



Improved understanding of gene expression regulation using systems biology

Robert S Kuczenski, Kunal Aggarwal and Kelvin H Lee[†]

This article reviews the current state of systems biology approaches, including the experimental tools used to generate 'omic' data and computational frameworks to interpret this data. Through illustrative examples, systems biology approaches to understand gene expression and gene expression regulation are discussed. Some of the challenges facing this field and the future opportunities in the systems biology era are highlighted.

Expert Rev. Proteomics 2(6), 915–924 (2005)

Gene expression is often considered to be the flow of biologic information from DNA, to mRNA and finally to functional protein products. However, this perspective does not explicitly consider genome-wide regulation of gene expression. Such regulation can occur at multiple levels. Some general examples of important types of regulation include competition for limiting substrates and local pathway control, including positive- and negative-feedback loops. These complexities contribute to a nonlinear trend observed in mRNA and protein expression profiles in various organisms [1–5]. Theoretical studies of gene expression networks have reinforced this observation [6,7]. These studies highlight that mRNA or protein expression information, by itself, is not sufficient to elucidate the relationship between genome sequence, gene expression and cellular dynamics. Due to the complex network of interactions involved, gene expression and its regulation are ideally suited to being studied using a framework that not only integrates expression data from multiple cellular levels, but also combines experimental and computational approaches.

When studying gene expression on a moderate or large system, a large number of measurements may be necessary to characterize the behavioral dynamics of the system. Computational tools are often necessary to extract the desired information from this large quantity

of data. These computational tools may also be used to develop predictive power to guide experimental design. Systems biology is defined here as an interactive collaboration between these computational and experimental efforts, to integrate system-wide information and to understand an otherwise incomprehensibly complex system, as illustrated in FIGURE 1. In this review, a system implies an organism, tissue, cell, cell compartment or any other biologic system of interest. The systems biology perspective is not a new concept [8–10], but the advent of new high-throughput experimental techniques has sparked a renewed interest in the field. For a more general review of the field of systems biology, the reader is directed to the currently available literature [11–13]. This article will focus on the application of systems biology to understanding gene expression and gene expression regulation. The article begins with a brief overview of some experimental and computational methods, as well as some data analysis and statistical tools relevant in these studies. Using illustrative examples, this review highlights current studies that benefited from utilizing a systems biology approach, tool development for these studies and a research area that may benefit from a systems biology approach. Finally, the article concludes with thoughts on the current status and future directions for this field.

CONTENTS

Experimental tools
& resources

Modeling & simulation of
gene expression networks

Systems biology
approaches

Expert commentary

Five-year view

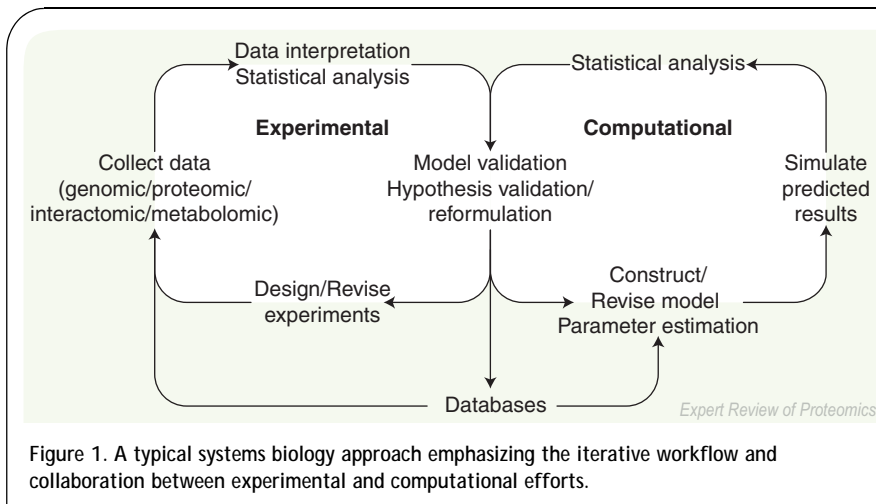
Key issues

References

Affiliations

[†]Author for correspondence
Cornell University, School of
Chemical & Biomolecular
Engineering, 120 Olin Hall,
Ithaca, NY-14853, USA
Tel.: +1 607 255 4215
Fax: +1 607 255 9166
KHL9@cornell.edu

KEYWORDS:
computational biology,
gene expression, systems biology



The proteome of a cell is commonly analyzed using gel-based approaches or shotgun proteomic methods. Both of these approaches involve separation of complex mixtures of proteins, either using 2D gel electrophoresis (2DE) [19–21] or using multidimensional chromatography [22–25], followed by a mass spectrometry (MS)-based analysis for protein characterization. New protein quantitation approaches involving isobaric labeling of proteolytic peptides resolve some of the limitations of the 2DE and traditional shotgun proteomic methods [5,26].

Metabolite target analysis, which involves analyses of samples for one or multiple compounds from a complex mixture,

is often carried out using gas chromatography (GC)/MS [27] and liquid chromatography (LC)/MS [28]. Nuclear magnetic resonance (NMR) has been used to generate metabolic fingerprints to detect effects of genetic perturbations in organisms [29]. It has also been used to determine carbon fluxes in known metabolic pathways [30].

Protein–protein and protein–small molecule interactions are more commonly studied using a yeast two-hybrid screening procedure [31]. Spotted protein arrays are also used to probe protein–small molecule interactions [32,33] and measure protein activity [34,35].

The quality of data collected using the techniques above can be affected by multiple factors, including technological, biologic and experiment-to-experiment variation. For example, technical variation may arise in microarrays due to variation in array manufacture, nonspecific hybridization to the arrays and so on. Biologic variation can be attributed to differences in the biologic make-up of different organisms from the same species. This situation can occur when cells from a single strain are cultured in different media types. Experiment-to-experiment variability is more user specific and can arise due to multiple factors, including differences in sample preparation. All of these factors of variation necessitate a good experimental design that includes technical and biologic replicates in the experiments. This design will help to estimate the variation in the expression measurements that results from the factors discussed above, and will improve the reliability of the data collected.

Data interpretation

A number of statistical tools have been developed for DNA microarray data processing and interpretation. Normalization procedures are available to remove the differences arising in measurements due to technical reasons such as labeling efficiency, scanning and other systematic biases in measurements [36]. Such procedures facilitate a direct comparison of multiple microarray measurements corresponding to different samples, with the goal of detecting changes in gene expression. Based on different normalization methods, a variety of algorithms have

Experimental tools & resources

Systems biology approaches to studying gene expression often involve genome-wide data collection in an organism subjected to different environmental and/or genetic perturbations, usually at multiple time points. Genetic perturbations could include gene mutation, overexpression, deletion, or post-transcriptional gene silencing. Genome-wide data collection may involve measurements of mRNA and protein expression, metabolite concentrations, and interactions between different macromolecules. Levels of mRNA and protein expression provide a direct measure of gene expression. The metabolic state of a system can also be used to gain further insights into gene expression, because this state is largely derived from global gene and protein expression. Further, metabolites can be considered to be the products of cellular regulatory processes, and the biochemical response of an organism can be characterized by its effect on the differential accumulation of individual metabolites. Information from measurements of gene and protein expression, and metabolite concentration, can be used to hypothesize possible regulatory pathways. However, a direct experimental measurement of the interactions between the different molecules involved can help validate the proposed mechanism of regulation. The following sections provide a brief overview of some of the high-throughput technologies that are currently being used to measure mRNA and protein expression levels, metabolite concentrations and protein–small molecule interactions. For a detailed description of these techniques, the reader is directed to appropriate references. A brief discussion of data analysis and storage paradigms currently available for interpreting and sharing the resulting data is also provided.

Data collection

Genome-wide parallel gene expression measurements are commonly carried out using spotted complementary DNA arrays [14] and high-density oligonucleotide arrays [15]. In addition to measuring the mRNA expression levels, DNA microarrays are also used to characterize spliced genes [16], resequence whole stretches of DNA [17], and identify single nucleotide polymorphisms (SNPs) [18].

been proposed to extract gene expression levels from signal intensities measured using DNA microarrays [37,38]. Although multiple methods have been proposed to process DNA microarray data, there is no consensus over the best data processing algorithm [37,38]. However, there are methods available to determine genes with statistically significant changes in expression, irrespective of the choice of the data-processing method [38,39].

A variety of data mining tools are also available to explore interactions between genes and to reveal patterns of expression. Clustering methods can be used to organize genes showing similar gene expression into meaningful groups that can be used to develop taxonomies. Principal component analysis (PCA) or singular value decomposition can be used to reduce the dimensionality of the gene expression data sets and to organize genes into different groups based on similar expression profiles [40,41].

Unlike the field of microarrays, few statistical tools are currently available to process data obtained using shotgun proteomic methods. However, data analysis involved in quantitation of protein expression by independent measurements of multiple peptides using MS is similar to the quantitation of mRNA expression using probe pairs on high-density oligonucleotide arrays. Some of the concepts involved in microarray data analysis, including noise reduction and normalization, are also relevant to the processing of data obtained using shotgun proteomic methods.

Metabolome data analysis is also similar to transcriptomic data analysis. Metabolite profiles of mutants of genes can be used to assign those genes into different functional categories. This kind of classification may be performed using statistical methods such as clustering and PCA on NMR spectra from cell extracts [42,43].

Data storage

Significant progress has been made recently to systematically store the data from biologic experiments in publicly available databases [44–50]. These databases store information related to mRNA expression profiles, metabolic pathways and chemical reactions, including the enzymes involved, macromolecular structural data

and protein–protein interactions (TABLE 1). A public database of MS spectra of various metabolites has also been established to facilitate unambiguous identification of metabolites in complex biologic samples [51]. These databases provide quick and easy access to currently available biologic information. This information may be required not only for judicious planning of new experiments, but also for formulating models.

To maximize the utility of these databases, it is important to store information in a systematic format so that it can be easily interpreted and efficiently searched. Standard formats for information storage have already been suggested for DNA microarray experiments (Minimum Information About A Microarray Experiment [MIAME]) [52], proteomics data [53] and metabolomic experiments (Minimum Information About A Metabolic Experiment [MIAMET]) [54]. It is also important to report data from experiments to databases in an unprocessed form so that the user can assess its quality and reliability. Furthermore, these databases should be regularly updated only with nonredundant information.

Modeling & simulation of gene expression networks

Mathematical modeling is a fundamental component of the systems biology approach, as illustrated in FIGURE 1. The modeling of gene expression networks can yield insights into their connectivity, control and stability. The product of these modeling efforts is often used to predict the response of a biologic system to a perturbation (e.g., change in environmental condition). By providing insight into system dynamics, model predictions can efficiently guide the design of future experiments to advance the understanding of gene expression networks. Numerous approaches to modeling exist; the reader is directed to more specific reviews on their implementation [55,56]. In this section, some of the important factors to be considered when constructing a mathematical model are reviewed. In addition, stability in gene expression networks and the differentiation of competing models will be covered. This section will conclude with a discussion of software applications and programming languages that are useful for mathematical modeling in systems biology studies.

Table 1. Commonly used repositories for biologic information.

Category	Database	Ref.
Genomes and analysis	National Center for Biotechnology Information (NCBI)	[101]
Expression profile-related databases	Gene Expression Omnibus (GEO) Stanford Microarray Database (SMD)	[44,45]
Enzymes and metabolic pathways	Kyoto Encyclopedia of Genes and Genomes (KEGG) Encyclopedia of <i>Escherichia coli</i> Genes and Metabolism (EcoCyc) Encyclopedia of Metabolic Pathways (MetaCyc)	[46–48]
Protein sequence and functional annotation	Swiss-Prot protein knowledgebase	[102]
Macromolecular structural data	Protein Data Bank (PDB)	[49]
Protein–protein interactions	Biomolecular Interaction Network Database (BIND) Database for Interacting Proteins (DIP)	[50,103]

Creating a mathematical model

The construction of a mathematical model begins with several decisions regarding the representation of the biologic system. These decisions relate to the scope of the network and the detail used therein to describe the biologic system in the model. The most basic information incorporated into a mathematical model of a gene network concerns the chemical species, such as mRNAs or proteins, and the connectivity of these species. This information may be analyzed using graph theory [57]. Large or poorly understood gene expression networks often use this analysis as a starting point to determine the possible connectivity and importance of different molecular species, as shown in FIGURE 2A. Ideker and coworkers used a graph-based approach to construct a physical interaction network that was consistent with a series of microarray experiments in yeast (an example of which is shown in FIGURE 2B) [58]. Using this network, the authors were able to predict a new regulatory mechanism and verify this mechanism experimentally. The inclusion of additional information, such as rate equations for the time evolution of the chemical species, creates a mechanism-based model that can be used to gain a greater insight into the roles of specific proteins, mRNAs and metabolites (FIGURE 2C). Smolen and coworkers built a detailed mechanism-based mathematical model of circadian rhythms in *Drosophila* utilizing a series of ordinary differential equations to represent the network [59]. They verified a number of experimental observations using this model, including adjustment of the phase of gene expression oscillation to light–dark cycles. Among other predictions, they proposed a mechanism by which the network maintains a constant period of oscillation over a wide range of temperatures.

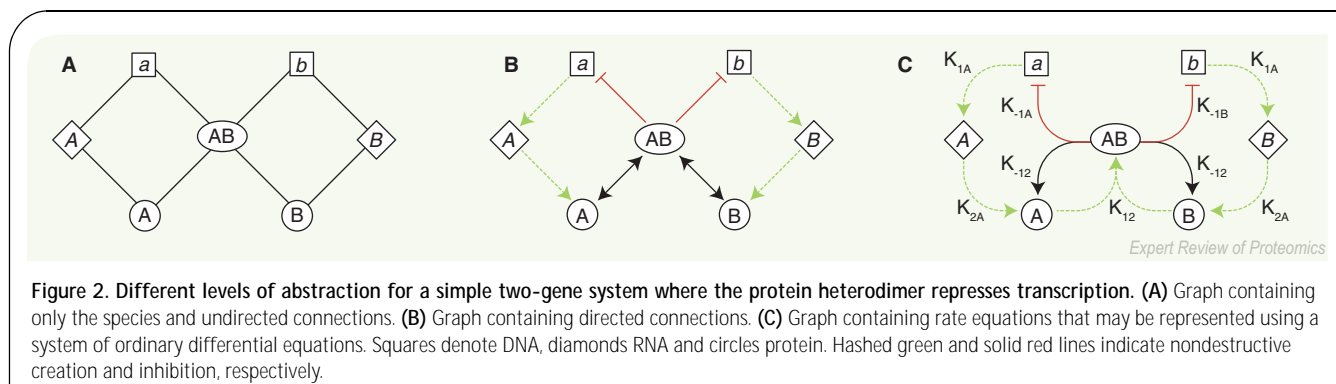
The scope and detail used in mechanism-based models is often limited by the computational resources available to the project, the type and amount of experimental data available for model validation, and the current understanding of the network topology and interactions. It is not always possible or necessary to include all levels of known detail, such as when this detail may overshadow the understanding of global gene expression. Using the available knowledge of the gene network from existing literature, several mathematical tools may aid in determining the sections of the network that are of greater importance to the description of the experimental data. One such tool is metabolic flux analysis (MFA), which calculates the

fraction of the overall flux through local pathways, thus providing information on the dominant pathways active in a biologic response [60]. Elementary mode analysis, an extension of MFA including structural analysis, divides a network into the simplest components and may also be used to determine important pathways in a gene network [61]. These tools can help reduce the scope or level of detail in the model without compromising the ability to describe available experimental observations.

An important decision in the use of mechanism-based models is the choice between a deterministic or stochastic solution. Deterministic solutions typically consist of a set of ordinary or partial differential equations that describe the connectivity and time-evolution of chemical species. These solutions may be preferred because they are computationally simpler than stochastic solutions. However, deterministic solutions do not represent systems effectively if the molecules are present in low numbers or concentrations. This is usually the case if the model incorporates information on transcription factors and cofactors. Stochastic solutions evolve the species' concentration by selecting reaction(s) based on the probability of their occurrence. These solutions can be used to gain insight into the response of gene expression networks to large fluctuations in species at low concentrations.

Tuning parameterized models

Mathematical models often contain adjustable parameters, which imply variables that are not explicitly determined in the solution (e.g., reaction rate constants). These parameter values can be obtained directly from experimental measurements or estimated using knowledge of similar reactions, and may be tuned to best describe observed experimental behavior. Computational algorithms used to tune parameter values require an objective function, which is a mathematical relation used to quantify the ability of a model to match experimental data. Common objective functions include the least-squares or χ^2 function for fitting quantitative experimental data (e.g., time-series concentration data) [62,63], and a cybernetic approach for fitting qualitative behavior (e.g., maximum growth) [8]. With an objective function defined, a number of different optimization algorithms can be used to find a set of parameters that best describe, or fit, the experimental data. Some common algorithms perturb the current parameter values to search for this best fit parameter set, but they do not assure



such a set [64]. Other algorithms examine all possible combinations of parameter values [65]. These global optimization algorithms can provide the best fit parameter set in a finite amount of time, but are not easily applicable to all types of mathematical models. With a best fit parameter set determined, a statistical test may be used to measure the ability of the model to describe the experimental data [63]. A model may or may not be able to accurately describe the experimental data. Misrepresentation of the biologic system in the model, such as the absence of a critical portion of the gene network, may be responsible for a model's inability to capture observed experimental behavior. Collaborative efforts between experimental and computational groups, a hallmark of systems biology, are key to discovering the sources of these inconsistencies between model predictions and experimental observations.

Most experimental data, including those from biologic systems, contain some error (as described previously). The propagation of experimental error to the estimated model parameters, and ultimately the model predictions, should be determined. Knowledge of the errors in model predictions may add confidence to the model predictions and enable a more accurate comparison between these predictions and experimental observations. A sensitivity analysis of a mathematical model is one such tool used to examine the effect of variations in parameter values on the model predictions [63]. Using an approximation of the objective function (described previously), the parameter variance may be determined from the extent of variation in parameter values allowed within the model while still maintaining its ability to fit the experimental data. An ensemble of parameter sets can also be used to propagate experimental error to both the estimated parameters and model predictions [66]. This ensemble can be generated by sampling parameter sets near the best fit parameter set, keeping only those that are consistent with the experimental data. Brown and coworkers used this method to generate novel predictions for a signaling network [67]. The authors demonstrated that, while the experimental data may not have constrained all model parameters, they could propagate the confidence in the experimental data to find precise predictions regarding the weight of different signaling modules in the response to two growth factors.

Statistically analyzing models

The ability to absorb moderate amounts of statistical noise without significant physiologic change, while maintaining the flexibility necessary to adapt to new conditions, is a fundamental property of biologic systems. Mathematical models of these systems should also have this property. Model solutions found using stochastic methods inherently generate information regarding the possible steady states of a system and the stability of these steady states. For models with a deterministic solution, a bifurcation analysis is a common method used to analyze the stability of different steady-state behaviors of a gene network [68]. This analysis examines the system stability to parameter variations and characterizes the boundaries between qualitatively different system behaviors (e.g., normal vs. cancerous cell growth).

Gene expression networks often have competing hypotheses to describe the observed experimental behavior. These hypotheses may be delineated by examining each models' ability to describe the experimental data. Several statistical analyses are available to determine which hypothesis, if any, is statistically better at describing the experimental behavior [63,66].

Software applications & programming languages

Several markup languages have been created to facilitate the transfer of models of gene networks between not only different software applications, but also various research groups. The most prominent of these languages is the systems biology markup language (SBML), which is designed for describing gene expression networks [69]. SBML is used in many systems biology applications and has been integrated with many common applications, such as Mathematica (Wolfram Research, Inc., USA) through MathSBML [70]. Another language used to describe models is CellML, which has greater flexibility in describing mathematical models than SBML, although it is not as widely supported by software applications. These and other markup languages will play an important role in the standardization of model descriptions, thereby enabling seamless exchange of models between different software applications in addition to collaboration between different research groups.

Many applications have been specifically developed to facilitate the modeling and simulation workflow typically found in systems biology, thus reducing the amount of time necessary to create and use a mathematical model. One example is the Exploratory Research for Advanced Technology (ERATO) systems biology workbench (SBW) [71], which is a modularized program designed to flexibly incorporate new computational tools as they are developed. A collaborative effort is currently underway to bridge SBW with the Biological Simulation Program for Intra- and Inter-Cell Processes (Bio-SPICE), another modular collection of biologic modeling tools focusing on spatiotemporal processes. This effort will create a large set of complimentary modeling tools [104].

For those who prefer to create a program to describe their model, interpreted languages provide a high-level platform including prepackaged mathematic, scientific and statistical tools. Python [105] and Perl [106] are currently among the most widely used interpreted languages. Python is an object-oriented scripting language and is well established in the scientific community. While Perl pre-dates Python, it lacks many of the scientific tools available to Python. Matrix Laboratory (MATLAB; The Mathworks, Inc., USA) is another interpreted language often used in scientific work that includes several core scientific packages. Since these languages have a higher computational cost than a compiled programming language, it is possible to incorporate core model components written in C/C++ or Fortran within these scripting languages. This feature allows the speed of C/C++ or Fortran to be combined with the extensive toolboxes available in these interpreted languages. The Simplified Wrapper and Interface Generator (SWIG) [107] and F2Py [108] are two examples of interface generators used to incorporate C/C++ and

Fortran modules into Python, respectively. These languages are easy to learn, and use C-like or prototype-like syntax to extend program functionality and increase productivity.

Systems biology approaches

The following section gives several examples of systems biology studies utilizing some of the aforementioned tools and resources to understand gene expression and its regulation. Examples of galactose (GAL) utilization in yeast, enhancing hemolysin secretion and profiling of colon cancer are used to illustrate the benefits of a systems biology approach. Since tool development remains at the forefront of systems biology, network component analysis is used to illustrate a new tool that has a great potential in systems biology studies. Finally, studies of circadian rhythms are provided as an example of a field that may benefit from a systems biology approach.

Galactose utilization in yeast

The pathway involved in GAL metabolism is turned on in yeast cells only in the presence of GAL. The genes, enzymes and metabolites involved in GAL metabolism in yeast have been well defined through various biochemical studies. Ideker and coworkers refined the understanding of the existing pathway of GAL utilization by performing a series of systematic perturbations to the critical components of the pathway, and verifying whether existing molecular interactions in the GAL network could account for observed gene expression changes [58]. Genetically altered yeast strains with a complete deletion of one of the nine GAL genes involved in transport, enzymatic or regulatory functions were grown in the presence and absence of GAL. Global mRNA expression in each perturbed state was compared with the expression in wild-type yeast grown in similar conditions to identify genes exhibiting a change in gene expression due to the introduced perturbation. Genes with similar expression responses were observed to belong to similar functional groups. The differences in protein abundances between the wild-type yeast in the presence and absence of GAL were also estimated using a shotgun proteomics approach involving isotopic labeling of proteins [20,22]. Using a combined transcriptomic and proteomic analysis, genes exhibiting a change in expression only at the protein level and not at the mRNA level were identified. Such gene products were hypothesized to be potential targets for post-translational modification.

A mathematical model of molecular interactions, which connected GAL with other metabolic processes in yeast, was constructed using information from public databases, as mentioned previously. The gene expression information from perturbation experiments was superimposed on this model to test its predictive power. The discrepancies between the predicted and observed expression responses were used to refine the model. Based on experimental data as well as model predictions, the authors suggested new hypotheses about regulation of GAL utilization, including a regulatory role of galactose-1-phosphate in controlling GAL gene expression, and validated these hypotheses using double gene-deletion experiments.

Enhanced hemolysin secretion in Escherichia coli

The secretion of recombinant protein products has the potential to simplify downstream product purification, which may lead to significant decreases in production costs. The secretion of hemolysin via the Type 1 hemolysin secretion pathway in *Escherichia coli* was increased using a series of experimental and computational efforts [72]. Initially, a parent *E. coli* strain was engineered to express the hemolysin (Hly) secretion pathway. This strain was then subjected to random mutagenesis and selection for increased HlyA secretion (the natural substrate of the Type 1 hemolysin secretion pathway). A mutant showing a fourfold increase in HlyA secretion was selected for further analysis. The mRNA and protein expression profiles of the parent and mutant strains were characterized using high-density oligonucleotide arrays and a gel-based proteomics approach. The resulting mRNA and protein expression profiles revealed simultaneous downregulation of all observed transfer RNA synthetases. The authors hypothesized that this downregulation was due to a decrease in the translation rate.

Using a mathematical model of the translation process (whose parameters were tuned to describe the data already collected), a new *hlyA* nucleotide sequence was found that decreases the translation rate by 38%, without altering the amino acid sequence [73]. Using site-directed mutagenesis, these changes were introduced to *hlyA*. This newly engineered plasmid demonstrated an eightfold increase in secretion of HlyA over the parent strain. The combination of experimental and computational efforts in this systems biology study led to a deeper understanding of the Type 1 hemolysin secretion pathway and an increase in the secretion of HlyA through site-directed mutagenesis.

Integrated profiling of cancer

Many researchers have taken a reductionist approach towards cancer research. However, a few large-scale integrated efforts combining the genome, proteome and metabolome have been reported [74–76]. The majority of these integrated approaches have combined either genomic and transcriptomic data or transcriptomic and proteomic data in an attempt to understand underlying mechanisms of cancer. Modeling efforts have also been undertaken to predict the response of cancer cells to various drugs. A mechanism-based human colon cancer model, consisting of interconnected signal transduction pathways and gene expression networks, has been constructed [77]. The information for the possible interactions and kinetic data used in this model was obtained from public databases. Time-course experiments conducted on Caco2 and HCT116 colon cell lines were used to collect mRNA abundance and protein activity data, which were then used to estimate unknown parameters in the model. The model was used to predict the change in protein concentrations with respect to time inside a single cell during a complete cell cycle. The simulation results were used to extract the proliferation rates of cells, which were compared with experimental results. The model has also been used to predict the cellular phenotype after a gene knockdown. Knockdown of

Cyclin D, a gene that regulates cellular growth, was predicted to cause cell cycle arrest after one cell division, as observed experimentally. Using a systems biology approach, the authors were able to build a human colon cancer model that can simulate the concentration profiles of different molecules in the cell and can thus be used to identify molecules with altered concentration in the disease state. The model can also help to determine the mechanism of a particular drug by examining its effect on the concentration of these molecules.

Network component analysis

Liao and coworkers have developed a new analysis method, network component analysis (NCA), which integrates gene expression data with known connectivity information between genes and their transcription factors (TFs) to predict transcription factor activities (TFAs) [78,79]. TFA is the strength of the signal transmitted by a TF to different promoters. NCA is based on the principle that a TF has to be in an activated form to bind DNA and regulate a promoter. This activation may require post-translational modification or the binding of specific ligands to the active gene product. Hence, the gene expression level of a TF does not necessarily correspond to its activity level. NCA also predicts the contribution of a particular TF in controlling a specific gene that is regulated by multiple TFs by calculating the control strength of that TF on that gene. This method has been used to predict dynamic activity profiles of various TFs related to cell cycle regulation in *Saccharomyces cerevisiae* [79] and to reconstruct regulatory signals in *E. coli* during transition from glucose to acetate [78]. The information on connectivity between TFs and genes in these studies was obtained from genome-wide location data available in public databases. This approach can be used to distinguish changes in gene regulation by various TFs in the diseased versus normal state. Furthermore, it can be used to identify the contribution of various TFs in regulating the expression of different genes in these states, thus providing a potential method of identifying prospective targets for therapy.

Circadian rhythm in *Drosophila*

A circadian cycle of gene expression has been found in organisms ranging from bacteria to insects and mammals. This periodic oscillation is closely linked to several important physiologic characteristics, such as sleep-wake patterns [80]. The central components of the core gene network have been elucidated over the past 30 years using reductionist approaches (reviewed in [81]). Modeling of this gene network has primarily focused on verifying the oscillatory experimental observations and the stability of these oscillations [82,83]. Current efforts are using detailed mathematical models to gain a more complete picture of the control and regulation of the network, such as light entrainment [59]. High-throughput experimental work has recently begun to focus primarily on microarray data to analyze the transcriptome in the brain [84]. The combination of these efforts in a systems biology approach may yield deeper insight into the mechanisms of the central pacemaker and the role of this system in higher organisms.

Expert commentary

Systems biology is a technology-driven field, and the development of experimental and computational tools will continue to be an integral part of the field for the foreseeable future. Significant developments in the ability to make genome-wide measurements have enabled a shift in the reductionist approach undertaken in gene expression studies towards an integrative methodology. Recent systems biology approaches to understand gene expression and regulation have been successful in improving the existing knowledge of the gene networks studied. Although there is no definitive paradigm for systems biology that always leads to new biologic insights, the attempts to integrate information from different 'omic' levels are necessary to move the field forward. Current attempts also involve mathematical modeling of the biologic processes under consideration. This integration is necessary because modeling provides an important avenue for comparing hypotheses for system behavior with the experimental data. Furthermore, selecting between competing hypotheses using a validated model of a biologic system to guide experimental design is often faster than a purely experimental approach. This reduces the number of experiments necessary to differentiate between the correct and incorrect hypotheses.

An increasing number of computational groups are using experimental data collected in different laboratories. This is an encouraging trend that not only brings together inputs from groups specializing in different areas, but also reduces the time invested by a single group in iteratively collecting data and refining models. The collaborative use of experimental data by these diverse groups requires the storage of experimental data in a fully defined manner in publicly available repositories. Furthermore, the data from experiments designed to perturb the cellular machinery (e.g., the individual components of metabolic and signaling pathways) should also be stored to reduce the redundancy in the experiments performed by groups working in the same field.

Successful systems biology studies will require the skills of experts from a variety of fields such as physics, mathematics and engineering, in addition to biochemistry and biology. Incorporation of biology coursework into disciplines such as physics and engineering, and the inclusion of mathematics and statistics exercises into life sciences programs, will foster improved communication and collaboration between the fields. In addition to enabling collaboration, these courses can create well-rounded students who will apply their knowledge to novel systems biology research as the next generation of scientists.

Five-year view

Before effective and widespread use of systems biology can be realized, a more complete and standardized set of tools needs to be developed for both the experimental and computational components. Although significant developments have been made in obtaining high-throughput data at different 'omic' levels, there is an immediate need for mature statistical tools to separate noise from complex experimental data. The computational research needs to increase the amount of modeling program

reuse by developing a standardized set of approaches to constructing models of biologic systems, simulating their behavior and statistically analyzing their results. The development of these experimental and computational tools will require the collaboration of experts across many fields, and must involve coordination with new educational curricula.

Acknowledgements

This work was supported in part by National Science Foundation (NSF) BES-0120315 and the New York State Office for Science, Technology, and Academic Research.

Key issues

- Systems biology requires collaborative experimental and computational efforts.
- Systems biology will continue to be a technology-driven field as high-throughput experimental techniques are created and developed.
- Mathematical modeling tools can provide insight into complex systems, and the construction of reusable modeling tools will maximize the impact and efficiency of computational research.
- Databases and other standardized methods for sharing information will be crucial for this collaborative effort to succeed.

References

Papers of special note have been highlighted as:

• of interest

•• of considerable interest

- 1 Anderson L, Seilhamer J. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* 18, 533–537 (1997).
- **First study that highlights the nonlinear relationship between mRNA and protein expression profiles.**
- 2 Griffin TJ, Gygi SP, Ideker T *et al.* Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* 1(4), 323–333 (2002).
- 3 Lee PS, Shaw LB, Choe LH, Mehra A, Hatzimanikatis V, Lee KH. Insights into the relation between mRNA and protein expression patterns: II. Experimental observations in *Escherichia coli*. *Biotechnol. Bioeng.* 84(7), 834–841 (2003).
- 4 Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* 19(3), 1720–1730 (1999).
- 5 Ross PL, Huang YLN, Marchese JN *et al.* Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* 3(12), 1154–1169 (2004).
- 6 Mehra A, Lee KH, Hatzimanikatis V. Insights into the relation between mRNA and protein expression patterns: I. Theoretical considerations. *Biotechnol. Bioeng.* 84(7), 822–833 (2003).
- 7 Hatzimanikatis V, Lee KH. Dynamical analysis of gene networks requires both mRNA and protein expression information. *Metab. Eng.* 1(4), 275–281 (1999).
- 8 Wiener N. *Cybernetics or Control and Communication in the Animal and the Machine*. The MIT Press, Cambridge, UK (1948).
- 9 Mesarovic MD. *System Theory and Biology*. Springer-Verlag, NY, USA (1968).
- 10 Bertalanffy LV. *General System Theory, Foundations, Development, Applications*. George Braziller, NY, USA (1969).
- 11 Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* 2, 343–372 (2001).
- **Provides an overview of the systems biology approach to studying a biologic system.**
- 12 Kitano H. Systems biology: a brief overview. *Science* 295(5560), 1662–1664 (2002).
- **Provides an overview of the systems biology approach to studying a biologic system.**
- 13 Aggarwal K, Lee KH. Functional genomics and proteomics as a foundation for systems biology. *Brief Funct. Genomic Proteomic* 2(3), 175–184 (2003).
- 14 Lockhart DJ, Winzler EA. Genomics, gene expression and DNA arrays. *Nature* 405(6788), 827–836 (2000).
- **One of the first published studies using DNA microarrays to measure mRNA expression.**
- 15 Selinger DW, Cheung KJ, Mei R *et al.* RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nature Biotechnol.* 18(12), 1262–1268 (2000).
- **One of the first published studies using DNA microarrays to measure mRNA expression.**
- 16 Fehlbauer P, Guihal C, Bracco L, Cochet O. A microarray configuration to quantify expression levels and relative abundance of splice variants. *Nucleic Acids Res.* 33(5), e47 (2005).
- 17 Garcia CK, Mues G, Liao YL *et al.* Sequence diversity in genes of lipid metabolism. *Genome Res.* 11(6), 1043–1052 (2001).
- 18 Cho RJ, Mindrinos M, Richards DR *et al.* Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. *Nature Genet.* 23(2), 203–207 (1999).
- 19 O'Farrell PH. High-resolution 2-dimensional electrophoresis of proteins. *J. Biol. Chem.* 250(10), 4007–4021 (1975).
- 20 Finehout EJ, Lee KH. An introduction to mass spectrometry applications in biological research. *Biochem. Mol. Biol. Edu.* 32(2), 93–100 (2004).
- 21 Lee KH. Proteomics: a technology-driven and technology-limited discovery science. *Trends Biotechnol.* 19(6), 217–222 (2001).
- 22 Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.* 17(10), 994–999 (1999).
- 23 Cagney G, Emili A. *De novo* peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nature Biotechnol.* 20(2), 163–170 (2002).
- 24 Conrads TP, Alving K, Veenstra TD *et al.* Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and ¹⁵N-metabolic labeling. *Anal. Chem.* 73(9), 2132–2139 (2001).
- 25 Yao XD, Freas A, Ramirez J, Demirev PA, Fenselau C. Proteolytic ¹⁸O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal. Chem.* 73(13), 2836–2842 (2001).

- 26 Thompson A, Schafer J, Kuhn K *et al*. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.* 75(8), 1895–1904 (2003).
- 27 Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L. Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry. *Plant J.* 23(1), 131–142 (2000).
- 28 von Roepenack-Lahaye E, Degenkolb T, Zerjeski M *et al*. Profiling of *Arabidopsis* secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiol.* 134(2), 548–559 (2004).
- 29 Fiehn O. Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* 48(1–2), 155–171 (2002).
- 30 Grivet JP, Delort AM, Portais JC. NMR and microbiology: from physiology to metabolomics. *Biochimie* 85(9), 823–840 (2003).
- 31 Bartel PL, Roecklein JA, SenGupta D, Fields S. A protein linkage map of *Escherichia coli* bacteriophage T7. *Nature Genet.* 12(1), 72–77 (1996).
- 32 Ren B, Robert F, Wyrick JJ *et al*. Genome-wide location and function of DNA binding proteins. *Science* 290(5500), 2306–2309 (2000).
- 33 Krylov AS, Zasedateleva OA, Prokopenko DV, Rouviere-Yaniv J, Mirzabekov AD. Massive parallel analysis of the binding specificity of histone-like protein HU to single- and double-stranded DNA with generic oligodeoxyribonucleotide microchips. *Nucleic Acids Res.* 29(12), 2654–2660 (2001).
- 34 Houseman BT, Huh JH, Kron SJ, Mrksich M. Peptide chips for the quantitative evaluation of protein kinase activity. *Nature Biotechnol.* 20(3), 270–274 (2002).
- 35 MacBeath G, Schreiber SL. Printing proteins as microarrays for high-throughput function determination. *Science* 289(5485), 1760–1763 (2000).
- 36 Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19(2), 185–193 (2003).
- 37 Irizarry RA, Hobbs B, Collin F *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2), 249–264 (2003).
- 38 Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 18(4), 546–554 (2002).
- 39 Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* 98(9), 5116–5121 (2001).
- 40 Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA* 97(18), 10101–10106 (2000).
- 41 Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.* 455–466 (2000).
- 42 Raamsdonk LM, Teusink B, Broadhurst D *et al*. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnol.* 19(1), 45–50 (2001).
- 43 Nicholson JK, Lindon JC, Holmes E. ‘Metabonomics’: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29(11), 1181–1189 (1999).
- 44 Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30(1), 207–210 (2002).
- 45 Gollub J, Ball CA, Binkley G *et al*. The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.* 31(1), 94–96 (2003).
- 46 Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28(1), 27–30 (2000).
- 47 Keseler IM, Collado-Vides J, Gama-Castro S *et al*. EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.* 33, D334–D337 (2005).
- 48 Krieger CJ, Zhang PF, Mueller LA *et al*. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* 32, D438–D442 (2004).
- 49 Berman HM, Westbrook J, Feng Z *et al*. The Protein Data Bank. *Nucleic Acids Res.* 28(1), 235–242 (2000).
- 50 Bader GD, Betel D, Hogue CWV. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 31(1), 248–250 (2003).
- 51 Kopka J, Schauer N, Krueger S *et al*. GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* 21(8), 1635–1638 (2005).
- 52 Brazma A, Hingamp P, Quackenbush J *et al*. Minimum Information About a Microarray Experiment (MIAME) – toward standards for microarray data. *Nature Genet.* 29(4), 365–371 (2001).
- 53 Taylor CF, Paton NW, Garwood KL *et al*. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nature Biotechnol.* 21(3), 247–254 (2003).
- 54 Bino RJ, Hall RD, Fiehn O *et al*. Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.* 9(9), 418–425 (2004).
- 55 de Jong H. Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.* 9(1), 67–103 (2002).
- 56 Gardner TS, Faith JJ. Reverse-engineering transcription control networks. *Physics Life Rev.* 2, 65–88 (2005).
- 57 Barabasi AL, Oltvai ZN. Network biology: understanding the cell’s functional organization. *Nature Rev. Genet.* 5(2), 101–113 (2004).
- **Review of the tools used to analyze cellular networks.**
- 58 Ideker T, Thorsson V, Ranish JA *et al*. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292(5518), 929–934 (2001).
- 59 Smolen P, Hardin PE, Lo BS, Baxter DA, Byrne JH. Simulation of *Drosophila* circadian oscillations, mutations, and light responses by a model with VRI, PDP-1, and CLK. *Biophys. J.* 86(5), 2786–2802 (2004).
- 60 Stephanopoulos GN, Aristidou AA, Nielsen J. *Metabolic Engineering*. Academic Press, NY, USA, (1998).
- 61 Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED. Metabolic network structure determines key aspects of functionality and regulation. *Nature* 420(6912), 190–193 (2002).
- 62 Arndt RA, MacGregor MH. Nucleon–nucleon phase shift analyses by chi-squared minimization. In: *Nuclear Physics*. Alder B, Fernbach S, Rotenberg M (Eds), Academic Press, NY, 253–296 (1966).
- 63 Bevington PR, Robinson DK. *Data Reduction and Error Analysis for the Physical Sciences. Third Ed.* McGraw Hill, NY, USA (2003).
- **Useful text for analyzing and modeling experimental data.**

- 64 Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical Recipes in C: The Art of Scientific Computing, Second Ed.* Cambridge University Press, NY, USA (1992).
- 65 Tawarmalani M, Sahinidis NV. Global optimization of mixed-integer nonlinear programs: a theoretical and computational study. *Math. Program.* 99(3), 563–591 (2004).
- 66 Brown KS, Sethna JP. Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E.* 68(2–1), 021904/021901–021904/021909 (2003).
- **Novel approach to analyzing mathematical models of biologic systems.**
- 67 Brown KS, Hill CC, Calero GA, Lee KH, Sethna JP, Cerione RA. The statistical mechanics of complex signaling networks: nerve growth factor signaling. *Phys. Biol.* 1, 184–195 (2004).
- 68 Kuznetsov YA. Elements of applied bifurcation theory. In: *Applied Mathematical Sciences*. Antman SS, Marsden JE, Sirovich L (Eds), Springer-Verlag, NY, USA, 112, (2004).
- 69 Finney A, Hucka M. Systems biology markup language: level 2 and beyond. *Biochem. Soc. T.* 31(6), 1472–1473 (2003).
- 70 Shapiro BE, Hucka M, Finney A, Doyle J. MathSBML: a package for manipulating SBML-based biological models. *Bioinformatics* 20(16), 2829–2831 (2004).
- 71 Sauro HM, Hucka M, Finney A *et al.* Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *Omics* 7(4), 355–372 (2003).
- 72 Lee PS, Lee KH. Engineering HlyA hypersecretion in *Escherichia coli* based on proteomic and microarray analyses. *Biotechnol. Bioeng.* 89(2), 195–205 (2005).
- 73 Shaw LB, Zia RK, Lee KH. Totally asymmetric exclusion process with extended objects: a model for protein synthesis. *Phys. Rev. E. Stat. Nonlin. Soft Matter Phys.* 68(2 Pt 1), 021910 (2003).
- 74 Beer DG, Kardia SLR, Huang CC *et al.* Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Med.* 8(8), 816–824 (2002).
- 75 Chen GA, Gharib TG, Wang H *et al.* Protein profiles associated with survival in lung adenocarcinoma. *Proc. Natl Acad. Sci. USA* 100(23), 13537–13542 (2003).
- 76 Wu R, Lin L, Beer DG *et al.* Amplification and overexpression of the L-MYC proto-oncogene in ovarian carcinomas. *Am. J. Pathol.* 162(5), 1603–1610 (2003).
- 77 Christopher R, Dhiman A, Fox J *et al.* Data-driven computer simulation of human cancer cell. In: *Applications of Bioinformatics in Cancer Detection*. Umar A, Kapetanovic I, Khan J (Eds), New York Academy of Sciences, NY, USA, 132–153 (2004).
- 78 Kao KC, Yang YL, Boscolo R, Sabatti C, Roychowdhury V, Liao JC. Transcriptome-based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis. *Proc. Natl Acad. Sci. USA* 101(2), 641–646 (2004).
- 79 Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA* 100(26), 15522–15527 (2003).
- 80 Toh KL, Jones CR, He Y *et al.* An hPeR2 phosphorylation site mutation in familial advanced sleep phase syndrome. *Science* 291(5506), 1040–1043 (2001).
- 81 Stanewsky R. Genetic analysis of the circadian system in *Drosophila melanogaster* and mammals. *J. Neurobiol.* 54(1), 111–147 (2003).
- 82 Stelling J, Gilles ED, Doyle FJ III. Robustness properties of circadian clock architectures. *Proc. Natl Acad. Sci. USA* 101(36), 13210–13215 (2004).
- 83 Smolen P, Baxter DA, Byrne JH. Modeling circadian oscillations with interlocking positive and negative feedback loops. *J. Neurosci.* 21(17), 6644–6656 (2001).
- 84 McDonald MJ, Rosbash M. Microarray analysis and organization of circadian gene expression in *Drosophila*. *Cell* 107(5), 567–578 (2001).

Websites

- 101 National Center for Biotechnology Information
www.ncbi.nih.gov
(Viewed November 2005)
- 102 Swiss-Prot Protein Knowledgebase
http://us.expasy.org/sprot
(Viewed November 2005)
- 103 Database of Interacting Proteins
http://dip.doe-mbi.ucla.edu
(Viewed November 2005)
- 104 Bio-SPICE: Open Source Systems Biology
www.biospice.org
(Viewed November 2005)
- 105 Python Programming Language
www.python.org
(Viewed November 2005)
- 106 Perl Programming Language
www.perl.com
(Viewed November 2005)
- 107 Simplified Wrapper and Interface Generator (SWIG)
www.swig.org
(Viewed November 2005)
- 108 Fortran to Python Interface Generator
http://cens.ioc.ee/projects/f2py2e
(Viewed November 2005)

Affiliations

- Robert S Kuczenski, BSc
Graduate Student, Cornell University, School of Chemical & Biomolecular Engineering,
120 Olin Hall, Ithaca, NY 14853, USA
Tél.: +1 607 255 3832
Fax: +1 607 255 9166
RSK23@cornell.edu
- Kunal Aggarwal, BTech
Graduate Student, Cornell University, School of Chemical & Biomolecular Engineering,
120 Olin Hall, Ithaca, NY 14853, USA
Tél.: +1 607 255 3832
Fax: +1 607 255 9166
KA62@cornell.edu
- Kelvin H Lee, PhD
Associate Professor, Cornell University, School of Chemical & Biomolecular Engineering,
120 Olin Hall, Ithaca, NY 14853, USA
Tél.: +1 607 255 4215
Fax: +1 607 255 9166
KHL9@cornell.edu