SYSTEMS ENGINEERING CHALLENGES AND OPPORTUNITIES IN COMPUTATIONAL BIOLOGY

Costas D. Maranas*, Gregory L. Moore, Anthony P. Burgard, Anshuman Gupta
The Pennsylvania State University
University Park, PA 16802

Abstract

In this paper, three examples of predictive as well as descriptive biological modeling and optimization frameworks abstracted by research work in our group are described. These examples provide the opportunity to review the current state of the art in the respective problems and to draw analogies and contrasts with algorithmic development and applications in process systems engineering. Specifically, we discuss (i) optimizing of DNA recombination in the context of directed evolution experiments for protein engineering; (ii) probing the performance limits of metabolic networks with optimization-based techniques and (iii) identifying gene regulatory networks from DNA microarray data by utilizing contrasting systems identification methods.

Keywords

DNA shuffling, codon optimization, metabolic engineering, flux balance analysis, microarray time series data, gene regulatory network.

Introduction

Systems engineering approaches are currently emerging as vital tools for deciphering the behavior of biological systems. They are used for a variety of applications ranging from multiple DNA sequence alignment to effective database mining. Systems engineering contributions in biology can broadly be categorized as descriptive and predictive. Descriptive or data-driven approaches utilize multivariate statistics and information theoretic concepts to explain and/or augment the information content of genomic and proteomic data as well as measurements in the context of biological systems. Predictive approaches, on the other hand, rely on modeling analyses to estimate or set bounds for the behavior of biological systems involved in protein structure and function, gene expression, metabolic flux distributions, genetic circuit regulation and in higher levels of organization.

First, an optimization based procedure (Moore and Maranas, 2002) is described for guiding/optimizing the sequence diversity spanned by combinatorial libraries generated by DNA recombination. The procedure builds on the predictive ***e***Shuffle framework (Moore et al., 2001) that was developed to quantify the statistics of library diversity generated via DNA recombination for a variety of protocol setups. In the optimization procedure, the fact that many amino acids have multiple codon representations (*i.e.*, different triplets of DNA nucleotides spell the same residue) is exploited. This codon optimization procedure (***e***CodonOpt) relies on a mixed-integer linear programming (MILP) representation and solution strategy.

The next section switches focus from protein engineering to analysis and discovery in metabolic pathways. Specifically, we examine how to identify the theoretical performance limits of metabolic networks, constrained only by stoichiometric and thermodynamic constraints, in the presence of gene additions or deletions. The recombination of new genes into a metabolic network and/or the deletion of existing ones is modeled using binary variables leading to an MILP representation (Burgard and Maranas, 2001; Burgard et al., 2001). The objective function to be optimized typically draws upon the maximization of biomass, ATP usage or specific biochemical secretions. Here a procedure (***ObjFind***) is discussed for rigorously identifying what objective function, if any, is consistent with a set of experimentally derived metabolic fluxes. This procedure gives rise to a bilevel optimization problem that by using LP duality is converted into a single level nonlinear optimization problem.

* To whom all correspondence should be addressed.
E-mail: costas@psu.edu, Phone: (814) 863-9958, Web page: http://fenske.che.psu.edu/faculty/cmaranas/

Finally, the last section highlights ongoing efforts aimed at identifying the underlying regulatory networks that govern gene expression. DNA microarray experiments can nowadays routinely provide gene expression data in a high-throughput manner (*i.e.*, thousands of genes probed simultaneously) for a time series or system perturbation study. Here we highlight two alternative approaches for inferring underlying regulatory networks defined as a gene-gene directed graph of yes/no interactions. The first approach is based on a descriptive method that constructs a statistical model of the conditional gene expression (Bayesian networks) while the later uses a predictive model-based description of gene expression as a function of the expression of other genes in the previous time points.

Modeling and Optimization in Protein Engineering

Directed evolution methods accelerate the process of Darwinian evolution and selection to generate proteins with improved function. These methods (see Figure 1) typically begin with the infusion of diversity into a limited set of parental nucleotide sequences through DNA recombination and/or mutagenesis. The resulting combinatorial DNA library is ligated into an expression vector and transformed into an appropriate host. A high-throughput screening or selection procedure is then used to identify the best variants for final sequencing or additional rounds of recombination or mutagenesis. The cycles of recombination/mutagenesis, screening and isolation continue until a protein with the desired level of improvement is found.

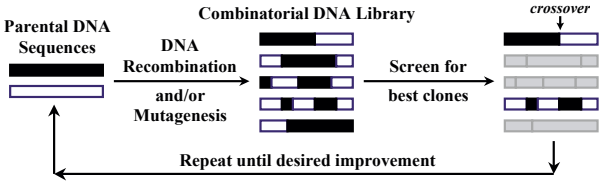

Figure 1: Schematic representation of the key steps of directed evolution experiments. Crossovers are defined as the junction points between segments from different parental sequences.

In the last few years, remarkable success stories of directed evolution have been reported. Many directed evolution studies have been highlighted in excellent reviews by Petrounia and Arnold, (2000), Schmidt-Dannert (2001) and Brakmann (2001). Despite these successes, directed evolution protocols have been developed and used largely based on empirical information and experience without a quantitative understanding of how diversity is distributed in the combinatorial DNA libraries and what crossover combinations are likely to give rise to functional recombinant protein hybrids.

A key challenge in directed evolution is that only an infinitesimally small fraction of the diversity afforded by DNA sequences can be examined regardless of the efficiency of the screening procedure. For example, a 500-nucleotide gene implies $4^{500} \approx 10^{301}$ alternatives, but even the most efficient *in vivo* screening methods can query only up to 10^7-10^8 DNA sequences (typically limited by transformation efficiency). In addition, only a few of the library members may be functional. Therefore, it is important to know *how diversity is generated and allocated in the combinatorial DNA library and what crossover patterns yield recombinant hybrids that are likely to be functional*. Earlier we developed a systematic computational framework named ***e***Shuffle (Moore et al., 2001) for identifying the statistics of library diversity for the DNA shuffling recombination protocol. Both experimental results and theoretical predictions have revealed that DNA shuffling cannot generate crossovers below a parental sequence identity threshold of approximately 60%. Therefore, the use of an optimized codon representation to increase the number of crossovers generated has been explored.

Here we discuss a method of designing DNA sequences that, upon DNA shuffling, generate an increased number of crossovers by exploiting the inherent redundancy in the codon representation. For example, isoleucine has the following three synonymous codon representations: ATA, ATC and ATT. The key motivation here is that it is possible to optimize the underlying parental DNA sequence codon representation for increasing and/or shaping diversity while at the same time preserving the parental amino acid encodings in the generated combinatorial protein libraries. This strategy is well suited in cases where parental sequences are synthetically generated (*e.g.*, through oligomer ligation). The utility of codon usage optimization has been recognized and exploited in a largely empirical way in the context of industrially developed directed evolution protocols such as oligo shuffling (Stemmer, 2000) and GeneReassembly (Short, 1999). Here a systematic computational framework named ***e***CodonOpt (Moore and Maranas, 2002) is described for exploring the limits of performance that can be achieved through codon optimization. Specifically, we discuss a constraint-based modeling framework that permits only nucleotide sequences encoding the underlying parental proteins as solutions. It utilizes 0-1 binary variables as on/off switches to model the presence of a specific codon choice in a given residue position. DNA shuffling (Stemmer, 1994ab) is used as the benchmark recombination method to showcase the framework. The ***e***CodonOpt framework (along with Figures 2 and 3) was introduced in Nucleic Acids Research, a publication of Oxford University Press.

The basic problem addressed in this work can be stated as follows. *Given a set of parental proteins, design the optimal nucleotide sequences encoding those proteins for a given diversity objective.* Below, the index notation, variables, parameters and constraints utilized in the basic ***e***CodonOpt model are listed:

Indices

$i \in \{1,2,\ldots,B\}$ = set of nucleotide positions

$k \in \{1,2,\ldots,K_{tot}\}$ = set of parental sequences

$n, n_1, n_2 \in \{A,T,C,G\}$ = set of nucleotides in positions i, $i+1$, $i+2$ in parental sequence k

Variable set

$$x_{ink} = \begin{cases} 1, \text{ if nucleotide } n \text{ is present at position } i \\ \quad \text{in parental sequence } k \\ 0, \text{ otherwise} \end{cases}$$

Parameters

$$a_{ink} = \begin{cases} 1, \text{ if nucleotide } n \text{ is permitted at position } i \\ \quad \text{in parental sequence } k \\ 0, \text{ otherwise} \end{cases}$$

$$b_{inn_1k} = \begin{cases} 1, \text{ if nucleotide pair } (n, n_1) \text{ is permitted at} \\ \quad \text{positions } (i, i+1) \text{ in parental sequence } k \\ 0, \text{ otherwise} \end{cases}$$

$$c_{inn_2k} = \begin{cases} 1, \text{ if nucleotide pair } (n, n_2) \text{ is permitted at} \\ \quad \text{positions } (i, i+2) \text{ in parental sequence } k \\ 0, \text{ otherwise} \end{cases}$$

Specifically, the proposed model utilizes the binary variable x_{ink} to represent the underlying nucleotide representation n = *(A, T, C, G)* at every sequence position i of the parental protein k. Parameter a_{ink} is equal to one only if there exists at least one codon representation that allows the use of nucleotide n at position i of parental sequence k. Parameter b_{inn_1k} is equal to one only if nucleotides (n,n_1) are both permitted at the first two codon positions, whereas parameter c_{inn_2k} is equal to one if nucleotides (n,n_2) are present at the first and third codon positions. These parameter values are determined by scanning the parental proteins against the codon translation table. See Tables 1-3 in the supplementary material of (Moore and Maranas, 2002) for a complete list of parameter values for all twenty amino acids.

Codon Constraints

Because only one nucleotide choice n can be present at each position i of sequence k, x_{ink} is allowed a non-zero value for only one of the *(A, T, C, G)* choices for n for every *(i,k)* pair (see constraint (1)). In addition, if a particular triplet *(i,n,k)* is not permitted ($a_{ink} = 0$) then variable x_{ink} is forced to zero (constraint (2)).

$$\sum_n x_{ink} = 1, \quad \forall\, i,k \tag{1}$$

$$x_{ink} = 0, \quad \forall\, i,n,k : a_{ink} = 0 \tag{2}$$

Constraints (1) and (2) suffice for residues with a single degenerate position (*e.g.*, alanine). Additional constraints are needed for residues with multiple codon redundancies such as serine, arginine and leucine.

Constraint For Serine Encoding Positions

Specifically for serine with degenerate first and second codon positions, if a consecutive pair (n,n_1) is disallowed ($b_{inn_1k} = 0$) then x_{ink} and x_{i+1,n_1k} cannot both be equal to one.

$$x_{ink} + x_{i+1,n_1k} \le 1, \quad \forall\, i,n,n_1,k : b_{inn_1k} = 0 \tag{3}$$

Constraint For Arginine, Leucine and Serine Positions

Similarly, for degeneracies in the first and third position for arginine, leucine and serine residues, the following constraint is needed.

$$x_{ink} + x_{i+2,n_2k} \le 1, \quad \forall\, i,n,n_2,k : c_{inn_2k} = 0 \tag{4}$$

Limiting the Number of Codon Manipulations

One may want to limit the number of codon representation changes (*i.e.*, silent nucleotide mutations) made to the wild-type DNA sequences. Specifically, the total number of silent nucleotide point mutations in the designed sequences could be set to be less than an upper limit P. This requires the definition of the following additional parameters:

$$\delta_{nn'} = \begin{cases} 1, \text{ if } n = n' \text{ (nucleotide identity)} \\ 0, \text{ otherwise} \end{cases}$$

w_{ink} = codon representation corresponding to the wild-type (original) nucleotide sequences

P = maximum number of point mutations permitted from wild-type nucleotide sequences

Constraint (5) establishes an upper bound to the total number of allowable silent point mutations.

$$\sum_k \sum_i \sum_{n,n'} (1 - \delta_{nn'}) x_{ink} w_{in'k} \le P \tag{5}$$

This constraint-based modeling framework allows the space of possible codon representations (codified in variable x_{ink} and subject to constraints (1-4)) to be searched for the one that optimizes a user defined diversity objective. Next, results for maximizing the number of crossovers generated for a pair of parental sequences are discussed. Additional results for and (i) minimizing bias in family DNA shuffling and (ii) maximizing the relative frequency of crossovers in specific structural regions are provided in (Moore and Maranas, 2002), and optimized sequences for each of the objectives are supplied in the supplementary material of the same reference.

Crossover statistics for different parental sequence codon representations can be estimated by the ***e***Shuffle

program, as discussed earlier (Moore et al., 2001). However, because the CPU of an ***e***Shuffle run can range from minutes to hours, utilizing ***e***Shuffle in the context of optimization loops is impractical for all but the simplest cases. Instead, two simple surrogate objectives for crossover generation are postulated and subsequently tested: (a) maximization of the pairwise sequence identity between the parental DNA sequences and (b) minimization of the total free energy change upon complete annealing of the two DNA sequences. Both of these surrogates for crossover generation capture the fact that crossover formation in DNA shuffling occurs predominantly within regions of near perfect sequence identity.

Surrogate (a): Maximizing Pairwise Sequence Identity

This intuitive surrogate for crossover generation implies that the degree of sequence identity between a pair of DNA sequences correlates with the number of crossovers generated. The calculation of the sequence identity is performed by counting the total number of matching nucleotides, $M_{k\tilde{k}}$, between two aligned parental sequences *k* and $\tilde{k}$.

$$M_{k\tilde{k}} = \sum_{i} \sum_{n,\tilde{n}} \delta_{n\tilde{n}} x_{ink} x_{i\tilde{n}\tilde{k}}, \quad \forall\, k, \tilde{k} > k \tag{6}$$

Note that the nonlinearity introduced by the product of binary variables ($x_{ink} x_{i\tilde{n}\tilde{k}}$) is eliminated (Moore and Maranas, 2002). Therefore, the first surrogate for maximizing crossover generation upon codon optimization involves *maximizing* $M_{k\tilde{k}}$ subject to constraints (1-4) and (6). Constraint (5) is added if a limit on the total number of silent nucleotide mutations is needed. This problem belongs to the class of mixed-integer linear programming (MILP) problems and is solved using CPLEX 7.0 (Brooke et al., 1998) accessed through the GAMS modeling environment (Brooke et al., 1998). Note without any additional restrictions such as (5), this problem decomposes over codons and can be solved in linear complexity. This decoupling, however, does not hold for the second surrogate.

Surrogate (b): Minimizing ΔG *of Annealing*

The second surrogate objective implies that crossover generation correlates with the total free energy change upon complete annealing of the recombining pair of DNA sequences. The free energy change is approximated using empirical nearest-neighbor parameters (SantaLucia Jr., 1998) that decompose the free energy calculation into the sum of the contributions of overlapping 2-nucleotide (nt) units (see Figure 2). Matching pairs contribute negative free energy terms lowering the total free energy change of annealing, whereas mismatches contribute positive terms increasing the free energy change. Parameter set $\Delta G^{pair}_{nn_1\tilde{n}\tilde{n}_1}$ stores the free energy change associated with the annealing of nucleotide pair *(n,n1)* with $(\tilde{n}, \tilde{n}_1)$. The total free energy change $\Delta G_{k\tilde{k}}$ upon complete annealing of two parental sequences $(k, \tilde{k})$ is calculated by summing over the contribution of all nucleotide pairs at positions *(i,i* + 1*)* along the entire sequence length.

$$\Delta G_{k\tilde{k}} = \sum_{i} \sum_{n,n_1,\tilde{n},\tilde{n}_1} \Delta G^{pair}_{nn_1\tilde{n}\tilde{n}_1} \left(x_{ink} \cdot x_{i+1,n_1 k} \cdot x_{i\tilde{n}\tilde{k}} \cdot x_{i+1,\tilde{n}_1\tilde{k}} \right), \quad \forall\, k, \tilde{k} > k \tag{7}$$

Note that the four-term product in the expression is subsequently expressed in an equivalent linear form (Moore and Maranas, 2002). Therefore, the second surrogate for crossover generation in DNA shuffling involves *minimizing* $\Delta G_{k\tilde{k}}$ subject to constraints (1-4, 7) and optionally (5).

$$\Delta \boldsymbol{G} = \Delta \boldsymbol{G}_{\mathtt{AT/TC}} + \Delta \boldsymbol{G}_{\mathtt{TC/CG}} + \Delta \boldsymbol{G}_{\mathtt{CG/GC}} + \Delta \boldsymbol{G}_{\mathtt{GA/CG}} + \Delta \boldsymbol{G}_{\mathtt{AT/GA}}$$

Figure 2: Calculation of annealing free energy change using overlapping nearest-nucleotide pairs.

These two surrogate choices are tested based on the DNA shuffling of two glycinamide ribonucleotide (GAR) transformylases. Specifically, the DNA shuffling of the *E. coli* and human versions of GAR transformylase is studied. The wild-type parental sequences share a very low nucleotide sequence identity of 49% even though the two enzymes share the same function and presumably the same structure (Ostermeier et al., 1999). In the absence of any codon optimization, DNA shuffling crossovers are extremely rare for this system as shown previously in (Moore et al., 2001); therefore, there is clearly a need to increase the number of crossovers generated.

First, surrogate objective (a), maximizing the sequence identity of the two GAR transformylases, M_{12}, is examined. The maximum sequence identity upon codon optimization is identified for an increasing number of allowed silent nucleotide mutations. These codon-engineered parental sequences are next fed to ***e***Shuffle to predict the total number of crossovers expected to be generated upon DNA shuffling. Crossover numbers are plotted in Figure 3 from zero (wild-type) to 320 permitted silent mutations. Interestingly, after 90-100 point mutations are accumulated, the total number of crossovers rapidly increases reaching a maximum value of about 1.5 crossovers per sequence. Beyond this point, sequence identity ceases to correlate with crossover generation leading to the plateau effect beyond 140 silent mutations as shown in Figure 3. The second surrogate objective, involving the minimization of the free energy change of annealing, ΔG_{12}, provides much better correlation with the extent of crossover formation. Almost twice as many

crossovers are formed compared with the previous surrogate (see Figure 3). The key difference is that, unlike sequence identity, the free energy change continues to correlate strongly with crossover formation even for very high numbers of silent mutations, preventing the onset of the plateau.

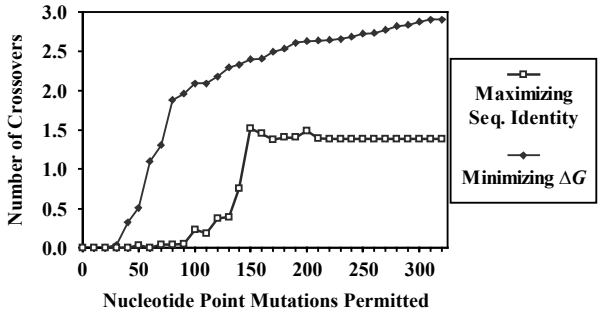

Figure 3: The total number of crossovers increases as more point mutations are permitted. Free energy change outperforms sequence identity as a surrogate.

The free energy change of annealing outperforms sequence identity as a surrogate for crossover formation because it appropriately weighs the thermodynamic contribution of different matches and mismatches. In addition, by considering the contribution of overlapping nucleotide pairs, it places a higher emphasis on blocks of sequence identity over isolated nucleotide matches. Sequence identity is not as successful as a surrogate because the matching nucleotides do not necessarily group into crossover-generating blocks of sequence identity. The qualitative trends in the result hold for a wide range of example problems studied so far implying that free energy of annealing appears to be universally superior to sequence identity as a predictor of crossover formation. This result has a direct implication on the way DNA shuffling studies are conducted and parental DNA sequences are engineered.

Optimization in Metabolic Modeling

The analysis and modification of metabolic pathways, known as metabolic engineering, has attracted significant interest in recent years catalyzed by the rapidly increasing number of sequenced microbial genomes. Overall, over sixty microbial genomes have been completely sequenced and many more sequencing projects are currently underway (Peterson et al., 2001). In addition, current bioinformatic tools will allow the functional assignment of 40-70% of these newly sequenced genomes (Eisenberg et al., 2000). This flood of genomic information coupled with gene annotation and metabolic reconstruction efforts is poised to revolutionize the understanding of microbial metabolic networks of biotechnological and biomedical importance. Improved understanding of metabolic networks may uncover the manipulations necessary to enhance the biochemical production capabilities of a microorganism or suggest promising targets in pathogenic microbes for gene therapy. Here we utilize recently acquired genomic information in conjunction with stoichiometric models of microbial metabolism to develop optimization techniques for enhancing their descriptive and predictive capabilities.

Metabolic models are particularly useful for investigating how raw materials (*i.e.*, glucose, oxygen, etc.) are converted to products (*i.e.*, desired biochemicals or cellular biomass precursors – amino acids, nucleotides, lipids, etc.) in individual cells. Stoichiometric (*i.e.*, flux balance analysis) models of cellular metabolism require only the stoichiometry of biochemical pathways and cellular composition information to identify boundaries for the flux distributions available to the cell. The underlying principle of flux balance analysis (FBA) is mass balances on the metabolites of interest. Fo r a steady-state metabolic network comprised of N metabolites and M metabolic reactions we have,

$$\sum_{j=1}^{M} S_{ij} v_j = b_i, \qquad \forall\ i \in N$$

where S_{ij} is the stoichiometric coefficient of metabolite i in reaction j, v_j represents the flux of reaction j, and b_i quantifies the network's uptake (if negative) or secretion (if positive) of metabolite i. For all internal metabolites, b_i is zero.

Typically, the resulting flux balance system of equations is underdetermined as the number of reactions exceeds the number of metabolites, and additional information is required to solve for the reaction fluxes. A popular technique for investigating metabolic flux distributions is linear optimization (Varma and Palsson, 1994). The key conjecture is that the cell is capable of spanning all flux combinations allowable by the stoichiometric constraints. Thus instead of predicting exactly how a metabolic network behaves, these models narrow the range of possible phenotypes (i.e., cellular behaviors) these systems can display in pursuit of a metabolic objective. Therefore, FBA predictions are typically treated as theoretical limits to the performance of metabolic networks. However, experimental evidence suggests that organisms have developed control structures to ensure optimal growth (*i.e.*, maximum biomass production) in response to certain environmental conditions (Edwards and Palsson, 2000; Edwards et al., 2001). The general flux balance analysis model for a steady-state metabolic network optimized for maximum biomass formation is

$$\max \qquad Z = v_{biomass}$$

subject to

$$\sum_{j=1}^{M} S_{ij} v_j = b_i, \qquad \forall\ i \in N$$

$$v_j \geq 0, \qquad \forall\ j \in M$$

Table 1: Model predictions for maximum theoretical yields of seven amino acids for growth on acetate and glucose

	Maximum Theoretical Yield (mmol / per 10 mmol Glucose)			Maximum Theoretical Yield (mmol / per 10 mmol Acetate)		
	Modified Keasling '97	Universal Model	% Increase	Modified Keasling '97	Universal Model	% Increase
Arginine	9.26	10.07	8.75%	2.43	2.65	9.05%
Asparagine	18.18	19.23	5.77%	4.66	4.91	5.45%
Cysteine	11.49	11.90	3.57%	3.29	3.42	3.80%
Histidine	9.77	9.80	0.23%	2.43	2.54	4.53%
Isoleucine	8.00	8.07	0.91%	2.13	2.13	-
Methionine	7.04	7.19	2.16%	1.81	1.85	2.46%
Tryptophan	4.67	4.73	1.28%	1.17	1.19	1.32%

Modified Keasling '97: Modified Pramanik and Keasling (1997) *E. coli* model
Universal Model: Modified Pramanik and Keasling (1997) model augmented with the non-*E. coli* reactions compiled from KEGG

where $v_{biomass}$ is a flux drain comprised of all necessary components of biomass (*i.e.*, amino acids, nucleotides, etc.) in their appropriate biological ratios (Neidhardt and Curtiss, 1996). Stoichiometric models have in some cases been successful in predicting the phenotypical characteristics of cells such as growth rates (Pons et al., 1996; Edwards et al., 2001), metabolic byproduct secretion rates (Varma et al., 1993), biochemical production rates (Jorgensen et al., 1995; Henriksen et al., 1996), and viability in the presence of gene deletions (Edwards and Palsson, 2000).

In silico Modeling of Gene Additions

The explosive growth of annotated genes associated with metabolism calls for a systematic procedure for determining the most promising recombination choices. A framework has been developed utilizing flux balance analysis and mixed-integer programming tools to select the mathematically optimal gene set for recombination into *E. coli* or other prokaryotes from a metabolic database encompassing many genes from multiple species. The resulting pathways need not lie directly on main production pathways, as they may enhance production indirectly by either redirecting metabolic fluxes into the production pathways or by increasing the energy efficiency of the present pathways.

A comprehensive stoichiometric matrix containing all known metabolic reactions from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) and Ecocyc (Karp et al., 2000), (a web review of Ecocyc is available by Maranas and Burgard (2001)), and other sources has been compiled and incorporated into the flux balance model of the model organism (*e.g.*, *E. coli*). Such a stoichiometric matrix, containing already 3,400 reactions, is currently under development (Burgard and Maranas, 2001). We refer to this multi-species stoichiometric matrix as the Universal stoichiometric matrix.

A detailed description of how to model gene additions or deletions within a MILP framework is provided in (Burgard and Maranas, 2001). Here, for the sake of simplicity of presentation, a one to one mapping of genes to reactions is assumed. This is relaxed in the detailed model description (Burgard and Maranas, 2001). Specifically, let binary variables y_j describe the presence or absence of each gene/metabolic reaction j.

$$y_j = \begin{cases} 0, & \text{if gene } j \text{ is not expressed in the host organism} \\ 1, & \text{if gene } j \text{ is present and functional} \end{cases}$$

Thus the following constraint,

$$0 \le v_j \le v_j^{\max} y_j$$

ensures that $v_j = 0$ if there exists no active gene capable of supporting reaction j. Alternatively, if such a gene, v_j is allowed to assume any value between zero and an upper bound v_j^{max}. Selecting up to h new genes to recombine into the host organism (*i.e.*, *E. coli*) so that a metabolic objective v^* is maximized can be formulated as an MILP problem. This is accomplished by augmenting the LP flux balance model with constraint $y_j = 1, \forall\, j \in E$ that ensures that all *E. coli* genes are available as well as constraint

$$\sum_{j \in NE} y_j \le h$$

that allows up to h foreign genes to be incorporated in *E. coli* out of the comprehensive list contained in the Universal matrix (*i.e.*, *NE*). Reactions chosen by the model but absent in *E. coli* (*i.e.*, all nonzero y_j elements of *NE*) provide routes for manipulating the cellular metabolism through recombinant DNA technology.

Preliminary results using the flux balance *E. coli* model (Pramanik and Keasling, 1997) modified (Burgard and Maranas, 2001) to include up to date information from Ecocyc (Karp et al., 2002) demonstrate that improvements to seven amino acid production pathways of *E. coli* are theoretically attainable with the addition of genes from foreign organisms (see Table 1). In most cases, only one or two genes were added to the original amino acid

production pathway even though the complete list of 3,400 reactions was available for selection. The mechanism of all identified enhancements is either by: (i) improving the energy efficiency and/or (ii) increasing the carbon conversion efficiency of the production route. Manipulation of the arginine pathway showed the most promise with 8.75% and 9.05% improvements for growth on glucose and acetate, respectively. Figure 4 shows the pathway modifications introduced in the recombined network for growth on glucose. Overall, the additional genes used by the Universal model save the original pathway three net ATP bonds increasing arginine production by 8.75%. Similar trends are revealed when other native and Universal amino acid production routes for glucose and acetate substrates are examined.

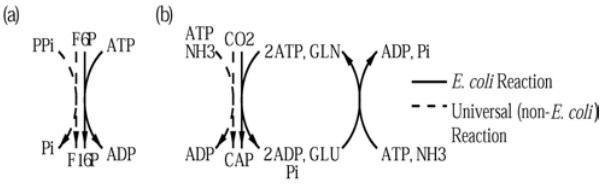

Figure 4: Difference between optimal E. coli and Universal arginine production pathways for growth on glucose: (a) The pyrophosphate dependent analog of 6-phosphofructokinase in the Universal model replaces the ATP dependent version present in E. coli. (b) Carbamate kinase in the Universal model replaces carbamoyl phosphate synthetase from the E. coli network.

Minimal Reaction Set Identification

The recent explosion of fully sequenced genomes has brought significant attention to the question of how many genes are necessary for sustaining cellular life. A minimal genome is generally defined as the smallest set of genes that allows for replication and growth in a particular environment (Cho et al., 1999), and the existence of multiple, quite different, species and environment specific minimal genomes has been speculated (Huynen, 2000). Here we discuss a computational procedure for testing this claim by estimating the minimum life-sustaining core of metabolic reactions required for given growth rates under different uptake conditions. We formulate this problem as the following optimization problem

$$\min \quad \sum_{j=1}^{M} y_j$$

subject to

$$\sum_{j=1}^{M} S_{ij} v_j = b_i, \qquad \forall\ i \in N$$

$$v_{biomass} \geq v_{biomass}^{target}$$

$$0 \leq v_j \leq v_j^{max} y_j\ , \quad \forall\ j \in M$$

$$y_j \in \{0,1\}\ , \qquad \forall\ j \in M$$

that solves for the smallest set of metabolic reactions that satisfies the stoichiometric constraints and meets a biomass target production rate $v_{biomass}^{target}$. Alternatively, instead of a biomass target, minimum levels of ATP production or lowest allowable levels of key components/metabolites could be incorporated in the model. The novel feature of this analysis is that whereas previous attempts utilized reductionist methodologies to extract the set of essential genes through a series of gene knock-outs, here we simultaneously assess the effect of all reactions on biomass production and select the minimal set that meets a given growth rate target (whole-system approach). A minimal gene set can then be inferred by mapping the enzyme(s) catalyzing these reactions to the corresponding coding genes.

Preliminary results based on the *E. coli* FBA model of Edwards and Palsson (2000), for the first time quantitatively demonstrated that minimal reaction sets and thus corresponding minimal gene sets are strongly dependent on the uptake opportunities afforded by the growth medium and the imposed growth requirements (Burgard et al., 2001). Figure 5 shows the number of reactions in each minimal set as a function of the growth demands placed on the network for growth on glucose. In addition, whereas it was found that an *E. coli* cell grown on glucose substrate requires at least 224 metabolic reactions, a cell cultured on an optimally engineered medium could support growth with as few as 122 metabolic reactions.

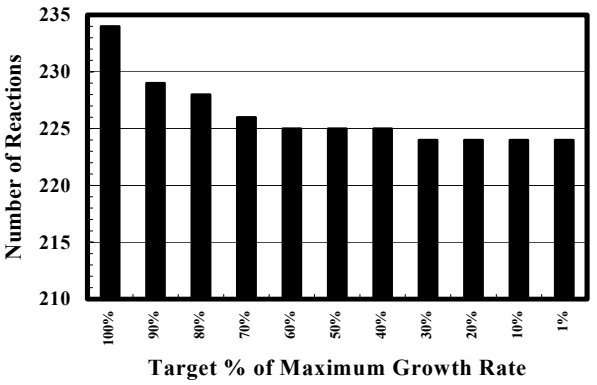

Figure 5: Number of reactions in each minimal set as a function of the imposed growth demands for a glucose uptake environment.

Furthermore, note that neither the minimal reaction sets nor their corresponding reaction fluxes were found to be unique. Even after excluding cycles and isoenzymes, hundreds of multiple minimal sets were identified providing a computational confirmation of the astounding redundancy and flux redirection versatility of the E. coli network. The existence and subsequent identification of multiple optima in undetermined stoichiometric models

could be used to examine metabolic regulation hypotheses or to aid the design of experiments aimed at discriminating between alternate flux distributions (Lee et al., 2000). Finally, for the glucose-only uptake study, it must be noted that all identified minimal reactions sets included 11 out of 12 reactions whose corresponding gene deletions were determined experimentally to be lethal for growth on glucose. Earlier analyses (Edwards and Palsson, 2000) based on single gene deletions conducted with this model using LP optimization were able to identify only 7 out of 12 lethal gene deletions motivating the importance of considering simultaneous gene deletions within an MILP framework.

Determination of Cellular Objective Functions

Here we describe a mathematically rigorous framework for testing whether experimental flux data are consistent with different hypothesized objective functions. Specifically, we identified weights or coefficients c_j for every reaction j that accept a set of experimental fluxes v_j^{exp} as an optimal solution to the following linear programming problem heretofore referred to as the Primal.

$$\max \qquad Z_P = \sum_{j=1}^{M} c_j v_j$$

subject to

$$\sum_{j=1}^{M} S_{ij} v_j = 0, \qquad \forall\ i \in N$$

$$v_{GLC} = \text{upt. rate}, \quad \forall\ j \in \text{glc. upt.}$$

$$v_j \geq 0, \qquad \forall\ j \in M$$

This mathematical framework, named ***ObjFind***, requires the solution of a bilevel optimization problem minimizing the squared deviations of identified fluxes v_j from experimental data v_j^{exp} while ensuring that the identified fluxes are the solution of the inner optimization problem.

$$\min_{c_j} \qquad \sum_{j \in E} (v_j - v_j^{\exp})^2$$

subject to

$$\left(\begin{array}{ll} \max\limits_{v_j} & \sum\limits_{j \in P} c_j v_j \\ \text{subject to} & \\ \sum\limits_{j=1}^{M} S_{ij} v_j = 0, & \forall\ i \in N \\ v_{GLC} = \text{upt. rate}, & \forall\ j \in \text{glc.upt.} \\ v_j \geq 0, & \forall\ j \in M \end{array} \right)$$

$$\sum_{j \in P} c_j = 1$$

$$c_j \geq 0, \qquad \forall\ j \in P$$

Here the coefficients c_j for the inner problem are adjusted by the outer problem so that the sum-squared difference between the experimental fluxes and the optimal inner solution v_j are minimized. A solution strategy founded upon duality theory concepts was employed. For example, the dual problem associated with the linear programming problem given by the Primal is

$$\min \qquad Z_D = (\text{uptake rate}) \cdot g$$

subject to

$$\sum_{i=1}^{N} u_i S_{ij} \geq c_j, \qquad \forall\ j \in P$$

$$\sum_{i=1}^{N} u_i S_{ij} \geq 0, \qquad \forall\ j \notin P, \text{ glc. upt.}$$

$$\sum_{i=1}^{N} u_i S_{ij} + g = 0, \quad \forall\ j \in \text{glc. uptake}$$

where u_i is the dual variable associated with the first set of constraints in the Primal, g is the dual variable associated with the glucose uptake constraint. The dual variables, u_i and g, indicate the change in the optimal value of Z_P per unit change in the right hand side of their associated constraint. Likewise, the reaction fluxes, v_j, are the dual variables associated with the constraints to the Dual problem. Strong duality implies that if the primal has an optimal solution, so does the dual, and their respective optimal objective values are equal. Furthermore, the primal and dual problems can be simultaneously feasible only at their respective optimal solutions. Therefore by constructing an optimization problem formulation that includes both the **Primal** and **Dual** constraints along with an equality constraint forcing their respective objective function values to be equal to each other, we ensure that any feasible solution (v_j, g, u_i) will be optimal to both the **Primal** and **Dual** problems. The solution of the following single level nonlinear optimization problem (NLP)

$$\min \qquad \sum_{j \in E} (v_j - v_j^{\exp})^2$$

subject to

$$Z_P = Z_D$$

$$\sum_{j=1}^{M} S_{ij} v_j = 0, \qquad \forall\ i \in N$$

$$v_{GLC} = \text{upt. rate}, \quad \forall\ j \in \text{glc. upt.}$$

$$\sum_{i=1}^{N} u_i S_{ij} \geq c_j, \qquad \forall\ j \in P$$

$$\sum_{i=1}^{N} u_i S_{ij} \geq 0, \qquad \forall\ j \notin P, \text{ glc. upt.}$$

$$\sum_{i=1}^{N} u_i S_{ij} + g = 0, \quad \forall\ j \in \text{glc. uptake}$$

$$\sum_{j \in P} c_j = 1$$

$$v_j \geq 0, \qquad \forall\ j \in M$$

$$c_j \geq 0, \qquad \forall\ j \in P$$

systematically characterizes the set of all possible c_j values consistent with the minimization of sum-squared difference between a subset of observed fluxes v_j^{exp} and an optimal solution to the Primal. It should be noted that the constraint ($Z_P = Z_D$) is nonconvex due to the bilinear $c_j v_j$ terms. Therefore, multiple starting points were used to identify many local optimal solutions in search of the global optimum. Problems containing as many as 200 variables were solved in seconds using MINOS 5.0 accessed via the GAMS modeling environment on an IBM RS6000-270 workstation.

In this study, coefficients of importance (CoI) were first assigned to each reaction flux that consumes, by either draining or dissipating, a resource in the network. The reaction fluxes associated with these coefficients are shown with bold arrows in Figure 6. ***ObjFind*** was then employed to identify CoI's consistent with the aerobic and anaerobic experimental flux distributions for each condition (see Figure 7). Remarkably, the CoI's for both growth conditions were strikingly similar even though the flux distributions for the two cases were quite different. This unexpected convergence is consistent with the presence of a single metabolic objective driving the flux distributions in both cases. It also appeared that fluxes with similar CoI's clustered within groups that are both topologically and functionally related as depicted by Figure 7.

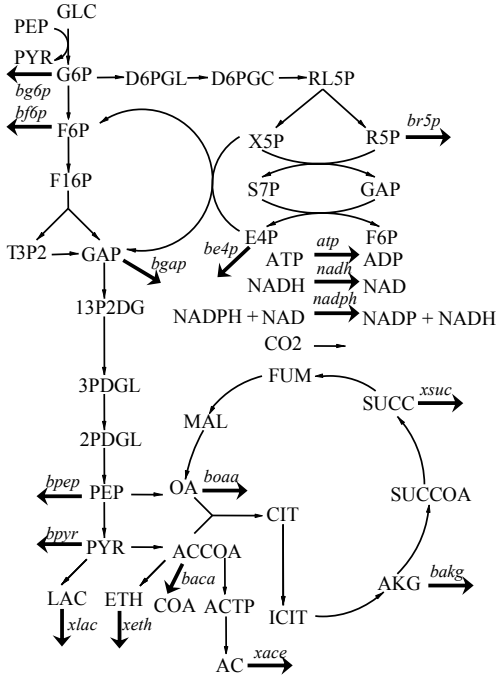

Figure 6: The reaction fluxes allowed to assume non-zero coefficients of importance are shown with bold arrows. Note the magnitudes of the CoI's are similar for both the aerobic and anaerobic growth conditions.

Next, we examined how close these coefficients of importance track the biomass maximization hypothesis. A biomass reaction flux, complete with energy and reducing power requirements, was included to drain metabolic precursors in their appropriate ratios as proposed by (Neidhardt and Curtiss, 1996) for biomass formation. A CoI was assigned to this aggregate biomass flux. We then identified its maximum value capable of explaining the flux distributions for the aerobic and anaerobic cases, respectively. We find that the maximum possible values of the biomass CoI's for the aerobic and anaerobic cases were 0.58 and 0.68, respectively, as shown by Figure 7. No other flux has a coefficient of importance nearly as high as the one identified for biomass formation.

As more complete flux distributions are deciphered through isotopomer experiments, we expect the ***ObjFind*** procedure to be used for testing/inferring hypothesized objective functions in metabolism.

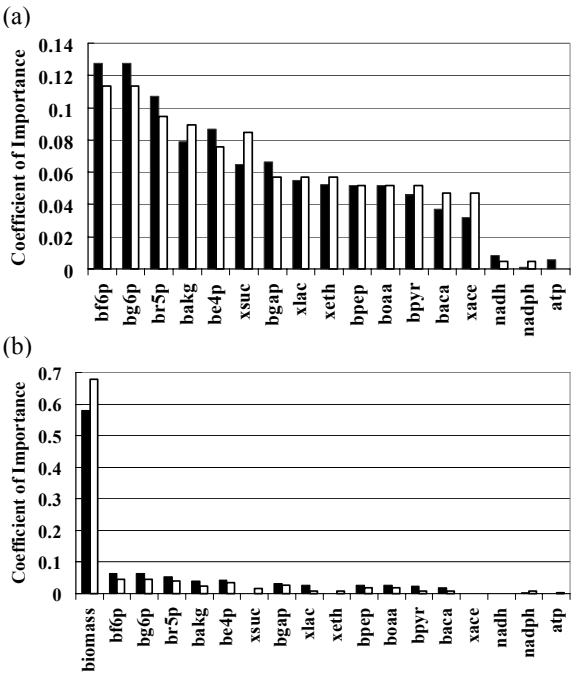

Figure 7: The values of the coefficients of importance for the aerobic (■) and anaerobic (□) experimental flux distribution when (a) no biomass flux is included and (b) a biomass flux is included.

Gene Regulatory Network Identification

Ongoing large-scale sequencing projects and subsequent gene annotation efforts are generating a "spare-parts" catalogue of the genes present in an organism as well as their likely functions with an ever accelerating pace. While the same set of genes is present in all cells of an organism, widely different temporal and

spatial gene expression profiles give rise to astoundingly diverse cellular morphology and functionality. Consequently, to understand how genes modulate intracellular and intercellular processes, entire regulatory gene interaction networks need to be deciphered so that questions such as which genes are expressed, when, where and to what extent, can be answered with relative certainty and confidence. Elucidation of such networks is essential not only for understanding the mechanics of various fundamental biological processes, like metabolism, cell growth and differentiation, but also for uncovering novel techniques/products that can be subsequently used to alter these processes.

To this end, DNA microarray technology has emerged as the enabling tool for rapidly measuring the spatio-temporal expression levels of genes in a massively parallel fashion (Spellman et al., 1998). However, most gene regulatory networks are comprised of a complex web of positive and negative feedback loops. This has motivated the development of formal systems-based network inference methods for unambiguously identifying the structure of gene regulatory networks given temporal gene expression data obtained through DNA microarray experiments (Figure 8).

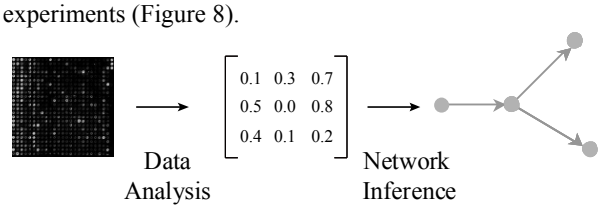

Figure 8. Two steps involved in uncovering gene regulatory networks from microarray data.

Most early network inference methods relied primarily on clustering genes on the basis of their expression profiles (Eisen et al., 1998; Wen et al., 1998; D'haeseleer et al., 2000). Lately, there has been considerable interest in developing computational tools that go beyond answering the question of whether two genes have similar expression profiles. Instead, the central question that is being raised is whether we can find, hidden within gene expression data, the signature, extent and directionality of interactions between different genes. In other words, rather than simply grouping genes with similar expression profiles, new methods attempt to learn gene regulatory patterns from expression data.

Broadly, these methods can be classified into two distinct categories based on their fundamental treatment of gene interactions. *Model-based* methods assume that there exists a formalism $Y = f(X)$ that captures the effect of expression level of gene X on that of gene Y. Different choices for the functionality of $f(\,)$ (*e.g.*, linear, sigmoidal, *etc.*) give rise to different versions of model-based methods (D'haeseleer et al., 1999; Weaver et al., 1999; Holter et al., 2001). On the other hand, *statistics-based* methods start from a subtly different assumption by postulating that the experimentally observed gene expression profiles correspond to samples drawn from an unknown, multivariate probability distribution which needs to be determined. Bayesian networks provide the means for achieving this objective by postulating a conditional probability model that explains the observed expression data (Friedman et al., 2000; Pe'er et al., 2001).

In light of these alternative approaches for uncovering gene regulatory networks, key research questions include:

1. Given a hypothesized gene regulatory network and a set of time series gene expression data, is it possible to identify arcs in the hypothesized network that are inconsistent with the observed data and suggest new ones that were missing in the original hypothesis?
2. Given a set of interacting genes, is it possible to identify whether there are missing genes (or actors with no DNA fingerprint) that are strongly modulating gene expression levels in the current gene set?
3. Finally, how can future forced gene expression/repression experiments be suggested for reducing the ambiguity/redundancy in the inferred gene regulatorynetworks?

Bayesian Networks

The Bayesian network formalism has emerged as one of the most promising statistics-based methods for analyzing gene expression data (Hartemink et al., 2001; Pe'er et al., 2001). In this approach, the mRNA expression levels of the genes are modeled as random variables. The state of the overall system (the underlying gene regulatory network) is subsequently modeled as a joint probability distribution over these random variables. Estimation of this probability distribution and its structural features is the basic goal of the Bayesian network approach (Friedman et al., 2000).

To this end, a genetic regulatory system is modeled as a directed acyclic graph $G = <V,E>$. The vertices $i \in V$, ($i = 1,\ldots,N$) represent the genes and the directed edges $(i,j) \in E$, ($i = 1,\ldots,N$; $j = 1,\ldots,N$) represent conditional dependencies between the expression levels X_i and X_j. Each gene i is associated with a conditional probability distribution $p\,(\,X_i \mid \mathrm{Pa}\,(\,X_i\,)\,)$ where $\mathrm{Pa}\,(\,X_i\,)$ is the set of parental regulators of gene i. The joint probability distribution is then obtained as follows (see Figure 9 for a representative example).

$$p(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} p\left(X_i \middle| \mathrm{Pa}(\mathrm{X_i})\right)$$

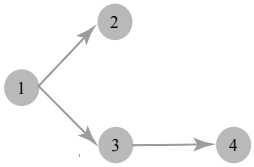

Figure 9. Bayesian network encoding the distribution $p(X_1,X_2,X_3,X_4) = p(X_1)\,p(X_2 \mid X_1)\,p(X_3 \mid X_1)\,p(X_4 \mid X_3)$.

The decomposable form of the above equation is obtained under the Markov assumption that states that the expression level of a gene is independent of all other genes *given* the expression level of its parents. The basic problem addressed in the Bayesian network inference methodology can be stated as follows.

Given a set of experimental time series expression data X_{it}^{e} *for i = 1,...,N genes at t = 1,...,T time points, determine (i) the parental set Pa* (X_i) *for each gene and (ii) the parameters describing the conditional distributions implied by the resulting network connectivity*.

In a typical microarray experiment dataset, the total number of genes is approximately 5,000-10,000. The expression of these genes is usually monitored at 10-20 time points. The frequency of sampling depends on the time-scale of the biological phenomenon that is under investigation. Given the experimental uncertainty associated with these experiments, 4-5 replicate expression levels are recorded at each time point.

The quality of fit of a particular network configuration to the data is assessed via scoring functions such as the Bayesian information criterion (BIC) or the Minimum Description Length criterion (MDL) (Heckerman, 1998). The BIC metric, which is adopted in our research, involves two components, an approximate log-likelihood term for the posterior probability of the model given the data, and a penalty term for preventing overfitting of the data with additional parameters. Maximization of the BIC over the space of alternative network structures and parameter settings results in the best fitting model while eliminating unnecessary arcs. Since the number of possible network structures is super-exponential in the number of genes, heuristic searches such as greedy-hill climbing and simulated annealing are employed to find local maximums in the structure space. The basic algorithmic steps are as follows:

1. Calculate the mutual information content for all possible gene pairs.
2. For each gene *i*, select *k* genes with the highest mutual information to form the parent set Pa (X_i).
3. Generate a random guess structure that satisfies the imposed parental relationships.
4. Compute the BIC score for the guess structure.
5. Consider all structures *one-move away* from the guess structure and calculate their BIC scores. A one-move away structure is obtained by either adding, deleting or reversing a single arc in the original guess structure.
6. Select the structure with the largest improvement in BIC over the original guess structure.
7. Repeat steps 3-6 over N (*e.g.*, $N \approx 200$) times. Select the top k (*e.g.*, $k \approx 20$) scoring models and average arcs.

Interpretation of the resulting Bayesian network is not always straightforward. Special attention needs to be paid to the statistical confidence that can be assigned to the uncovered gene interactions. Bootstrap resampling of the data and subsequent relearning of the structure is adopted as the mean for determining the confidence level of an interaction (Friedman et al., 2000). Those interactions that tend to be robust to resampling (present in most resampled networks) are assigned higher confidence levels. Even for interactions with high confidence, there is often uncertainty about their directionality. For instance, an interaction may indicate a causal relationship in which the expression of gene A leads to activation/repression of gene B (directed arc), or may simply indicate co-expression of the genes (undirected arc). Follow up gene perturbation experiments would be necessary to confirm if a particular gene has causal link to another gene in the network.

The developed procedure is applied to a 21 gene yeast time series data set. The regulatory network obtained is shown in Figure 10. The two main advantages of the Bayesian network approach are (i) the ability to handle noisy microarray data and (ii) the incorporation of previous biological knowledge in the learning algorithm by postulating a prior structure distribution (*e.g.*, if a certain causal relationship is generally accepted in the literature, then structures that have this relationship can be assigned a higher probability) (Hartemink et al., 2001). However, the key drawback of this approach is that it cannot identify feedback loops in the regulatory network due to the acyclic nature of the network structure. This provides the motivation for investigating model-based approaches as described next.

Model-based Methods

A variety of model-based approaches have been investigated for deciphering genetic regulatory networks from gene expression data. A Boolean network representation was among the first formalisms proposed to model gene interactions (Somogyi and Sniegoski, 1996; Akutsu et al., 1999; Ideker et al., 2000). In this approach, genes are assumed to be either ON or OFF and the input-output relationships between them are modeled through deterministic logical functions (such as AND, OR, NOT, *etc.*). More recently, an extension of this approach to account for uncertainty in expression data has been proposed in the form of probabilistic Boolean networks (Akutsu et al., 2000; Shmulevich et al., 2002). The strong simplifying assumptions on which these approaches are based, such as Boolean gene expression and synchronous network dynamics, enable the analysis of relatively large regulatory networks. However, in most real gene expression settings, these underlying idealizations may not be appropriate since (i) genes are frequently expressed at intermediate expression levels and (ii) time delays are observed in state transitions (Jong, 2002). In the light of these limitations, more general approaches have been proposed. These include linear weight modeling (D'haeseleer et al., 1999; Weaver et al., 1999; Zak et al., 2001), ordinary differential equations (Chen et al., 1999) and S-systems (Savageau, 1998; Akutsu et al., 2000; Maki et al., 2001).

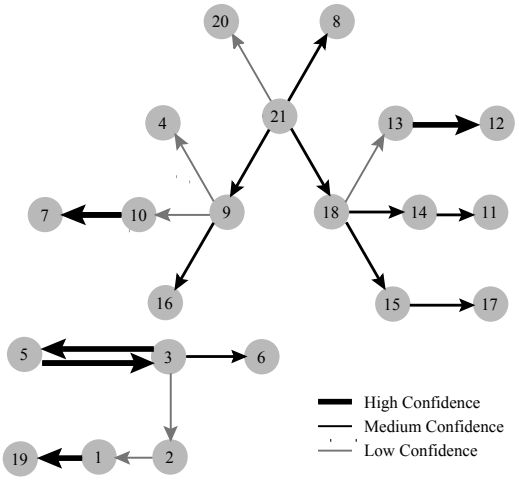

Figure 10. Bayesian network inferred from yeast microarray data.

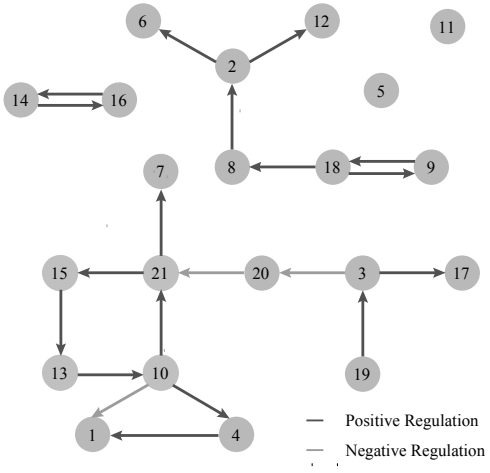

Figure 11. Regulatory network uncovered through the model-based approach.

We are currently exploring a linear, additive model embedded within a constrained optimization framework for extracting the underlying regulatory network from gene expression time-series data. Specifically, binary (0-1) variables Y_{ij}^{+}/Y_{ij}^{-} are used to model whether gene *j* enhances/represses the expression of gene *i*. The strength of these regulatory interactions and their temporal variation is captured through the regulatory coefficient variable r_{ijt}. A linear additive relationship is assumed between the gene expression levels as given by

$$X_{it} = \sum_{j=1}^{N} X_{jt}^{e} r_{ijt} \left(Y_{ij}^{+} - Y_{ij}^{-} \right)$$

where X_{it} is the predicted expression level of gene *i* based on the experimentally measured expression levels X_{jt}^{e} (*j* = 1,…,*N*) at time point *t*. Subsequently, the total absolute error between the predicted and the experimental gene expression levels is minimized to determine the optimal network connectivity through the solution of the following optimization model.

$$\min \sum_{i=1}^{N} \sum_{t=1}^{T} \left| X_{it} - X_{it}^{e} \right|$$

subject to

$$X_{it} = \sum_{j=1}^{N} X_{jt}^{e} r_{ijt} \left(Y_{ij}^{+} - Y_{ij}^{-} \right)$$

$$Y_{ij}^{+} + Y_{ij}^{-} \le 1$$

$$0 \le r_{ijt} \le r_{ij}^{\max} \left(Y_{ij}^{+} + Y_{ij}^{-} \right)$$

$$Y_{ij}^{+}, Y_{ij}^{-} \in \{0,1\}$$

The second constraint in the above model formulation enforces exclusivity on the type of regulatory interaction that can be observed between any two genes. Bounds on the regulatory effect of this interaction are provided by the third constraint. The proposed model-based approach has the following key advantages.

1. Additional biological knowledge in terms of limits on the maximum number of total interactions M^{tot} and individual gene interactions M_i is systematically incorporated through additional constraints such as $\sum_{i=1}^{N} \sum_{j=1}^{N} \left(Y_{ij}^{+} + Y_{ij}^{-} \right) \le M^{tot}$ and $\sum_{j=1}^{N} \left(Y_{ij}^{+} + Y_{ij}^{-} \right) \le M_i$.
2. Time delays and temporal variations in the network interactions are identified through the time-dependent regulatory coefficient. Also, changes in the network connectivity over time can be readily accommodated by introducing a time index on the binary variable.
3. The proposed modeling approach has the flexibility to account for uncertainty in experimentally measured expression levels. Specifically, this can be achieved by includinga probabilistic description of the experimental data in terms of gene expression *scenarios* within the model setting in conjunction with a stochastic programming framework.

For highlighting the applicability of the proposed approach, it is applied to the 21 gene data set previously analyzed with the Bayesian network approach. The regulatory network obtained is shown in Figure 11. The total number of allowed interactions is limited to 20, the number of interactions identified in the Bayesian network (Figure 10), and a maximum limit of 4 parents per gene is enforced.

Comparison of the networks obtained by the two alternative approaches indicates that even though there are significant differences in the network connectivity, certain broad features are conserved. For instance, gene 21 emerges as a central gene in both networks. Also, some gene interactions are either modulated with additional intermediate genes (for instance, the interaction between

gene 7 and gene 10 in the Bayesian network appears with gene 21 as an intermediate in the model-based network) or with reversed directionalities (*e.g.*, interaction between genes 20 and 21). The key advantage of the model-based approach is the identification of feedback loops, such as the one consisting of genes 10, 21, 15 and 13 in Figure 11.The fact that only one of the four high confidence links that was identified in the Bayesian network (10 to 7) is also identified in the model-based approach, highlights the need for exploring multiple network identification techniques in order to uncover a final "consensus" model.

Summary

In this paper, three problems of a biological nature studied in our group were addressed. Computational frameworks motivated by methodologies commonly used in process systems engineering were discussed for each of the problems. Both predictive and descriptive methods were included, illustrating the convergence of different methods and modeling techniques required in the fields of protein engineering, metabolic engineering and regulatory network identification.

It is becoming increasingly apparent that the systems paradigm is most appropriate for biological analysis since organisms are much more than the sum of their individual parts (*i.e.*, genes, proteins, etc.). Due to the exponential increase in sequencing and proteomic data, the use of formal systems engineering algorithms and methodologies in computational biology will become increasingly vital in the upcoming years. Challenges will be plentiful due to the necessity of connecting the genome/proteome with phenotypic characteristics, and systems engineering is well positioned to capitalize on these opportunities.

Acknowledgements

Financial support from National Science Foundation Awards BES0120277 and CTS9701771 and hardware support by the IBM-SUR program are gratefully acknowledged.

References

Akutsu, T., Miyano, S. and Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Proc. Pac. Symp. Biocomput. (PSB 1999), 4*, 17.

Akutsu, T., Miyano, S. and Kuhara, S. (2000). Inferring qualitative regulations in genetic networks and metabolic pathways. *Bioinformatics, 16*, 727.

Brakmann, S. (2001). Discovery of superior enzymes by directed molecular evolution. *ChemBioChem, 2(12)*, 865-871.

Brooke, A., Kendrick, D., Meeraus, A. and Raman, R. (1998). GAMS: A User's Guide. Washington, D.C., GAMS Development Corporation.

Burgard, A. P. and Maranas, C. D. (2001). Probing the performance limits of the Escherichia coli metabolic network subject to gene additions or deletions. *Biotechnol. Bioeng., 74*, 364-375.

Burgard, A. P., Vaidyaraman, S. and Maranas, C. D. (2001). Minimal Reaction Sets for *Escherichia coli* Metabolism under Different Growth Requirements and Uptake Environments. *Biotechnol. Prog., 17*, 791-797.

Chen, T., He, H. L. and Church, G. M. (1999). Modeling Gene Expression with Differential Equations. *Proc. Pac. Symp. Biocomput. (PSB 1999), 4*, 41.

Cho, M. K., Magnus, D., Caplan, A. L. and McGee, D. (1999). Policy forum: genetics. Ethical considerations in synthesizing a minimal genome. *Science, 286(5447)*, 2087, 2089-2090.

D'haeseleer, P., Liang, S. and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics, 16(8)*, 707.

D'haeseleer, P., Wen, X., Fuhrman, S. and Somogyi, R. (1999). Linear modeling of mRNA expression levels during CNS development and injury. *Proc. Pac. Symp. Biocomput. (PSB 1999), 4*, 41.

Edwards, J. S., Ibarra, R. U. and Palsson, B. O. (2001). In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat. Biotechnol., 19(2)*, 125-130.

Edwards, J. S. and Palsson, B. O. (2000). The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA, 97(10)*, 5528-5533.

Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA, 95*, 14863.

Eisenberg, D., Marcotte, E. M., Xenarios, I. and Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature*, 405, 823-826.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). Using bayesian networks to analyze expression data. *J. Comp. Biol., 7*, 601.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S. and Young, R. A. (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Proc. Pac. Symp. Biocomput. (PSB 2001), 6*, 422.

Heckerman, D. (1998). A tutorial on learning with bayesian networks. In M. I. Jordan, ed. *Learning in Graphical Models, Kluwer*, 301.

Henriksen, C. M., Christensen, L. H., Nielsen, J. and Villadsen, J. (1996). Growth energetics and metabolic fluxes in continuous cultures of Penicillium chrysogenum. *J. Biotechnol., 45(2)*, 149-164.

Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V. and Banavar, J. R. (2001). Dynamic modeling of gene expression data. *Proc. Natl. Acad. Sci. USA, 98*, 1693.

Huynen, M. (2000). Constructing a minimal genome. *Trends Genet., 16*, 116.

Ideker, T. E., Thorsson, V. and Karp, R. M. (2000). Discovery of regulatory interactions through pertubation: Inference

and experimental design. *Proc. Pac. Symp. Biocomput. (PSB 2000), 5*, 302.

Jong, H. D. (2002). Modeling and Simulation of Genetic Regulatory Systems: A Literature Review. *J. Comp. Biol., 9*, 67.

Jorgensen, H., Nielsen, J. and Villadsen, J. (1995). Metabolic Flux Distributions in Penicillium chrysogenum During Fed-Batch Cultivations. *Biotechnol. Bioeng., 46*, 117-131.

Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., Pellegrini-Toole, A., Bonavides, C. and Gama-Castro, S. (2002). The EcoCyc Database. *Nucleic Acids Res., 30(1)*, 56-58.

Lee, S., Phalakornkule, C., Domach, M. M., Grossmann, I. E. (2000). Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comp. Chem. Eng.*, 24, 711-716.

Maki, Y., Tominaga, D., Okamoto, M., Watanabe, S. and Eguchi, Y. (2001). Development of a system for the inference of large scale genetic networks. *Proc. Pac. Symp. Biocomput. (PSB 2001), 6*, 446.

Moore, G. L. and Maranas, C. D. (2002). eCodonOpt: a systematic computational framework for optimizing codon usage in directed evolution experiments. *Nucleic Acids Res., 30(11)*, 2407-2416.

Moore, G. L., Maranas, C. D., Lutz, S. and Benkovic, S. J. (2001). Predicting crossover generation in DNA shuffling. *Proc. Natl. Acad. Sci. USA, 98(6)*, 3226-3231.

Neidhardt, F. C. and Curtiss, R. (1996). Escherichia coli and Salmonella : cellular and molecular biology. Washington, D.C., ASM Press.

Ostermeier, M., Shim, J. H. and Benkovic, S. J. (1999). A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotechnol., 17(12)*, 1205-1209.

Pe'er, D., Regev, A., Elidan, G. and Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. *Bioinformatics, 17(S1)*, S215.

Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K. and White, O. (2001). The Comprehensive Microbial Resource. *Nucleic Acids Res., 29(1)*, 123-125.

Petrounia, I. P. and Arnold, F. H. (2000). Designed evolution of enzymatic properties. *Curr. Opin. Biotechnol., 11(4)*, 325-330.

Pons, A., Dussap, C. and Pequignot, C. (1996). Metabolic flux distribution in Cornybacterium melassecola ATCC 17965 for various carbon sources. *Biotechnol. Bioeng., 51*, 177-189.

Pramanik, J. and Keasling, J. D. (1997). Stoichiometric Model of *Escherichia coli* Metabolism: Incorporation of Growth-Rate Dependent Biomass Composition and Mechanistic Energy Requirements. *Biotechnol. Bioeng., 56*, 398-421.

SantaLucia Jr., J. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA, 95(4)*, 1460-1465.

Savageau, M. A. (1998). Rules for the evolution of gene circuitry. *Proc. Pac. Symp. Biocomput. (PSB 1998), 3*, 54.

Schmidt-Dannert, C. (2001). Directed evolution of single proteins, metabolic pathways, and viruses. *Biochemistry, 40(44)*, 13125-13136.

Shmulevich, I., Dougherty, E. R., Kim, S. and Zhang, W. (2002). Probabilistic Boolean networks: a rule based uncertainty model for gene regulatory networks. *Bioinformatics, 18*, 261.

Short, J. M. (1999). US5,965,408: Method of DNA Reassembly by Interrupting Synthesis.

Somogyi, R. and Sniegoski, C. A. (1996). Modeling the complexity of genetic networks: Understanding multigenic and pleitropic regulation. *Complexity, 1(6)*, 45.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P. and Botstein, D. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell, 9*, 3273.

Stemmer, W. P. C. (1994a). DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc. Natl. Acad. Sci. USA, 91*, 10747-10751.

Stemmer, W. P. C. (1994b). Rapid evolution of a protein in vitro by DNA shuffling. *Nature, 370*, 389-391.

Stemmer, W. P. C. (2000). US6,132,970: Methods of Shuffling Polynucleotides.

Varma, A., Boesch, B. W. and Palsson, B. O. (1993). Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Appl. Environ. Microbiol., 59(8)*, 2465-2473.

Varma, A. and Palsson, B. O. (1994). Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Bio/Technology, 12*, 994-998.

Weaver, D. C., Workman, C. T. and Stormo, G. D. (1999). Modeling Regulatory Networks with Weight Matrices. *Proc. Pac. Symp. Biocomput. (PSB 1999), 4*, 112.

Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Baker, J. L. and Somogyi, R. (1998). Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA, 95*, 334.

Zak, D. E., Doyle III, F. J., Gonye, G. E. and Schwaber, J. S. (2001). Simulation Studies for the Identification of Genetic Networks from cDNA Array and Regulatory Activity Data. *ICSB 2001 Proceedings*, 231-238.