OMICS A Journal of Integrative Biology
Volume 7, Number 3, 2003

A Sense of Life: Computational and Experimental Investigations with Models of Biochemical and Evolutionary Processes

BUD MISHRA,[1,3] RAOUL-SAM DARUWALA,[1] YI ZHOU,[1,2] NADIA UGEL,[1] ALBERTO POLICRITI,[8] MARCO ANTONIOTTI,[1] SALVATORE PAXIA,[1] MARC REJALI,[1] ARCHISMAN RUDRA,[1] VERA CHEREPINSKY,[1] NAOMI SILVER,[1] WILLIAM CASEY,[1] CARLA PIAZZA,[12] MARTA SIMEONI,[12] PAOLO BARBANO,[1] MARINA SPIVAK,[1] JIAWU FENG,[1] OFER GILL,[1] MYSORE VENKATESH,[1] FANG CHENG,[1,2] BING SUN,[1] IULIANA IONIATA,[1] THOMAS ANANTHARAMAN,[6] E. JANE ALBERT HUBBARD,[2] AMIR PNUELI,[1,3] DAVID HAREL,[9] VIJAY CHANDRU,[5] RAMESH HARIHARAN,[5] MICHAEL WIGLER,[3] FRANK PARK,[7] SHIH-CHIEH LIN,[3] YURI LAZEBNIK,[3] FRANZ WINKLER,[10] CHARLES R. CANTOR,[11] ALESSANDRA CARBONE,[4] and MIKHAEL GROMOV[4]

ABSTRACT

We collaborate in a research program aimed at creating a rigorous framework, experimental infrastructure, and computational environment for understanding, experimenting with, manipulating, and modifying a diverse set of fundamental biological processes at multiple scales and spatio-temporal modes. The novelty of our research is based on an approach that (i) requires coevolution of experimental science and theoretical techniques and (ii) exploits a certain universality in biology guided by a parsimonious model of evolutionary mechanisms operating at the genomic level and manifesting at the proteomic, transcriptomic, phylogenic, and other higher levels. Our current program in "systems biology" endeavors to marry large-scale biological experiments with the tools to ponder and reason about large, complex, and

[1]Department of Computer Science and Mathematics, Courant Institute of Mathematical Sciences, New York University, New York, New York.
[2]Department of Biology, New York University, New York, New York.
[3]Cold Spring Harbor Lab, Cold Spring Harbor, New York.
[4]Institut des Hautes Etudes Scientifiques, Le Bois-Marie, Bures-sur-Yvette, France.
[5]Strand Genomics, Bangalore, India.
[6]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin.
[7]School of Mechanical and Aerospace Engineering, Seoul National University, Kwanak-Gu, Seoul, Korea.
[8]Dipartimento di Matematica ed Informatica, Universitá degli Studi di Udine, Udine, Italy.
[9]Faculty of Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot, Israel.
[10]Research Institute for Symbolic Computation, University of Linz, RISC-Linz, Linz, Austria.
[11]Sequenom, Inc., San Diego, California.
[12]Dipartimento di Informatica, Universitá Cá Foscari di Venezia, Venice, Italy.

subtle natural systems. To achieve this ambitious goal, ideas and concepts are combined from many different fields: biological experimentation, applied mathematical modeling, computational reasoning schemes, and large-scale numerical and symbolic simulations. From a biological viewpoint, the basic issues are many: (i) understanding common and shared structural motifs among biological processes; (ii) modeling biological noise due to interactions among a small number of key molecules or loss of synchrony; (iii) explaining the robustness of these systems in spite of such noise; and (iv) cataloging multistatic behavior and adaptation exhibited by many biological processes.

INTRODUCTION

THE INTRODUCTION OF INFORMATION TECHNOLOGY into biology and biotechnology has already transformed and accelerated the nature of biological research. The impact of this transformation has been felt in the areas of agriculture, pharmacology, disease diagnosis and prognosis, forensics, defense against biowarfare, biometry, and, ultimately, in the manner in which we interact with our immediate natural environment. The biggest impact of information technology thus far has been at the descriptive level, enabling the collection of the bits and pieces of biological structures or part-lists that can be computed from the slew of data generated by modern high-throughput instruments. However, the contribution of information technology to biological research is still relatively minor when compared to what can be achieved if we are able to understand the functional properties of these part-lists.

Recent progress in our observational and experimental abilities has allowed us to understand the largely unobservable transparent structures of the cell. We are now able to catalog the genomic sequence of an organism, quantify the transcriptional states of a cell through microarrays, or even track photo-labeled single molecules *in vitro* or *in vivo*.

More recently, we have also become familiar with novel "computational" approaches that rely on simultaneous progress on many fronts: (a) vast amounts of computing power (through distributed or tightly-coupled parallel computers) have become available; (b) accurate physical models at kinetic mass-action, stochastic, spatio-temporal, and hybrid discrete/continuum levels can be created; (c) algorithmic efficiency can be achieved through symbolic and qualitative computation; and (d) logical reasoning systems and other analysis tools at multiple resolutions can be constructed with relative ease. These approaches borrow ideas from computational theory and logic, systems and engineering sciences, and applied mathematics.

A determined drive to realize large-scale computational tools for systems biology has come through DARPA's BioComp/BioSPICE project involving several investigators. As part of this research effort, several of us have been creating computational tools (e.g., Simpathica, NYU BIOSim, NYU BIOWave, and XS-System), integrating these tools with the other tools in the larger effort (www.biospice.org), and participating in design of the systems, languages, and experiments involved in this effort.

Thus, we posit that it is possible to create powerful simulation, analysis, and reasoning tools for working biologists that can be used in deciphering the functional properties of genomes, proteomes, cells, organs, and organisms by drawing upon mathematical and computational approaches developed in the fields of dynamical systems, kinetic analysis, computational theory, and logic. Thus, creating accurate and integrated tools for this purpose has become one of the grand challenges in computer science today. Coevolution of experimentation technology and design methods is crucial, lest we repeat the unfortunate history of "theoretical biology" from the early part of last century, which attempted to develop an abstract mathematical theory for biology without recourse to experimentation (works of D'arcy Thompson, Alan Turing, and Nicholas Refshavsky)

At present, there is no clear way to determine if the current body of biological facts is sufficient to explain the phenomenology. In the biological community, it is not uncommon to assume certain biological problems to have achieved a cognitive finality without rigorous justification. In these particular cases, rigorous mathematical models with automated tools for reasoning, simulation, and computation can be of enormous help in uncovering cognitive flaws, qualitative simplifications, or overly generalized assumptions. Ideal candidates for such study would include: the prion hypothesis, cell-cycle machinery (e.g., DNA repli-

cation and repair, chromosome segregation, cell-cycle period control, and spindle pole duplication), muscle contractility, processes involved in cancer (e.g., cell-cycle regulation, angiogenesis, DNA repair, apoptosis, cellular senescence, and tissue space modeling enzymes), signal transduction pathways, circadian rhythms (especially the effect of small molecular concentration on their robustness), and many others.

We believe a modern computational and systems theory that provides a solid mathematical foundation for biological systems is needed. While traditional systems theory focuses on simple behavioral attributes (such as reachability or robustness) of small, idealized systems (e.g., linear time-invariant systems), and classical computational theory focuses on the evolution of discretized, synchronous systems (e.g., finite state automata or Boolean networks), the methods of systems biology must be based on many complex and interconnected attributes.

We concentrate on four focus areas of research.

(a) Biochemical Process Theory. Biochemical process theory seeks to create a unified framework in which one can understand biochemical pathways and the evolutionary processes that shaped them. There are several interrelated components. These include models derived from kinetic mass-action that focus on concentrations of chemicals; models derived from discrete events such as commitment, differentiation, and self-renewal; models derived from cell-signaling; models derived from large populations of cells; and models accounting for cellular compartmentalization and transport across membranes and between compartments. The demands that these models impose on computer science go well beyond the biological areas and one quickly finds oneself grappling with diverse issues related to the numerical, probabilistic, logical, and symbolic aspects of computation, and, in particular, how they can be implemented on modern computer architectures.

(b) Evolutionary Processes, Genomes, and Pathway Models. The biochemical processes that have evolved in nature are governed by a set of simple processes that continue to alter genomes gradually. Duplication, translocation, deletion, and mutation are examples of such genomic processes. The main effort here is to develop a set of models with solid mathematical and biological foundations that allow for the systematic understanding of these probabilistic processes and the constraints they impose on biological systems, as well as the deciphering of the modular structure of large-scale biochemical processes. Furthermore, we exploit this structure through metamodeling and metalanguages for composing, transforming, and validating the basic building blocks of the models.

(c) Advanced Tool Architectures. Our toolkits comprise a set of reusable, inter-operating software modules integrated within a bioinformatics language and environment. The system is freely distributed as open-source software (as part of www.biospice.org) and allows the users to augment and share improved software within the user community. The main tools address simulation, reasoning, and analysis of biochemical processes, but also integrate with genome and genome-evolution analysis tools.

(d) Experimental Research. The program is founded upon large-scale *in vivo* and *in vitro* experiments that are able to reveal cognitive flaws, incompleteness, and false assumptions that may have been incorporated into the model. In order to achieve this, we concentrate on time-course gene-expression data that can be obtained through microarray analysis or similar methods based upon mass-spectroscopy. We also augment our data with proteomic data whenever possible.

Thus, the coevolving experimental research guides the theory, tool development, and model validation. The main emphasis is naturally placed on providing biologists and biotechnologists with the capability to analyze large and complex biological systems and devise intelligent experiments without being forced to deal with the mathematical details and complexity of the system. Hierarchy and composition are the basic cornerstones of our tools, methods, and models.

BIOCHEMICAL PROCESS THEORY

Several biological and biochemical mechanisms can be modeled with relatively simple sets of differential algebraic equations (DAE's). In the past, we have constructed and demonstrated to biologists the utility of a powerful computational tool with the ability to query massive sets of numerical data obtained from

in silico experiments on complex biological systems. The initial design of the computational tool derives its expressiveness, flexibility, and power from integrating many well-established and time-tested approaches in numerical analysis, symbolic computation, temporal logic, model checking, and visualization. The basic system has been successively augmented into a new system, dubbed "XS-system" (Antoniotti et al., 2002, 2003), as it extends the basic foundations provided by the "S-system models of biochemical processes." (See Voit, 2000, and also Antoniotti et al., 2002, Antoniotti et al., 2003, Cornish-Bowden, 1999, and Savageu, 1976.)

Our core design principles for XS-systems are based upon: (a) an elegant structure founded upon an extendable set of building blocks or models: for example, syntheses, degradations, reversible reactions, enzymatic reactions, reactions modulated by coenzymes, and reactions constrained by stoichiometric conditions; (b) flexibility in terms of compositionality and hierarchy; (c) expressibility; and finally, (d) the capability to create consistent semantics. Thus, in some sense, our rudimentary XS-system can be thought of as "the RISC (reduced instruction set computer) of systems biology" and provides the foundation upon which a complex language for systems biology can be built. For the same reasons, XS-systems also provide an elegant pedagogic tool for computer scientists and biologists to understand key biological processes as well as the algorithms used to describe them.

We have extended the core XS-system with many useful modules—in fact, a key design criterion of our system was the ease with which such new modules could be introduced. Obvious examples of these key modules include Hill equations, saturation effects, concentration changes due to cell volume growth, the effect of small numbers of molecules, and detailed models of transcription, RNA stability, and protein degradation. In each case, the extended system also allows for various representations such as models based on ODE's (Ordinary Differential Equations), SDE's (Stochastic Differential Equations), timed automata, or hybrid automata for each new module introduced, sometimes allowing for multiple representations with instructions on how a representation may be selected.

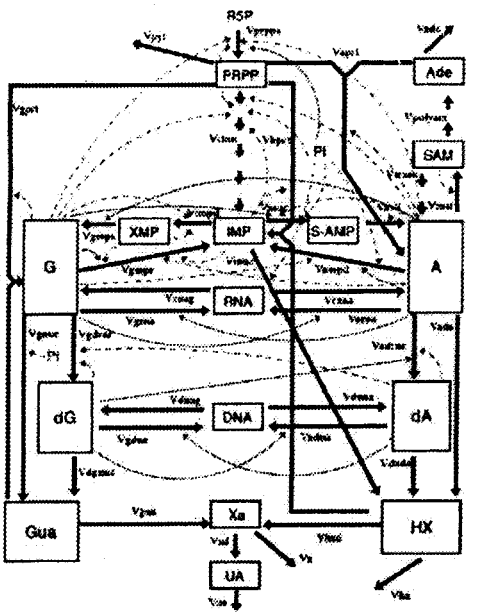
Our approach can be justified by simply noting that a vast number of models of interest to working biologists can be expressed and analyzed by the simplest core system. It is at this simplest core level that the most interesting biological mechanisms reveal themselves and provide the most useful insights for more experiments and more detailed or more complex models. As an anecdote, our group recently analyzed a Caspase cascade model for apoptosis with Lazebnik-lab at Cold Spring Harbor Lab: using the XS-system, a rough topological model was constructed and analyzed in less than half an hour. The model exhibited a quadratic growth of C3A when Caspase-9 was active and a linear growth in the absence of Caspase-9—an effect that was observed in the laboratory experiments (unpublished results). The simplest model also pointed out the possible presence of several unmodeled molecules as the core model could not explain various well-known thresholding effects. Thus, we contend that while we have left ample room for complexity in the system, it is for the simplicity, elegance, and compositionality that we have strived.

A natural completion for the XS-system is an automaton summarizing the states along which the simulated biochemical system evolves in time. The automaton, so generated, allows the user to view, manipulate, and reason about the system, using a well-integrated set of tools. As an example, we may consider the case study of purine metabolism (Antoniotti et al., 2002; Voit, 2000) which illustrates the fact that the "right" cellular behavior is often difficult to capture with an initial abstraction model (Fig. 1).

Our existing tools for studying biological processes are augmented with simple interfaces that allow a biologist to input a model, visualize the functions of the biological processes, iteratively improve the model, and construct "what-if" *in silico* experiments. Moreover, our tools have the ability to identify missing features and partial or incomplete models through the model-checking algorithms for temporal logic.

Biological preliminaries

The genome of an organism, the genetic core of a cell, is a collection of genes in its DNA. The role of a gene is to encode for protein structure. The sequence of amino acids, specified by DNA through the transcription and translation processes, determines the three-dimensional structure and biochemical properties of the proteins as well as the nature of their interactions. The proteins, in the form of transcription factors and other operons, may in turn regulate gene expression. Other factors, such as mRNA stability, protein degradation, post-translational modifications, and many other biochemical processes, tightly regulate the




```
Always (PRPP > 50 * PRPP1)
  implies
    steady_state()
    and Eventually(IMP1 > IMP2)
    and Eventually(HX < HX1)
    and Eventually(Always(IMP = IMP1))
    and Eventually(Always(HX = HX1))
```

The main metabolite in purine biosynthesis is 5-phosphoribosyl-α-1-pyrophosphate (PRPP). A linear cascade of reactions converts PRPP into inosine monophosphate (IMP). IMP is the central branch point of the purine metabolism pathway. IMP is transformed into AMP and GMP. Guanosine, adenosine, and their derivatives are recycled (unless used elsewhere) into hypoxanthine (HX) and xanthine (XA). XA is finally oxidized into uric acid (UA). In addition to these processes, there appear to be two "salvage" pathways that serve to maintain IMP levels and thus adenosine and guanosine levels as well. In these pathways, adenine phosphoribosyltransferase (APRT) and hypoxanthine-guanine phosphoribosyltransferase (HGPRT) combine with PRPP to form ribonucleotides.

FIG. 1. The pathway to the left depicts the classical (but incomplete) model of purine metabolism. Using the XS-system and the approaches described earlier, one can automatically create the model and perform model checking on the resulting qualitative automaton to show that the model fails to satisfy the temporal logic formula encoding robustness (bottom left). This reasoning process and manipulations with the XS-system ultimately led to a more complete model that does satisfy the robustness formula.

time-constants involved in the resulting biochemical machinery. (See Alberts et al., 1995, and Cantor and Smith, 1999.)

An enzyme, E, is a protein which can enhance the activity of a chemical reaction by attaching to a substrate, A, and making the formation of the product, P, energetically easier:

$$E + A \rightleftharpoons EA \rightarrow E + P$$

In general, equations of this kind take the form

$$A + B \rightleftharpoons_{K_-}^{K_+} C + D \quad \text{and} \quad [\dot{A}] = K_-[C][D] - K_+[A][B]$$

where the rate of change of A's concentration is given by the difference of the "synthesis rate" ($K\_[C][D]$) and the "dissociation rate" ($K_+[A][B]$).

XS-systems

Using a system of first-order differential equations (in explicit form), one can construct a general model of a complex biochemical reaction involving many genes and proteins.

The basic ingredients of our XS-system are the n dependent variables $X_1, \ldots, X_n$ and the m independent variables $X_{n+1}, \ldots, X_{n+m}$. Let $D_1, \ldots, D_{n+m}$ be the domains where the $n + m$ variables take values. We augment the form described in Voit, 2000, with a set of algebraic constraints which serve to characterize the conditions under which a given set of equations is derived from a set of maps.

The basic differential equations constituting the system take the following power law form:

$$\dot{X}_i = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n+m} X_j^{h_{ij}} \tag{1}$$

and

$$C_l(X_1(t), \ldots, X_{n+m}(t)) = \sum_j \left(\gamma_{lj} \prod_{k=1}^{n+m} X_k^{f_{ljk}} \right) = 0 \tag{2}$$

where the α_i's and β_i's are called rate constants. These rate constants govern the positive and negative contributions to a given substance. The γ_{lj}'s called rate constraints; these act concurrently with the exponents f_{ljk} to delimit the evolution of the system over a specified $(n + m)$–dimensional surface.

The set of differential equations in our XS-system can always be rewritten (recast) in a special canonical form by purely algebraic transformations and a further inclusion of a set of algebraic constraint equations. Canonical forms have several advantages over more general forms of equations, since they can be more easily manipulated, integrated, and interpreted in mathematical terms. Thus, the XS-system model we use has a simple canonical form: a list of expressions, each describing the rate of change of a given quantity in a model (say, the concentration of a reactant), plus a set of equations describing some constraints on the relationships among some of the parameters characterizing the model. Each of the rate-describing expressions has a very simple form as well: it is simply binomial—the difference between two algebraic power-products (or monomials), one representing synthesis and the other, dissociation.

A system of differential equations of the form described above can be integrated either symbolically or by numerical approximation. Frequently, standard approximation techniques suffice; however, in some cases, novel algorithmic techniques are required. Both symbolic and numerical cases are considered in our program on an equal footing, as we believe that without the reliance upon symbolic computational methods and composition properties, a system of this kind will ultimately fail when scaled to more realistic situations.

Traces and trace automata

We describe the semantics of an XS-system using an automaton that captures the qualitative features of the biological system. The automaton is constructed as follows: a snapshot of the system variables constitutes a possible state of the automaton. Transitions are inferred from traces of the system variables' evolution. Final states are those states in which the simulation reaches a recognizable end, and are supposed to represent the equilibrium points for the metabolic pathway being modeled.

More interesting states (with deteriorating computational efficiency implications) are constructed by grouping together several time instants according to simple rules, such as a linearization rule that groups states where the rate of change is within a user-defined parameter.

A major focus of our research is the question of how to tame the complexity of such automata by "collapsing the models" symbolically or modularly representing the models. This approach allows for studying multiple evolutions and experiments differing in rate constants and kinetic orders in parallel (within a single structure).

Note that the automata proposed here are not necessarily unique, and one may consider more complex automata with different semantics that are amenable to different kinds of logical analysis. For instance, we may extend the theory to: timed automata, hybrid automata, communicating automata, and "algebraic differential automata," where the states and state transitions are described by algebraic rules only. We aim to exploit these symbolic structures to create better reasoning systems: for example, variants of CAD (Cylindrical Algebraic Decomposition) to create qualitative automata, Morse-theoretic models of transitions be-

tween steady states, or differential algebraic elimination theory for input-output behavior and their usage in collapsing (Mishra, 2000, and Mishra, 2002).

Reasoning with temporal logic

A simple and natural example of a logical property of many biological systems is the one describing the existence of a steady-state. Informally, a system is in a "steady-state" when nothing "changes" in the system as time passes. Very often, the biologist knows not only that, in the absence of external stimuli, such a state must be reached sooner or later, but also the relative values of substances involved in such a state. Another natural property involves the boundedness of the reactant concentrations involved in a biological process and may need to be ascertained as a precondition to other interesting properties, such as the existence of a limit-cycle, multiple stable solutions, metastability, or hysteresis (Bhalla and Iyengar, 1999, and Wigler and Mishra, 2002).

A simple reasoning tool, based on the notion of model checking, may analyze a single trace of the automaton results, where the single trace is created from only one "set-up" (e.g., initial conditions, a set of values for the parameters, and a set of signaling events) for the system being analyzed. In order to allow more than one trace, it is necessary to consider different "set-ups," for example, many possible values for the parameters, such as rate constants and kinetic orders. These multiple "set-ups" may be defined parametrically (symbolically), deterministically sampled over a cartesian grid, or randomly sampled with respect to some reasonable distribution.

Fundamental to temporal logic (Gabbay et al., 1994, Mishra, 2002, and Mishra and Clarke, 1985) is the notion that time-dependent terms from natural language, such as "eventually" and "always," can be given a precise meaning (semantics) in terms of the abstract behavior of the system under discourse. As an example, consider the following sentence: "*The concentration of guanosine triphosphate* (*GTP*) *is equal to x*." Given a biological system in equilibrium, the above sentence may or may not be true at any or all instants of time. In particular, we can easily construct sentences (in a suitable natural language) that express the fact that, given a certain set of initial conditions, the above sentence will eventually hold true. Temporal logic precisely formalizes the meaning of the term "eventually" (and other such "modes": "always," "infinitely often," and "almost always"), and the resulting semantics lead to a precise model-checking algorithm for determining the validity of temporal logic sentences in the context of our trace automaton. We have also augmented the traces with time-frequency analysis using Linear Discriminant Bases (LDB), Local Karhunen-Loeve (LKL), and other techniques based on multi-resolution wavelet analysis. Temporal logic, thus enriched, allows us to understand the dynamic properties of the biological processes and to "cluster" different components of the system that may be co-regulated or anti-regulated.

Reasoning with statecharts

We have also been collaboratively investigating how to use the visual formalism of statecharts (Damm and Harel, 2001, Efroni et al., 2003, Harel, 1987, Harel and Gery, 1997, Harel and Marelly, 2003, Harel et al., 2003, Kam et al., 2001, and Kam et al., 2003) to address the challenge of analyzing biochemical processes. For instance, under this formalism, we have presented a detailed model for T cell activation using statecharts within the general framework of object-oriented modeling. Encouraged by our early successes, we have embarked upon a far more ambitious project—applying the same methodology to constructing a full detailed model of the developmental processes that lead to the formation of the egg-laying system in *C. elegans.*

We have demonstrated that the statechart-based approach is applicable to the challenge of modeling large numbers of biological processes in a hierarchical fashion and relating the various cellular events within the spatial and temporal context in which they occur to each other. Such models are useful for investigating the behavior of a system under many given scenarios, raising questions that were not thought of before and confronting questions which, because of their complexity, cannot be addressed by standard laboratory techniques and/or pure intuition alone. We believe that the statechart formalism integrates elegantly with formal verification methods and allows one to test whether the formal representation of the model fulfills the requirements that emerge from existing biological data.

Open questions

Several interesting questions remain to be further explored:

- **Reactions Models:** We have primarily focused on a simple ODE model using DAE's, and narrowed this even further to a model based on XS-systems. Does this imply a significant deviation from reality (Bhalla and Iyengar, 1999, and Guat et al., 2002)? How can a stochastic model representing small numbers of molecules that interact pair-wise and randomly be incorporated (Wigler and Mishra, 2002)? We have already devised an efficient spatial Gillespie-like algorithm to perform stochastic computations and used it successfully to understand a stem-cell model. We need to further explore how to combine the SDE models with the algebraic-constrained DAE models.
- **Hybrid Systems:** Certain interactions are purely discrete, and after each such interaction, the system dynamics may change. For a hybrid model of this kind the underlying automaton must be modified for each such mode. How do these enhancements modify the basic symbolic model?
- **Spatial Models:** The cellular interactions are highly specific to their spatial locations within the cell. How can these be modeled with symbolic cellular-automata? How can we account for the dynamics due to changes in the cell volume? The time constants associated with diffusion may vary from location to location; how can that be modeled?
- **State Space (Product Space):** A number of interacting cells can be modeled by product automata. In addition to the classical "state-explosion problem," we also need to pay attention to the variable structure due to (a) cell division, (b) apoptosis, and (c) differentiation.
- **Communication:** How do we model communication among cells mediated by interactions between extra-cellular factors and external receptors?
- **Hierarchical Models:** As we go to more and more complex cellular processes, a clear understanding can be obtained only through modularized hierarchical models. What are the ideal hierarchical models? How do we model a population of cells with related statistics?
- **Symbolic Verification:** If a biologist wishes to reason about a system with logical queries in an appropriate query language (e.g., temporal logic), what are the best query languages? What are the best algorithms that take advantage of the symbolic structures (Mishra, 2000, and Mishra, 2002)? What are the correct ways to solve problems associated with (a) model equivalence, (b) experimental analysis, and (c) reachability analysis?

EVOLUTIONARY PROCESSES, GENOMES, AND PATHWAY MODELS

Regulatory and metabolic processes in biology do not occur in isolation, nor are they static in nature. Hence, a better understanding of biology is hinged on a deep information-theoretic study of evolving genomes and their roles in governing metabolic and regulatory pathways.

Various biochemical and cellular processes—including point mutation, recombination, gene conversion, replication slippage, DNA repair, translocation, imprinting, and horizontal transfer—constantly act on genomes and drive the genomes to evolve dynamically. These alterations in the genomic sequences can further lead to the corresponding changes in the higher-level cellular information (transcriptome, proteome, interactome), and are crucial in explaining the myriads of biological phenomena in the higher-level cellular processes. (See Bailey et al., 2002, Buldyrev et al., 1993, Gerstein et al., 2001, Lucito et al., 2000, Ohno, 1970, Paxia et al., 2002, Peng et al., 1992, and Vulic et al., 1999.)

We have created a parsimonious mathematical model to explain various observations on the statistical structure of genomes (e.g., mer-frequency distribution in genomes) and its implications to the topology of the regulatory processes. Our model, based on the "evolution by duplication" theory originally proposed by Ohno in the 1970's (Ohno, 1970), is an extension of a Polya's urn model and considers genome evolution as a stochastic process with three main events: substitution, deletion, and duplication. A simpler model, based on evolving Eulerian graphs, fits nicely with real-world data for mer-frequency distributions. These results suggest that despite the highly diversified evolutionary environment for different organisms, the essential composition of the evolutionary dynamic, metabolic, and regulatory processes is commonly shared.

Genome evolution processes

Ohno's theory is well supported by molecular biology. There are various molecular mechanisms that can cause gene duplication. These include (a) unequal crossing over, (b) DNA polymerase slippage, and (c) heterologous recombination. If we assume that the target gene of every duplication is randomly chosen from the genes that are already in the genome, then we have a realization of a Polya's urn model. Therefore, under the "evolution by duplication" theory, genome evolution can be viewed as a stochastic duplication process that leads to a highly correlated structure in the genomes and repeated motifs in the regulatory network topologies.

In our Eulerian graph model, each mer-species of a particular length is represented by a node. Whenever two non-overlapping mers are immediately adjacent to each other in the genome, they are connected by a directed edge. Therefore, the in- and out-degree of a node indicates the copy number of the corresponding mer in the genome. Genome evolution is modeled by graph evolution, which is composed of three possible processes (Fig. 2).

This simple model possesses enough expressive power to explain the genome structures in many species as well as how they can be connected in a phylogeny. Furthermore, essentially the same structures lead to higher level models for protein-protein interaction and regulatory networks, without obscuring the essence and parsimony of the evolution processes.

Genome grammar

In order to study the evolution processes of the kind just described in more detail, we have also developed an efficient backend for low-level simulation tasks in genomic evolution (Paxia et al., 2002). The current system can be easily enhanced to also model and derive statistics for protein–protein interaction and pathway structure. The simulated genomes and their statistical analysis also play a central role in designing and validating bioinformatics tools as well as in population studies in relation to linkage analysis. Our simulation environment uses advanced programming techniques such as lazy computation, efficient indexing, metadata manipulation, and deterministic randomization to achieve a spectacular improvement over similar naive implementations. This environment addresses the following concerns:

- **Large Data Sizes:** The sequences we deal with are huge, ranging from a few hundred mega-bases (a single chromosome) to a few giga-bases (multiple genomes). Furthermore, when performing popula-

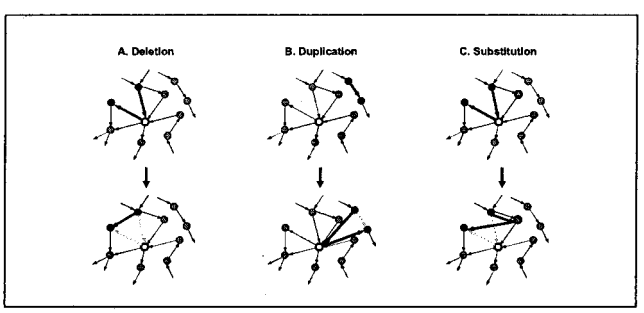


FIG. 2. The three processes during graph evolution: *deletion*, *duplication*, and *substitution*. In each process, the target node (clear circle) is chosen with preference for nodes with larger degrees: if the i-th node has degree k_i, the probability of it being chosen is proportional to $\frac{k_i}{\sum_{j=1}^{N} k_j}$. In deletion (**A**), a pair of edges of the target node (thick black arrows), one incoming and one outgoing, is randomly chosen and deleted, and a new edge (thick black arrow) is added between the ascendent and descendent nodes (black filled circles). In duplication (**B**), new edges are added between the target node and the ascendent/descendent nodes (black filled circles) of an edge (thick black arrow) randomly chosen to be deleted. In substitution (**C**), a randomly chosen pair of edges of the target node (thick black arrows), one incoming and one outgoing, is rewired to a randomly chosen substitute node (gray filled circle with a thick boundary). Note that all the processes during graph evolution preserve the equality of the in-degree and out-degree of each node.

tion simulation studies, we deal with multiple (potentially, a few billion) copies which change at every generation.

- **Large Data Movements:** The mutations we consider (e.g., transpositions and deletions) are expensive to simulate in a naive way. This is a serious problem, which forces researchers to abandon detailed simulation at a base-pair resolution, and opt instead to simulate the changes in the positions of just a few markers on the sequence.

Genome evolution, using the simulation engine described above, is studied through an "abstract machine" (the genome grammar; Fig. 3) that provides a quick means of specifying sequences with specific statistical properties. This simple programming language, with basic types consisting of sequences, transformations (i.e., mutations), numbers (encoding both deterministic and probabilistic events), and primitive operations, allows one to generate sequences with very specific probability distributions.

Language for interaction

The genome of an organism admits a simple mathematical description and a convenient representation in a computer (with augmented annotation). In contrast, there is no simple way of describing the cell, either statically or temporally (dynamically). We have been creating a formalism of combinatorial and numerical (entropic) structures on spaces of sequences which reflect, to some degree, the organization and functions of DNA and proteins. This formalism, called genoplex, distinguishes specific subsets of segments (e.g., exons within an ORF or CIS-regulatory elements) and assigns to each subset labels indicating the nature of the relationship among the elements within that subset.

Open questions

Several interesting questions remain to be further explored:

- **Models of Biochemical Processes Involved in Evolution:** The key genome duplication processes and their biology remain poorly understood. How can we accurately model such processes as (a) unequal crossing over, (b) DNA polymerase slippage, and (c) heterologous recombination? Do the evolution dynamics drive the statistical structures of the genomes to a limit distribution, or do they themselves acquire other derived dynamic properties? How do these processes interact with processes involved in selection? How are they modified in a disease? For instance, in a polyclonal tumor, oncogenes tend to acquire higher copy numbers, while tumor suppressor genes tend to be deleted (Lucito et al., 2000). How can we understand these processes in the disease state or in mutants?
- **Evolutionary Selection at Proteome Level:** Our genome models and genome grammar need to be further modified to model duplications of the protein domains. How can we use these properties to create a precise understanding of the protein-protein interaction data obtained from "two-hybrid" experiments?

```
n = | sequence |;
fragment_len = poisson(1000);
fst_pos = uniform(n - fragment_len +  1);
lst_pos = fst_pos + fragment_len -1;
repeat_num = poisson(5);
t = #(fst_pos, lst_pos, repeat_num);

sequence@t;

n = | sequence | ;
pos = uniform(n);
mut = point_mutation ( pos, {!} );

sequence@mut[0.3] ;
```

In this simple Genome Grammar script, the variable 'sequence' is a variable of the type *sequence* and is imported into the grammar from the outside. The first block of statements defines a transformation 't' which is either a *repeat* or a *delete*, depending on the value of a Poisson variable of mean 5. If the result of the Poisson sampling is 0, the transformation is a *delete*, otherwise it is a *repeat*. The fragment that gets deleted or repeated is chosen uniformly from the sequence and its length is a Poisson variable of mean 1000. The symbol '#' denotes a *repeat* or a *delete* with exactly the meaning above; the symbol '@' denotes the application of a transformation to a sequence. In the second block of statements, we apply a point mutation with a probability of 0.3.

FIG. 3. A genome grammar example.

- **Evolutionary Selection at Interactome Level:** How can we take our models to the next level to understand the relations among the transcription factors, regulatory elements, and genes? These models will be essential for interpreting gene-expression profiles obtained from microarray experiments.
- **Motifs and their Robustness:** What are the common motifs in regulatory pathways? Why is there a preponderance of a few motifs, and not all the possible ones? What are the common modules? How are they organized? Are topological descriptions of these motifs sufficient for understanding their functions? How can we study pathways, such as RAS/PKC/MAPK or the ones involved in cell cycle, to understand their multiple modes? Why are these motifs and their modes so robust?
- **Hierarchical Organization:** How can we design a language for describing, modifying, and transforming biological models that can take advantage of these intriguing but elegant structures? Can we use these properties to design mutants, double-mutants, "knock-outs," and other experimental systems to validate and understand our models?

ADVANCED TOOL ARCHITECTURES

In the domain of most immediate interest, namely, post-genomic biology, the conventional concept of a distributed set of tools, made available through a web-browser interface, fails to adequately respond to the challenges, complexity, and exponentially-growing demands. Thus, we aim at applying our expertise to create a tool architecture that has the following properties: (a) flexibly composable software-modules, (b) supports for multiple scripting languages, software libraries, and multiple computer architectures, (c) free-format databases, inter-operating between multiple formats, (d) capabilities for rapid prototyping to handle new experiments, and (e) easy integration of domain-knowledge.

Below, we describe some of the tools that address a few of the issues described above.

Valis

Valis is an environment for exploring problems in bioinformatics. The core components of the Valis project are the underlying database structure and the algorithmic development platform. The Valis database allows the user to analyze very large genetic sequences. Data structures that allow fast string matching to support analyses like mer-frequency analysis have been limited to sequences of approximately 100 million base pairs (Mbp). The Valis database allows for the creation of annotated sequences that are much larger than this limit. Many similar database systems rely on large binary object support in standard relational databases and on fixed formats of limited varieties to represent the necessary annotations for this analysis. It is our experience, however, that in exploring genetic data, annotations are often of arbitrary size and format. The underlying Valis database allows the use of sets, types, and reference counting in the annotation scheme while keeping both the storage requirements and the run-time cost of manipulating the annotated sequences low. The algorithmic development platform is another innovative area of the Valis system.

The Valis environment can be reviewed as a novel cross-language scripting platform. Algorithms implemented by a research group in one language can be utilized as building blocks in scripts by others. Valis currently supports scripting in many languages: Perl, Python, ECMAScript (JavaScript), Visual Basic, R (public domain S-plus), and Octave. Scripts can leverage one of the many public domain software libraries that we have incorporated into Valis. It also incorporates the Gnu Scientific Library (GSL) for additional support with standard numerical algorithms. Valis provides numerous visualization tools that allow the user to quickly display sequences, maps, microarray data, tables, graphs, and annotations. These widgets can be customized from the scripts.

Simpathica

Simpathica and the related XS-system allow the user to describe and interact with biochemical reactions. The system consists of three major components: the frontend, used for describing the model concisely; the core, used for creating internal representations of the model; and the backend, used for deriving the properties of the model of interest.

Using a simple graphical or textual representation like SBML (Systems Biology Markup Language) or MDL (Model Description Language), a mathematical description in terms of DAE (differential algebraic description), ODE, PDE, SDE, or Stochastic interactions is implemented in a Gillespie-like algorithm. Simpathica also creates an alternate description of the system and related data. When the system consists of many modules, they are represented individually, at many levels of abstraction, and with rules for compositionality.

The system supports a wide range of analysis tools: model checking with a propositional branching-time temporal logic, time-frequency analysis tools with wavelets, linear discriminant bases, Karhunen-Loeve analysis, clustering using wavelet-bases, information-theoretic approaches to measure the significance of the system components, and approaches based on statistical learning theory. All these systems, and a database for trajectory (with an architecture similar to NYUMAD) were developed under the larger BioComp/BioSPICE (supported by DARPA) effort and are available on the web (www.biospace.org).

NYUMAD

We have also developed a system to maintain and analyze biological abundance data (e.g., microarray expression levels or proteomic data) along with the associated experimental conditions and protocols. The prototype system is called the NYU MicroArray Database (NYUMAD) and is in the process of being expanded to deal with many other related experiments. It uses a relational database management system for data storage and has a flexible database schema that has been designed to store any type of abundance data along with general research data, such as experimental conditions and protocols. The database schema is defined using standard SQL (Structured Query Language) and is therefore portable to any SQL database platform. NYUMAD supports the MAGE-ML standard (www.mged.org/Workgroups/MAGE/mage-ml.html) for the exchange of gene expression data, defined by the Microarray Gene Expression Data Group (MGED), and is accessible via the web.

NYUMAD is a secure repository for both public and private data. Users can control the visibility of their data, so that initially the data might be private, but after the publication of the results the data can be made visible to the larger research community. Data analysis tools are supplemented with visualization tools. The goal is to not only provide a set of existing techniques but to continually incorporate increasingly sophisticated and mathematically robust methods in data analysis and to provide links and integration with our other tools, such as the Valis system.

Further development

Other areas that need to be further developed in our tool architecture include:

- **Extension of the Database:** As the systems described here evolve to encompass more complex and general models, they must acquire abilities to integrate other novel data sources: for example, proteomic data, ChIP (Chromatin Immuno-Precipitation) data, and mass-spectroscopic data. In response to these demands, the database schemas must be generalized and new XML-based formats introduced. Both Valis and Simpathica systems must be able to integrate new sources of experimental data as well as provide more cogent forms of visualization.
- **Integration:** We expect new bioinformatics systems to become continually available either in the public domain or from our collaborators. We need to develop a framework to integrate them into the larger system effortlessly.
- **Dissemination:** Finally, there are many issues involved in how other users and developers can access and improve the system collaboratively.

EXPERIMENTAL RESEARCH

A computational system developed for biological sciences and biotechnology will fail to be relevant if it does not recognize the observational and experimental nature of its sister fields since its very conception. “Experimentation without imagination and imagination without recourse to experimentation” would yield either myopic anecdotes or barren theories.

We also note that the current demands from industrial and governmental applications (e.g., homeland security, rational drug discovery, and population studies) cannot simply rely on *in silico* tools without concomitant trials on real biological substrates. Our experience with such trials on real biological systems will inspire confidence (from the other practitioners in the community) that the toolkit being made available by our team has indeed been field-tested.

Time-Course in vitro and in vivo data

The most interesting data points for our study will be time-course data, describing the genome, transcriptome, and proteome within a single cell, or an even more detailed picture, if the technology to perform compartmentalized single cell analysis becomes available. In the absence of such technology, we have to make do with mRNA collected from a small population of cells, where individual cells within the population may be moving through the cell cycle in an unsynchronized manner. Of course, without the proteomic data, transcriptomes tell less than half the story. The dynamics within the cell cannot be adequately described by the abundance data. We need to know all the important "time-constants," taking into account such effects as mRNA stability, protein degradation, multi-merization, and modification. Nonetheless, the experimental systems considered here can be easily extended with further technological improvements.

Microarray and gene-chip technologies provide an approach (to monitor the whole genome on a single chip) for studying interactions among thousands of genes simultaneously under many different experimental conditions. However, in many applications the key problem has been statistical noise in the data, attributable to non-specific hybridization, cross-hybridization, competition, diffusion of the target on the surface, or base-specific structural variations of the probe. A better understanding of this noise will come from the kinetic analysis of the base-pairing, denaturing, and diffusion processes; these processes are extensively studied by us, in order to deconvolve the transcriptional dynamics associated with cellular functions.

One can replace the hybridization-based gene expression analysis by any technology that allows rudimentary processes for "resequencing," that is, checking whether a particular DNA strand is composed of a given known sequence of bases. Sequenom has developed one such gene expression technology using their mass-spectroscopy based approach to "resequencing." (See Ding and Cantor, 2003.) The technology has many advantages: it creates less noisy data as it does not suffer from various kinds of error processes listed earlier for microarrays; it could potentially calibrate for mRNA stability; and finally, as more and more genes are identified, it could be easily adapted to account for the new transcriptomes.

It is quite obvious to us that our tools can be used in other applications too: (a) gene discovery, (b) disease diagnosis, (c) drug discovery (pharmacogenomics), or (d) toxicological research. Hence, they are developed with a flexible structure.

Experimental systems

Below, we describe two representative experimental systems that we are studying at length. Additional systems we are exploring include whole-genome ChIP assays, processes involved in apoptosis, circadian clocks, RAS/PKC/MAPK pathways, cell cycle models, the immune system, positional-information based models and their effects on patterning and segmentation, processes involved in developmental biology, processes involved in DNA replication, repair, and recombination, and many others.

- **Co-Cultivation Experiments:** Cells have a complex system of interacting signal transduction pathways by which external stimuli, such as hormones, growth factors, or cell–cell and cell–matrix contacts, direct the function of intracellular proteins and gene expression. Until recently it has been impossible to view as an ensemble the complex responses cells make to these various stimuli. However, with the advent of DNA microarrays and because of the development of inducible gene expression systems and short interfering RNA systems, it is now possible to perturb the expression of one specific gene at a time in the cell and measure the cell's response. For instance, in order to obtain a clearer picture of how the RAS oncogene (a gene that, when mutated, can contribute to cancerous growth) behaves, we can study how it induces various transcriptional changes (Wigler, 1990). RAS is a central component of many signal transduction pathways, and the induction of RAS causes a large number of changes within cells, the exact changes depending on the host cell in which it is induced. We can ask

whether the changes in gene expression induced by RAS are the indirect consequence of the production of extracellular factors, which then act upon the cell producing them and its neighbors. One way of differentiating direct from indirect changes is by their temporal order, with early responses being direct, and some later responses being indirect. To examine this hypothesis more clearly, we look at the response of mixed populations of cells. One cell component is engineered to induce RAS expression following stimulation with an artificial insect hormone. The other component is of the same cell type, but not RAS inducible. After induction, the two populations are separated, and their responses examined by the analysis of microarray data. This may be a generally applicable technique to discover the existence of factors involved in cell–cell communication.

- **Stem-Cell Experiments:** The dynamics of stem cell proliferation are poorly defined, yet stem cells are vital for growth, healing, and the general homeostasis of many animal and plant tissues. Thus, studies to address the basic biology of stem cells are paramount. The nematode *C. elegans* provides a convenient and well-characterized system to study stem cells (Hubbard and Greenstein, 2000). The adult *C. elegans* germ line is a polarized tissue with a distal stem-cell population at the end of the tube-like gonad and differentiated cells located further proximally. Cell division in the distal zone gives rise to germ cells that enter the meiotic pathway and eventually form gametes (eggs and sperm). Actively dividing cells (a mitotic process) are observed some 20–25 cell-diameters away from the distal tip, despite the membrane-bound nature of the ligand hypothesized to be responsible for signaling through a Notch-like pathway. A plausible model for such zone-division between mitotic and meiotic cells in the gonad has not been experimentally confirmed and has remained elusive.

  We are building and testing a rigorous computational model with analysis tools for *C. elegans* germ line stem-cell growth, based on real observations of cell division patterns in the distal mitotic zone. For each possible hypothesis, a probabilistic model capturing the "spatio-temporal, hybrid, and stochastic" nature of this problem is created and can be verified through the Simpathica system and *in vivo* experiments involving wild types and several mutants.

Further development

Other areas that need to be further developed include the following:

- **A Taxonomy of Models and Experiments:** Clearly, the example experiment systems that we, as a group, can adequately address represent only a small fraction of the many diverse systems that biologists are interested in. We must pay special attention to educating current and future generations of biologists in how to extend these examples. In particular, we must show by example how experiments and theories can coevolve in biology.
- **Novel Experimental Systems:** We must expect the biochemical techniques to continue to make significant improvements in the coming decades, perhaps enabling very fast whole-genome sequencing (Lai et al., 1999), high-throughput measurements of transcriptional and translational profiles, and even real-time monitoring of activities of single cells in a population. How can these new experimental systems be interfaced to our project seamlessly?
- **Generating Falsifiable Experiments:** As we gather many isolated models addressing different aspects of biology, a coherent picture will emerge, and yet, it will point to new mysteries and paradoxes. These can be resolved by proposing experiments that will either strengthen hypotheses consistent with the existing theories or falsify certain beliefs. Thus, a major component of our work must also focus on creating a knowledge base that can allow the researchers to access our current knowledge and quickly assess if there are inconsistencies.

ACKNOWLEDGMENTS

The work reported in this paper was supported by grants from DARPA's BioComp project ("Algorithmic Tools and Computational Frameworks for Cell Informatics") and AFRL contract (contract F30602-01-2-0556).

REFERENCES

ALBERTS, B., BRAY, D., LEWIS, J., et al. (1995). *Molecular Biology of the Cell*, 3rd ed. (Garland, New York).

ANTONIOTTI, M., POLICRITI, A., UGEL, N., et al. (2002). xS-systems: eXtended S-systems and algebraic differential automata for modeling cellular behavior. Presented at the International Conference on High-Performance Computing, Bangalore, India.

ANTONIOTTI, M., PARK, F., POLICRITI, A., et al. (2003). Foundations of a query and simulation system for the modeling of biochemical and biological processes. Presented at the Pacific Symposium on Biocomputing, Hawaii.

ANTONIOTTI, M., PARK, F., POLICRITI, A., et al. (2003). Simulating large biochemical and biological processes and reasoning about their behavior. Presented at the International Conference on Systems Biology, Sweden.

ANTONIOTTI, M., POLICRITI, A., UGEL, N., et al. (2003). Model building and model checking for biochemical processes. Cell Biochemistry and Biophysics (CBB) **38,** 271–286.

BAILEY, J.A., GU, Z., CLARK, R.A., et al. (2002). Recent segmental duplications in the human genome. Science **297,** 1003–1007.

BHALLA, U.S., and IYENGAR, R. (1999). Emergent properties of networks of biological signaling pathways. Science **283,** 381–387.

BULDYREV, S.V., GOLDBERGER, A.L., HAVLIN, S., et al. (1993). Fractal landscapes and molecular evolution: modeling the myosin heavy chain gene family. Biophysics **65,** 2673.

CANTOR, C., and SMITH, C. (1999). *Genomics: The Science and Technology Behind the Human Genome Project* (Wiley, New York.)

CORNISH-BOWDEN, A. (1999). *Fundamentals of Enzyme Kinetics* (Portland Press, London.)

DAMM, W., and HAREL, D. (2001). LSCs: breathing life into message sequence charts. Formal Methods in System Design **19,** 45–80.

DING, C., and CANTOR, C.R. (2003). A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight. MS. Proceedings of the National Academy of Sciences USA **100,** 3059–3064.

EFRONI, S., HAREL, D., and COHEN, I.R. (2003). Towards rigorous comprehension of biological complexity: modeling, execution, and visualization of thymic T cell maturation. Genome Research (special issue on systems biology).

GABBAY, D.M., HODKINSON, I., and REYNOLDS, M. (1994). *Tempral Logic: Mathematical Foundations and Computational Aspects, Volume 1* (Clarendon Press, Oxford.)

GERSTEIN, M., QIAN, J., and LUSCOMBE, N.M. (2001). Protein family and fold occurrence in genomes: power-law behavior and evolutionary model. Journal of Molecular Biology **313,** 673–681.

GUET, C.C., ELOWITZ, M.B., HSING, W.H., et al. (2002). Combinatorial synthesis of genetic networks. Science **296,** 1466–1470.

HAREL, D. (1987). Statecharts: a visual formalism for complex systems. Science of Computer Programming **8,** 231–274.

HAREL, D., and GERY, E. (1997). Executable object modeling with statecharts. Volume 30, Computer 31–42.

HAREL, D., EFRONI, S., and COHEN, I.R. (2003). Reactive animation. *Lecture Notes in Computer Science* (Springer-Verlag, New York).

HAREL, D., and MARELLY, R. (2003). *Come, Let's Play: Scenario-Based Programming Using LSCs and the Play-Engine* (Springer-Verlag, New York).

HUBBARD, E.J.A., and GREENSTEIN, D. (2000). The *Caenorhabditis elegans* gonad: a test tube for cell and developmental biology. Developmental Dynamics **218,** 2–22.

KAM, N., HAREL, D., and COHEN, I.R. (2001). Modeling biological reactivity: statecharts vs. Boolean Logic. Presented at the Second International Conference on Systems Biology (ICSB2001).

KAM, N., HAREL, D., KUGLER, H., et al. (2003). Formal modeling of *C. elegans* development: a scenario-based approach. Presented at Computational Methods in Systems Biology, First International Workshop.

LAI, Z., JING, J., ASTON, C., et al. (1999). A shotgun optical map of the entire *Plasmodium falciparum* genome. Nature Genetics **23,** 309–313.

LUCITO, R., WEST, J., REINER, A., et al. (2000). Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. Genome Research **10,** 1726–1736.

MISHRA, B. (2002). A symbolic approach to modelling cellular behavior. Presented at the International Conference on High-Performance Computing, Bangalore, India.

MISHRA, B. (2000). Computational differential algebra. *Geometrical Foundations of Robotics* (World-Scientific, Singapore, 111–145.

MISHRA, B., and CLARKE, E.M. (1985). Hierarchical verification of asynchronous circuits using temporal logic. Theoretical Computer Science **38,** 269–291.

OHNO, S. (1970). *Evolution by Gene Duplication* (Springer-Verlag, New York).

PAXIA, S., RUDRA, A., ZHOU, Y., et al. (2002). A random walk down the genomes: DNA evolution in VALIS. Computer **35,** 73–79.

PENG, C.-K., BULDYREV, S.V., GOLDBERGER, A.L., et al. (1992). Long-range correlations in nucleotide sequences. Nature **356,** 168–170.

SAVAGEAU, M.A. (1976). *Biochemical System Analysis: A Study of Function and Design in Molecular Biology* (Addison-Wesley, New York).

VOIT, E.O. (2000). *Computational Analysis of Biochemical Systems* (Cambridge University Press, Cambridge).

VULIC, M., LENSKI, R.E., and RADMAN, M. (1999). Mutation, recombination, and incipient speciation of bacteria in the laboratory. Proceedings of the National Academy of Sciences USA **96,** 7348–7351.

WIGLER, M. (1990). Oncoproteins. GAPs in understanding Ras. Nature **346,** 696–697.

WIGLER, M., and MISHRA, B. (2002). Wild by nature. Science **296,** 1407–1408.

Address reprint requests to:
Dr. Bud Mishra
715 Broadway, Rm. 1002
New York, NY 10003

E-mail: mishra@nyu.edu