MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Towards Computational Formulation of Complex Theories through Model Interchange and Integration

Guy Tsafnat[a], Enrico W. Coiera[a] and Mike Bain[b]

[a] *Centre for Health Informatics, University of New South Wales, Sydney, NSW, Australia*
[b] *School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, Australia*

Abstract

Traditionally the time it took researchers to collect high-quality data through experimentation was longer than the time it took to formulate theories that explain them. In recent years, high-throughput data gathering methods have reversed this. Advances in computer technology and a systems approach to biology has led to the adoption of computer models as a means to formulate and test theories in silico faster and more cheaply than previously possible. This has led to a proliferation of model representations and ontologies, creating a complexity that is unwieldy to its human authors. Based on these trends and state-of-the-art practices evidenced in the literature, a new trend of computer-generated, integrated models of high complexity is predicted. In this paper we explain the trend and the challenges that it is likely to raise.

Keywords:

computer aided knowledge discovery, multi-models, multi-scale models, multi-method models

Introduction

In recent years, changes have been emerging in the way life-scientists make novel discoveries. New high-throughput measurement technologies such as mass spectrography, flow cytometry, microarray chips and biobanks are so widespread that more data is gathered than can be analysed [1]. Computer models are nowadays well established as tools required for gaining profound insights from data. The amounts of data and complexity of models is becoming a research area [2-5].

Another emerging research area, in which models are used in scientific discovery, is called "hypothesis generation". Briefly, an algorithm generates models that don't contradict known data [6,7]. In this paper we explain and give examples of two integrated biomedical models. We then discuss the requirements for computer generated integrated biomedical models.

Materials and methods

Many distinct methods have been shown to be effective in modelling life science systems. Some of the more common methods are differential equations [8], numerical methods [9], system dynamics [10], Bayesian networks [11], -calculus [12] and petri-nets [13].

Computer aided knowledge discovery is a relatively new sub-discipline of Machine Learning that is concerned with supporting the discovery of scientific theories. In biomedicine, machine learning has only been used for about 15 years [6,7,14-18]. The on-line Journal of Biomedical Discovery and Collaboration, the first peer-reviewed journal dedicated to biomedical knowledge discovery, was launched in March, 2006 [19].

Multi-modelling

Model construction is a tedious manual task even when assisted by authoring software. Recently, automatic generation of models was shown to be a promising way of constructing hypotheses and suggesting experiments that can falsify them and thus lead to new discoveries [15,20]. These successes exemplify a potentially substantial change in the way we discover natural phenomena.

Complex models are generally limited by several factors. The following issues are of interest to this paper:

- *Two or more modelled interacting systems might not have a common representation suitable to all of them.*
- *The data available for different modelled systems or even different parts of a system are collected in a variety of experimental settings and can't easily be consolidated.*
- *Models' complexity is hindering their intelligibility.*

These limitations can usually be alleviated by modelling each system independently and then interchanging information between the models. We call this method **multi-modelling**.

We distinguish between two kinds of multi-models: multi-scale models, and multi-method models.

Multi-scale models

These are multi-models in which at least one pair of interacting systems operate in different spatio-temporal scales (*e.g.* a cell model integrated with a tissue model) or ontologies *(e.g.* a protein model integrated with a metabolic model). While the models need not be spatial themselves, the interaction crosses scales.

Multi-method models

These are multi-models in which at least one pair of interacting systems are simulated using different representations. Many representations are agnostic to scale. The representation method determines how modelled systems are simulated, visualised (or otherwise explained) and how the model interacts with- or is compared to other models.

Typically, multi-models are either multi-scale, multi-method or both. It is possible to construct multi-models that are neither, assembling components at the same scale and using the same representation. A model of the interactions between two separately developed metabolic pathway models that use the same representation is an example of a multi-model that is neither multi-method nor multi-scale.

Model integration through spatial representation

A representation that is to interchange data between models that vary in method and scale will need one or more mathematical entities to be present in both models. Since spatial relationships are very common in biomedical modelling across all scales, we created a spatio-temporal representation language called the Field Representation Language (FRL)[21] to facilitate interchange of spatial information between different models.

FRL is a representation of algebraic fields which are constructs that vary over a domain defined in space and/or time. The language uses analytic expressions in predefined spatio-temporal coordinates (*e.g.* X, Y, Z) to define fields analytically, represent new interpolation methods and define field domains. FRL supports fields that are composed of one or more other fields. The composition is defined by mathematical relationships given as analytic expressions.

To facilitate integration between different simulation environments we created the Abstract Field Layer (AFL) – a mathematical library that can read and manipulate FRL files and perform numeric interpolation, derivation and integration on them. AFL is released under the General Public Licence (GPL) and is available from http://visiome.sourceforge.net.

Results

Multi-method model example

The work presented in this section was previously published [22]. This section summarises the work for completeness. The emphasis here is on the interchange of information between models of different representations. The interchange is illustrated in Figure 1.

The multi-model described is for treatment planning of progressed liver cancer. Ferromagnetic microspheres are introduced locally and then lodge in tumour capillaries. When the patient is subjected to an alternating magnetic field the microspheres emit heat, preferentially heating and destroying tumour tissue.

We developed a three-dimensional fractal model of microvasculature within the peripheral shell of a tumour (Figure 2a). Our model construction starts with four initial cylindrical segments representing arterial vessels. These initial segments then branch into smaller daughter segments according to probabilities derived from the literature. The segments are further constrained to a spherical shell. In turn, each segment branches and the process is repeated until the segments are too thin to allow microspheres through. The resulting vasculature covers more than an octant of the tumour.

We constructed 21 different vasculature models using different random seeds. We simulated microsphere infusion into the four initial segments of each model and recorded the proportional size and location of each cluster of microspheres in the vasculature in a FRL field. One of these fields is shown in Figure 2b.

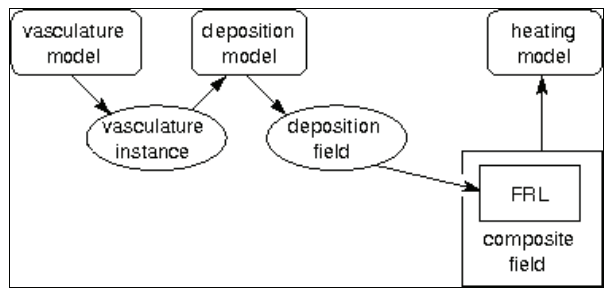

Figure 1 – Interchange between the three component models of the liver cancer treatment multi-model

Separately, we modelled an octant of the tumour with tetrahedral finite elements and assigned a volumetric heat generation rate (VHGR) to each element. The values for each element's VHGR are calculated based on the number of microspheres lodged in that element's space.

We used a composite field to scale and orientate the microsphere probabilities models so that the modelled regions were aligned. We then used AFL to compute the integral of the probabilities field over each tetrahedral element of the heating model. Listing 1 is a part of the FRL representation of the composite field that transforms the coordinate systems of the microsphere deposition field "depo5", the fifth instance of the 21 model instances generated.

The integral for each element of the composite field gives the proportion of microspheres lodged in the volume of the element and from that we computed the VHGR. The temperature field resulting from the heating simulation is shown in Figure 2c.

Listing 2 is a pseudo-code illustration of the algorithm that assigns the VHGR to each element. The integral of the field gives a number between 0 and 1 that represents the fraction of the microspheres in the entire tumour that are lodged in the volume of the current element. The VHGR is linearly proportional to that fraction.

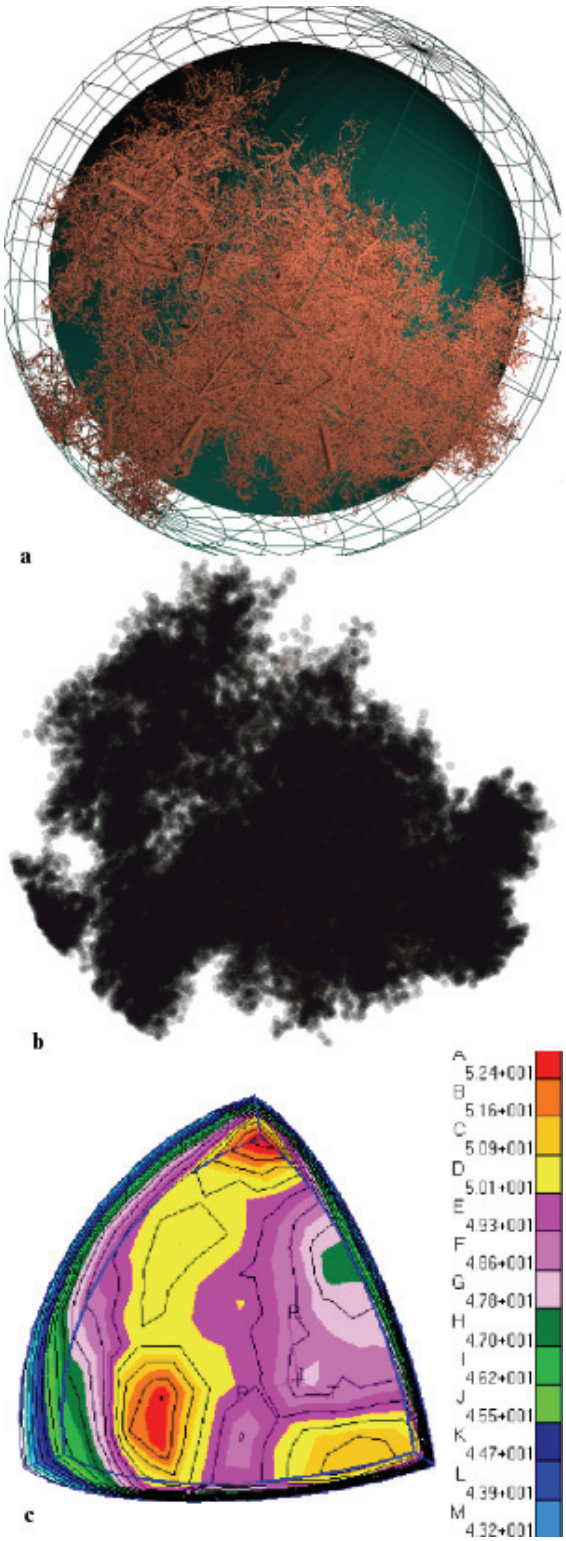

Figure 2 - Four arterial trees constrained to the spherical shell of the tumour periphery (a), their corresponding microsphere distribution field (b) and the resulting temperature field showing a heterogeneous heating pattern (c)

```
<composite name="depo5c">
  <fieldref uri="file:///depo5.xml"/>
  <function>
     depo5( -81.65*(X+Y) + 163.3*Z,
            -115.47*(X+Y+Z) + 11.25,
            -141.42*(X-Y) - 30)
  </function>
</composite>
```

Listing 1 – The composite FRL field that scales and reorientates the field "depo5" from the coordinates system of the deposition model, to that of the heating model. The coefficients were worked by hand

```
depo := load AFL field ("depo5c")
for each element e in mesh do:
    Tetrahedra t := e.getVertices()
    fraction f := depo.integral(t)
    e.assignVHGR(totalVHGR * f)
```

Listing 2 – a pseudo-code of the algorithm that assigns the VHGR to each element in the heating model based on information obtained from the composite field shown in Listing 1

Multi-scale model example

The work summarised in this section was previously presented in [23]. We created a 2D finite-difference model of activation propagation in a rabbit sinoatrial node (SAN). The model was composed of two models of specialised SAN myocytes and 231 spatio-temporal models of the extra-cellular environment (which can be thought of as the extra cellular fluid or ECF) of the tissue (*e.g.* a field of membrane potential). The myocyte models were developed separately [24] as a series of ODEs. The environment models were represented as FRL fields, of which 29 change over time as part of the interaction between individual myocytes. Individual cells interact with each other through a feedback mechanism embedded in their models. The feedback process is facilitated by changes to the ECF. For example, if one cell emits Ca^{2+} ions to the ECF, the Ca^{2+} concentration level around it goes up and this affects nearby cells.

Figure 3 illustrates the interaction between three myocytes and the environment. At each iteration T of the simulation, each myocyte reads the membrane potential's spatial derivative from the ECF at the end of T-1 (B) uses an ODE solver to compute the state of the cell at T, and updates the ECF (A). Listing 3 shows the algorithm that controls this process.

The result is a tissue model in which activation of individual myocytes was synchronised through the extra-cellular environment (Figure 4).

Discussion and conclusion

Coming up with new hypotheses is still normally done manually and is often called "model identification", "system identification" or the "inverse" problem. Existing hypothesis generation algorithms search through a space of unknown variables and identify points in the space that are consistent with the known data and first-principle rules.

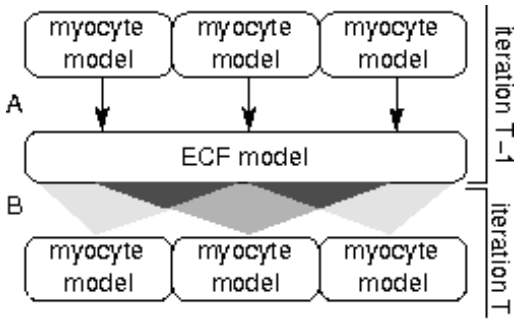

Figure 3 – Interactions in the SAN multi-model. The values set by each arrow in A is at a particular point in the ECF but affects larger regions (in grey) in B of the next iteration

```
ECF_T := set of AFL fields
ECF_T-1 := set of AFL fields
for each iteration T do:
    ECF_T-1 := ECF_T
    clear(ECF_T)
    for each myocyte m do:
        m.read(ECF_T-1)
        m.solveToNextIteration()
        m.update(ECF_T)
```

Listing 3 – Pseudo code of the algorithm that causes the cells to interact and thus drives the simulation

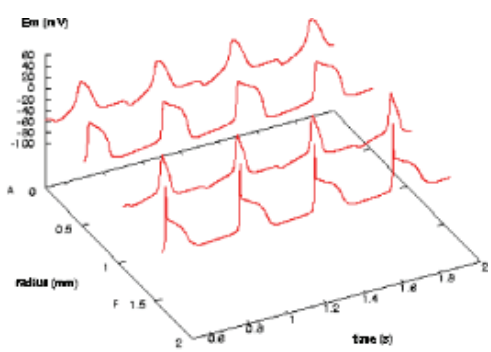

Figure 4 – A plot of membrane potential (Em) at 4 points along a radial slice of the membrane potential field in the SAN tissue model. The regular activation pattern indicates that the cells are interacting through their environment

To be able to search through knowledge spaces of multiple scales and using multiple methods, model-generating algorithms would need to be able to consider several representation languages, use a variety of modelling methods and present complex models in a human-readable fashion [25].

In addition to spatio-temporal interchange we argue for a **logical** setting for model integration in hypothesis generation and propose a machine learning approach based on inductive logic programming (ILP) [26]. Consider, a multi-model composed of three

for example, the problem of discovering metabolic networks in eukaryotic cells. Normally this process relies on quantitative methods such as flux analysis [27]. King *et al.* [17] demonstrated for the first time a successful ILP multi-model that integrated metabolic pathways from KEGG [28] with data from auxotrophic growth experiments in the budding yeast *S. cerevisiae.* Hypothesis generation in this case was the search for new experiments to maximize the refutation rate of potential genetic regulatory interactions in a multi-stage process, at the end of which a robot scientist had converged on a network to explain the data.

In a hypothetical model of pheromone response: the hypothesis space would contain a model of the cell membrane sensors, a protein interaction pathway, a transcriptional regulatory network and a set of metabolic pathways. These different types of models all have to be integrated to generate a hypothesis. ILP is the *only* machine learning approach that can integrate arbitrary background knowledge and hence different model types. A logical interface in AFL could facilitate this integration.

Standard representation languages are essential for interchange between models developed independently. However, machine learning algorithms that can reason in multiple scales and interact with other algorithms require a more profound solution, of which representation is only a part.

The current research in the area of multi-modelling is predominantly limited to standardisation of representation languages. Among those, quantitative languages seem to be leading the charge [21,29,30]. However, as of yet, there are no standard representation languages that are recognised by *any* international standards organisation.

In contrast with the current model identification algorithms which are constrained to a particular modelling method, multi-method hypothesis generation algorithms should be capable of choosing appropriate modelling methods. For this task, ontology of the available methods is required [4].

Given this ontology, a calculus of model interactions based on their model type would allow intelligent algorithms to decide which methods to use and how to construct the interactions between the modelled systems. Finally, with an ontology and a calculus, new machine learning techniques need to be formulated that are able to deduce knowledge in multiple representations, using multiple methods and support a variety of inter-model interaction models.

High-throughput biological methods necessitate high-throughput analysis methods. Computer aided knowledge discovery is an emerging field that shows promise in interpreting information from large volumes of complex data. Computational multi-modelling techniques are relatively new techniques that make complex models wieldy. As theories become more detailed, supported by more data and harder to formulate, an amalgamation of computer aided knowledge discovery and computational multi-modelling will become necessary.

In this paper we presented two examples of manually constructed multi-models and referred to several discoveries in the life-sciences that were directly aided by computer generated hypotheses. Finally, we identified several enabling areas of computer generated multi-models that are still deficient.

Acknowledgments

This work is supported by a NSW Health Capacity Building Infrastructure Grant.

References

[1] Smalheiser NR. Informatics and hypothesis-driven research. EMBO Reports 2002;3(8):702.
[2] Tomita M. Whole-cell simulation: a grand challenge of the 21st century. Trends in Biotechnology 2001;19(6):205-210.
[3] Bassingthwaighte JB. Strategies for the physiome project. Annals of Biomedical Eng. 2000;28(8):1043-1058.
[4] McCulloch AD, Huber G. Integrative biological modelling in silico. Novartis Found Symp 2002;247:4-19; discussion20-5, 84-90, 244-52.
[5] Rubin DL, Grossman D, Neal M, Cook DL, Bassingthwaighte JB, Musen MA. Ontology-based representation of simulation models of physiology. In: AMIA 2006 Symposium Proceedings; 2006. p. 664-668.
[6] Langley P. The computational support of scientific discovery. Int'l J. of Human-Comp. Studies 2000;53(3):393-410.
[7] King RD, Garrett SM, Coghill GM. On the use of qualitative reasoning to simulate and identify metabolic pathways. Bioinformatics 2005;21(9):2017-2026.
[8] Alon U, Surette MG, Barkai N, Leibler S. Robustness in bacterial chemotaxis. Nature 1999;397:169.
[9] Fernandez JW, Mithraratne P, Thrupp SF, Tawhai MH, Hunter PJ. Anatomically based geometric modelling of the musculo-skeletal system and other organs. Biomechanics and Modeling in Mechanobiology 2004;2(3):139 - 155.
[10] Akutsu T, Miyano S, Kuhara S. Inferring qualitative relations in genetic networks and metabolic pathways. Bioinfor-matics 2000;16(8):727-734.
[11] Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol 2000;7(3-4):601-620.
[12] Regev A, Silverman W, Shapiro E. Representation and simulation of biochemical processes using the -calculus process algebra. In: Pacific Symposium on Biocomputing. vol. 6; 2001. p. 459-470.
[13] Matsuno H, Inouye SIT, Okitsu Y, Fujii Y, Miyano S. A new regulatory interaction suggested by simulations for circadian genetic control mechanism in mammals. J Bioinform Comput Biol 2006;4(1):139-153.
[14] Hau DT, Coiera EW. Learning Qualitative Models of Dynamic Systems. Machine Learning 1993;26:177-211.
[15] King RD, Muggleton SH, Srinivasan A, Sternberg MJ. Structure-activity relationships derived by machine learning:the use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. Proc Natl Acad Sci U S A 1996;93(1):438-442.
[16] Langley P. Machine learning and its applications: Advanced Lectures. vol. 2049. Springer Berlin / Heidelberg; 2001.p. 230-248.
[17] King RD, Whelan KE, Jones FM, Reiser PGK, Bryant CH, Muggleton SH, Kell, DB, Oliver, SG. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 2004;427(6971):247-252.
[18] Butte AJ, Chen R. Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. In: AMIA 2006 Symposium Proceedings; 2006. p. 106-110.
[19] Smalheiser NR. Launching the "Journal of Biomedical Discovery and Collaboration". Journal of Biomedical Discovery and Collaboration 2006;1(1).
[20] Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO. Systems approach to refining genome annotation. Proc Natl Acad Sci 2006;103(43):17480-17484.
[21] Tsafnat G. The Field Representation Language. Journal of Biomedical Informatics 2006;Submitted.
[22] Tsafnat N, Tsafnat G, Lambert TD, Jones SK. Modelling heating of liver tumours with heterogeneous magnetic microsphere deposition. IOP Journal of Physics and Biology 2005;50(12):2937-2953.
[23] Tsafnat G, Kua KG, Cloherty SL, Lambert TD. Field solver framework and reference implementation. In: IASTED Applied Simulation and Modelling; 2005.
[24] Cloherty SL, Dokos S, Lovell NH. Qualitative Support for the gradient model of cardiac pacemaker heterogeneity. In: The 27th Annual Conf. Biomed. Eng. IEEE Engineering in Medicine and Biology Society. Omnipress; 2005.
[25] Muggleton S, Michie D. Machine intelligibility and the duality principle. BT Technology Journal 1996;14:15-23.
[26] Muggleton S, de Raedt L. Inductive Logic Programming: theory and methods". J. Logic Prog. 1994;19,20:629-679.
[27] Schilling CH, Schuster S, Palsson, BO, Heinrich R. Metabolic pathway analysis: nasic concepts and scientific applications in the post-genomic era. In: Biotechnol Prog 1999;15:296-303.

[28] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. In: Nucleic Acids Res. 2000;20:27-30
[29] Hucka M, Bolouri H, Finney A, Sauro HM, Doyle JC, Kitano H, et al. The Systems Biology Markup Language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 2003;19(4):513-523.
[30] Lloyd CM, Halstead MDB, Nielsen PF. CellML: Its future, present and past. Prog Biophys Mol Biol 2004; 85(2-3):433-450.

Address for correspondence

Guy Tsafnat, Centre for Health Informatics, Coogee Campus, University of New South Wales, Sydney, NSW 2052, Australia.
guyt@unsw.edu.au