Virtual experiments for reusable models

Jonathan Cooper[1], Gary Mirams[1], Mark Slaymaker[1], Andrew Simpson[1], Jon Olav Vik[2], Dagmar Waltemath[3]

[1]Dept. of Computer Science, University of Oxford [2]Centre for Integrative Genetics, Norwegian University of Life Sciences [3]Dept. of Systems Biology and Bioinformatics, University of Rostock

Correspondence: jonathan.cooper@cs.ox.ac.uk, Wolfson Building, Parks Rd, Oxford, OX1 3QD, UK

Introduction

One of the key issues in computational biology is the way we integrate models and data, e.g. in model fitting, validation, or selection. Current approaches are typically ad-hoc and disconnected, and a more formal, integrated approach is urgently needed. A quantitative model should provide an unambiguous and testable description of a proposed mechanism. However, today the results obtained from model simulation and analysis are often not reproducible, and the models therefore are hard to re-use. Tasks such as comparing different hypotheses against experimental data, determining a model's suitability or limitations for a particular study, or incremental development of models, are still challenging and often performed inadequately. Various community standards for representing models themselves exist [e.g. 1,2] and so the models can be exchanged. However more information is needed. *Virtual experiments* are required in order to simulate in the models precisely the same protocols employed in generating the experimental data used to develop or test the models. Furthermore these protocol descriptions must also be sharable in standard formats (e.g. building on SED-ML [3]) with the models, in order to achieve effective re-use.

We argue here that building up a repository of models and protocols, curated and linked with the corresponding experimental data, is fundamental to improving both the quality of our models, and the ease with which they can be re-used. *Functional curation* [4] demands that when computational models are being developed and curated the primary goal should be the continuous evaluation of model predictions against experimental data; this is also envisioned by the model development framework in the VPH-FET roadmap [5]. As discussed there, both technical and societal challenges to achieving such a vision exist. Given the broad range of models and experiments studied in the life sciences, developing standard formats to represent them is a significant and on-going task. Much effort is then needed to develop the tool infrastructure around such standards. Perhaps even more challenging however is achieving community uptake. It requires significant progress on usability of solutions, including good documentation and training materials. It also requires that solutions provide added value to researchers, with demonstrator projects by "eager adopters" manifesting the potential impact.

This paper describes our approach to achieving the above-mentioned goals, a selection of the progress made thus far in conception and implementation, and some future directions.

Use-cases

Development of both software tools and standard formats needs to be driven by specific scientific applications if the results are to be pragmatic, useful, usable, and hence taken up by the community. The benefits need to be demonstrated to potential users through concrete use-cases. Our investigation of the requirements for virtual experiments, and our tool developments, are therefore grounded in several application areas.

Our primary application area, also considered in [4], is cardiac electrophysiology. It has several features making it well suited to our purposes. Firstly it is a well-developed field, with a large variety of models, built using the same paradigm, and representing essentially the same system, which may be compared fruitfully. This variety is also expressed through variations in modelling convention (e.g. units, positive directions of flow) [6], which provide challenges for applying a single experiment to multiple models. The post-processing required in typical electrophysiology experiments is also often complex, yielding strong requirements for language design.

Other areas are now starting to be considered, in order to ensure a wider applicability for our approach. These include the cell-cycle, immunology, synthetic biology, visual psychophysics and neuro coding. A particularly interesting case is discrete cell-based modelling within Chaste [7], where the model is encoded by an executable program, rather than in a markup language. This provides additional challenges in interfacing, but

may yield a useful approach to bridging with legacy or unusual models.

Describing virtual experiments

In defining virtual experiments there is a balance to be struck between a standardised language that is reasonably concise, and hence providing support in many tools is not too difficult, and allowing flexibility for researchers to represent new and varied kinds of experiments. The benefits of standard formats for exchange have been discussed repeatedly [e.g. 8], but there is often also an overhead associated with their use [5]. The major consequence that we see is that in order for standards to be taken up by the end users, use of standards must be made easy. In addressing this problem we consider the following aspects.

Firstly, through examining the kinds of experiments required by our scientific applications, we are determining the minimal set of semantic constructs required in a "protocol language" that still allows the largest possible set of common experiments to be encoded. Note that we are not seeking to encode every possible experiment – unusual or especially complex cases may well be better expressed using general purpose programming languages and/or workflow systems.

Secondly, we argue that there is great value in the protocol language supporting the definition of common generic components that may be parameterised, and hence instantiated for specific scenarios. A library of such components may then be built up, facilitating the creation of new experiment descriptions. For example, a common experiment type in cardiac electrophysiology is the voltage clamp, where a potential is applied to the cell membrane, and the current response analysed. This generic protocol is used with different transmembrane currents and different applied voltage traces – these would become inputs to a parameterised protocol. Any voltage clamp experiment could then be specified quickly and easily.

Related to this, we have recently added the concept of *nested protocols* to the prototype described in [4]. In a nested protocol, one protocol may reference another as though it were a model, wrapping model preprocessing, simulation, and post-processing within an outer experiment.[1] This supports uses such as a protocol performing a single pace of a myocyte being embedded within a dynamic steady-state simulation – the single pace is performed repeatedly until some post-processed quantities converge.

As the SED-ML standard [3] being developed in the systems biology community already offers a frame for the exchange of simulation setups, we are investigating to what extent SED-ML can already support the use-cases we identify, and where it cannot we will submit extension proposals for review by the community.

Supporting tools and infrastructure

Standardised descriptions of virtual experiments are not sufficient in themselves. If there is to be widespread uptake of this approach, these standards need to be embedded within usable tools to provide added benefit to modellers. Building on our prototype tools for executing experiments [4] we are investigating several options.

The experiment descriptions have to be maintained and made accessible to end-users. We are therefore linking our work to on-going research on Simulation Experiment Management Systems (SEMS[2]). We provide our experiment descriptions to the management system, making them thereby searchable and comparable with other existing experimental setups.

To provide a link with experimental data, we are using the middleware framework *sif* (service-oriented interoperability framework) [10]. This was designed to facilitate the sharing and aggregation of data from distributed, heterogeneous data sources, and developed originally to support healthcare applications (see, e.g., [11]); in recent years, the focus has turned to Systems Biology applications [12]. The *sif* framework is currently being used to manage the distributed execution of experiments, and to curate these along with the associated data (both experimental and simulated) and models from multiple sources.

Future work will also look at developing plugins for the OpenCOR modelling environment[3] to allow design-

[1] This is a distinct concept from nested simulations [9], in which only the simulation phase may be nested.
[2] http://www.sbi.uni-rostock.de/research/research-projects/single/33/
[3] http://www.opencor.ws/

ing and running virtual experiments as an integral part of model development.

Discussion and conclusions

The description of virtual experiments in standard formats is essential to future progress in VPH research. **Only sharing the experimental setups along with the models themselves makes the modelling work truly useful to others, and provides the crucial link to data.** The acceptance and use of realistic and validated models will be increased, allowing researchers from different disciplines to share resources and develop new knowledge.

The proposed framework will also enable a richer characterization of the behavioural repertoire of models. By automatically and comprehensively testing multiple models and protocols users can have confidence that the model they have chosen or developed provides a good approximation of the desired physiology. Model comparison under multiple protocols allows the impact of changes in parameter values or model structure to be ascertained in greater detail. Furthermore, the model development process will be facilitated. Generic experiment descriptions will allow faster setup of simulation experiments. Continually appraising models against collections of protocols and desired outputs will ensure that desired functionality is not lost. As models, protocols, and data are curated together in open repositories, the store of global knowledge and understanding is increased.

We plan further to refine, enhance, and implement the ideas introduced in this manuscript through a collaborative EU project.

References

1. Hucka, M. *et al,* 2004 Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project. *IEE Syst Biol* **1**, 41–53.

2. Lloyd, C. *et al.* 2004 CellML: its future, present and past. *Prog Biophys Mol Biol* **85**, 433-450.

3. Waltemath, D. *et al*. 2011 Reproducible computational biology experiments with SED-ML - The Simulation Experiment Description Markup Language. *BMC Syst Biol* **5**, 198. (DOI 10.1186/1752-0509-5-198.)

4. Cooper, J., Mirams, G. & Niederer, S., 2011 High throughput functional curation of cellular electrophysiology models. *Prog Biophys Mol Biol* **107**, 11-20. (DOI 10.1016/j.pbiomolbio.2011.06.003.)

5. VPH-FET Consortium, 2011 In *VPH-FET Research Roadmap – Advanced Technologies for the Future of the Virtual Physiological Human*, pp. 40-49 [Online] https://www.biomedtown.org/biomed_town/VPHFET/reception/vphfetpublicrep/plfng_view

6. Cooper, J. *et al.* 2011 Considerations for the use of cellular electrophysiology models within cardiac tissue simulations. *Prog Biophys Mol Biol* **107**, 74-80. (DOI 10.1016/j.pbiomolbio.2011.06.002.)

7. Pitt-Francis, J. *et al*. 2009 Chaste: a test-driven approach to software development for biological modelling. *Comput Phys Comm* **180**, 2452-2471. (DOI 10.1016/j.cpc.2009.07.019.)

8. Cooper, J. *et al*. 2010 The Virtual Physiological Human ToolKit. *Phil Trans R Soc A* **368**, 3925−3936. (DOI 10.1098/rsta.2010.0144.)

9. Bergmann, F. 2012 SED-ML: Nested Simulation Proposal. *Nature Precedings*. (DOI 10.1038/npre.2012.4257.2.)

10. Simpson, A. *et al*. 2008 A healthcare-driven framework for facilitating the secure sharing of data across organisational boundaries. *Stud Health Tech Inform* **138**, 3-12.

11. Simpson, A. *et al*. 2010 GIMI: the past, the present, and the future. *Phil Trans R Soc A* **368**, 3891-3905.

12. Simpson, A. *et al*. 2010 On the secure sharing and aggregation of data to support systems biology research. In *Proc7th International Conference on Data Integration in the Life Sciences,* pp. 58-73.