sciences and to implement a new biology curriculum that can both influence and be adopted by other universities. Such scientists will be critical to the success of new approaches to biology, exemplified by the DOE Genomics:GTL program in which comprehensive datasets will be assembled with the goal of enabling predictive modeling of the behavior of microbes and microbial communities, as well as the biological components of life, such as multiprotein machines.

56

Biomic Approach to Predictive Cell Modeling

P. J. Ortoleva* (ortoleva@indiana.edu), L. Ensman, J. Fan, K. Hubbard, A. Sayyed-Ahmad, F. Stanley, K. Tuncay, and K. Varala

Indiana University, Bloomington, IN

Replication, transcription, translation and metabolism, as well as physiological dynamics, are all strongly coupled and must be accounted for if predictive cell modeling is to be attained. Key barriers to doing so are that many processes are not yet understood in detail, and that the phenomenological parameters in a kinetic cell model are not yet calibrated. In this presentation we show how multiplex data (notably cDNA microarray time series) and incomplete cell models can be integrated via information theory to overcome these difficulties.

A kinetic cell model is expressed as $\frac{\partial \Psi}{\partial t} = F\left[\Psi, \Lambda, \Phi\right]$ (1)

where Ψ is a set of descriptive variables (RNA populations, metabolite, concentrations, etc.), Λ is a set of model parameters and Φ is a set of cell descriptive variables for which we do not have governing equations (e.g. concentrations of species for which reactions have not yet been delineated). The problem is that (1) we cannot run such an incomplete model as the timecourse $\Phi(t)$ is required; and (2) therefore we cannot calibrate the model parameters Λ.

To overcome these difficulties, in our procedure we use information theory to construct the probability ρ that is a function of Λ and a functional of the timecourse $\Phi(t)$. We use available data (notably cDNA microarray, proteomics, NMR) in steadv-state or time series to construct ρ. With this we seek the most probable values of $\Lambda,\Phi(t)$ by solving $\partial\rho/\partial\Lambda = 0$, $\delta\rho/\delta\Phi(t) = 0$, the latter being a functional differential equation for the timecourse $\Phi(t)$. Regularization is also used to insure that there are no unphysically short timescale effects in $\Phi(t)$ created by the sparseness of, or noise in, the data.

The above methodology has been implemented for the case of transcription/translation modeling integrated with cDNA microarray time series analysis. The input of our software is microarray time series and a putative transcription control network (the latter being incomplete and possibly error-

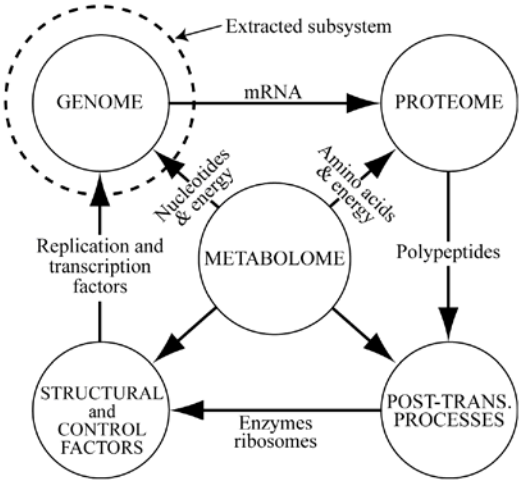
* Presenting author

prone). Output is calibrated transcription factor (TF) binding constants and transcription rate coefficients for all genes. Degradation rate constants for each type of RNA are also provided as are the timecourse of TF thermodynamic activities.

Our methodology has been tested on *E. coli* where we have identified some likely mistakes in existing *E. coli* networks and have delineated the TF timecourse that coordinate the change in transcription patterns accompanying a change in carbon source in the surroundings. The method is found to be very robust to omnipresent noise in the microarray data and to allow one to discover errors in the proposed regulatory network or to discover new TF/gene interactions. Possible interactions identified using sequence analysis are screened and their up/down regulatory function is identified.

The implications of our approach are far-reaching. The approach is applicable to large systems (thousands of genes, hundreds of transcription factors). Its multiplex and automated character will greatly accelerate the delineation of the gene regulatory network. The many rate and thermodynamic constants calibrated will make cell biomic modeling feasible. As suggested in Fig. 1, our approach allows for the piecewise development and calibration of a cell biomic model.

Figure 1. An extracted subsystem (here the genome) can be run and calibrated using probability functional information theory.



The hierarchical nature of the organization of intracellular structure and their multiple dimensional character are key to cell function. A cell must be understood in terms of its specialized zones wherein reaction and transport occur and molecules are exchanged among these zones. In summary, a cell is a very complex molecular processor that involves dynamic on fibrils (1D), membranes (2D) and with bulk medium (3D). The multiple dimensional character of intracellular dynamics is accounted for in CellX which accounts for reaction-transport dynamics along membranes embedded in bulk media, and the exchange among these zones via boundary conditions. Recent experimental studies suggest that MinC, MinD, and MinE proteins play a key role in the location of the Z-ring. The absence of Min dynamics results in location of Z-rings near the poles and imprecise cell division. MinC is an inhibitor to the formation of the Z-ring. MinC and MinD oscillations are observed to be in phase whereas MinE oscillation is coupled to MinC and MinE dynamics. Results for the autonomous localization of the division plane and the segregation of two daughter chromosomes will be presented as example applications of CellX.

57

The BioWarehouse System for Integration of Bioinformatics Databases

Tom Lee, Valerie Wagner, Yannick Pouliot, and Peter D. Karp* (pkarp@ai.sri.com)

SRI International, Menlo Park, CA

BioWarehouse [1] is an open-source toolkit for constructing bioinformatics database (DB) warehouses. It allows different users to integrate collections of DBs relevant to the problem at hand. BioWarehouse can integrate multiple public bioinformatics DBs into a common relational DB management system, facilitating a variety of DB integration tasks including comparative analysis and data mining. All data are loaded into a common schema to permit querying within a unified representation.

BioWarehouse currently supports the integration of Swiss-Prot, TrEMBL, ENZYME, KEGG, BioCyc, NCBI Taxonomy, CMR, and the microbial subset of Genbank. Loader tools implemented in the C and Java languages parse and load the preceding DBs into Oracle or MySQL instances of BioWarehouse.

The presentation will provide an overview of BioWarehouse goals, architecture, and implementation. The BioWarehouse schema supports the following bioinformatics datatypes: chemical compounds, biochemical reactions, metabolic pathways, proteins, genes, nucleic acid sequences, features on protein and nucleic-acid sequences, organism taxonomies, and controlled vocabularies.

BioWarehouse is in use by several bioinformatics projects. An SRI project is developing algorithms for predicting which genes within a sequenced genome code for missing enzymes within metabolic pathways predicted for that genome [2]. BioWarehouse fills several roles within that project: it is used to construct a complete and nonredundant dataset of sequenced enzymes by combining protein sequences from the UniProt and PIR DBs, and by removing from the resulting dataset those sequences that share a specified level of sequence similarity. Our current research involves extending the pathway hole filling algorithm with information from genome-context methods such as phylogenetic signatures, which are obtained from BioWarehouse thanks to the large all-against-all BLAST results stored within CMR. Another SRI project is comparing the data content of the EcoCyc and KEGG DBs using BioWarehouse to access the KEGG data in a computable form.

BioWarehouse is supported by the Department of Energy and by DARPA through the DARPA BioSPICE program for biological simulation.

1. BioWarehouse Home Page http://bioinformatics.ai.sri.com/biowarehouse/
2. Green, M.L. and Karp, P.D., "A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases," *BMC Bioinformatics* 5(1):76 2004 http://www.biomedcentral.com/1471-2105/5/76.

 * Presenting author

58

Building Large Biological Dynamic Models of *Shewanella oneidensis* from Incomplete Data

Ravishankar R. Vallabhajosyula[1*](rrao@kgi.edu), Sri Paladugu[1], Klaus Maier[2], and Herbert M. Sauro[1]

[1]Keck Graduate Institute, Claremont, CA and [2]University of Stuttgart, Stuttgart, Germany

The objective of this study is to investigate certain issues related to building large dynamic models of reaction networks. In particular, emphasis is placed on the performance of computational methods, both in terms of accuracy and computational time. In addition, we are also interested in methods for approximating reaction rate laws when kinetic information is not readily available.

To investigate these questions we have constructed a hybrid model that is a first attempt at a model of *Shewanella oneidensis MR-1* that includes Glycolysis and TCA cycle. Understanding how energy pathways function in *S. oneidensis* is very important to modelling the interaction of this organism with its environment.

The SBML files for Glycolysis and TCA Cycle for *Shewanella* are available from KEGG [1]. These were used to assist in the construction of a test model. However, there is a lack of kinetic information for the dynamics of the underlying metabolic reactions in the KEGG database. To overcome this problem, estimates were made by comparing data from similar pathways in *Escherichia Coli*. While we were able to adapt a kinetic model of glycolysis from work published by [2], the data for the TCA cycle was still lacking. A hybrid model using Linlog kinetics was constructed for the unknown reaction kinetics in the TCA Cycle.

Approximating Rate Laws

Linlog kinetics [3] were recently developed as an attractive alternative to the commonly used Michaelis-Menten like kinetics. The underlying theoretical framework for this approach is described using an example of a branched pathway in [4]. Linlog kinetics provides a very good approximation to Michaelis-Menten kinetics and requires fewer parameters. The equations for Linear, Power-Law and Linlog approximations to Michaelis-Menten kinetics, carried out about an operating effector concentration S_0 are shown in Fig. 1.

Fig. 1 Equations for Michaelis-Menten and other Approximation Methods

$$\text{Michaelis-Menten} \quad V = \frac{V_{max}S}{K_m+S}$$

$$\text{Linear Approximation} \quad V = mS + c, \qquad \text{where } m = \frac{V_{max}K_m}{(K_m+S_0)^2} \text{ and } c = \frac{V_{max}S_0}{(K_m+S_0)^2}$$

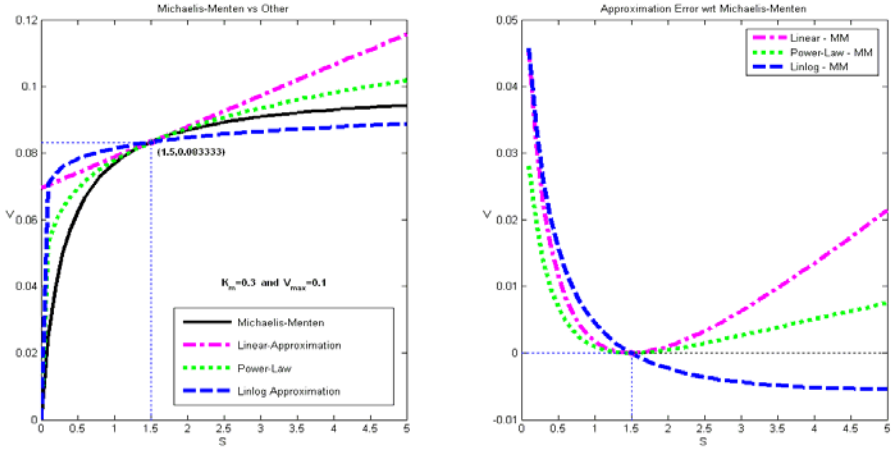
$$\text{Power-Law Approximation} \quad V = \frac{V_{max}S_0}{K_m+S_0}\left(\frac{S}{S_0}\right)^{\beta}, \qquad \text{where } \beta = \frac{K_m}{K_m+S_0}$$

$$\text{Linlog Approximation} \quad V = \frac{V_{max}S_0}{K_m+S_0}.\left[1+\epsilon\, ln\left(\frac{S}{S_0}\right)\right], \qquad \text{where } \epsilon = \frac{K_m}{K_m+S_0}$$

The equation describing the Linlog kinetics is a combination of a linear term with a logarithmic component involving the effector concentrations, scaled appropriately with the respective elasticities.

A comparison of Linear, Power-law and Linlog approximations with Michaelis-Menten kinetics shows that Linlog has the least error over a wide range of effector concentrations. This comparison is shown in Fig. 2.

Fig. 2 Linear, Power-Law and Linlog Approximations vs. Michaelis-Menten kinetics



Performance Issues

The *Shewanella oneidensis* model was simulated using various software tools to investigate their performance given a complex network. These include publicly available simulators SCAMP and Jarnac, simulators built with languages such as FORTRAN, Java, C and C#, as well as simulators derived from commercially available software packages which included Mathematica and Matlab.

This study will provide useful insights into building a more powerful simulator that can handle especially complex networks necessary for carrying out large scale simulations. In this regard, it is essential to understand how existing simulators are structured internally. Since each of these simulators has a different approach to generating the solutions, one may be better than others at simulating a given network. For example, SCAMP and Jarnac generate optimized internal byte-code. Simulators for C and FORTRAN are based on compiled code. The C# and Java simulators are based on byte-code interpretation. The Java simulator code was generated in two ways, 1) The model equations were interpreted at run-time, and 2) A Java class was generated to function as a solver. Matlab performance was tested by generating a standard Matlab ODE function from the SBML of the *Shewanella* model. Mathematica code was similarly generated. All code other than Jarnac and SCAMP was generated using SBW SBML translator modules.

The results from the simulation of the combined Glycolysis and TCA cycle pathway in *Shewanella oneidensis* will be presented in the poster at the next GTL conference in 2005.

References

1. KEGG: Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg
2. C. Chassagnole, N. Noisommit-Rizzi, J.W. Schmid, K. Mauch, M. Reuss. "Dynamic Modeling of the central carbon metabolism of Escherichia Coli". *Biotechnol. Bioeng.*, **79**(1), pp.53-73, (2002)
3. L. Wu, W. Wang, W.A. van Winden, W.M. van Gulik and J.J. Heijnen. "A new framework for the estimation of control parameters in metabolic pathways using lin-log kinetics", *Eur. J. Biochem.*, **271**, pp.3348-3359, (2004).

* Presenting author

4. D. Visser, J.J. Heijnen. "Dynamic Simulation and metabolic re-design of a branched pathway using Linlog kinetics", *Metabolic Engineering*. **5**, pp.164-176, (2003).

59

A Bayesian Method for Identifying Missing Enzymes in Predicted Metabolic Pathway Databases

Michelle L. Green* (green@ai.sri.com) and Peter D. Karp

SRI International, Menlo Park, CA

The PathoLogic program constructs Pathway/Genome databases by using a genome's annotation to predict the set of metabolic pathways present in an organism. PathoLogic determines the set of reactions composing those pathways from the enzymes annotated in the organism's genome. Many enzymes in a genome may be missed during the initial annotation effort or may be assigned a non-specific function (e.g., "thiolase family protein"). These missing or incomplete annotations can result in *pathway holes*. Pathway holes occur when a genome appears to lack the enzymes needed to catalyze reactions in a pathway. If a protein has not been assigned a specific function during the annotation process, any reaction catalyzed by that protein will appear as a missing enzyme or pathway hole in a Pathway/Genome database.

We have developed a method [1] that efficiently combines homology and pathway-based evidence using Bayesian methods to identify candidates for filling pathway holes in Pathway/Genome databases. Our program, which is now part of the Pathway Tools software, identifies potential candidate sequences for pathway holes, and combines data from multiple, heterogeneous sources to assess the likelihood that a candidate has the required function. By considering not only evidence from homology searches, but also genomic and functional context (e.g., are there functionally-related genes nearby in the genome?), our algorithm emulates the manual sequence annotation process to determine the posterior belief that a candidate has the required function. The method can be applied across an entire metabolic pathway network and is generally applicable to any pathway database. We achieved 71% precision at a probability threshold of 0.9 during cross-validation using known reactions in computationally-predicted pathway/genome databases.

After applying our method to 255 pathway holes in 99 pathways from the CauloCyc database, the predictions from this program completed fourteen additional pathways. The program made putative assignments to 53 pathway holes, including annotation of 2 sequences of previously unknown function. The newly completed pathways include "fatty acid oxidation pathway", "oxidative branch of the pentose phosphate pathway", "peptidoglycan biosynthesis", "pyridine nucleotide biosynthesis", "pantothenate and coenzyme A biosynthesis", "de novo biosynthesis of pyrimidine ribonucleotides", "de novo biosynthesis of purine nucleotides II", "histidine biosynthesis I", "tyrosine biosynthesis I", "phenylalanine biosynthesis II", and "alanine biosynthesis I".

1. Green, M.L. and Karp, P.D., "A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases," *BMC Bioinformatics*, 5:76 2004. http://www.biomedcentral.com/1471-2105/5/76.

60

Does EcoCyc or KEGG Provide a Preferable Gold Standard for Training and Evaluation of Genome-Context Methods?

Peter D. Karp* (pkarp@ai.sri.com) and Michelle L. Green

SRI International, Menlo Park, CA

Motivations: Genome-context methods such as phylogenetic profiles infer functional associations between genes and constitute a new approach to prediction of gene function. Most past developers of genome-context methods have trained and validated their algorithms against metabolic pathway DBs: primarily KEGG [1] and EcoCyc [2]. This work addresses the question of which of these two DBs is the optimal resource for training and evaluation of genome-context methods. Our hypothesis is that EcoCyc is the preferable resource to use because: (a) KEGG pathway maps contain many false-positive functional associations because KEGG maps tend to be much larger than EcoCyc pathways, and therefore contain genes from many different biological pathways, (b) KEGG pathways are less accurate because they are computationally predicted whereas EcoCyc pathways are curated from the biomedical literature, and (c) EcoCyc contains other types of functional relationships besides the metabolic and two-component signal transduction pathways that KEGG contains, such as descriptions of regulatory relationships between transcription factors and other genes.

Method: We evaluated this hypothesis by randomly choosing pairs of genes from the same KEGG and EcoCyc pathways, and counting the frequency with which those gene pairs show chromosomal adjacency, or similar phylogenetic profiles, since these basic methods for predicting functional associations are shown to have utility by both EcoCyc and KEGG.

Results: The hypothesis is validated by the following results. Two genes chosen at random from the same EcoCyc pathway were 4.7 times more likely to be adjacent on the chromosome than two genes chosen at random from the same KEGG map. Furthermore, gene pairs chosen from the same KEGG map that are not in the same EcoCyc pathway are even less likely to be chromosomally adjacent or to exhibit similar phylogenetic profiles. Similar results were obtained for the BioCyc and KEGG datasets for seven other organisms. In addition, two genes chosen at random from the same EcoCyc pathway were 3.0 times more likely to have similar phylogenetic profiles than two genes chosen at random from the same *E. coli* KEGG map. In addition, we find that transcription factors and the genes that they regulate show significant chromosomal adjacency and similar phylogenetic profiles.

Summary: EcoCyc and the BioCyc DBs for other organisms are preferable resources for training and evaluation of genome-context methods because their pathways are more biologically meaningful, and because they contain a wider range of biological functional associations, such as those between transcription factors and the genes they regulate, thus allowing genome-context methods to recognize a wider set of biological relationships.

1. M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya, "The KEGG databases at GenomeNet," Nucleic Acids Research 40:42-6 2002.
2. I.M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I.T. Paulsen, M. Peralta-Gil, and P.D. Karp, EcoCyc: a comprehensive database resource for *E. coli, Nucleic Acids Research*, Database Issue, 2005 (in press).

* Presenting author

61

Towards a Physics and Systems Understanding of Ion Transport in Prokaryotes

Shreedhar Natarajan[1], Asba Tasneem*[1], Sameer Varma[1], Lakshminarayan Iyer[2], L. Aravind[2], and Eric Jakobsson*[2] (jake@ncsa.uiuc.edu)

[1]University of Illinois, Urbana, IL and [2]National Institutes of Health, Bethesda, MD

Ion transport mechanisms play three fundamental roles in biological systems: 1) generation and sensing of electrochemical signals, 2) generation of osmotic force for regulating water flow, and 3) energy transduction. It is useful to study these functions in an integrated manner because: 1) Ion transporters as they perform all three functions pose similar issues in understanding the physical bases of those functions, and 2) it is common for the same transporter to be critical for more than one function. Indeed, it is universally true that transporters which are critical for one function will affect the functioning of networks of transporters critical to other functions, because every transporter in a membrane affects osmotic, chemical, and electrical driving forces for every other transporter in the same membrane.

In this paper we report on our efforts along three lines: 1) By bioinformatics means we have discovered homologues of chemo-sensitive postsynaptic ion channels in prokaryotes, including in *Synechococcus,* which is the organism whose GTL project we are associated with. This discovery may point to a previously unknown molecular mechanism for electrochemical signaling between prokaryotes, and may also facilitate determination of structures for the ligand-gated channel family of gene proteins. 2) We have developed improved methods for assigning protonation states of electrically interacting titratable residues in the lumen of bacterial porins. This is critical in order to make realistic models for ion permeation through these channels. 3) We are adapting phylogenetic profiling methods to infer transport and regulatory networks that govern ion and water homeostasis in prokaryotes.

Supported by grant # 0235792 from the National Science Foundation, the U.S. Department of Energy's Genomics: GTL program (www.doegenomestolife.org) under project, "Carbon Sequestration in *Synechococcus Sp.*: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org), and the intramural program of the National Center for Biotechnology Information.

62

OptStrain: A Computational Framework for Redesign Microbial Production Systems

Priti Pharkya and Costas D. Maranas* (costas@psu.edu)

Pennsylvania State University, University Park, PA

In this talk, we will discuss the hierarchical computational framework ***OptStrain*** aimed at guiding pathways modifications, through reaction additions and deletions, of microbial networks for the overproduction of targeted compounds. A comprehensive database of biotransformations, referred to as the Universal database (with over 5,700 reactions), is compiled and regularly updated by downloading and curating reactions from multiple biopathway database sources. Combinatorial optimization is then employed to elucidate the set(s) of non-native functionalities, extracted from this Universal database, to add to the examined production host for enabling the desired product formation. Subsequently, competing functionalities are identified and removed to ensure higher product yields coupled with growth. This work establishes an integrated computational framework capable of constructing stoichiometrically balanced pathways, imposing maximum product yield requirements, pinpointing the optimal substrate(s), and evaluating different microbial hosts. The range and utility of ***OptStrain*** is demonstrated by addressing a variety of product molecules and experimental verifications.

63

DEMSIM: A Discrete Event Based Mechanistic Simulation Platform for Gene Expression and Regulation Dynamics

Madhukar Dasika and Costas D. Maranas* (costas@psu.edu)

Pennsylvania State University, University Park, PA

The advent of high-throughput technologies has provided a major impetus for developing sophisticated computational frameworks to unravel the underlying regulatory circuitry that governs the response of biological systems to environmental and genetic perturbations. A systems engineering view reveals that gene expression dynamics are governed by processes that are essentially event driven, i.e., many events have to take place in a predetermined order with uncertain start and execution times to accomplish a certain task. There are many parallels between gene expression and manufacturing systems. In analogy to a manufacturing facility which produces a certain amount of finished product at a particular time with a certain probability, the transcription process produces mRNA transcripts with probability determined by the cellular environment and availability of required components. Similarly, accumulating mRNA and protein levels in the cell are akin to product inventory held in warehouses in a manufacturing system. Motivated by the numerous parallels between these two seemingly different settings, we have used discrete event simulation, which is a powerful tool employed to model and simulate supply chains and manufacturing systems, to model and simulate gene expression systems. In this talk, we will describe the DEMSIM tool that we have developed to test and hypothesize putative regulatory interactions.

* Presenting author

The key feature of the DEMSIM platform is the abstracting of underlying transcription, translation and decay processes as stand-alone modules. A module is further characterized by a sequence of discrete events. For example, the transcription module is composed of the following sequence of discrete events: (i) binding of RNA polymerase to promoter sequence (ii) transcription elongation and (iii) transcription termination. Each module is described by a set of physical and model parameters. Physical parameters correspond to parameters which are known *a priori* from literature sources and are fixed within the simulation framework (e.g. length of gene, transcription rate, etc.). In contrast, model parameters are regression parameters that are fitted using the available experimental data. Subsequently, the simulation is driven by the communication between these modules in accordance with the specifics of the regulatory circuitry of the biological system being investigated. The stochasticity inherent to all events is captured using Monte Carlo based sampling.

The DEMSIM software implementation consists of the following three key components: (i) an event list that contains all the events that need to be executed along with their respective execution times, (ii*)* a global simulation clock that records the progress of simulation time as events are sequentially executed, and (iii) a set of state variables that characterize the system and which are updated every time an event is executed. At every time step, events corresponding to all active (non–terminated) modules in the system are included in the event list. Subsequently, the event list is sorted and the event having the smallest execution time is executed. The simulation clock is advanced and the execution time of all other events is updated. Such a sequential procedure prevents the occurrence of causality errors by ensuring that an event with a later time stamps is not executed before an event with an earlier time stamp. Furthermore, since the execution of certain events leads to the creation of new modules and the termination of existing ones, the number of active modules in the system is updated and new events are included in the event list. This procedure is then repeated for the duration of the simulation horizon and state variables such as number of mRNA and protein molecules are recorded.

In this talk we will present the results for three biological systems of different levels of complexity that we have used to benchmark the DEMSIM platform. Simulation results for the relatively simple *lac* operon system of *E. coli* will demonstrate that the parameters embedded in the framework can indeed be trained to reproduce experimental data. Subsequently, the ability of the framework to serve as a predictive tool will be highlighted with reference to the SOS response system of *E. coli*. Simulation results will focus on DEMSIM's ability to accurately predict the *de novo* response of the system to externally imposed perturbations. Finally, simulation studies for the *araBAD* system will demonstrate the framework's ability to distinguish between different plausible regulatory mechanisms postulated to explain observed gene expression profiles. Overall, the presented results will highlight the broad applicability of the discrete-event paradigm, on which DEMSIM is based, to gene regulatory systems.

64

On the Futility of Optima in Network Inferences and What Can Be Done About It

Charles (Chip) E. Lawrence* (lawrence@dam.brown.edu)

Brown University, Providence, RI

Inference of metabolic and regulatory networks has become a hot topic over the last few years, and a number of reports of efforts to infer such networks using genome scale data have appeared. One factor often overlooked in these fledgling efforts stems from the large size of network solution spaces, which grow exponentially with the number of nodes in graphical representation of the network. The difficulty of drawing inferences in high dimensional setting is a well-recognized problem in statistical learning theory that is often enunciated as the "curse of dimensionality". Most explanations of this curse don't convey well its connection with inferences and tend to be abstract. Here I'll use a Bayesian framework to directly illustrate how difficulties in network inference grows rapidly with network size, illustrate how the curse of dimensionality can so easily render even guaranteed optimal solutions unlikely, and show what can be done about it. Specifically, we will see that the curse stems from the large number of terms in the normalizing constant in Bayes rule, and that the individual terms define opportunities for limiting the curse's ill effects. I'll use a special class of planar networks, whose structure allows for guaranteed optimal solutions and a guaranteed representative sampling, to show how optimal are often misleading, but nevertheless likely network connections and excluded connections can be strongly inferred based on samples from the solution space. I'll also show how incompleteness or inaccuracy of an underlying model of a network can yield optimal solutions that are not even close to right. Background material on Bayesian inference and its connection with inferences based on optimality methods will be given as a aid for those who may not be fully familiar with statistical inference methods.

* Presenting author