

The current cellular models range from relatively small models of prokaryotes up to yeast, eukaryotic and mammalian cells. There are many models of small pathways systems, mostly focused on receptor kinetics or signal transduction. The problem is not always that the connections aren't known, but that usually the kinetics is poorly understood. The largest models appear to be based on *E. coli* and there are several being developed in different countries (US, Japan, UK, and Canada). Recently an exciting development has been the creation of the International Alliance on *E. coli* models (IECA) was announced (Holden 2002). IECA will bring together the global community of experimentalists and modelers to create an all-embracing model of *E. coli*. The current generation of cellular and pathway models are being used not only in the academic community, but have found industrial applications for process design in biotechnology, metabolic bioengineering, pharmaceutical target validation, strain selection and different types of screening procedures.

As more and more data is being generated, the pathways that can be reconstructed become larger as do the resulting mathematical models with both more entities and interactions included. Soon expect that many of the current computational approaches will no longer be adequate. These fall into the following categories:

- knowledge management and representation
- analytical techniques for analysis of the model
- visualization of the pathway and the simulation results
- numerical simulation
- analysis of the simulation results
- new knowledge discovery.

The theoretical gaps arise from a lack of adequate theory for these large meso-scale problems. For example, there is need to understand what should be the constraints for guaranteeing positive, semi-definite asymptotic and bounded solutions. There are efficiency gaps that arise from the scaling up of the problems - both for the numerical simulations and for graph theoretical approaches. Lastly, there is a need for better software that scales well and is efficient. In this paper, we will focus on opportunities for artificial intelligence techniques and not dwell on the larger set of computational challenges.

The model building process

Constructing the mathematical models is a form of knowledge management - combining biological knowledge with experimental data in a formal, consistent manner. The mathematical representation of the kinetics of the interactions, and any mass balances, present a rigorous approach to providing formality and consistency. Nonetheless, there is a need to pick reasonable interactions with plausible parameters.

Modeling building is not a singular event, but recurs as new data and new biological knowledge accumulate triggering a need to revisit the model and make sure it is still accurate and to incorporate the new information. Hence, there is a need to develop a systematic process to facilitate and automate model development. In addition, given a particular knowledge representation, there is an additional need to understand how to trigger this process and understand the resulting differences in the old and new models. The triggering process should be as automatic as possible with regard to the integration or checking of data, and to put the pieces together to create a new pathway model *de novo*. This automation requirement implies the use of inference agents examining the existing model, the new data or information, and deciding on what needs to be done. If model building were a relatively expensive process, one would not want to trigger new versions for minor data changes. A rather mundane consequence of building and re-building models is the need to have versioning, much as is done in software engineering.

Model building consists of several straightforward steps. The first is a bioinformatics exercise to identify which entities are involved - DNA, RNA, proteins, peptides and small molecules - from interactions with each other, in various biological processes like transcription, translation, signal transduction, catalytic cycles of enzymes, and membrane transport. The next step - **static model development** - finishes with the identification of the interactions and, at this point, the model contains only connectivity (topology) of the interactions, and no kinetic information. At this point, some of the connections may be speculative or tenuous and there is a need to include this state of affairs in the knowledge representation.

Automatic collection processes are used to peruse special databases, even the literature, to find pathway information. Because diverse sources are used, ambiguity and inconsistency can result from the particularly daunting problem of the naming of the various species. While biological names are often standardized, they do change from time to time, and there are no guarantees that standardized names are used. Different names, spelling variants, even misspellings, are a serious problem with potentially drastic consequences to the quality of the model. Biological names are usually unstructured, but the names of chemical compounds may have more structure - which is both a curse and blessing. Depending upon context dichlorobenzene and 1,2-dichlorobenzene may be considered synonymous. There is a need for a fast and context-dependent fuzzy spelling dictionary to identify such situations.

The second major stage is to create the **kinetic models** by adding the mechanistic ordinary differential equations that determine the dynamical state of the system at any moment

of time. The next step is to use experimental data, in an inverse numerical approach to determine the rate constants in the model. This is largely a mathematical optimization problem and while there are serious scalability problems, they will not be dwelled upon here. The third stage of modeling is to generate the corresponding mathematical equations for the whole pathway and several XML-based standards have emerged for this step. SBML(Hucka, Finney et al. 2002) and CELLML (Hedley, Nelson et al. 2001) are two of the better known ones. After the equations are generated, numerical simulations based upon integrating the differential equations and a variety of mathematical analyses are possible, including asymptotic stability, parameter sensitivity, and linear stability. An analysis of the simulation results can be used to look for 'new' biology.

We have successfully applied this multi-step model-building strategy to successfully create and integrate more than 14 kinetic sub-pathways of the *E. coli* metabolism(Goryanin, Demin et al. in press).

AI challenges in pathway modeling

Many of the computational challenges arise from the innate mathematical nature of the modeling. Some are inherent from the scalability issue. Yet, if one takes the view that the modeling process is also a knowledge manipulation problem, then many other problems emerge. These fall into several broad categories:

- knowledge acquisition and representation
- simulation of the model
- analysis and visualization of the simulation results
- discovering new biology

The modeling process can be viewed as acquiring knowledge, reformulating into a new representation (the mathematical model) and extracting new information from that model (the simulation). These efforts have been traditionally cast in either a mathematical or a biological perspective. It is our view, that by augmenting the traditional approaches with artificial intelligence viewpoints will only enhance the modeling effort.

The knowledge acquisition aspect arises from the need to mine the literature and biological databases to automatically collect kinetic information for pathways. In many ways, this is a text mining and automated parsing problem. The difference arises in the need to be able to supply domain specific information to improve generic text mining approaches.

Recently there has been increasing interest in developing biological ontologies (Consortium 2000; Consortium). However, they have largely been focused on genetic information and there is a need to extend them to the larger realm of cellular functioning. Cellular function encompasses genes, proteins and other kinds of entities. Genes are spatially inert, but their functioning varies temporally.

Proteins are temporally active and, in many cases, have complex spatial activity within the cell. For example, a cell surface receptor may bind a ligand, activate a signal transduction pathway, then internalize into the interior of the cell, lose its ligand, and then reappear on the cell surface to repeat the cycle. Both temporal and spatial logics would be employed.

The resulting network representation will not be perfect. There will be inconsistencies arising from multiple knowledge sources. There will be incompleteness since we have only imperfect knowledge of the biology. This may manifest itself as unconnected segments of the topology or parts of the model that lack kinetic information. Some parts of the model may be speculative or poorly understood. Such ambiguity will always be inherent in the models and it is better to adapt to it than to try to ignore it. Clearly, this puts a requirement on the knowledge representation and acquisition to incorporate any ambiguity. However, it is not ambiguity alone that must be considered - the quality of the data comes into play as well. Quality information comes from several sources - the accuracy of the biology or experimental error from the kinetics experiments.

Numerical simulation has a long history and is a well established discipline within applied mathematics. Nonetheless, in the modeling of such large systems, not all the problems are mathematical. A planning problem arises from a subtle aspect of how simulations are performed and understanding what is truly intrinsic to the models. Models are driven by parameters some of which can be considered as being internal to the model. Internal parameters occur directly to describe the kinetics and the relationships in the model. The external parameters describe either the simulation environment of the situation being simulated or the biological environment to which the model will be subjected. For example, kinetic rate constants are internal to the model, but environmental factors such as temperature, type of nutrient available (carbohydrates or lipids) or pH of the medium surrounding a bacterium would be external. Implicit external parameters occur as well - integration time step sizes and ranges, choices of thresholds for integrators, etc. From one simulation run to the next, some of them may change. The instantiation of the internal and external parameters with explicit values creates a simulation instance. Parameter choices may not only be scalar, but vector as well. For example, Monte Carlo sampling of parameter space may be used, or parameters may be 'swept' over a range of values for sensitivity calculations. The choice of one or more parameter values and the methods of those choices, form the basis of the concept of simulation 'scenarios.' A scenario is the *raison d'être* for a simulation run and may constitute more than one related numerical simulation. Since scenarios have purposes, they encode the goals of the resulting analyses. In this way, they act as automated planning for the following analyses.

There is a need for a formal description language for developing scenarios, automated analysis planning, and the execution of the analyses.

To date, most simulations have been numerical integrations of differential equations - continuous, discrete, or stochastic. In the language of fuzzy logic, they depend upon 'crisp' quantities. Yet, often our knowledge of the equations is not crisp, but ambiguous. It would be desirable to represent the ambiguity throughout the integration from all the sources of the uncertainty - error, biological knowledge, rate equations, rate constants, and initial conditions. There are few good techniques for incorporating such uncertainty in numerical integrations. Integrators using interval analysis (Moore 1966; Kearfott and Kreinovich 1996) are limited to only handling uncertainty in the initial conditions (reference). Much more is needed. An extension of fuzzy arithmetic to fuzzy integration is needed. A numerically stable, robust and accurate integrator based on fuzzy arithmetic would be a major advance. It would also represent an intriguing blend of artificial intelligence and numerical analysis.

There is yet another synthesis needed - integrators that blend qualitative and quantitative simulation. Such an approach would be a tool to get around the incompleteness issue in adding kinetics to static models. Oftentimes, rough kinetics can be 'guessed', but must be expressed in exact quantitative ways as if there were no qualitative aspects. Having a hybrid integrator that combines quantitative integration with qualitative simulation would allow models to be larger than if just 'pure' quantitative approaches were used. It would be obvious which parts of the model and results were qualitative. Let's go further and ask for an integrator that does it all:

- exact quantitative - continuous, discrete, stochastic
- fuzzy quantitative
- qualitative

This is clearly a challenging and ambitious goal - one that will not only advance pathway modeling, but other fields of biological simulation as well.

One of the scalability issues that arise is the ability of humans to analyze the results of the simulations. As pathways model become larger, it no longer is feasible to scan the resulting time curves either for validation, comparison or knowledge discovery purposes. Tens of thousands of time curves present a daunting problem. Comparing that many time curves across multiple simulations makes the problem even more harrowing. Better tools are needed, in conjunction with the scenario concept, for automating these searches. There are elements of pattern recognition and data mining, but more is needed. The scenarios set the goals, the model or the larger biology sets the context, for the examination. In addition, these knowledge domain

constraints must somehow be integrated to build better analyzers.

Perhaps the most intriguing aspect of simulation analysis, is the search for new knowledge - *new* biology. If we take the view that a model is good, then we should be able to find knowledge that has escaped our attention so far. In fact, one can construct a gedanken experiment along these lines. Suppose we had a perfect model - complete, consistent, no error, and no ambiguity. Could we rediscover biological pathways? Could we rediscover the known phenomena associated with these systems? Let's take the experiment further and drop all labels. We know the entities in the model only by their actions. Could we rediscover the concepts of protein or gene? Our 'real' models are not perfect - far from it. So our attempts at knowledge discovery are going to be harder. Nonetheless, we should still be willing and able to make the attempt. This is where good tools for this type of a knowledge discovery enterprise are sorely needed.

Similar to knowledge discovery is the model validation problem. Typically, models are validated through the parameter determination process. To find the kinetic parameters, an inverse problem is solved - parameters are chosen such that some type of optimization objective function is minimized. The most common objective function is usually an L_p norm between the model predictions and experimental kinetic data. One of the many potential problems is that many interactions are not pairwise, and many entities simply could be missing. As a result, many hypothetical variables would be needed to fit the data, leading data overfitting and the inability to interpret the hypotheses in a biological meaningful way. So, the model is validated in the sense that its predictions predict results 'close' to the experiments. There are several problems with this approach. It is a 'necessary' method, but by no means a 'sufficient' one. Part of the problem is that experiments measure only a very small subset of all the entities in the model. Therefore, while a good fit may be obtained, there is no a priori guarantee that all the entities are well represented. Additional problems like noise and numerical ill-posedness come into play as well. Therefore, model validation is not as accurate or complete as desired.

As a rule if the model is good, then the knowledge that went into its building should be extractable from it. The fitting of experimental data does not address this. So we need tools for validating the model by comparing the biological knowledge that went into its building with the simulation results. In one sense, this is very similar to the problem of software validation and the concept of assertion testing for proving correctness. Validation can be considered as a form of forward chaining inference :

input_biology \Rightarrow simulation_results

The challenge is to be able to represent the biology in a formal representation in order to make the inference checking work. The biology is both temporal and spatial and may use rules that encompass all spatial positions or all times. The representation must be expressive for events occurring on these two (concentration and time) or five (concentration, position, time) dimensional surfaces. The basic constructs involve surface shape characteristics (e.g. extrema, saturation, decay, etc.) or time events (e.g. positions of extrema or zeroes). A trivial example would be to require that all enzyme concentrations, at all positions, at all times would be positive semi-definite. These formal assertions can be used for a variety of purposes:

- plausibility - the biology behaves as expected
- diagnostic - finding out why the system behaves as it does
- speculative - finding phenomena that is postulated or rarely observed

In this way the assertion checking serves as the eyes-and-ears of the modeler to find inconsistencies, validation failures, or to discover new biology.

A lot remains to be done in the computational toolkit for modeling biological pathways. the challenges are in many disciplines - biology, computational biochemistry, mathematical analysis, software development, and in the general understanding of the knowledge system. It will take multi-disciplinary efforts to tackle this difficult, but ultimately rewarding area. Development of these modeling systems for virtual experimentation will take on greater importance as the biological systems become more complex. Currently the development of whole cell *E. coli* has already strained the existing computational tools (Goryanin, Demin et al. in press). Such models will become more sophisticated and will become ever more important in pharmaceutical, medical, and biological research.

ACKNOWLEDGMENTS

The authors would like to thank all of our collaborators who are involved in the *E. coli* whole cell modeling effort at GSK - Nick Juty, Hugh Spence, Serge Dronov and Oleg Demin at Moscow State University.

REFERENCES

The Gene Ontology Consortium, (2000). "Gene ontology tool for the unification of biology." *Nature Genetics* 25: 25-29.

The Gene Ontology Consortium, (2001). "Creating the gene ontology resource: design and implementation." *Genome Research* 11: 1425-1433.

Goryanin, I., O. Demin, et al. (in press). *Applications of Whole Cell and Large Pathway Mathematical Models in the Pharmaceutical Industry. Future Perspectives in Metabolic Profiling*. B. N. Kholodenko. Wymondham, United Kingdom, Horizon Press.

Hedley, W. J., M. R. Nelson, et al. (2001). "A short introduction to CellML." *Philosophical Transactions of the Royal Society of London A* 359: 1073-1089.

Holden, C. (2002). *Alliance Launched to Model E. coli*. *Science*. 297: 1459-1460.

Hucka, M., A. Finney, et al. (2002). "The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models." *Bioinformatics*.

Kearfott, R. B. and V. Kreinovich (1996). *Applications of Interval Computations*, Kluwer Academic Publishers.

Moore, R. E. (1966). *Interval Analysis*. Englewood Cliffs, New Jersey.