Title
SEMANTIC ANALYSIS AND SYNTHESIS OF COMPLEX BIOLOGICAL SYSTEMS

Permalink
https://escholarship.org/uc/item/6kp625wn

Journal
International Journal of Software Engineering and Knowledge Engineering, 15(03)

ISSN
0218-1940

Authors
XIE, FEI
SHEU, PHILLIP C-Y
LANDER, ARTHUR
et al.

Publication Date
2005-06-01

DOI
10.1142/s0218194005002415

Peer reviewed

International Journal of Software Engineering
and Knowledge Engineering
Vol. 15, No. 3 (2005) 547–569

World Scientific
www.worldscientific.com

SEMANTIC ANALYSIS AND SYNTHESIS OF COMPLEX BIOLOGICAL SYSTEMS*

FEI XIE†, PHILLIP C.-Y. SHEU‡

Department of EECS and Biomedical Engineering,
University of California, Irvine, USA
†*fxie@uci.edu*
‡*psheu@uci.edu*

ARTHUR LANDER

Department of Developmental Cell Biology,
University of California, Irvine, USA
adlander@uci.edu

VITTORIO CRISTINI

Department of Biomedical Engineering,
University of California, Irvine, USA
cristini@math.uci.edu

In general biologists are not accustomed to formulating biological problems in the precise mathematical terms that are required to solve the problems analytically or numerically. Although many computational tools for systems biology have been developed recently, our observations indicate that many of these tools are powerful only in the hands of those who know a lot about how to use them. For most biologists, the tools have a protracted learning curve and unfriendly user interface that often diminish their likelihood of being used.

Our long-term goal is to build a knowledge system that allows biologists to synthesize complex biological systems via natural language interactions, and the system is able to generate the corresponding mathematical descriptions so that the often cumbersome communication process between biologists and mathematicians/engineers in formulating complex biological problems in mathematic terms can be performed more easily.

To focus, the first goal in this research is to build a knowledge system prototype that focuses on transport related biological problems that occur from the cellular to tissue level. We address specifically two inter-related problems: (1) Provision of an intelligent system that is capable of automatically synthesizing smaller components into more complex systems; Provision of a user-friendly and natural language interface.

Keywords: Systems biology; complex biological systems; model synthesis; natural language processing; intelligent system; knowledge base.

*This research is supported in part by the National Institute of Health and the National Science Foundation.

1. Introduction

The science of biology has developed to a point that one must study complex biological systems consisting of multiple fields of studies to supplement the traditional reductionist approach. However, very often three problems emerge:

- In general biologists are not accustomed to formulating biological problems in the precise mathematical terms that are required to solve the problems analytically or numerically. Although collaborations between biologists and mathematicians/engineers can overcome this problem, differences in language, knowledge base, and "culture" between the two groups often make it difficult to initiate such collaborations.
- Although many computational tools for systems biology have been developed recently, our observations indicate that many of these tools are powerful only in the hands of those who know a lot about how to use them. For most biologists, the tools have a protracted learning curve and unfriendly user interface that often diminish their likelihood of being used.
- The current information technology has not been able to enable efficient knowledge sharing. Researchers not only find it difficult to locate and utilize existing work on a given problem but also find it difficult to make their work available for others to use.

This research addresses the fundamental information technologies required to address, or at least alleviate, the above problems at different levels of abstraction. Specifically, a computer system is being developed so that biologists are able to describe a complex biological system in ways that they are comfortable with. This system in the meantime is able to generate the corresponding mathematical descriptions, so that the often cumbersome communication process between biologists and mathematicians/engineers in formulating complex biological problems in mathematic terms can be performed more easily.

Our long-term goal is to build a knowledge system that can solve biological problems across many levels. However, we realize that it is a rather ambitious goal and it is not likely to be accomplished in a short period of time. Therefore, our current goal is to build a knowledge system that focuses on transport related biological problems that occur from the cellular to tissue level. Our research includes especially the following objectives:

1. Provision of an intelligent system that is capable of automatically synthesizing smaller components into a more complex system. In our preliminary studies we have focused on a set of common biological processes, such as diffusion and molecular interactions, as the "building blocks" of more complex systems. Although most biologists would know what kind of basic processes are involved in their systems, they often do not know how these processes interact with each other to create complex behaviors, while in most cases these interactions are the results of known physical laws. It is important not only to identify the physical laws that "glue" together a complex system, but also to devise a knowledge representation scheme that enables computer programs to perform the synthesis automatically and efficiently.

2. Provision of a natural language interface. It is important to have an intuitive and interactive interface so that biologists who have little training can learn smoothly and rapidly how to construct the "building blocks" mentioned in objective #1 to synthesize complex biological systems. The most effective interface will be a natural language interface. Building a natural language interface is in fact strongly related to the previous objective as natural language processing is very dependent on its knowledge domain.

This paper is organized as follows. Section 2 describes the background and related work, especially the background on computational systems biology, tumor growth modeling and simulation, and the related information technologies. Section 3 introduces our overall approach. Section 4 discusses the formulation of the morphogen gradient problem using our approach, and Sec. 5 discusses the formulation of the tumor growth problem. Section 6 describes the structure of our knowledge base system and our current implementation. Section 7 describes our approach to a natural language interface for our system. Section 8 concludes the paper.

2. Background and Related Work

Modern biology often needs to deal with complex biological systems. They are complex in the sense that many biological processes occur at the same time. In many cases, the traditional reductionism approaches have shortfalls, which mostly arise from "information overload" and "over-simplification" [1]. Therefore, it has become increasingly necessary to study complex biological systems whose behavior cannot be explained simply in terms of individual parts.

In order to study complex biological systems, biologists often need to collaborate with mathematicians, because even if the individual processes in a complex system are easy to understand and model, the behavior of the whole system may involve complex interactions among its parts. However, one of the largest obstacles hindering the effective collaboration between biologists, physical scientists and mathematicians is the very difference among the languages used by these various disciplines. Concepts like "boundary value problem", "isotropy", "flux" and "cross-correlation" are not really foreign to biologists, but the ways they are treated by a biologist are very different from the ways that, say, a mathematician would treat them. Hence, each needs to spend many hours with the other before they can effectively communicate. Although much can be done to facilitate such one-to-one interactions, the communication is likely to continue to be a difficult problem.

2.1. *Computational systems biology*

The core of systems biology is molecular biology. One of the subjects molecular biologists study at the cellular level is "signaling", which refers to how cells communicate with each other. A cell sends signals, usually in the form of chemical substances, to other cells. A cell receives and responds to a variety of signals from outside (e.g., activate genes, change shape) typically by inducing a chain of chemical reactions inside itself, a so-called “intracellular signaling pathway”, and it is the interactions of such pathways that allow the integration of multiple signaling inputs. In addition to systems of signaling pathways, cells also possess metabolic pathways which can be quite complex. These consist of series of enzymatic steps that transform molecules of one type into another, and underlie both the breaking down of food into energy, and the manufacturing of all cellular components.

Computational schemes can be applied to model cell signaling, metabolic and gene expression networks. Moreover, quite a few software packages are available to help biologists analyze such models. A partial list includes Cellerator [2], Virtual Cell [3], E-Cell [4], Gepasi [5] and Jarnac [6]. Most of them are deterministic. Typically, a user first defines pathways in certain ways. In Virtual Cell, for example, the user defines chemical reactions by drawing; in Cellerator, as another example, the user defines reactions using a Mathematical Palette. These programs translate the reactions specified by the user into a set of differential equations and solve them using Matlab, Mathematica, LSODE, and/or custom solvers. On the other hand, some software packages employ stochastic approaches to simulate intracellular pathways. These include Gibson’s Next Reaction Algorithm [7], StocSim [8], MCell [9], and the “Stochceller” module for Cellerator. While all of these software packages are powerful, they are mostly suitable for use by biologists who already have a model in hand, and who have enough mathematical sophistication to know exactly what kinds of ancillary information (e.g., boundary conditions) must be provided to make a solution or simulation possible.

Cell signaling, metabolic and gene expression pathways and networks can often be dealt with as time-dependent but space-independent processes to simplify the analysis. The problems become much more complicated when time-and-space-varying processes, such as molecular diffusion, cell migration, growth of cells within a tissue, etc. must be taken into account. The problems also become more complicated when processes of different time- and length-scales must be considered together. Formulating models of such processes is non-trivial. Moreover, the above software packages can provide a support for solving only a subset of such models. Nonetheless, significant progress has been made by groups of biologists and modelers working together. For example, recent research has discovered how embryonic patterns are established by gradients of diffusible molecules and their inhibitors [10–12], how cell migration is controlled (e.g. [13]), and how tumors grow and change shape (see references later). The latter area provides a particularly good example of a complex biological system in which processes occurring at many levels — from the molecule up to the organ level — are interdependent. Some of the current issues in modeling tumor growth are discussed below.

Although the community of systems biology has created powerful tools, there are several shortcomings based on many biologists’ experiences. First, most of these software packages focus on a particular level (e.g., molecular interaction level or tissue level) or a particular subject domain. Much needs to be explored to see the ways to integrate different biological models into a larger, more complex system. More importantly, in order to ensure that a program captures all the relevant behaviors to generate valid simulations, the existing software packages demand a rather high level of biophysical and mathematical sophistication on the biologist interacting with the program. As a result, many biologists do not know how to begin using these software packages and feel frustrated. Lastly, the existing systems biology tools do not provide an efficient knowledge querying and sharing infrastructure. Work done by the Systems Biology Work Bench Group has taken the first step to define the XML-based Systems Biology Markup Language (SBML), which is aimed as a common representation to store biochemical models [14]. However, SBML in itself does not guarantee that biologists can efficiently search for biological models defined by others or can efficiently compare and integrate different biological models. A knowledge base system on top of SBML is needed to accomplish this task.

2.2. *Tumor growth — Modeling and simulation*

The modeling and simulation of the tumor growth problem is one focus of our preliminary studies. In fact, the tumor progression problem provides an excellent example of both the challenges and the benefits of integrating biological models across levels. For this reason it is an ideal problem around which we can build a generally applicable technology to foster the formulation, integration and solution/simulation of complex biological models.

The biophysical, biochemical and mechanical processes characterizing tumor progression — which include both the growth of tumor cells and new blood vessels (vascularization or angiogenesis) to nourish the tumor — are very complex. The initial growth of tumors (presumably following genetic mutations of a cell or cells) is avascular (without blood supply) and relies on the diffusion transport of nutrients and oxygen through the extracellular matrix into the tumor. This stage presumably also relies on the production and transport of growth factors (e.g., insulin, PDGF, EGF, etc.) either locally or from distant sources. This stage of growth usually leads to a dormant state of millimeter size, due to limitations on the supply of nutrients to the bulk of the tumor cells.

To support further growth, angiogenesis must be triggered by the release of tumor angiogenic factors (e.g., vascular endothelial-cell growth factor, VEGF) that can lead, through a number of mechanisms, to tumor vascularization. Essential cell biological processes at this stage include: transport of angiogenic factors through the extracellular matrix, endothelial cell responses such as basement membrane degradation, migration and proliferation, and anastomosis (fusing of blood vessels) with a resulting blood micro-circulation.

At the same time, growth of the tumor continues to be influenced by a number of other processes in and around it. These include: appearance of new mutations that alter the growth properties of individual cells; the production, transport and competition for autocrine growth factors; the production of factors by peritumoral stroma; alterations in cell adhesion and motility that lead to cell migration; infiltration by immune and inflammatory cells which release a variety of factors; and remodeling of the tissue architecture around the tumor by molecules released by tumor cells (e.g., matrix metalloproteinases). Moreover, because angiogenesis (i.e., neo-vascularization) provides the tumor cells with an unlimited nutrient supply, it is uniquely poised to promote invasive growth, characterized by complex tumor morphology with fingering and separation of metastases. This complex morphology couples in nontrivial ways to all of the processes listed above. For example, it increases the surface-to-volume ratio of the tumor and thus is expected to enhance the spatial flux of growth factors, angiogenic factors and further vascularization, and influences infiltration by inflammatory cells and their secreted products.

In the mathematical biology community, research on the origins and development of tumors has received ever-increasing attention (e.g., see the recent review papers [15–17] and the numerous references therein). Mathematical models of tumor growth utilize systems of continuum reaction-diffusion equations as well as discrete random-walk theory are now sophisticated enough to describe tumor progression (e.g., [18–22]) including the effects of genetic mutations (e.g., [23–28]), autocrine signaling (e.g., [29]), and angiogenesis (e.g., [30–33]) including the underlying biochemistry [31, 32] (a complete list of references for all aspects of Tumor Modeling can be found in [36]). Very recently, multidimensional and multi-scale computer simulations of tumor progression have been presented [39, 44, 45].

2.3. *Information technology*

In principle, there should be a way for biologists to describe a complex system in biological terms and to interact with a software system that, through appropriate prompts and queries, can make a first pass effort to distill a system into one or more biological models. The mathematical equations underlying these biological processes can be generated automatically if the physical laws and mathematics governing these biological processes are understood thoroughly by the computer. Currently there is already a wealth of knowledge regarding the physics and mathematics of many biological processes, although more is being discovered everyday. Therefore, besides identifying the kinds of knowledge needed, how to store and use such knowledge is an IT challenge.

Unfortunately, most of the existing "biological databases" have been archive systems. Examples can be found at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) and the European Bioinformatics Institute (http://www.ebi.ac.uk/services/index.html). An extensive list of other molecular biology databases can be found at http://nar.oupjournals.org/cgi/content/full/27/

1/1/DC1/43. In addition, biologists have begun to develop a variety of anatomical and morphological databases, including the Digital Anatomist Project (University of Washington) [34] and the National Library of Medicine's Visible Human Project [35]. A fundamental difference between a biological database addressed in this research and a traditional biological database is that, in this research, the targets of a query are models and active objects that interact with each other. To our knowledge biological databases of this type do not yet exist.

As a brief overview of the database technology, most of the biological applications today have employed a relational database for storage. A relational database organizes data into tables that include fields. Two tables that include a same field are related to each other. Compared to the "flat file" approach that stores all data into a single file, the relational approach using tables is more flexible. Most relational database systems conform to the Structured Query Language (SQL) standard.

Object-oriented databases organize data into objects. An object can have attributes, which can also be objects. The recursive nature of an object permits ease of manipulation. Objects can inherit characteristics from other objects, making it easier to create new objects based on existing ones. An object can be associated with a set of procedures (methods) to manipulate its data. Attempts have been made to combine a relational database and an object-oriented database to create an object relational database. The SQL-99 standard extends the conventional SQL query language to allow tables and fields to be manipulated as objects. In addition, it allows any Boolean function to be used as a qualification for data retrieval. With such, the scope of SQL-99 becomes wider than the scope of conventional SQL.

Unfortunately, SQL-99 is not sufficient to express many queries needed in composing or analyzing a complex biological system. In addition, a user who lacks programming skills typically cannot compose conventional SQL or SQL-99 queries, and must rely on programs written by programmers to search and display data. Therefore the user's options are frequently very limited.

3. Overall Approach

Figure 1 summarizes our approach. Each "biological object" in our knowledge system has associated with itself a set of methods including predicates and actions that can be applied to the object. On top of the objects are rules that further define the

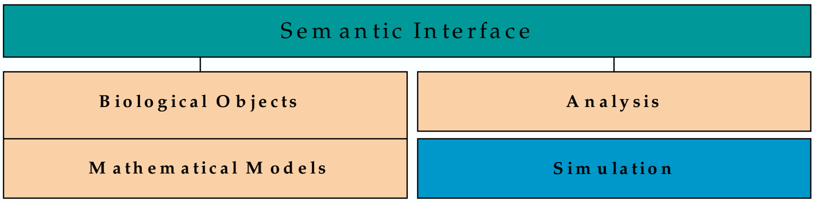


Fig. 1. Architecture.

behaviors of the objects. We consider a "model" to consist of two parts: the biological part (biological model) and the mathematical part (mathematical model), each is described in terms of a set of biological objects (variables) and rules. On top of objects is a natural language interface that allows the user to compose (program) a complex biological (model) object in terms of existing semantic building blocks. The biological models are translated into mathematical models for the purpose of analysis and simulation. Finally the results of analyses and simulations, which are instances of biological objects, together with the models themselves are stored in an object relational database so that biologists can query against (a) the models, and (b) the results from previous simulations and analyses. In the remaining sections, our system is referred to as a "BioFactory".

4. Morphogen Gradient Formulation

We have studied the modeling of morphogen gradients and developmental systems in which large scale patterns arise as a result of diffusion gradients of molecules (morphogens) which instruct cells to do different things at different levels of morphogen-receptor occupancy. Collaborating with applied mathematicians, we were able to explore numerical solutions for models involving morphogen synthesis, diffusion, reversible binding, reversible internalization and degradation [10].

The Biological Problem: Morphogenesis is the process by which embryonic cells develop into organisms with a set of 3-dimensional patterns of structures. The special class of molecule "morphogens" serves as the chemcial signal that differentiates cells from a cluster of equivalent cells into structured tissues. A so-called morphogen field consists of a morphogen source, and a group of cells. Typically the morphogen source produces morphogens, which are then distributed over the group of cells in the field. Each cell in the field has receptors that are capable of capturing the morphogens. How the morphogens are transported has been controversial, but when one takes into account processes such as degradation and internalization, appropriate mathematical analysis and simulation show that the diffusion-reaction of morphogens alone can produce morphogen gradients with the characteristics of those observed in experiments [10].

Biological Processes Involved: The basic biological processes involved in morphogen gradient formation are diffusion and reaction. The following describes the equations corresponding to each individual process.

1. First of all, the morphogens diffuse. Let the concentration of morphogens be $[M]$, the corresponding equation is (1). Here "$f([M])$" is a general function for the source term and the sink term of the morphogens. Without further conditions, $f([M]) = 0$. Of course, to completely describe diffusion mathematically, we also need to define the boundary conditions, initial conditions, and the external sources and sinks of the diffusing morphogens. Since these are very problem

dependent, we omit the details here.

2. Reversible bindings occur between receptors and morphogens. The chemical equation describing this reversible binding process is $R + M \leftrightarrow RM$. Let k_1 be "on" rate and k_2 be the "off" rate. Let the concentration of the receptors in question be $[R]$, and the concentration of the morphogen-receptor complex be $[RM]$, then the equations that describe reversible bindings are (2), (3), and (4).
3. Finally, the complex RM degrades: $RM \to \emptyset$. Let k_{deg} be the rate of degradation. The equation that describes this process is (5).
4. Since morphogens are bound to the receptors and released from the receptor-morphogen compounds, we can combine Eqs. (1) and (2) to get (6).

$$\partial[M]/\partial t = D\nabla^2[M] + f([M]) \tag{1}$$
$$\partial[R]/\partial t = -[R] * [M] * k_1 + [RM] * k_2 \tag{2}$$
$$\partial[M]/\partial t = -[R] * [M] * k_1 + [RM] * k_2 \tag{3}$$
$$\partial[RM]/\partial t = [R] * [M] * k_1 - [RM] * k_2 \tag{4}$$
$$\partial[RM]/\partial t = -[RM] * k_{\text{deg}} \tag{5}$$
$$\partial[M]/\partial t = D\nabla^2[M] - [R] * [M] * k_1 + [RM] * k_2 \tag{6}$$

Although in the above equations, only one type of morphogen binds to one type of receptor, in real biological systems there are many types of morphogens and receptors involved, and the receptors may have their own dynamics such as receptor production or receptor internalization.

Observations about the Basic Biological Processes: Having looked at this very simple problem, we can already answer some of the questions posed earlier in this section. The type of "knowledge" we are looking for here actually consists of these three types of biological processes (diffusion, reversible binding and degradation), and they are the "building blocks" of our morphogen gradient problem. Once our knowledge system stores the relevant meta-information associated with the three biological processes, a biologist who is interested in obtaining the mathematical equations that describe a morphogen gradient problem based on diffusion-reaction can do so by only providing the biological description of the diffusion, binding, and degradation processes to our system. He or she does not need to look through the literature or consult a mathematician to obtain the mathematical equations for the individual processes, because the mathematical equations corresponding to these three biological processes are already known to the system.

Observation about Synthesis: Diffusion, binding and degradation are not only observed in morphogen gradients. They occur in many other biological problems as well. Since we treat these three processes as individual building blocks, they need not always be present in the same biological model. For instance, a biologist who is only observing the diffusion of a type of molecule in a tissue may not need to be aware of any reversible binding process. He or she therefore can build a biological model consisting of the diffusion process only. However, it is possible that the biologist later discovers there is a reversible binding between the molecule of interest and receptors. It is natural for him/her to wonder what the effect of the binding on the diffusion process is. How can we "synthesize" these two things? In this case, because the left-hand sides of both equations (1) and (2) are the rates of change (of the morphogen concentration), we only need to add the right-hand sides of (1) and (2) to obtain Eq. (6). This process of superposition is the simplest example of how to "synthesize" two different models.

5. Tumor Growth

While the building blocks of simple models, like the morphogen gradient problem, may be linked together by superposition, we may not be able to synthesize more complex biological systems (e.g., in which diffusion and reaction couple with each other) as easily. Our preliminary study shows, however, that it is still feasible.

Biological Processes Involved: Let us summarize the basic processes in this complex tumor growth model.

1. *Diffusion of the nutrients.* Let the concentration of the nutrients be σ, then we have Eq. (7). However, since the diffusion process occurs in a much slower time scale than the other processes of the system, we can further modify the diffusion equation to (8). This time we need to list the boundary conditions as well. Since there are three domains of interest, the three boundary conditions are specified in (9), (10), and (11).
2. *Absorption of the nutrients by the cells.* Let λ be the rate of absorption, we have Eq. (12).
3. *Supply of nutrients from the outer source.* Let λ_B be the rate of blood-to-tissue transfer and σ_B be the nutrient concentration in the outer source (blood stream in this case), we have Eq. (13).
4. *Growth.* Here we choose a linear growth model. Therefore, the tumor tissue is assumed to grow in proportion to the nutrient concentration. Let μ be the velocity field, λ_M be the rate of cell mitosis, λ_A be the rate of cell apoptosis, λ_N be the rate of volume loss in necrotic core, the linear growth model has Eqs. (14), (15), and (16).
5. Since processes 2 and 3 in fact define the source and sink of the nutrient concentration, we need to combine Eqs. (12) and (13) with (8) to get Eq. (15).

$$\partial\sigma/\partial t = D\nabla^2\sigma + f(\sigma) \tag{7}$$

$$0 = D\nabla^2\sigma + f(\sigma) \tag{8}$$

$$\sigma = \sigma_\omega \tag{9}$$

Nutrition concentration is a constant on the outer boundary of the healthy cells,

$$[\sigma] = 0 \tag{10}$$

Nutrition concentration has zero jump on the boundary of health and tumor domains,

$$\sigma = \sigma_N \tag{11}$$

Nutrition concentration is a constant on the outer boundary of tumor and necrotic domains,

$$\partial\sigma/\partial t = -\lambda_\sigma \tag{12}$$

$$\partial\sigma/\partial t = +\lambda_B(\sigma_B - \sigma) \tag{13}$$

$$\nabla \cdot \mu = \lambda_M \sigma/\sigma_\omega - \lambda_A \text{ (in the tumor cell domain)} \tag{14}$$

$$\nabla \cdot \mu = 0 \text{ (in the healthy cell domain, where tissues does not grow)} \tag{15}$$

$$\nabla \cdot \mu = -\lambda_N \text{ (in the necrotic core domain)} \tag{16}$$

$$0 = D\nabla^2\sigma + \lambda\sigma + \lambda_B(\sigma_B - \sigma) \tag{17}$$

Observations about the Basic Biological Processes: We were able to dissect the complex process involved in the biological model of concern into individual biological processes. We have encountered one of the processes before, i.e., diffusion. There are also new processes, such as growth. In fact the linear growth Eqs. (7)–(9) apply to not only the tumor cell growth problem, but also many other growth problems. The way to "synthesize" these building blocks is straightforward. For instance, in the linear growth equation, the growth of cells is always proportional to some driving force; in this case it is the nutrient.

The Biological Problem of the Angiogenesis Model: Angiogenesis is another complex biological process that is important to tumor growth. As the tumor cells grow, it makes nutrients harder to get to the inside of the tumor, resulting in the death of many tumor cells at the core of the tumor and the formation of the necrotic core. However, the dead tumor cells produce Tumor Angiogenesis Factor (TAF), which attract endothelial cells. Endothelial cells are the cells that form blood vessels. As a result, blood vessels are developed inside the tumor, which supply nutrients to help the tumor grow.

Biological Processes Involved

1. *Diffusion, absorption, and degradation.* Let c be the concentration of the TAF, n be the concentration of the endothelial cells, λ_C be the rate of degradation of TAF, and λ_{CN} be the rate of absorption of TAF by the endothelial cells, the equations are (18) and (19).
2. *Advection-Diffusion.* Let χ_c be the chemotaxis coefficient for TAF, we have the Eq. (20).

$$0 = D\nabla^2 c - \lambda_C c - \lambda_{\mathrm{CN}} cn \tag{18}$$

$$c = c_0 \text{ (on the necrotic core boundary)} \tag{19}$$

$$\partial n/\partial t + \nabla \cdot ((\chi_c \nabla c + \mu)n) = D_n \nabla^2 n \tag{20}$$

Synthesis of the Tumor Growth Model and the Angiogenesis Model: In the previous sections, we have identified the basic biological processes for two biological problems: the tumor growth model and the angiogenesis model. Obviously, these two processes affect each other. On one hand, the formation of blood vessels from endothelial cells brings in the nutrients needed for tumor growth. On the other hand, the TAF released by the necrotic cells attracts endothelial cells into the tumor tissue. Now, suppose that two biologists have built a growth model and an angiogenesis model respectively, are we able to synthesize them based on an automatic procedure?

In the diffusion-advection Eq. (20), we have μ, which represents the field velocity of the tumor in growth. This naturally corresponds to μ in the growth equations (14)–(16) of the growth model.

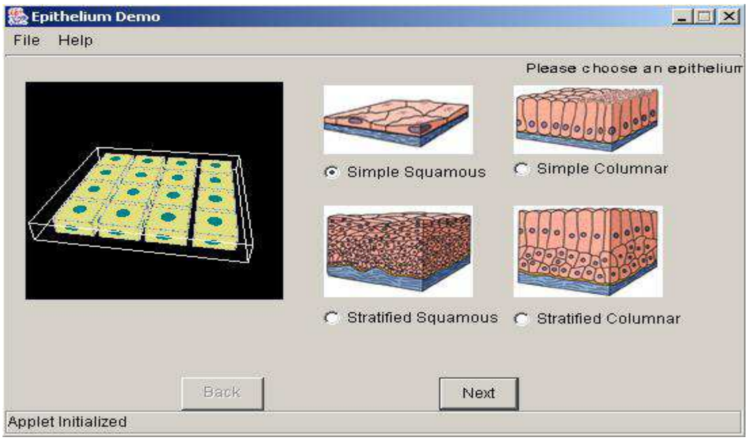
6. Knowledge System

Once the knowledge needed is identified, it needs to be stored and retrieved efficiently. The nature of our software system posts several unique requirements.

- The system must be user-friendly to biologists; otherwise it does not serve to alleviate the basic problem we are trying to solve. A combination of natural language and Graphical User Interface is probably needed to fulfill this requirement.
- The knowledge base system should be an open system to accommodate new knowledge.
- The knowledge representation scheme should enable efficient ways for model synthesis and model simplification as well as ways to query missing pieces of data.

There are at least two possible points of interest. In the diffusion-advection Eq. (20), we have μ, which represents the field velocity of the tumor in growth. This naturally corresponds to μ in the growth equations (14)–(16) of the growth model. Another connection point is that the change of the necrotic core boundary actually affects the boundary condition of the concentration of TAF in Eq. (20), and we can describe this influence mathematically. Through these two connections, we may then combine the two mathematic descriptions together.

Preliminary Implementation: We have implemented a demo application that allows users to define diffusion, reversible binding, degradation and catalysis processes. This demo application is a GUI-based system that uses *SemanticObjects* as the underlying engine. Through semantic building blocks (e.g., Verb: "Compose a model", Adjective: "that includes diffusion", "that includes reaction", etc.) in *SemanticObjects*, the GUI interface for defining the four types of biological processes can be brought up.

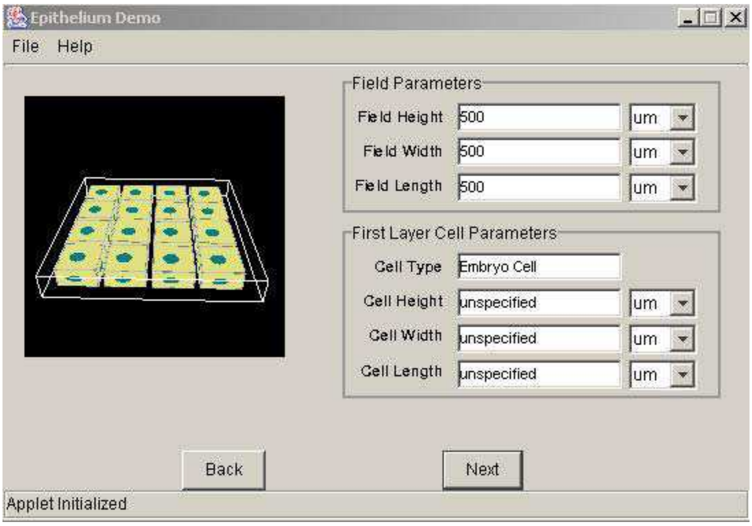


Please choose an epitheliur

● Simple Squamous ○ Simple Columnar

○ Stratified Squamous ○ Stratified Columnar

Fig. 2. Choose epithelium type.



Field Parameters

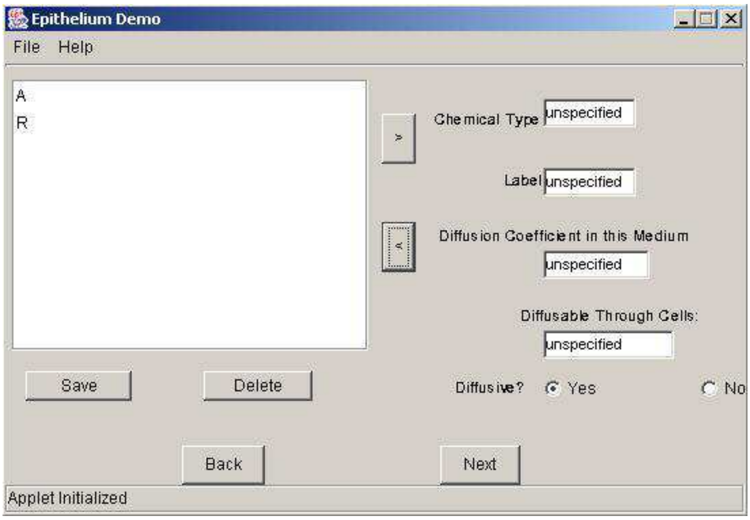
Field Height	500	um
Field Width	500	um
Field Length	500	um

First Layer Cell Parameters

Cell Type	Embryo Cell	
Cell Height	unspecified	um
Cell Width	unspecified	um
Cell Length	unspecified	um

Fig. 3. Define parameters.

For the diffusion process, the user first chooses the type of medium via which the diffusion occurs. Users can choose from four types of epitheliums (simple squamous, stratified squamous, simple columnar, and stratified columnar). A 3-D visual model is generated (the left most canvas as shown in Fig. 2). Then the user specifies different types of parameters associated with the geometry for the epithelium chosen. For instance, the user can specify the height, width and length of the whole epithelium or can leave them unspecified in Fig. 3. If the user chooses not to specify a parameter because she/he is not certain what is to be entered, the value of the unspecified parameter will be generated automatically.



A
R

Chemical Type: unspecified

Label: unspecified

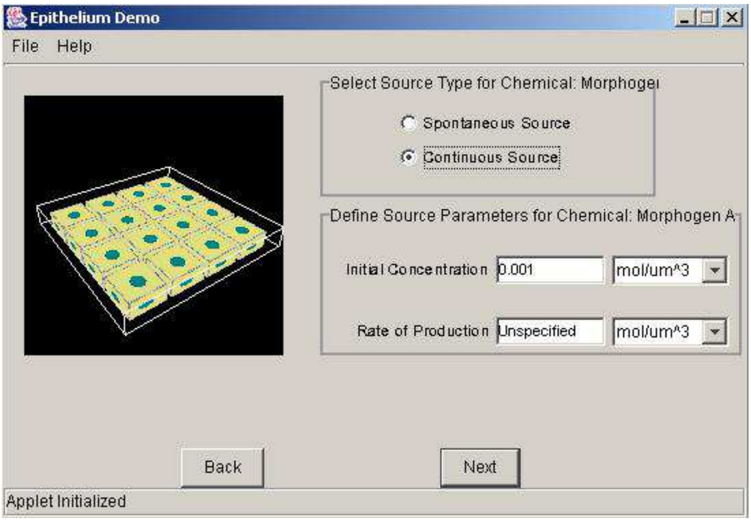
Diffusion Coefficient in this Medium: unspecified

Diffusable Through Cells: unspecified

Diffusive? ● Yes ○ No

Applet Initialized

Fig. 4. Specifying molecules.



Select Source Type for Chemical: Morphoge

○ Spontaneous Source
● Continuous Source

Define Source Parameters for Chemical: Morphogen A

Initial Concentration: 0.001 mol/um^3

Rate of Production: Unspecified mol/um^3

Applet Initialized

Fig. 5. Specifying boundary conditions.

The user can also define the type of molecules that diffuse. He/she can define the name of the molecule, the variables used in the equations, the diffusion coefficient, etc. (see Fig. 4). Then the boundary conditions are defined (as shown in Fig. 5). Currently, three types of boundary conditions can be selected for the apical surface and the basolateral surface of the epithelium. Therefore, she/he may choose the (1) "Reflective" condition if the diffusing molecules are reflected back at the

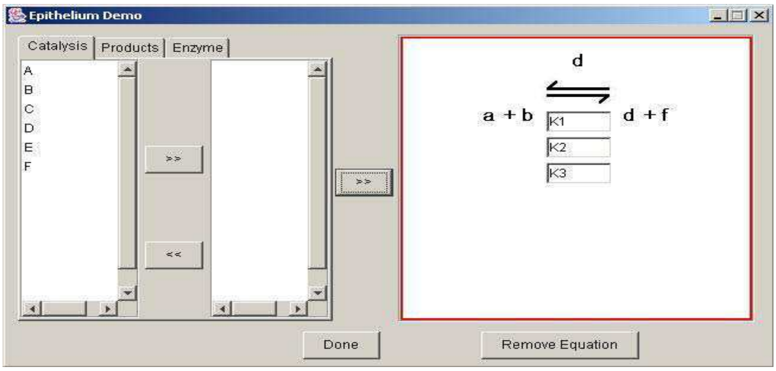


Fig. 6. Catalysis.

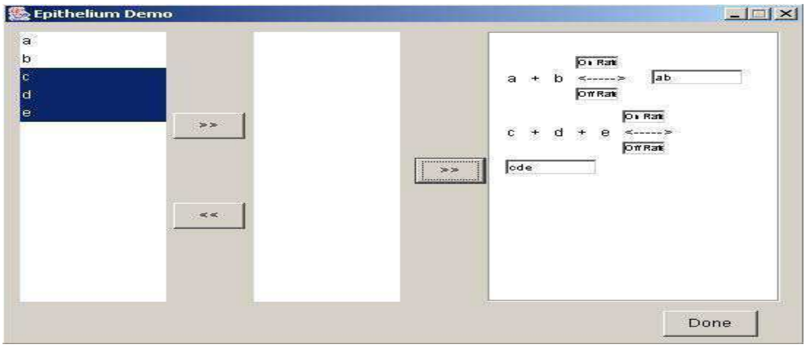


Fig. 7. Querying existing reaction models.

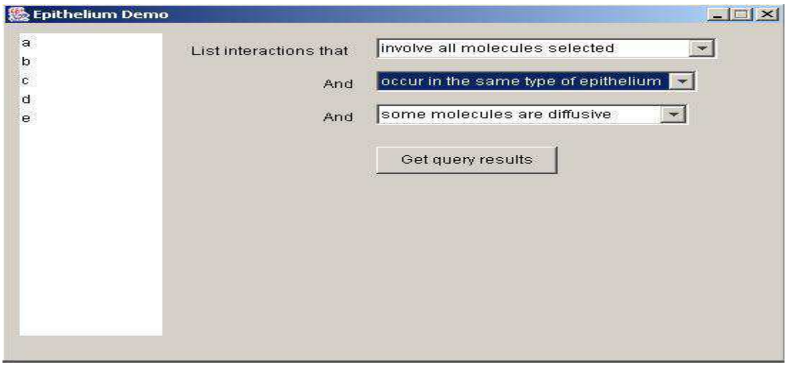
boundary, (2) "Absorptive" condition if the molecules are absorbed at the boundary, or (3) "Permissive" condition if the molecules can pass the boundary freely. In future implementations, more flexible ways will be provided for defining various types of geometry using a set of geometry templates.

For the three types of molecular interactions, i.e., reversible binding, degradation and catalysis, the user selects the types of the molecules involved in these reactions, and enters the chemical equations with parameters. For catalysis, he/she needs to specify the reactants, products, and enzymes. The demo system will generate the corresponding chemical equations, as shown in Fig. 6. Users also need to specify the rate constants.

For reversible bindings, the molecules involved have to be selected and the on/off rate constants have to be specified, as shown in Fig. 7.

Alternatively, the user can search the database for any existing reaction models that involve any of the molecules of interest, as shown in Fig. 8.

Once the simple diffusion and reaction models are built, the demo can automatically generate the corresponding diffusion-reaction equations (Fig. 9).

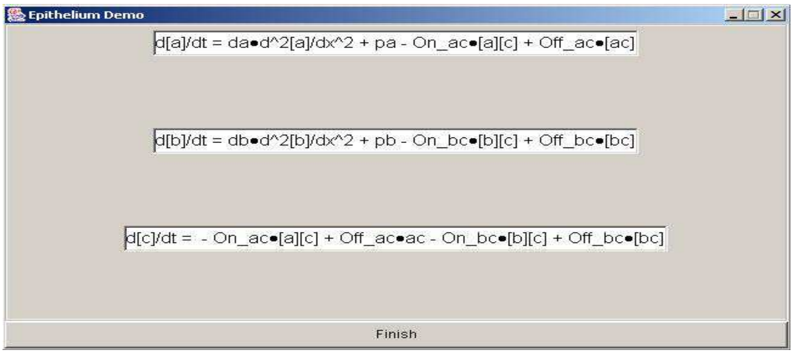


List interactions that: involve all molecules selected

And: occur in the same type of epithelium

And: some molecules are diffusive

Fig. 8. Reversible binding.

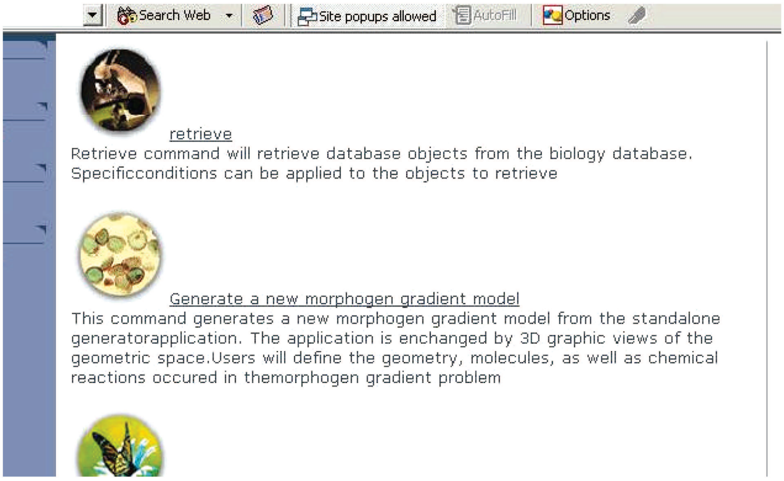


```
d[a]/dt = da•d^2[a]/dx^2 + pa - On_ac•[a][c] + Off_ac•[ac]

d[b]/dt = db•d^2[b]/dx^2 + pb - On_bc•[b][c] + Off_bc•[bc]

d[c]/dt =  - On_ac•[a][c] + Off_ac•ac - On_bc•[b][c] + Off_bc•[bc]
```

Fig. 9. Equations.



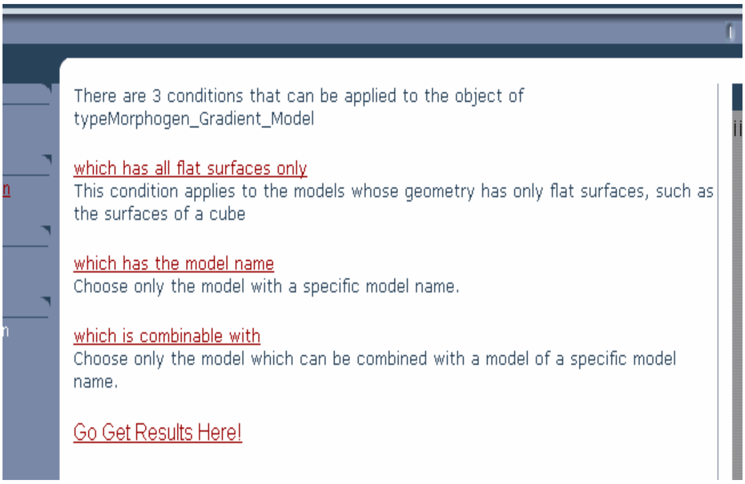
retrieve

Retrieve command will retrieve database objects from the biology database. Specificconditions can be applied to the objects to retrieve

Generate a new morphogen gradient model

This command generates a new morphogen gradient model from the standalone generatorapplication. The application is enchanged by 3D graphic views of the geometric space.Users will define the geometry, molecules, as well as chemical reactions occured in themorphogen gradient problem

Fig. 10. Select verb (operation).



There are 3 conditions that can be applied to the object of typeMorphogen_Gradient_Model

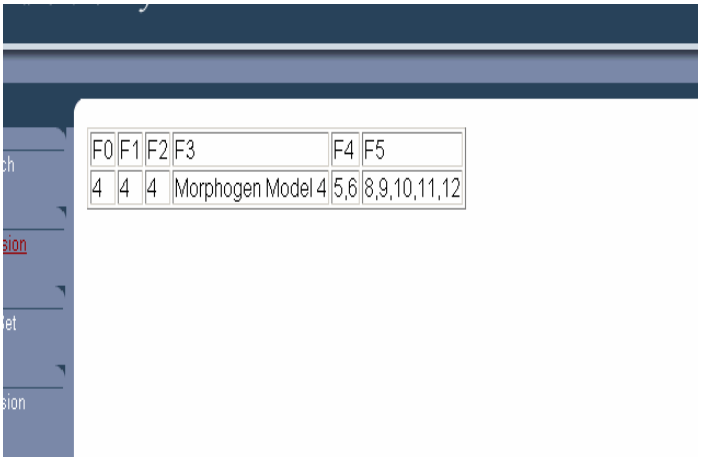
which has all flat surfaces only
This condition applies to the models whose geometry has only flat surfaces, such as the surfaces of a cube

which has the model name
Choose only the model with a specific model name.

which is combinable with
Choose only the model which can be combined with a model of a specific model name.

Go Get Results Here!

Fig. 11. Define condition.

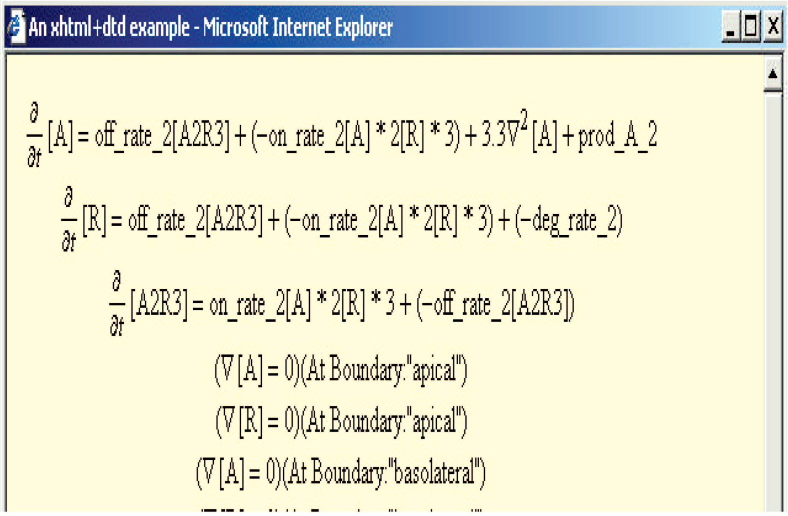


F0	F1	F2	F3	F4	F5
4	4	4	Morphogen Model 4	5,6	8,9,10,11,12

Fig. 12. Retrieve output.

A web interface for the Morphogen Gradient Knowledge Base has been created. Users can search a Morphogen Gradient based on some search criteria, generate the mathematical equations for a biological model, and synthesize a set of smaller models into a lager, more complex model. Each type of operation is represented as an English verb. Therefore, users use the web interface to choose a verb (an operation), for instance, "retrieve" (Fig. 10). "Retrieve" will find a set of models that satisfy certain conditions, so the next step is to have users define these conditions (Fig. 11).

Having specified the operation "retrieve" and the conditions, the web interface will return the data of the models that satisfy the conditions (Fig. 12). If the opera-



$$\frac{\partial}{\partial t}[A] = off\_rate\_2[A2R3] + (-on\_rate\_2[A] * 2[R] * 3) + 3.3\nabla^2[A] + prod\_A\_2$$

$$\frac{\partial}{\partial t}[R] = off\_rate\_2[A2R3] + (-on\_rate\_2[A] * 2[R] * 3) + (-deg\_rate\_2)$$

$$\frac{\partial}{\partial t}[A2R3] = on\_rate\_2[A] * 2[R] * 3 + (-off\_rate\_2[A2R3])$$

$$(\nabla[A] = 0)(\text{At Boundary:"apical"})$$

$$(\nabla[R] = 0)(\text{At Boundary:"apical"})$$

$$(\nabla[A] = 0)(\text{At Boundary:"basolateral"})$$

Fig. 13. Equation generation output.

tion is model mathematical equation generation, the web interface will generate the corresponding equations in a new window (Fig. 13). If the operation is synthesizing two models, a new model will be created in the database, and the web interface output will be similar to that of Fig. 13.

This demo has served as an experiment to show how our concepts may be carried out in real life. For instance, although we have identified diffusion as a common biological process that could be treated as a building block for biological models, we want to extend the demo to see exactly what steps are needed to completely define a diffusion process, how to store all the information pertinent to diffusion into the database, and how to generate equations from the defined models. For diffusion alone, this task may seem trivial. However, via interactions with biologists, many hidden issues come to the surface. For instance, the problem of how to represent boundary conditions turns out to be quite a challenge. Also, in a situation in which a biologist leaves many parameters unspecified simply because he/she does not know the values, the question of whether to check external sources for the values, or to automatically assign values to the unspecified parameters so that we can still generate mathematical equations remains to be solved systematically.

7. Natural Language Interface

Our web interface uses Structure Natural Language (SNL) to facilitate the querying processes. As we have seen, SNL requires users to choose a verb, a noun and conditions from a set of pre-built natural language building blocks. A further advancement of this limited natural language capability is to provide a true natural language interface where users can type in natural language commands and pose

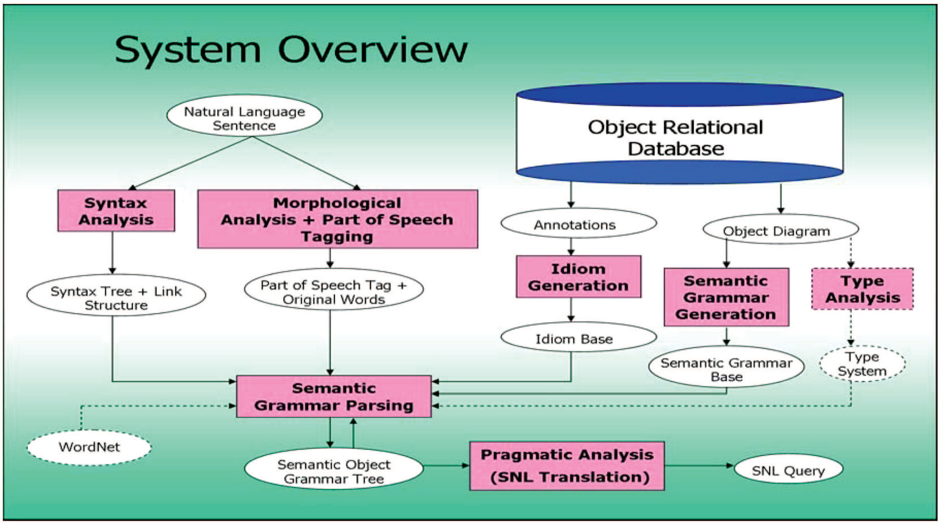

System Overview
Natural Language Sentence
Object Relational Database
Syntax Analysis
Morphological Analysis + Part of Speech Tagging
Annotations
Object Diagram
Idiom Generation
Semantic Grammar Generation
Type Analysis
Syntax Tree + Link Structure
Part of Speech Tag + Original Words
Idiom Base
Semantic Grammar Base
Type System
Semantic Grammar Parsing
WordNet
Semantic Object Grammar Tree
Pragmatic Analysis (SNL Translation)
SNL Query

Fig. 14. Semantic talk architecture.

questions. Specifically, we want to develop a natural language interface that allows biologists to ask questions about a model database in the following fashion:

Q: Retrieve the morphogen gradient model which has all flat surfaces.
Q: Synthesize "model 2" and "model 3".
Q: Which morphogen gradient model is compatible with model 4?

Accordingly we have proposed Semantic Talk, a natural language interface for SemanticObjects. Semantic Talk pre-builds the semantic grammar [40] of the model database, and can expand the default grammar to cover a wider range of natural language expressions through interactive and machine learning processes.

Semantic grammar is a set of grammar rules that can be used by computer programs to parse a limited set of natural language expressions. The set of natural language expressions a semantic grammar can parse is limited to the set of natural language expressions the semantic grammar can generate. Figure 14 shows how Semantic Talk generates and parses the semantic grammar for our biological model database.

The input to Semantic Talk is a natural language sentence. This sentence is simultaneously passed to the syntax analysis unit and the morphological and part-of-speech unit. The syntactic analysis process retrieves the English syntactic structure of the sentence, while the morphological and part-of-speech process tag each word in the sentence with morpheme and POS information. The next stage is to perform semantic analysis on this sentence. This is the most important processing unit in Semantic Talk. The Semantic Grammar Parsing unit uses a semantic grammar base, which stores all production rules and non-terminals of the grammar, to parse the input natural language sentence. The semantic grammar is generated

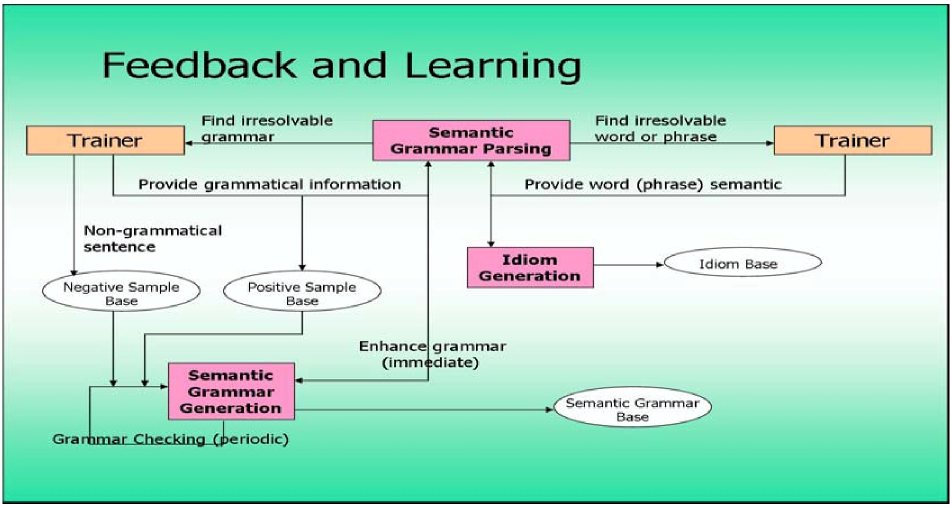

Feedback and Learning
Trainer
Find irresolvable grammar
Semantic Grammar Parsing
Find irresolvable word or phrase
Trainer
Provide grammatical information
Provide word (phrase) semantic
Non-grammatical sentence
Idiom Generation
Idiom Base
Negative Sample Base
Positive Sample Base
Enhance grammar (immediate)
Semantic Grammar Generation
Semantic Grammar Base
Grammar Checking (periodic)

Fig. 15. The learning architecture.

from our model database by analyzing the structure of our model database off-line. This grammar generation process is partially automatic and partially manual. After the Semantic Grammar Parsing stage, a SemanticObjects grammar tree is built and fetched to the Pragmatic Analysis unit, which will translate the grammar tree into SemanticObjects' query language.

No matter how carefully the set of semantic grammar is generated, there will always be new natural language expressions articulated by users that a fixed set of semantic grammar cannot parse and understand. Therefore, we also have the capability of learning in Semantic Talk.

Figure 15 shows the learning architecture of Semantic Talk. Semantic Talk can learn interactively and directly from trainers. Trainers are different from normal users of the system in that a trainer should be fairly familiar with the knowledge base, including what types of objects, verbs and adjectives are in the knowledge base; and they should also be familiar with the general principle about how the NLP system works. In most cases, trainers will directly tell the computer program how to generate a new semantic grammar rule to handle an unseen natural language utterance. Another type of interactive learning method that Semantic Talk employs involves normal users. Semantic Talk keeps a positive sample base for all the sentences that have been successfully parsed, as well as a negative sample base. Once the sample database is adequately large, we can employ hypothesis forming methods that are similar to [41] and [42], based on the semantic and syntactic features, to guess what a new natural language expression which the current semantic grammar fails to parse should mean. Then the system will ask normal users the validity of the guesses. Finally, Semantic Talk would employ grammatical inference methods such as that of [42] to learn new grammar rules from the positive and negative samples. This type of machine learning method does not require interactions with users at all.

8. Conclusions

This paper addresses a fundamental scientific problem as well as a fundamental information technology problem: to search for a mechanism that integrates biological parts into a complex biological system and to efficiently store, retrieve and analyze such biological models.

The goal of our research is to eliminate, or at least alleviate, the obstacles between biologists, mathematicians and engineers. In this paper, we have described the construction of a prototype "BioFactory" that allows a biologist to compose a biological system in terms of biological building blocks presented with a user-friendly interface. Two specific applications, namely the morphogenesis problem and the tumor growth problem, were chosen, because of their implications on the methods to undertake the research.

The information technologies developed by this research will not only address heretofore unanswered questions in, say tumor biology, but will also generate the building blocks that will enable the application of this approach to a wide range of problems in biology.

Acknowledgments

Fei Xei and Phillip Sheu like to thank Mr. Hiroshi Yamaguchi and Mr. Atsushi Kitazawa of NEC Soft, Japan for providing many technical advices and assistance about *SemanticObjects*. Vittorio Cristini acknowledges partial support from the National Science Foundation.

References

1. R. Gallagher and T. Appenzeller, Beyond reductionism, *Science* **284**(5411) (1999) 79.
2. B. E. Shapiro and E. D. Mjolsness, Developmental simulation with cellerator, *Proc. 2nd Int. Conf. on Systems Biology* (*ICSB*), 2001, pp. 342–351.
3. L. M. Loew and J. C. Schaff, The virtual cell: A software environment for computational cell biology, *Trends Biotechnol.* **19**(10) (2001) 401–406.
4. M. Tomita, K. Hashimoto, K. Takahashi, Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K. Yugi, J. C. Venter, and C. Hutchison, E-CELL: Software environment for whole cell simulation, *Bioinformatics* **15**(1) (1999) 72–84.
5. P. Mendes, GEPASI: A software package for modelling the dynamics, steady states and control of biochemical and other systems, *Comput. Appl. Biosci.* **9** (1993) 563–571.
6. H. M. Sauro, Jarnac: A system for interactive metabolic analysis. Animating the cellular, *Map 9th Int. BioThermoKinetics Meeting* **33** (2000) 221–228.
7. M. A. Gibson, Computational methods for stochastic biological systems PhD Thesis, Calif. Inst. Technology (2000).
8. N. Le Novère and T. S. Shimizu, StochSim: Modeling of stochastic biomolecular processes, *Bioinformatics* **17** (2001) 575–576.

9. T. M. Bartol, Jr., J. R. Stiles, M. M. Salpeter, E. E. Salpeter and T. J. Sejnowski, MCELL: Generalized Monte Carlo computer simulation of synaptic transmission and chemical signaling, *Society for Neuroscience Abstract* **22** (1996) 1742.
10. A. D. Lander, Q. Nie and F. Y. Wan, Do morphogen gradients arise by diffusion?, *Dev. Cell* **2** (2002) 785–796.
11. A. Eldar, R. Dorfman, D. Weiss, H. Ashe, B. Z. Shilo and N. Barkai, Robustness of the BMP morphogen gradient in Drosophila embryonic patterning, *Nature* **419** (2002) 304–308.
12. S. Y. Shvartsman, C. B. Muratov and D. A. Lauffenburger, Modeling and computational analysis of EGF receptor-mediated cell communication in Drosophila oogenesis, *Development* **129** (2002) 2577–2589.
13. S. P. Palecek, A. F. Horwitz and D. A. Lauffenburger, Kinetic model for integrin-mediated adhesion release during cell migration, *Ann. Biomed. Eng.* **27** (1999) 219–235.
14. M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, A. P. Arkin, B. J. Bornstein, D. Bray, A. Cornish-Bowden, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, J.-H. Hofmeyr, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, N. Le Novere, L. M. Loew, D. Lucio, P. Mendes, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D. Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang, The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models, *Bioinformatics* (2002).
15. H. S. Wiley, S. Y. Shvartsman, and D. A. Lauffenburger, Computational modeling of the EGF-receptor system: A paradigm for systems biology, *Trends Cell Biol.* **13** (2003) 43–50.
16. N. Bellomo and L. Preziosi, Modelling and mathematical problems related to tumor evolution and its interaction with the immune system, *Math. Comput. Modelling* **32** (2000) 413–452.
17. M. A. J. Chaplain, Avascular growth, angiogenesis and vascular growth in solid tumours: The mathematical modelling of the stages of tumour development, *Math. Comput. Modelling* **23**(6) (1986) 47–87.
18. H. P. Greenspan, On the growth and stability of cell cultures and solid tumors, *J. Theor. Bio.* **56** (1976) 229–242.
19. D. L. S. McElwain and L. E. Morris, Apoptosis as a volume loss mechanism in mathematical models of solid tumor growth, *Math. Biosciences* **39** (1978) 147–157.
20. A. Friedman and F. Reitich, Symmetry-breaking bifurcation of analytic solutions to free boundary problems and application to a model of tumor growth, *Trans. Amer. Math. Soc.* **353** (2001) 1587–1634.
21. C. Y. Chen, H. M. Byrne, and J. R. King, The influence of growth-induced stress from the surrounding medium on the development of multicell spheroids, *J. Math. Biol.* **43** (2001) 141–220.
22. G. J. Pettet, C. P. Please, M. J. Tindall, and D. L. McElwain, The migration of cells in multicell tumor spheroids, *Bull. Math. Biol.* **63** (2001) 231–257.
23. W. S. Kendal, The role of multiple somatic point mutations in metastatic progression, *Math. Biosci.* **108** (1992) 81–88.
24. T. L. Jackson, Vascular tumor growth and treatment: Consequences of polyclonality, competition and dynamic vascular support, *J. Math. Biol.* **44** (2002) 201–226.
25. M. Ohtaki and O. Niwa, A mathematical model of radiation carcinogenesis with induction of genomic instability and cell death, *Radiat. Res.* **156** (2001) 672–677.

26. J. H. Mao, K. A. Lindsay, R. J. Mairs, and T. E. Wheldon, The effect of tissue-specific growth patterns of target stem cells on the spectrum of tumors resulting from multistage tumorigenesis, *J. Theor. Biol.* **210** (2001) 93–100.
27. K. W. Kinzler and B. Vogelstein, Gatekeepers and caretakers, *Nature* **386** (1997) 761–763.
28. F. D. Alfano, A stochastic model of cellular transformation and its relevance to chemical carcinogenesis, *Math. Biosci.* **149** (1998) 95–106.
29. S. Y. Shvartsman, H. S. Wiley, W. M. Deen, and D. A. Lauffenburger, Spatial range of autocrine signaling: Modeling and computational analysis, *Biophysical J.* **81** (2001) 1854–1867.
30. A. R. A. Anderson and M. A. J. Chaplain, Continuous and discrete mathematical models of tumor induced angiogenesis, *Bull. Math. Biol.* **60** (1998) 857–900.
31. H. A. Levine, S. Pamuk, B. D. Sleeman, and M. Nilsen-Hamilton, Mathematical modeling of capillary formation and development in tumor angiogenesis: Penetration into the stroma, *Bull. Math. Biol.* **63** (2001) 801–863.
32. H. A. Levine, B. D. Sleeman, and M. Nilsen-Hamilton, Mathematical modeling of the onset of capillary formation initiating angiogenesis, *J. Math. Biol.* **42** (2001) 195–238.
33. M. J. Holmes and B. D. Sleeman, A mathematical model of tumor angiogenesis incorporating cellular traction and viscoelastic effects, *J. Theor. Biol.* **202** (2000) 95–112.
34. http://sig.biostr.washington.edu/projects/da/
35. http://www.nlm.nih.gov/research/visible/visible_human.html
36. L. Preziosi, *Cancer Modeling and Simulation* (CRC Press, 1̣8 June, 2003).
37. H. M. Byrne and M. A. J. Chaplain, Growth of nonnecrotic tumors in the presence and absence of inhibitors, *Mathematical Biosciences* **130** (1995) 151–181.
38. H. M. Byrne and M. A. J. Chaplain, Growth of necrotic tumors in the presence and absence of inhibitors, *Mathematical Biosciences* **135** (1996) 187–216.
39. X. Zheng, S. Wise+ and V. Cristini, Nonlinear simulation of tumor necrosis, neovascularization and tissue invasion via an adaptive finite-element/level-set method, *Bulletin of Mathematical Biology* **67** (2005) 211–259.
40. R. R. Burton, Semantic grammar: A technique for efficient language understanding in limited domains, Doctoral Thesis, University of California at Irvine, January 1976.
41. M. Gavaldà and A. Waibel, Growing semantic grammars, in *Proc. COLING/ACL-1998*, 1998.
42. D. Gildea and D. Jurafsky, Automatic labeling of semantic roles, *Computational Linguistics* **28**(3) (2002) 245–288.
43. K. Nakamura and M. Matsumoto, Incremental learning of context free grammars, *Proc. 6th Int. Colloquium on Grammatical Inference: Algorithms and Applications*, September 2002.
44. J. Sinek, H. Frieboes, X. Zheng and V. Cristini, Two-dimensional chemotherapy simulations demonstrate fundamental transport and tumor response limitations involving nanoparticles, *Biomedical Microdevices* **6** (2004) 297–309.
45. V. Cristini, J. Lowengrub, and Q. Nie, Nonlinear simulation of tumor growth, *Journal of Mathematical Biology* **46** (2003) 191.