Research proposal

Title:
Development and Implementation of a Data Model for Pathway Mapping

Student:
Linghao Yi

Supervisors:
Kevin Robertson (GTI),
Muriel Mewissen (GTI),
Peter Ghazal (GTI),
Douglas Armstrong (Informatics)
1. Introduction and Background

With the availability of whole genome sequences, research attention has shifted from gene sequences and genome content to protein functions and systems biology. Systems biology is a rapidly evolving field and is concerned with the study of the integrated functions of the cell regulatory network, genes, messenger RNA, proteins, protein interactions and the measurement of their relationships. In the past, classical biology focused on the function of the genes and proteins one at a time. Systems biology is an attempt to look at all the elements of the cell and their interrelationships. It may be distinguished from genomics in that genomics alone is a single dimension of information; instead we can look at many dimensions of biological information.

A systems level approach is a long term objective which will help in our understanding of biology and ultimately provide powerful approaches to drug discovery. As is known, the more different types of data can be integrated together, the deeper the insights are into the biology of the system. With systems biology we can bring together biological information in order to start a new exploration. It has the promise to help us to understand the causes of genetic diseases and will aid in the testing of new drugs. Diseases arise either as a consequence of gene defects or a combination of gene defects and/or pathologic or environmental cues. Systems biology will be crucial to our understanding of illness and will enable us to understand how diseases alter biological processes and circumvent feedback control. It will also enhance robust drug development and will aid in the evolution of personalized medicine.

21st century biology will gradually move from studying individual cellular components to addressing how individual modules interact to make the cellular system function as a whole. Thus, there will be a migration towards a system-wide view of cellular function based on an understanding of the molecular mechanisms of individual processes. Therefore, how to put all knowledge together is the key to the problem. The traditional way is to describe information is by written text or by traditional cartoon-like diagrams. However, written text is inherently ambiguous, and results often have to be reinterpreted by each reader of the article. Traditional diagrams are informal, often confusing, and much information lost. Thus, it is necessary to provide a set of
notations that have powerful expression capability and are highly readable for biochemical and gene regulatory networks. Molecular interaction maps (MIM) provide a potential solution to this problem.

A molecular interaction map is a graphical representation of protein interactions, modifications and transcriptional regulation events defined in the literature and a variety of public databases. Figure 1 shows an example of a MIM.

MIM are popular for several reasons. Firstly, we can not keep so many known interactions information in our mind. Secondly, molecular interaction maps can provide us with new interpretations and directions for future exploration. Third, they provide a short-hand for recording complicated findings or hypotheses.

Several models for pathway mapping have been proposed. Kohn designed symbols and mapping conventions that allow the graphical description of protein interactions and modifications as well as some aspects of transcriptional regulation and the integration of these events into large networks. Kitano extended this model by combining it with “logic gating” and developed the integration of multiple signal pathways. These studies have taken a first step towards a systems level understanding of the biological processes inside the cell. Recent work at the Scottish Centre for Genomic Technology and Informatics (GTI) has focused on the production of a molecular interaction map describing the signal transduction pathways leading to the up-regulation of antigen presentation and the promotion of cell survival in macrophages exposed to interferon-gamma. This map has been termed a “Subcellular Logic Interaction Map”. Some languages and software also are developed, such as, SMBL, CellDesigner. However, all these do not present biological examples as we expected. Thus some further work has to be done.

2. Problems & Deliverable

Several groups are currently active in the development of pathway graphical notation and a variety of approaches have been taken to representing protein-protein interactions, gene transcription,
protein modification etc. Over the past 12 months, the GTI has developed an extended graphical notation for the representation of pathways involved in immune responses to viral infection. The GTI has also developed a Java-based drawing tool for the construction of pathway diagrams. A review of this first-generation tool has revealed several shortcomings which pose interesting problems to the researcher. Perhaps the most significant of these relates to the data-model underpinning diagrams generated by the tool. At present, the XML generated by the editor (on construction of a diagram) contains little biological meaning. As a result, the output from the tool cannot be used for simulation or modeling studies.

### 3. Research plan

The GTI has compiled and collected information and data about several hundred individual components of the interferon pathway and how these components interact. Information recorded includes:

i) A variety of Unique ID’s for a range of public databases (e.g. Unigene, OMIM).
ii) Gene and protein sequence.
iii) A Protein functional description.
iv) Information on Protein – protein interactions.
v) Information on Protein – gene interactions.

This data has been compiled from external databases, internal experiments and published articles and is stored in spreadsheet format. This is not an efficient method for data storage since spreadsheets are not easy to query or to update.

Significantly, no systematic attempt has been made to characterize and formally describe signaling events, components, interactions and modifications occurring in the interferon pathway.
The aim of this project is to research and apply methods for the production of data model for cell signaling pathways. This data model will then be implemented in the form of XML which will be used to accurately represent a small section of interferon γ-signaling pathway.

Figure 1  A sub-section of the Interferon-γ signalling molecular interaction map showing protein-protein, protein-gene interactions, translocation, phosphorylation and protein-complex formation. (A submodule of the Interferon Gamma signaling Pathway showing protein and gene interactions. Taken from 'A Subcellular Logic Interaction Map (SLIM) representation of the Interferon Gamma Signaling Pathway' produced by Frankenburg et al. Unpublished data. Copyright GTI 2005)
Figure 2. A simplified diagram of protein complex formation and transcription. It is likely that even simple interactions will require relatively complicated or detailed descriptions.

Project Implementation

The project will be implemented as follows:

Step 1: Definition of the project requirements

Interview GTI staff;
Write requirement documents that include:

i) Description of the sub-section of the γ-interferon pathway to be studied

ii) Description of data types required

iii) A formal summary of the project requirements
Step 2: Research of methods for the production of a data model

i) Will involve a general review of data modeling methods and will provide answers to the following questions:
   What is data modeling?
   Why is it important?
   How has biological systems data-modelling been undertaken in the past?
   How does data modeling apply to system biology?
   What are the features, advantages and disadvantages of UML, BioUML, SBML?

The output from this stage will be a specification document outlining and justifying the technology to be applied to the production of the data model.

Step 3: Definition of cell signaling data model

Using the technology defined in step 2 a signaling pathway data model will be produced. The data model will describe and define all possible data types, relationships, dependencies, events, and interaction between molecules in the small section of the \(\gamma\)-signal pathway.

Step 4: Application and implementation the data model

i) Production of the cell signaling markup language, which can be used to unambiguously describe cell signal events.

ii) Application of the markup language to the description of a the small section of Interferon- \(\gamma\) signaling pathway.

4. Methods and techniques

The methods and techniques to be used will be defined after review of the requirements and study of other previous projects. Data modeling is the act of exploring data-oriented structures. Modeling is an essential part of large software systems, and helpful to medium and even small project as well. There are some methods that may be used, such as UML, BioUML.
UML is a language for modeling object systems based on a unification of Booch, Rumbaugh and Jacobson’s popular object-oriented modeling methods. UML provides many of the diagrammatical modeling techniques found in most modern object methods, such as object diagrams, state diagrams and object interaction diagrams. UML helps us specify, visualize, and document models of software systems, including their structure and design, in a way that meets all of these requirements.

UML is also widely used in Biology System design because it guides more formal descriptions. PEDRo is a good application of UML. The PEDRo UML class diagram provides a conceptual model of proteomics experiment data, which form the basis for XML and relational schemas.

BioUML is designed as a common purpose framework for systems biology providing formalized graphic notation to describe the structure and function of biological systems, their visualization and simulation as well as access to databases with relevant experimental data.

XML is a Markup language much like HTML. Because of its highly flexible, Internet-oriented, rich capabilities for linking data and an open framework for defining standard specification, XML is emerging as a standard for structuring documents. Many commercial and academic actors are adopting it as a standard for their data management. XML is also used as a general framework for annotating sequence data such as, BSML and BioML.

Because XML allows the definition of a set of tags to be applied to one or many documents, and these tags generally identify different types of elements in the document, with the possibility for recursion and referencing, it is very useful for us to define protein, gene, interaction, etc.

Once the data model is produced, it will be used for create an XML document which describes a small subsection of the Interferon-gamma signaling pathway. Some consideration will also be made on how to apply the previously defined data-model to the production of a database.

5. Evaluation
The success of this project will be assessed by the application of the previously generated data-model and XML to a new subsection of the Interferon-gamma signaling pathway.

6. Results

As described earlier, the purpose of this proposed study will be the development of a new and refined data model/markup language, which will accurately capture, describe and store pathway mapping information. It is hoped this data model may be applied to the production of database and can eventually be integrated with the existing pathway drawing application to produce a more effective research tool.

7. Time Schedule

May, 18 – May, 31 interview GTI staff and write requirements
June, 1 – June, 14 do research on the methods
June, 15 – July, 19 project implementation
July, 20 – Aug., 3 project evaluation
Aug., 4 – Aug., 20 writing final report

8. Backup Plan

A significant proportion of this project will be dedicated to researching published and potentially novel methods for the representation of cellular signaling pathways. It is anticipated that at each stage of the project a range of methods and approaches will be considered and applied.

References


