Research Statement

My research is focused on the development of rigorous mathematical and practically feasible approaches, algorithms and computational frameworks to analyze 1) the complexity observed in biology, 2) mathematical models to explain details of the cellular behavior, and 3) raw data available from literature, databases and the Web.

1. Modeling genome-scale metabolic networks

Metabolic reconstructions and stoichiometric matrices. This year the 50th anniversary of the discovery of the double helix was celebrated and the Human Genome Project announced its completion of a "final draft" of the DNA sequence for *Homo Sapiens*. In the future, the rapid accumulation of large amounts of genomic data will persist and the analysis of genomes will continue to represent one of the most significant areas of science.

Given a particular sequenced organism, the organism's metabolic reaction network can be reconstructed from its annotated genome. The process of mapping genomes into metabolic networks is called *metabolic reconstruction*. Metabolic reconstructions provide solid knowledge of the organisms' reaction networks topology and stoichiometry. The organism's stoichiometry is represented in a *stoichiometric matrix* with matrix elements quantifying the stoichiometric coefficients of the individual metabolites in each reaction.

Up to now hundreds of metabolic reconstructions have been generated and this process will exponentially accelerate in academia and industry. Given the increasing volume of metabolic information, the following two important questions arise: 1) Can stoichiometric matrices provide new insights into characterization of underlying genomes? and 2) Can such matrices unravel novel knowledge of the cellular metabolism? Below I will address these and other issues related to genome-scale stoichiometric modeling.

Genome-scale stoichiometric modeling. Kinetic modeling of the cellular metabolism has been successfully used for a long period of time (Reich and Selkov, 1981). However elegant, kinetic modeling approaches do not scale well to genome-scale models of complex microorganisms. The two fundamental obstacles are the lack of kinetic data and the size of models. Indeed, values of thousands of kinetic constants are unknown. Besides, kinetic mechanisms and regulation are also unknown for many enzymes. Even if all the data were known, additionally a typical genome-scale model would include thousands of variables and equations to analyze. The analysis of such huge models is a computationally challenging task.

As a remedy to this, a genome-scale modeling approach, termed *flux balance analysis* or *stoichiometric modeling*, has been recently developed and reviewed (Reed and Palsson, 2003). Achievements in stoichiometric modeling include simulation of cellular responses to addition and deletion of genes, events in external media, yield improvements of useful biochemicals, *etc.* (Edwards *et al.*, 2002).

A stoichiometric model encompasses: 1) a set of mass balances to describe distributions of steady state fluxes in the network, 2) thermodynamics to pinpoint irreversible reactions, and 3) "boundary conditions" to set transport fluxes across the cell's systemic boundary. Stoichiometric models are usually highly undetermined and, so, they are often formulated as linear programs where a specific biomass growth is optimized.

Since in stoichiometric models fluxes are viewed as independent variables, due to the omission of the functional dependence on concentrations and kinetic data, the models provide the broadest feasible boundary of flux distributions potentially available to the cell. In contrast, if all kinetic data were known, the fluxes and concentrations would be determined uniquely.

Complexity of genome-scale stoichiometric modeling. During the last decade, the following two approaches to the flux analysis of metabolic networks have been primarily developed: 1) The linear algebra approach, based on the matrix rank and null spaces (Heinrich and Schuster, 1996), and 2) The convex analysis approach, based on elementary flux modes (Heinrich and Schuster, 1996) and extreme pathways (Schilling *et al.*, 2000). Extreme pathways are the edges of the convex cone generated by the mass balances in the flux space. Elementary modes correspond to minimal sets of reactions that can simultaneously operate with nonzero fluxes.

Any flux distribution in the whole network can be represented as a sum of extreme pathways or elementary modes taken with non-negative coefficients.

Stoichiometric matrices are usually rank deficient, *i.e.* the rank is less than the number of rows and columns. Computation of ranks and null spaces of such big matrices is a nontrivial exercise due to numerical roundoff errors (Golub and van Loan, 1996). It is also known that elementary flux modes and extreme pathways result in combinatorial explosion. For instance, the central metabolism of *Escherichia coli*, containing only 110 reactions, gives rise to 43,279 elementary modes (Stelling *et al.*, 2002). At the same time, the genome-scale model of *Escherichia coli,* a typical workhorse procariotic organism, includes about 800 metabolites and 1000 reactions (Reed *et al.*, 2003). Given that, new approaches to the analysis of stoichiometric models should be developed.

Flux-coupling analysis. In a recent study we have analyzed how the general design principles of genome-scale metabolic networks can introduce meaningful relationships between fluxes (Burgard, Nikolaev et al., 2003) and metabolites (Nikolaev et al., 2003). As a result, a new approach based on the computation of maximal and minimal ratios for each pair of flux variables was developed. Mathematically, a series of equivalent linear programs was formulated to model responses of a metabolic network to perturbations of a single reaction flux. While a unit flux for one reaction is set, minimal and maximal values of the other reaction fluxes are computed. Flux ratios can be classified as follows: 1) If the ratio between two fluxes ranges from zero to infinity, the fluxes are *completely uncoupled*; 2) If the ratio ranges from zero to a finite constant or from a finite constant to infinity, the fluxes are *directionally coupled*; 3) If the flux ratio varies between two finite constants, then the fluxes are *partially coupled*; and 4) If the ratio is constant, the fluxes are *fully coupled.*

The flux coupling analysis provides surprisingly diversified information on stoichiometry-driven events in the cellular metabolism. Some of these include the structural analysis of pathways and their functions, equivalent knockouts and prediction of operons, *etc.* For instance, the three major metabolic pathways, glycolysis, the pentose phosphate (PPP), and the TCA cycle show significant internal coupling while being completely decoupled from one another. In other words, no glycolysis flux is capable of forcing flux through the TCA cycle or PPP based on stoichiometry alone. Directionally coupled fluxes can reveal reactions that should be suppressed to prevent or block the nonzero flux through a particular reaction (equivalent knockouts). Also, information on coupled fluxes can be used to estimate intracellular fluxes coupled to transport fluxes that can be measured and controlled.

Computational requirements for the flux-coupling analysis are in the order of 15-40 minutes for genome-scale models involving as many as 1,173 reactions upon utilizing the LINDO optimization solver (Lindo Systems, Inc.), accessed via C++ on an Intel Pentium IV, 2.4 GHz, 512 MB RAM computer. CPU times for the computation of one minimal or maximal flux ratio are in the order of milliseconds.

Prediction of operons. In prokaryotes, operons are two or more genes whose expression can be co-ordinated. The genes organized in operons are often – though not necessarily – functionally related (*e.g.*, they may encode enzymes of a particular metabolic pathway). The organization of genes into an operon may allow their expression to be co-coordinately "turned on" (*induced*) or "turned off" (*repressed*) according to the cell's needs.

Partially and fully coupled reactions can bear nonzero fluxes only if the enzymes catalyzing these reactions are simultaneously active. This suggests that enzyme subsets corresponding to flux-coupled reactions may be encoded by genes from common operons. As a result, we have developed a new computational technology and software to identify subsets of coherently regulated enzymes (Maranas, Nikolaev, and Burgard, 2003).

Comparisons of enzyme subsets with operons available from the Regulon DB database (Salgado *et al.*, 2001) has revealed that about 30% of the enzyme subsets, identified from the model for *Escherichia coli*, include two or more genes from common operons. Almost half of such subsets correspond exactly to operons. One can hope that when the quality of metabolic reconstructions improves, predictions from modeling studies will also improve.

Scale-free nature of flux centered metabolic graphs. To evaluate integrity and complexity of stoichiometric couplings between reaction fluxes the number $N(k)$ of nodes/fluxes implying k active fluxes was computed.[1]

[1] If nonzero flux v_1 through reaction R_1 assumes that reaction R_2 must have nonzero flux v_2, then flux v_1 is said to imply flux v_2.

Specifically, our goal was to see if directional coupling among metabolic fluxes is scale free, characterized by a relatively small number of well connected nodes, $N(k) \propto k^{-d}$, or random, where the number of arcs associated with each node follows a Poisson distribution. Scale-free distributions for three different organisms were found, *Escherichia coli*, a simple prokaryotic parasite inhabiting the human stomach (*Helicobacter pylori*), and Baker's yeast, a eukaryotic cell (*Saccharomyces cerevisiae*).

Unlike numerous previous studies (Jeong *et al.*, 2000; Dorogovtsev and Mendes, 2003) the nodes in flux-centered graphs denote metabolic functionalities (fluxes) and not metabolites or reactions. Two nodes from flux-centered graphs are connected if the corresponding fluxes are fully, partially, or directionally coupled. Genes corresponding to reactions with the most connected fluxes can be crucial in a sense that mutations or knockouts of such genes can significantly alter the whole cell's functioning or even lead to cell death. Therefore, the vulnerability of the cellular metabolism to genetic alterations can now be more directly assessed.

Genome-scale metabolic pools. An organism's stoichiometric backbone establishes certain barriers and limits on both reaction fluxes and metabolic concentrations. In particular, these can easily thwart any biotechnological or medical objectives if not well understood. For instance, often, the aim of metabolic engineering or medical treatment is to increase concentrations of some chemicals inside the cell. If the changes in the needed concentrations are constrained by conservation relationships, linear combinations of metabolites that do not change over time, the concentrations may reach internal steady states long before a desired metabolic shift occurs. A simple example is the conservation relationship [ATP] + [ADP] + [AMP] = const corresponding to a pool of adenine nucleotides ATP, ADP, and AMP. An increase in the concentration of any metabolite will be compensated by draining of the other metabolites from the pool.

In the ongoing research (Nikolaev et al., 2003) we have developed a new optimization-based approach to infer meaningful information on conserved metabolic pools. Whereas steady state fluxes can be found from the analysis of the right null space of a stoichiometric matrix, conservation relationships can be obtained from the analysis of the left null space of the same matrix. The previously developed approaches are based on computation of *all* elementary conserved moiety vectors (Heinrich and Schuster, 1996) and extreme concentration pools (Famili and Palsson, 2003). Combinatorial explosion of such solutions in genome-scale models is well known. However, in many practical cases, only particular metabolic pools can present interest. For instance, in biotechnology disrupting pools constraining concentrations of useful amino acids can help increase their yields. The new approach allows one to locate 1) all metabolites coupled within the same conserved pools; 2) metabolites absent from all conserved pools; and 3) pools of a particular interest. In cases 1) and 2) none of actual metabolic pools is computed.

StoichPro (eCell): A stoichiometric modeling computational engine. At present there are few computational tools that allow one to carry out stoichiometric modeling of metabolic networks. Some of these tools are developed to solve specific optimization problems and in most cases the source code is closed. To efficiently develop new algorithms and experiment with real data, I have implemented an STL/C++ stoichiometric modeling environment which can be viewed as a computational "dry lab." It includes 1) an electronic cell (eCell), based on recursive principles; 2) administrative routines to efficiently control data flows, memory allocation, error exceptions, *etc.*; 3) auxiliary utility routines, and 4) C++ abstract data types or wrappers to incorporate different optimization engines and codes written by third parties. A complete stoichiometric matrix is never stored and, instead, is dynamically assembled from stoichiometric equations of particular reactions. Also, a three vector representation of sparse matrices is used. This allows one to efficiently cope with typical stoichiometric matrices which can include up to 10^6 matrix elements with just less than 0.7% of nonzero elements.

2. Future plans and research directions

At present, the bioinformatics stage of using large amounts of accumulated data to form an inventory of genes and metabolic subsystems is progressing rapidly, and we are entering a period in which dynamic behavior of organisms will be explored by modeling. Based on this objective data-driven trend, I plan to bridge my current research, related to genome-scale modeling, with dynamic modeling of intracellular processes. While detailed

research plans can be influenced and shaped by academic surroundings, the following three research directions will help me establish my own independent research.

Stoichiometric modeling. Stoichiometric modeling can be viewed as a preliminary step in the analysis of a more complex intracellular dynamics based on kinetic properties. Metabolic regulation is a key issue in the understanding of the cellular metabolism and I plan to investigate the role that stoichiometry and reaction network connectivity can play in coordination of particular reaction fluxes. More precisely, I believe that stoichiometric models can help determine and predict barriers/limits under which amplification or expression of particular enzymes can or cannot result in a desired metabolic flux at steady state conditions.

Metabolic Control analysis. Modeling of biochemical transformations is complicated by metabolic regulation unknown in many cases. To remedy this, a theoretic framework, termed *metabolic control analysis* (MCA), has been developed to characterize the role of particular reactions in the control of concentrations and fluxes (Kacser and Burns, 1973; Heinrich and Rapoport, 1974). Control of particular reactions differs from metabolic regulation that is closely related to the biological function of metabolic pathways. Nevertheless, MCA has already proven extremely useful in quantifying metabolic regulation and has been used in the study of signaling pathways, biotechnology and drug discovery. I plan to extend MCA to the case of periodic metabolic oscillations to study effects of positive and negative feedbacks, responsible for emergence and stabilization of periodic regimes, respectively.

Simulation of interacting cellular populations. The study and simulation of cellular populations is another important example of dynamical interactions in biology. This study is significant because of the necessity to provide an adequate interpretation of population observations in terms of events occurring in single cells. In many cases stable asynchronous regimes (*e.g.*, with phases of intracellular oscillations uniformly distributed over the whole population) may not be observed in small population models, while these are abundant in wet lab experiments. Apparently, substantial attraction basins of such regimes can emerge in big population models. Since each cellular oscillator is usually "stiff," it can be quite difficult to numerically integrate and analyze the whole big population model with the phases uniformly distributed over all oscillators. Indeed, in such a regime at every time moment there will be an oscillator that undergoes a sharp transient process resulting in a very small integration step. To overcome such difficulties, new averaging approaches and techniques should be developed.

References

[1] Burgard*, A.P., **E.V. Nikolaev***, C.H. Schilling, and C.D. Maranas. 2003. Flux-coupling analysis of genome scale metabolic network reconstructions. *Genome Research* (*accepted*).[2]

[2] Dorogovtsev, S.N. and J.F.F. Mendes. 2003. *Evolution of networks: from biological nets to the Internet and WWW*, Oxford University Press, Oxford.

[3] Edwards, J.S., M. Covert, and B.O. Palsson. 2002. Metabolic Modeling of Microbes: the Flux Balance Approach. Mini-review. *Environmental Microbiology*, **4**(3), 133-140.

[4] Famili, I., and B.O. Palsson. 2003. The Convex Basis of the Left Null Space of the Stoichiometric Matrix Leads to the Definition of Metabolically Meaningful Pools. *Biophysical Journal*, **85**, 16-26.

[5] Golub, G.H. and C.F. van Loan. 1996. *Matrix Computations.* Johns Hopkins University Press, Baltimore.

[6] Heinrich, R., and S. Schuster. 1996. *The Regulation of Cellular Systems.* Chapman & Hall, New York.

[2] equal authorship

[7] Heinrich, R., and T.A. Rapoport. 1974. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength, *Eur. J. Biochem.*, **42**, 89-95.

[8] Jeong, H., B. Tombor, R. Albert, Z.N. Oltvai, and A.-L. Barabasi. 2000. The large-scale organization of metabolic networks. *Nature*, **407**, 651-654.

[9] Kacser, H. and J.A. Burns. 1973. The control of flux. *Symp. Soc. Exp. Biol.*, **27**, 65-104.

[10] Maranas, C.D., **E.V. Nikolaev**, and A.P. Burgard. 2003. Optimization-based analysis of topological features of genome-scale metabolic networks, The Pennsylvania State University, PSU Inventory Discovery No. 2003-2820 (*Provisional Patent Application*).

[11] **Nikolaev, E.V.**, A.P. Burgard, and C.D. Maranas. 2003. The structural analysis of genome-scale metabolic pools. *Biophysical Journal* (*in preparation*).

[12] Reed, J.L., and B.O. Palsson. 2003. Thirteen Years of Building Constraints-Based in silico Models of Escherichia coli. *Journal of Bacteriology*, **185**(9), pp. 2692-2699.

[13] Reed, J.L., T.D. Vo, C.H. Schilling, and B.O. Palsson. 2003. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.*, **4**, R54.

[14] Reich, J.G., and E.E. Selkov. 1981. *Energy Metabolism of the Cell: A Theoretical Treatise.* Academic Press, London.

[15] Salgado, H., A. Santos-Zavaleta, S. Gama-Castro, D. Millan-Zarate, E. Diaz-Peredo, F. Sanchez-Solano, E. Perez-Rueda, C. Bonavides-Martinez, and J. Collado-Vides. 2001. RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **29**, 72-74.

[16] Schilling, C.H., D.Letscher, and B.O.Palsson. 2000. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.*, **203**, 229-248.

[17] Stelling, J., S. Klamt, K. Bettenbrock, S. Schuster, and E.D. Gilles. 2002. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, **420**, 190-193.

Teaching Statement

I believe that research and teaching dynamically complement each other. I love to teach and always try to find an opportunity to interact with students. As a result, I have a diverse teaching experience and constantly try to improve my teaching skills and methodology.

1. Teaching experience. As a graduate student, I had frequent meetings with up to 10-15 students at their requests to discuss homework and exams. At the same time, I was a member of Lectureship Society "Znanie," ("Knowledge"), where I delivered scientific talks intended for a general audience.

While at the Institute for Mathematical Problems in Biology (IMPB), Biological Research Center near Moscow, I taught advanced mathematical courses in the School of Natural Sciences for high school students. My students were admitted to the Moscow State University and the Institute for Physics and Technology (PhyzTech), some of the highest ranked educational institutions of Russia. I frequently tutored undergraduate students from several Moscow educational institutions in formal mathematical courses in which the students struggled. After our meetings, the students considerably improved their grades.

While at the IMPB, I co-advised Dr. Gennady Cymbalyuk, who at that time was a graduate student studying computational neuroscience, and trained him in the theory of dynamical systems, the theory of bifurcations and numerical analysis. I formulated a research project to study dynamics in a system of coupled neuronal oscillators and guided Gennady through all stages of the project, starting from the formulation of a mathematical problem through modeling, data collection, and analysis. We completed the project with a publication in an internationally recognized journal [1]. Also, for this work Gennady was awarded a Soros' Scholarship in Mathematics by the International Science Foundation in Washington. Recently, I has been proud to learn that Dr. Cymbalyuk was awarded an Assistant Professorship in the Department of Physics and Astronomy at the Georgia State University.

While at the Integrated Genomics, Inc. in Chicago, I led a team of five professional programmers to develop a complex computer system to model cellular organisms. Although highly skillful in their professions, the programmers had relatively little experience in science, and I taught them elements of mathematics, biology, and numerical analysis within a limited time and under strong pressure to get the job done.

While at the Pennsylvania State University, I have co-advised and trained several graduate students in research, applied mathematics, and programming. As a result, we have prepared several papers accepted for publication in high profile journals [2-4]. Previously, the students would spend weeks manually preparing data for computer simulations. I held several training sessions to teach them Perl, a high level programming language designed for efficient manipulation of arbitrary text files where a predefined data pattern can be recognized. As a result, Tony Burgard and Priti Pharkya can now parse huge data sets (as, *e.g.*, the KEGG metabolite and reaction database), extract subsets of the needed data, and then transform the raw data into the form suitable for modeling. I also discussed with Madhukar Dasika basic elements of the spectrum theory for the differential-delay equations used in his microarray studies. I explained to Gregory Moore the limited applicability of contracting maps which he tried to use in his protein optimization studies.

2. Teaching philosophy. I firmly believe that fundamental knowledge, including universal approaches and techniques, should be taught first along with basic skills like problem motivation, logical reasoning, and abstraction. I consider a formal education as a starting point in a life-time learning process of professional growth. This is why I believe that the students must be given tools for learning independently.

Students need to digest great volumes of information covered in lectures, readings, and assignments, so they must be taught about time management, goal prioritizing, and technical writing. I view quizzes and exams not only as a means to evaluate students but also as a way to build students' confidence in solving problems under a limited time. Working on problems individually or in groups can also help students develop such socially important skills as patience, self-control and leadership.

Following a commonly accepted teaching practice, I intend to teach through examples and gradually move from simple ideas to higher levels of abstraction and generalization. In addition to a formal evaluation, I believe in having frequent informal conversations with students to understand their real capabilities and stumbling blocks.

Often such informal conversations can show that students know more than they demonstrated at a certain hour on a certain day. This is a good time to become better acquainted with students and to show sincere interest in each individual. Students who feel that a teacher is fair, approachable, and sincere will usually make a more determined effort to excel in their assignments.

3. Teaching interests. Most of all I wish to be seen as a resource for my students. Certainly, I am naturally interested in teaching courses related to my research interests. This would provide me with the opportunity to return to the classroom with new ideas inspired by current trends and achievements in modern science and technology. I always try to use computers inside and outside of the classroom and enjoy seeing my students gain a thorough understanding of the art of scientific computing.

With my background in applied mathematics, numerical methods, and modeling, I can teach required undergraduate courses including Calculus, Linear Algebra, Differential and Integral Equations, Equations of Mathematical Physics, Numerical Methods, *etc.* Also, I can teach both abstract and less formal graduate courses such as Theory of Dynamic Systems and Bifurcations, Bifurcations in Dynamic Systems with Symmetry, Mathematical Modeling in Biology, and Mathematical Biology.

References

[1] Cymbalyuk, G.S., **E.V. Nikolaev**, and R.M.Borisyuk. 1994. In-phase and anti-phase self-oscillations in a model of two electrically coupled pacemakers, *Journal of Biological Cybernetics* **71**, 153-160.

[2] Pharkya, P., **E.V. Nikolaev**, and C.D. Maranas. 2003. Review of the Brenda Enzymes Database, *Metabolic Engineering* **5**, 71-73.

[3] Burgard*, A.P., **E.V. Nikolaev***, C.H. Schilling, and C.D. Maranas. 2003. Flux Coupling Analysis of Genome Scale Metabolic Network Reconstructions, *Genome Research* (*accepted*).[3]

[4] **Nikolaev, E.V.**, A.P. Burgard, and C.D. Maranas. 2003. The structural analysis of genome-scale metabolic pools, *Biophysical Journal* (*in preparation*).

[3] equal authorship