A Distributed Computational Architecture for Integrating Multiple Biomolecular Pathways

V. A. Shiva Ayyadurai and C. F. Forbes Dewey, Jr.

M.I.T.

Abstract—**Biomolecular pathways are building blocks of cellular biochemical function. Computational biology is in rapid transition from diagrammatic representation of pathways to quantitative and predictive mathematical models, which span time-scales, knowledge domains and spatial-scales. This transition is being accelerated by high-throughput experimentation which isolates reactions and their corresponding rate constants. A grand challenge of systems biology is to model the whole cell by integrating these emerging quantitative models. Current integration approaches do not scale. A new parallel computational architecture, CytoSolve, directly addresses this scalability issue. Results are presented in the solution of a concrete biological model: the Epidermal Growth Factor Receptor (EGFR) pathway model published by Kholodenko. The EGFR pathway is selected since known solutions exist for this problem thus enabling direct comparison of the CytoSolve approach. Results from this effort demonstrate that CytoSolve provides a core platform for addressing a grand challenge of Systems Biology to model the whole cell by integrating multiple biomolecular pathway models.**

Index Terms—**systems biology, computational models, signaling networks.**

I. Introduction

BIOMOLECULAR pathways are building blocks of cellular biochemical function. Computational biology is in rapid transition from diagrammatic representation of pathways to quantitative and predictive mathematical models, which span time-scales, knowledge domains and spatial-scales. This transition is being accelerated by high-throughput experimentation which isolates reactions and their corresponding rate constants. A grand challenge of systems biology is to model the whole cell by integrating multiple biomolecular pathways. Current integration approaches do not scale. A new parallel computational architecture, *CytoSolve*, based on earlier work on integrating multiple molecular pathways by Ayyadurai *et al* (1), directly addresses this scalability issue.

Results from CytoSolve are presented in the solution of a concrete biological model: the Epidermal Growth Factor Receptor (EGFR) pathway model published by Kholodenko et al (2). The EGFR pathway is selected since known solutions exist for this problem thus enabling direct comparison of the CytoSolve approach. Snoep *et al* (3) have instantiated the Kholodenko EGFR model into a software language known as Systems Biology Markup Language (SBML), which can be computed on software programs including Cell Designer designed by Kitano (4) which, like other solvers, takes a *monolithic approach* towards modeling and solving biomolecular pathway models.

By monolithic approach, it is meant the construction of a biomolecular pathway must be done *exclusively* within the framework of that one system. Thus, if one considers all the pathways representing all biochemical functions of the whole cell, modeling of the whole cell, in a monolithic approach, will require all individual pathways to be loaded and integrated within this one system. If individual pathways were developed in other computing systems, monolithic systems either do not support such integration or make such integration extremely onerous, at best.

The EGFR model was first constructed using Cell Designer's monolithic approach to solve for the various species concentration levels, as predicted by Kholodenko. CytoSolve was then used to solve the same EGFR problem but in a distributed fashion to yield near exact results as that of Cell Designer. These results demonstrate the viability of CytoSolve's unique distributed approach not only to solve problems that monolithic approaches are capable of solving but also demonstrates CytoSolve's flexibility and scalability in integrating multiple biomolecular pathway models, which monolithic approaches are incapable of doing. In CytoSolve, any *one* pathway can exist in any format and there is no need to manually load, understand and interconnect each individual pathway, as is required in monolithic systems. The CytoSolve approach, therefore, provides a core platform for addressing one of Systems Biology's grand challenges: modeling the whole cell by integrating multiple biomolecular pathway models which span time-scales, knowledge domains and spatial-scales.

II. Methodology

The goal of this research is to validate the distributed

Manuscript received November 25, 2006. This work was supported in part by the Singapore-MIT Alliance.

V. A. Shiva Ayyadurai is with Massachusetts Institute of Technology, Cambridge, MA 02138 USA (phone: 617-233-6353; fax: 617-258-8559; e-mail: vashiva@ mit.edu).

C. Forbes Dewey, Jr., is with Massachusetts Institute of Technology, Cambridge, MA 02138 USA (e-mail: cfdewey@mit.edu).

approach of CytoSolve to integrate and compute multiple biomolecular pathway models and contrast this approach to extant monolithic approaches. To perform this evaluation, two elements are required:

I. A concrete biomolecular pathway for which there exists both diagrammatic and mathematical representations along with known solutions; and,

II. A proven monolithic approach for integrating and solving multiple biomolecular pathway models.

Relative to (I), the EGFR pathway model of Kholodenko *et al* (1), as shown in Fig. 1 is selected. Fig. 1 represents the *whole* EGFR model.

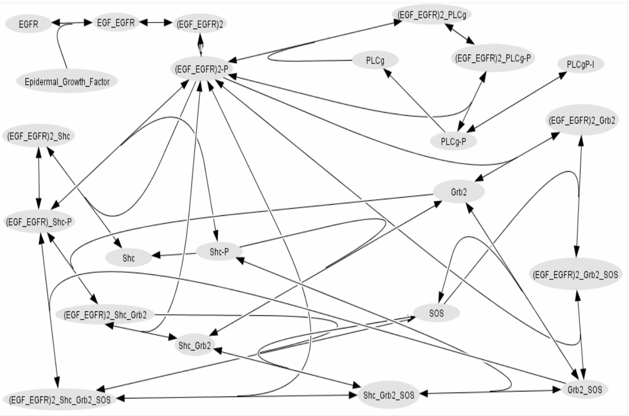

Fig. 1. Diagrammatic description of the whole EGFR pathway as published by Kholodenko *et al* (2).

Relative to (II), Cell Designer by Kitano *et al* (4) is selected as the monolithic method. There are many other systems such as Cell Designer that could have been selected; this tool was selected primarily based on its current popular use in the systems biology community. Cell Designer provides both a graphical mechanism for constructing the pathway diagram shown in Fig. 1 as well as an ordinary differential equation (ODE) solver for calculating the various species concentrations values over time. In Fig. 1, the creator of this pathway in Cell Designer had to "by hand" draw each and every species and then connect the species and instantiate the rate equations. Cell Designer requires the entire pathway to be coded into the Cell Designer system exclusively using the Cell Designer program.

Our methodology is to demonstrate that CytoSolve can integrate multiple biomolecular pathways without having to perform such "hand wiring". To demonstrate this, the following key steps are involved:

Step 1: EGFR Model Decomposition – Decompose the original Kholodenko *et al* (1) model into four sub-models, which will serve as the elements which need to be integrated to build the whole EGFR model in Fig.1;

Step 2: Sub-Model Solutions - Solve each of the four sub-models using both Cell Designer and CytoSolve to test the accuracy and compare computation time of each approach; and,

Step 3: Whole EFGR Model Solution - Enable CytoSolve to integrate all four sub models, each of which is distributed on four independent computers, with no human intervention and compare the integrated whole EFGR solution from CytoSolve with the whole EGFR solution calculated by Cell Designer.

Cell Designer and CytoSolve's central *controller* are executed on a Pentium 4 CPU 3.00 GHz Dell Workstation with 2 GB of RAM running Windows XP with Service Pack 2. In CytoSolve, each pathway model is treated as an independent entity, and is activated by communication with a central controller that insures mass conservation and other constraints on the aggregate system. Each of the individual sub-models (per Step 3), in the CytoSolve case, are also executed on a Pentium 4 CPU 3.00 GHz Dell Workstation with 2 GB of RAM running Windows XP with Service Pack 2.

III. Results

There are three sets of results. The first set of results represents the break-up of the EGFR pathway into its four sub-models. The second set of results provides the comparison of each sub-model executed in Cell Designer and in CytoSolve. The third set of results provides the entire EGFR model executed in both Cell Designer and CytoSolve.

A. *EGFR Model Decomposition*

The EGFR model of Kholodenko shown in Fig. 1 can also be considered to be derived by integrating a set of smaller pathways.

There are many such smaller pathways. In Fig. 2, Fig. 3, Fig. 4 and Fig. 5, diagrammatic representations of one set of such smaller pathways are created, and denoted as Model 1, Model 2, Model 3 and Model 4, respectively, which, if integrated would derive the whole EGFR pathway shown above in Fig. 1.

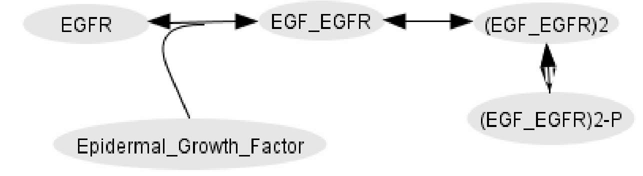

Fig. 2. Diagrammatic description of Model 1, one portion of the whole EGFR model.

In reviewing Model 1, Model 2, Model 3 and Model 4, one will recognize that the species **(EGF_EGFR)2-P** is shared by all four models. Model 3 and Model 4 share the common species **SOS**.

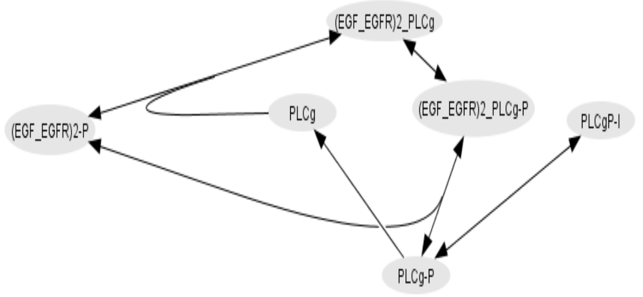

Fig. 3. Diagrammatic description of Model 2, second portion of the whole EGFR model.

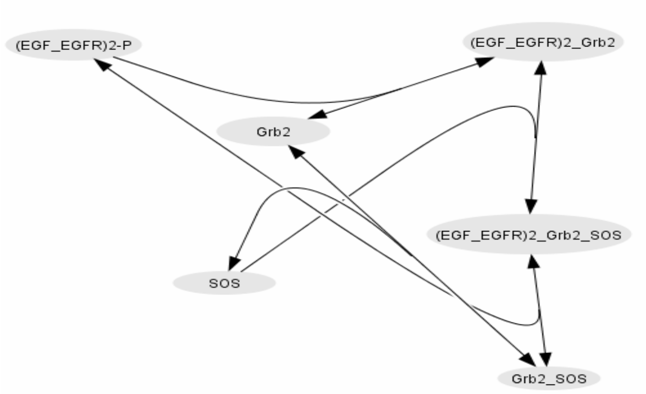

Fig. 4. Diagrammatic description of Model 3, third portion of the whole EGFR model.

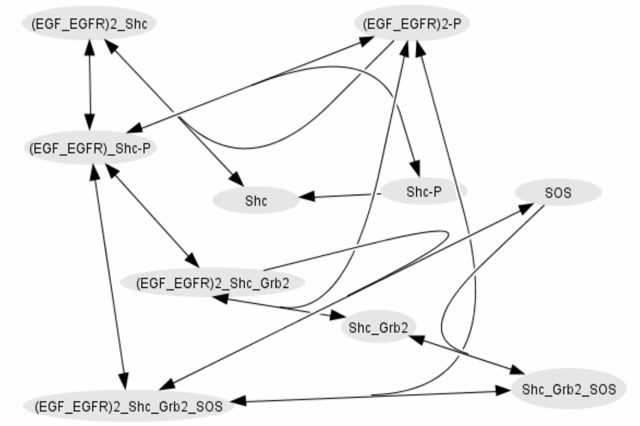

Fig. 5. Diagrammatic description of Model 4, fourth portion of the whole EGFR model.

B. Sub-Model Solutions

Below in Table I, the results of executing each of the four sub-models: Model 1, Model 2, Model 3, Model 4, first in Cell Designer then in CytoSolve individually are presented.

TABLE I
CELL DESIGNER AND CYTOSOLVE RESULTS
FOR COMPUTING EACH SUB-MODEL

Model	Cell Designer	CytoSolve	Difference
Model 1	1310 ms	4271 ms	0.021%
Model 2	1752 ms	4615 ms	0.034%
Model 3	1763 ms	4714 ms	0.015%
Model 4	2133 ms	5102 ms	0.017%

For Cell Designer, each model was loaded in one at time and then executed. For CytoSolve, CytoSolve's central controller was implemented on one server and each model was implemented on another server. The results in Table I for columns 2 and 3 are a result of averaging five different test runs. The Difference is calculated as the RMS average across those five test runs for various species concentrations in each sub-model. The difference in compute times is primarily due to network latency required for CytoSolve's central controller to contact and receive information back from each model. Cell Designer has no network latency since each model runs on the same server as Cell Designer.

C. Whole EGFR Model Solution

In this case, the full integration of all four models is performed to derive the whole EGFR model in Fig. 1. For Cell Designer, all four models (depicted in Figs. 2-5) were loaded into the Cell Designer system and had to be connected by hand to recreate the diagram in Fig. 1. This process took several hours to perform and ensure consistency and accuracy of the pathway as described by Kholodenko. For CytoSolve, the central controller was run on one machine and four separate computers were setup, each running one independent model. Recall, the goal in this exercise was to evaluate the difference in solution between CytoSolve and Cell Designer as well as computational time differences for deriving the whole EGFR model.

The results are shown in Table II.

TABLE II
CELL DESIGNER AND CYTOSOLVE RESULTS
FOR WHOLE EGFR MODEL

Cell Designer	CytoSolve	Difference
3217 ms	9685 ms	0.026%

IV. CONCLUSION

The results demonstrate the viability of CytoSolve's unique distributed approach not only to solve problems that monolithic approaches are capable of solving but also to provide greater flexibility and scalability in integrating multiple biomolecular pathway models, which monolithic approaches are incapable of doing. In CytoSolve, any *one* pathway can exist in any format on any computer, and there is no need to manually load, understand and interconnect each individual pathway, as is required in monolithic systems.

CytoSolve generated near similar results to Cell Designer; more importantly, the integration of the four sub-models in CytoSolve did not require any manual "wiring" as is needed by Cell Designer. CytoSolve's compute time was greater than Cell Designer; however, most of this compute time was due to network latency. Since CytoSolve works in a distributed parallel fashion, its compute time is a direct function of the compute time of the largest pathway plus the associated network latency. For Cell Designer, the compute time will be the compute time of the whole integrated pathway. Thus, as the number of pathways (sub-models) increase, Cell Designer's compute time will continue to increase, while CytoSolve's compute time will asymptotically reach a value equivalent to the compute time of the longest pathway.

Initial results from the EGFR example has demonstrated that CytoSolve can serve as an alternative to the monolithic approaches for integrating and solving biomolecular pathways. Most important is CytoSolve's core feature for integrating multiple pathway models, which can be distributed across multiple computing systems, without "hand wiring" of each model. While such a manual approach may be viable for a handful of models, it will not scale to support the integration of all pathway models necessary to model the whole cell. In addition, there are several other reasons why the monolithic approach will not scale.

First, scaling to, for example, over a hundred pathways and something like 10,000 equations – the level required to describe a single cell - would require a massive effort beyond the research expended to obtain the original individual pathways. A monolithic approach such as Cell Designer would not be able to effectively scale the integration of that many pathways.

Second, each pathway represents a knowledge domain, and it would be essentially impossible to have one person sufficiently knowledgeable in all the scientific areas to understand each of these domains well enough to manually construct a single monolithic program.

Third, the monolithic approach does not provide a means for pathways from proprietary models to be used with other models that are open source. An architecture such as CytoSolve's is needed that will allow people to contribute the output of their pathways to an external dynamic network of models without revealing the details of their internal structure.

Fourth, there has been no research to show that monolithic pathways can be distributed between machines for computational scalability. The CytoSolve approach parallelizes the computations from the beginning, making computational parallelization automatic.

Fifth, managing a monolithic model, composed of other sub-models, is a change management nightmare. Consider a small example of a monolithic model "cut and pasted" or concatenated from the four sub-models of EGFR, aforementioned, and each model being published and created by different authors. Now, suppose once the monolithic model has been constructed, that many months later, the authors of each of these models changes rate constants, pathway connections, etc., at that point the author of the monolithic model would have to rebuild the entire monolithic model, by instantiating changes from each author's model, which may be tenable for four sub-models (possibly based on the complexity and domain specificity of each model). Modeling the whole cell while managing such changes across a suite of hundreds of such sub-models will be untenable.

In summary, CytoSolve provides a core platform for addressing one of the grand challenges of Systems Biology: modeling the whole cell by integrating multiple biomolecular pathway models which span time-scales, knowledge domains and spatial-scales.

V. Future Work

Based on the results on applying CytoSolve to the EGFR pathway model, the following key areas of future work will be pursued:

(1) Advancements to the existing controller within the CytoSolve architecture. Such advancements will result from optimizing the time sequencing of how and when to call each sub-model during computation along with replacing the current event loop structure which has a fixed wait queue before processing the next computation.

(2) Specification of ontology standards for each pathway. Currently, each pathway model may use the same species; however, they may be named differently within each pathway model. There is a need to standardize the naming of species or provide an intermediate translation dictionary for fully automating the resolution of species names across pathway models.

(3) Demonstration of CytoSolve using other pathway models. New work is underway to use CytoSolve to solve a heretofore unsolved problem integrating extant pathway models. Currently, the authors are exploring integrating multiple pathway models involved in inflammatory response using CytoSolve's approach.

Acknowledgment

The authors would like to thank Ceryen Tan for providing software programming efforts for the distributed solver as well as Gene Deans and Prasad Jayraman for helping to setup the Semantic Web interface for remote distributed communication.

References

[1] V. A. Shiva Ayyadurai, C. F. Dewey, Jr., "Computing unsteady phenomenon across multiple molecular pathways," Poster presented at the 2005 BMES Annual Fall Meeting, Washington, D.C. USA, 2005.

[2] B. N. Kholodenko, O. V. Demin, G. Moehren, J. B. Hoek, "Quantification of short term signaling by the epidermal growth factor receptor," *Journal of Biological Chemistry*, vol. 274, no. 42 , pp. 30169-30181, 1998.

[3] J. L. Snoep, L. Lu, "Epidermal growth factor receptor signaling pathway," SBML software program file. European Bioinformatics Institute, 2005. December 5, 2005. <http://www.biomodels.net>.

[4] H. Kitano, *Cell Designer Version 3.2.0,* Software program, July 5, 2006.

V.A. Shiva Ayyadurai (SBEE'86–SMVS'89–SMME'90) holds a SBEE '86 in electrical engineering from M.I.T., a MSVS '89 in visual studies from the M.I.T. Medial Laboratory, a MSME '90 in mechanical engineering from M.I.T. He is currently a doctoral candidate at M.I.T. in systems biology within the M.I.T. Department of Biological Engineering.

C. F. Dewey, Jr. is a Professor at M.I.T. with join appointments in the M.I.T. Department of Mechanical Engineering and M.I.T. Department of Biological Engineering.