A Semantic Framework for Systems Biology Modeling: Integrative Bioinformatics for Data, Models and Experimental Evidence

Synopsis

The amount of health care and life sciences data available on the web has been growing at an almost exponential rate. Simultaneously, increasing demands for personalized medicine and pharmacokinetics has led to the availability of a large number of bio-models, which are often challenging to integrate due to inconsistencies in the base assumptions. This project will focus on creating an integrative bioinformatics framework for the semi-automated integration and evaluation of bio-models, by ranking biological interaction assertions based on experimental evidences and keeping track of provenance of relations used in the models. The novelty in this approach will be the use of linked data and semantic web technologies for data integration, continuous model evaluation and usage of logic frameworks for checking model consistency. The framework will be evaluated with a model of the MAPK-ERK signaling pathway, a biological mechanism highly relevant in cancer research.

Background

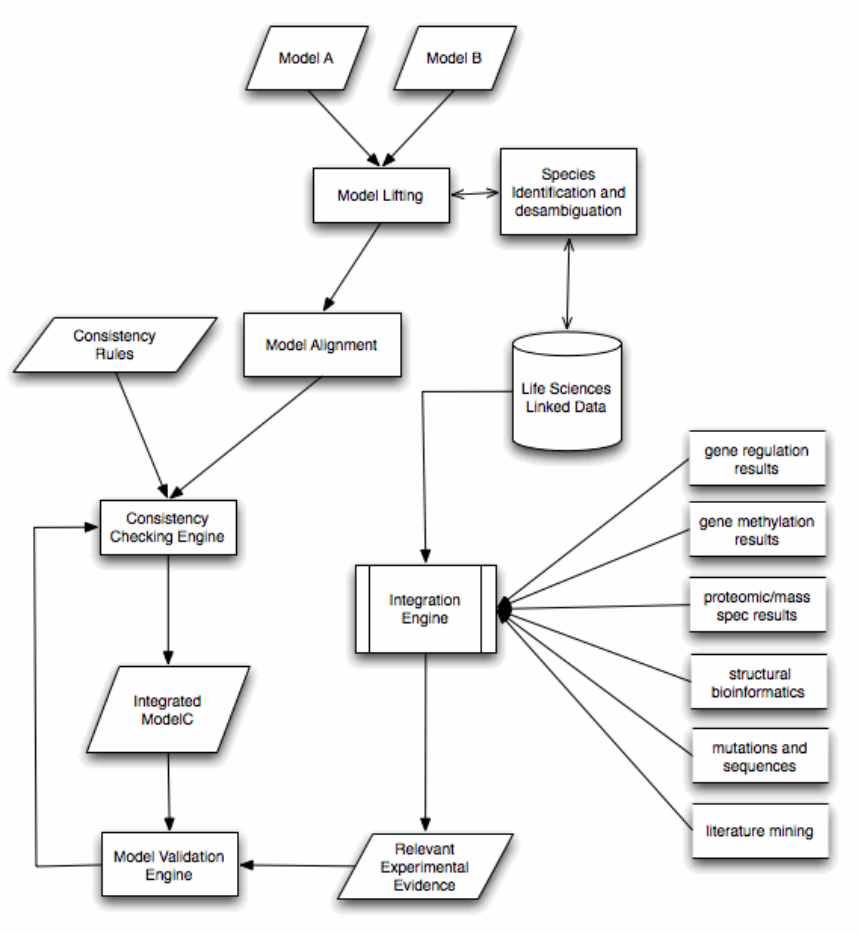
Modelling biological networks has several applications, from pharmacogenomics [1] to personalized medicine [2]. Biological models are devised for gene regulatory or metabolic networks, signal transduction and protein-protein interaction. The amount and quality of integrated information used as input affects heavily the accuracy and predictive power of the models. The traditional approach to modelling biological networks has been to split complex systems into individual parts which are assembled by gathering and manually integrating information from the literature. However, because this process does not take into account experimental data and provenance of information, models are often built on incorrect assumptions which may have to be revised when the state of the art changes [3]. As recently reported [4], even high profile publications can be dubious if provenance of raw experimental data and bioinformatics analysis cannot be traced back to the scientific conclusions used in the models. The time consuming and error prone process of modelling biological networks can be greatly improved by using the deluge of data from “omics” technologies that has been made available in the past decade. In order to use “omics” experimental data for modelling, the first critical step is to dynamically integrate it with existing knowledge [5,6]. This is still a problem since most datasets don’t share a common representation model. Semantic web and linked data technologies constitute a new paradigm in data representation which have matured to a point of providing a scalable solution for integrative bioinformatics [7–9], even when data is highly heterogeneous. Simultaneously, the publication of executable bioinformatics workflows for processing “omics” datasets is becoming increasingly relevant as a means to automatically process raw data into biological interaction assertions [10]. One final component core for enabling the integration and validation of biological models is the availability of logic frameworks such as answer set programming that can be applied for checking consistency based on a set of rules [11]. In this project, a framework will be devised

for semi-automatic assembly of models of biological networks by making use of raw experimental “omics” data made available by the cancer genome atlas project [5]. The framework will be applied towards the integration and validation of models of the ERK-MAPK signalling pathway, a process that is highly relevant in cancer progressing as it is involved in triggering apoptosis of tumorigenic cells[12].

Research plan method

In this task force, we will focus on the development/identification of the requirements for a framework for integration, validation and evaluation of models of biological networks. A possible architecture for the framework to validate and evaluate biological network models is presented in Fig 1. This architecture will be serve as as starting point to discuss methods and strategies.

Models will first be integrated using semantic lifting, which consists of abstracting the details of each model in order to identify common elements (e.g. common molecular species which may be named differently). Given that multiple models may use different terminologies to describe the same molecular species, a process of species identification and disambiguation will be applied. Species disambiguation will be enabled by using publicly available datasets in the linked open data (LOD) cloud as a lookup method. After model lifting, the multiple models will be aligned to ensure that the biological processes described are matched in all models, and a process of consistency checking will take place. Consistency checking will rely on verifying if both models respect a set of well known predefined biochemical and biological rules (e.g. two proteins need to be in the same cellular compartment in order to physically interact). Included also in the consistency rules are sets of known interactions, obtained from experimental evidence, which have been extensively tested and validated.



The process of consistency checking will produce a single integrated model, which will then be validated through the model validation engine. Model validation can rely on previous work for integrating multiple experimental methods (e.g. gene regulation, methylation, proteomics, etc) and using the LOD cloud as a lookup mechanism. Relevant experimental evidence collected through the integration engine will be used to validate the integrated model and, if applicable, to predict new interactions to the model. If new assertions are created, consistency checking will be applied to verify if the model remains valid.

For evaluating the platform, we will apply it in the development and validation of a model representing the MAPK-ERK signalling pathway. Models relevant for this pathway will be collected from the BioModels database [1] and experimental evidence will be collected from the public cancer genome atlas repository[2] and the international genomics consortium [3]. For model consistency checking, data from the curated databases such as the STRING database [4] will be used. These public datasets

[1] http://www.ebi.ac.uk/biomodels-main/
[2] https://tcga-data.nci.nih.gov/tcga/
[3] http://www.intgen.org/
[4] http://string-db.org/ua

will be exposed through a SPARQL endpoint prior to application in the integration engine using the methodology described in [13]. Finally, a set of semantic web services will be developed to ensure that the necessary statistical and data processing operations are executed prior to integration.

Expected outcomes / Impact

The primary outcome of the proposed project is a detailed and dynamic model of how the ERK-MAPK signaling pathway modulates apoptosis in tumor cells that will be continuously improved and validated against new experimental evidence. Such a model will have multiple applications at ongoing research in Life Sciences institutes, including in the study of cancer genetics as well as drug discovery since it will enable the testing of hypothesis *in silico* prior to its attempt *in vitro*.

The methodology used in the modelling process will make extensive use of cutting edge integrative bioinformatics techniques including semantic web and linked data technologies, which are the focus of the Task Force. As such, a secondary outcome of this work will be the availability of a framework for semi-automatic model assembly and validation. This framework will be applicable for simulation of multiple models of interacting biological entities and its continuous evaluation.

The resulting simulated models will be applicable in several domains including clinical decision support, pharmacogenomics and translational research. Semantic web technologies have only recently reached a maturation state and as such have rarely been applied in the simulation of biological networks. As such, we expect that one of the side effect of developing such a framework will be a significant improvement of existing semantic web technologies and the development of new ones, thus contributing to the advancement of the field of computer science in addition to that of biomedical research both important areas of active research. Finally, a long-term outcome of applying the proposed framework in modeling several aspects of the systemic biology puzzle will be the identification of methodologies and standards for experimental measurement and reporting that assure highest quality data. As an example, if a certain assertion in the model (e.g. A inhibits B) has been selected as highly likely to occur through several validation iterations, experimental evidence that contradicts such assertion may in fact be an artifact and indicate low quality in the experimental results. Multiple biomedical research institutions can directly benefit from these results as they will enable saving time and reducing costs in acquiring high quality data for biomedical research and simulation.

Given the novelty and applicability of the proposed approach towards improving biomedical discovery via semi-automated integration and validation of simulation models, each of the challenges addressed can potentially result in quality publications in high impact journals. Furthermore, this approach can also result in the first continually validated model of the ERK-MAPK signalling pathway, potentially enabling the translation of experimentally acquired knowledge into medical applications.

Major references

[1] W.E. Evans, Science 286 (1999) 487-491.

[2] G.S. Ginsburg, J.J. McCarthy, Trends in Biotechnology 19 (2001) 491-6.

[3] J.S. Almeida, C. Chen, R. Gorlitsky, R. Stanislaus, M. Aires-de-Sousa, P. Eleutério, J. Carriço, A. Maretzek, A. Bohn, A. Chang, F. Zhang, R. Mitra, G.B. Mills, X. Wang, H.F. Deus, Nature Biotechnology 24 (2006) 1070-1.

[4] S. Hutson, Nature Medicine 16 (2010) 618-618.

[5] H.F. Deus, D.F. Veiga, P.R. Freire, J.N. Weinstein, G.B. Mills, J.S. Almeida, Journal of Biomedical Informatics 43 (2010) 998-1008.

[6] H.F. Deus, E. Prud'hommeaux, M. Miller, J. Zhao, J. Malone, T. Adamusiak, J. McCusker, S. Das, P.R. Serra, R. Fox, M. Scott Marshall, Journal of Biomedical Informatics (2012).

[7] H. Chen, T. Yu, J.Y. Chen, Briefings in Bioinformatics (2012).

[8] H.F. Deus, R. Stanislaus, D.F. Veiga, C. Behrens, I.I. Wistuba, J.D. Minna, H.R. Garner, S.G. Swisher, J.A. Roth, A.M. Correa, B. Broom, K. Coombes, A. Chang, L.H. Vogel, J.S. Almeida, PloS One 3 (2008) e2946.

[9] H. Deus, M. Correa, R. Stanislaus, M. Miragaia, W. Maass, H. de Lencastre, R. Fox, J. Almeida, BMC Bioinformatics 12 (2011) 285.

[10] Y. Gil, V. Ratnakar, J. Kim, P. González-Calero, P. Groth, J. Moody, E. Deelman, IEEE Intelligent Systems 26 (2010) 62-72.

[11] M. Gebser, T. Schaub, S. Thiele, P. Veber, (2010) 36.

[12] W. Kolch, Nature Reviews Molecular Cell Biology 6 (2005) 827-837.

[13] H.F. Deus, E. Prud, J. Zhao, M.S. Marshall, M. Samwald, in:, ISWC 2010 SWPM, 2010.