The IGIPI Ontological Framework: Towards Integrating Gene Interactions with Protein Interactions

Bill Andreopoulos

Technical Report CS-2001-07

December 2001

Department of Computer Science and Engineering

4700 Keele Street North York, Ontario M3J 1P3 Canada

THE IGIPI ONTOLOGICAL FRAMEWORK: TOWARDS INTEGRATING GENE INTERACTIONS WITH PROTEIN INTERACTIONS

BILL ANDREOPOULOS
Department of Computer Science, York University
Toronto, Ontario, Canada M3J 1P3
billa@cs.yorku.ca

Combining the results of different genomic and proteomic studies poses a problem for researchers, especially when such studies produce conflicting results. The *IGIPI framework* is an ontological framework that addresses the problem of combining data from multiple genomic experiments. IGIPI views different genomic experiments as pieces of a puzzle that if positioned properly will create a more complete model of the cell. IGIPI can also be used as an *ontology-based clustering* technique, to predict gene function more effectively than traditional clustering techniques.

1. Introduction to Problem

Researchers in biological sciences often face the problem of combining the results of different genomic and proteomic studies. Development of novel bioinformatics tools for this purpose could provide researchers with many benefits, such as predicting gene function more reliably, or supporting the evolution of current knowledge by integrating it with new genomic data[5,11]. The field of software engineering provides software developers with modeling tools for solving analogous problems faced when building software[1]. We have addressed an important genomics problem by drawing a correspondence from modeling techniques used in software engineering to knowledge representation (KR) techniques used in molecular biology. The first section introduces the problem domain and explains why an analogy can be drawn between problems in the fields of software engineering and genomics. The second section describes the *IGIPI*[a] *ontological framework*, the solution we propose for KR problems in molecular biology. The next two sections explain how IGIPI can be utilized as a clustering and function prediction tool, and describe the software we have built for these purposes. We conclude by describing future work to be done.

[a] IGIPI stands for "Integrating Gene Interactions with Protein Interactions".

1.1 Motivation

We would like to think of the terms "protein function" and "gene function" as referring to similar concepts, since genes encode proteins in the first place[11]. Unfortunately, reality becomes complicated by what happens at the higher cellular level of proteins. For instance, protein interactions produced from two-hybrid studies may often not be mapped directly to gene interactions from synthetic mutant lethality (SML) studies[b]; this adds fuzziness to predicting the gene functions[5]. Sometimes a two-hybrid study may detect a protein interaction, although an SML study fails to detect an interaction between the corresponding genes. Reasons may include:

- **Suppressor mutation**: A mutation in one gene may restore (partially or fully) the function impaired by a mutation in a different gene, or at a different site in the same gene[11].
- **Nonallelic noncomplementation**: Mutations in two genes may fail to complement, because the gene products are subunits of the same multi-protein complex[11].
- **Conditional-lethal mutation**: Gene mutations may result in lethality under one environmental condition (e.g., high temperature) but not under another condition (e.g., lower temperature)[11].

Alternatively, if two genes exhibit synthetic lethality, this may not necessarily mean that their proteins also interact (and thus the genes may not have the same function). A reason for this discrepancy could be that the gene mutations affect two different protein pathways, which perform different functions but lead to death when combined[11].

Thus, it is necessary to create a complete picture of the cell, by combining the results of different genomic and proteomic studies. We have to combine the protein interactions derived from two-hybrid studies with the gene interactions from SML studies[11]. Furthermore, it is necessary to represent the experimental conditions under which any type of interaction was observed[5]. For this purpose, we propose building an *ontology*[c] , showing the experimental conditions under which the protein interactions can be mapped to the gene interactions[2,6,7,8,9]. By mapping the phenomena occurring at the higher cellular level of protein interactions to the SML data, we could assess the biological meaning of the observed synthetic lethalities with greater confidence[7,8,9]. Then we could draw more reliable conclusions about the gene functions.

We have addressed the above challenges by creating the *IGIPI ontological framework*; IGIPI stands for "Integrating Gene Interactions with Protein Interactions". Representing information derived from multiple genomic experimental studies requires solving the following KR problems:

[b] The purpose of SML studies is to identify interactions between genes in the genome[11].
[c] An ontology can be defined as a hierarchical taxonomy of the concepts in our domain[6].

1) Ability to represent time[7].
2) Ability to represent the fact that some genes may repress or affect negatively a biological function, while simultaneously inducing other biological functions.
3) Ability to represent all ways in which biological functions are manifested, including the specific group (module) of genes involved in each manifestation.
4) Ability to represent the biochemical processes responsible for a change in the module of genes inducing a biological function (e.g., by attracting more genes to join the currently active module or repelling other genes from the module[7]).

1.2 Background: Functional and Non-Functional Requirements in Software Engineering and Previous Bio-Ontologies

The IGIPI framework presented in this paper has its roots in a technique called the NFR framework, which was developed by Lawrence Chung et al. at the University of Toronto[1]. The NFR framework was originally designed for software engineers for the purpose of modeling *non-functional requirements* (NFRs) in software design[d].

Since the ideas for developing the IGIPI framework originated in the field of software engineering, our ontological framework can be viewed as creating a correspondence from software engineering to biology in the following two ways:
1) By drawing an analogy from *functional requirements*[e] imposed on software systems, to the *biological functions* of protein interactions. In our framework a biological function usually refers to a specific protein interaction, although it could also refer to a general function occurring by means of multiple pathways and complexes[5]. Our model makes the assumption that an experimenter's ultimate goal is to observe the biological function by any means available.
2) By drawing an analogy from *non-functional requirements* (NFRs) imposed on software systems, to the *means* by which biological functions can be observed experimentally. Given a specific biological function, there may exist many possible methods by which an experimenter can observe the function to occur in a genomic experiment[11]. Our framework offers abstractions so that each such method can be represented clearly and unambiguously. Two substantially different methods by which a biological function may be observed in an experiment would be represented as two different NFRs in our model.

Previous ontologies developed for the biological domain have focused mainly on gene function, by modeling how different gene functions relate to each other, as well as the

[d] A non-functional requirement (NFR) in software specifies a constraint on the design of a software system, by describing *not what* functions the software must perform but *how* the software must perform its functions.
[e] A functional requirement in software specifies *what* functions the software must perform.

constraints placed on gene functions by their surrounding environment[2]. Examples of projects developing ontologies for functional classification purposes are geneontology.org[2], EcoCyc, MIPS, YPD, KEGG and WIT[9]. Other ontologies have been defined for the purpose of modeling general concepts in bioinformatics (such as "gene" and "protein") and generic terms describing biological objects. Examples of ontologies proposed for this purpose are TAMBIS[8] and the OMB[13]. All these ontologies offer the benefit of defining semantics for biological concepts used with identical labels but different meanings across databases. Then different databases can be integrated, on the basis of the semantic repositories provided by these ontologies[6,13].

Previous functional classification ontologies assist in annotation of gene functions, a practice that usually involves depositing data in databases while publishing the experimental methods and results. Our ontology advances upon previous bio-ontologies by integrating experimental procedures in the functional annotation process; information on the experimental means can be described formally, so that knowledge can evolve as new experiments produce different and/or contradictory results. We were given many ideas for our work by the publications of Hafner and Fridman[7], who examined problems concerning representing knowledge on complex biochemical substances and transformations of such substances into different forms. Our method for representing transformations of biochemical substances (see Section 2.2) addresses the KR problems described by Hafner and Fridman[7], by allowing the modeling of relationships between inputs and outputs of a transformation, as well as how the semantic category of the inputs changes after a transformation occurs.

2. Description of the IGIPI Framework

The IGIPI framework is an ontological framework used for combining data produced by multiple genomic experiments on a specific biological function. For the interconnection of data from multiple experiments, an experimenter's aim is not to represent the biological functions themselves, since all functions occur at some point of time in a cell, under different experimental conditions or environmental stimuli[2]. The IGIPI framework is rather used to represent knowledge about the various *means* by which a biological function can be observed to occur in an experiment. If a function can be observed by means of various experimental methods (e.g., expression studies or two-hybrid studies) then an experimenter's goal should be to model the conditions (environmental or experimental) which distinguish the results of one method from another. This way, IGIPI allows an experimenter to interconnect the results from different biological experiments. Subsequently, this permits more reliable interpretation of genomic data and supports the

evolution of current biological knowledge, by allowing its easy integration with new data[3,6].

IGIPI offers semantic modeling abstractions for modeling the conditions that may lead to different experimental techniques producing different results. These conditions usually boil down to biological processes occurring in the interval between the protein interactions at a high cellular level and the genotypic behaviors observed at a low level. In this Section we describe the abstractions offered by the IGIPI framework, which address some of the unique challenges presented when interconnecting data from different genomic experiments. These modeling abstractions also provide a device for analyzing similarity between genes, as discussed in Section 3.

2.1 Timegoals: NFRs and Gene Observations

The IGIPI framework represents requirements on an experiment as *timegoals*. A timegoal is a goal that needs to be satisfied at a specific point of time in an experiment, in order for a biological function to be observed (e.g., a network of protein interactions). Timegoals are goals with no clear-cut criterion for their fulfilment. Instead, a timegoal may only contribute positively or negatively towards achieving another timegoal. By using this logic, a timegoal can be *satisficed* or not. In the IGIPI framework, *satisficing* refers to satisfying at some level a goal or a need, but without necessarily producing the optimal solution.

The IGIPI framework represents information about timegoals using a graphical representation called the *timegoal interdependency graph,* or *TIG*. An example of a TIG is given in Figure 1. A TIG records all timegoals being considered and the interdependencies between them. Each timegoal in a TIG is represented as an individual node and edges correspond to the interdependencies between them.

In a TIG each timegoal is represented as an individual node (or cloud). The IGIPI framework supports two types of timegoals: *NFRs* (high-level goals) and *gene observations* (low-level goals). The term NFR is derived from the software engineering term "non-functional requirement"; in our context NFR refers to a high level requirement placed on a biological experiment, without stating anything about the precise means by which this requirement will be achieved in the experiment. A developer can construct an initial TIG by identifying the top-level NFR that is expected to be observed and sketching a timegoal for it. Figure 1 shows observing the "yeast's adaptation to a heat shock" in an experiment as a root NFR timegoal at the top of the graph. The TIG provides a hierarchical arrangement of all the different timegoals; a general parent timegoal can be decomposed into more specific offspring timegoals at lower levels. To represent the requirements that would need to be satisfied for the "yeast's adaptation to a heat shock" to be observed

experimentally, the root NFR timegoal is decomposed into the NFR timegoals "gene expression study", "two-hybrid study" and "synthetic mutant lethality study". This means that performing any of these studies may lead to observing the yeast's adaptation to a heat shock. It is important to note that none of these NFRs make any explicit statements about the precise means by which the specific function could be achieved at a genomic level.

Timegoals are connected by interdependency links, which show *decompositions* of parent timegoals downwards into more specific offspring timegoals. In some cases the interdendency links are grouped together with an arc; this is referred to as an *AND* contribution of the offspring timegoals towards their parent timegoal, and means that both offspring timegoals must be satisficed to satisfice the parent. In other cases the interdendency links are grouped together with a double arc; this is referred to as an *OR* contribution of the offspring timegoals towards their parent timegoal and means that only one offspring timegoal needs to be satisficed to satisfice the parent. Figure 1 shows that either timegoal for one of three studies must be satisficed to satisfice the "yeast adaptation to a heat shock" timegoal.

The bottom of a TIG consists of the *gene observations* that represent requirements on the method(s) that need to be implemented at a low experimental level, to achieve one or more high-level NFRs. Usually a gene observation represents a manipulation or observation concerning a gene or protein that needs to be implemented at a low experimental level. Gene observations are also considered timegoals and thus may be decomposed into more specific gene observations at a lower level. For example, Figure 1 shows a gene observation representing the general action of observing the Msn2 gene; this gene observation gets decomposed into the timegoals of overexpressing the Msn2 gene and observing the Msn2 gene at its normal expression level at a low experimental level.

Gene observations make a positive or negative contribution towards achieving one or more high level NFR timegoals. Figure 1 shows how interdependency links are used to represent a gene observation timegoal's contribution towards achieving a high level NFR timegoal; as shown, such a contribution can be positive ("+" or "++") or negative ("-" or "--"). Since an NFR may receive both positive and negative contributions from many other gene observation timegoals, it is difficult to draw a strict line between whether an NFR is achieved or not achieved. To cope with this challenge, we use the concept of an NFR that is satisficed, as described above, to represent an NFR that is satisfied to a level high enough that the person carrying out the experiment can consider the timegoal to be achieved[1].

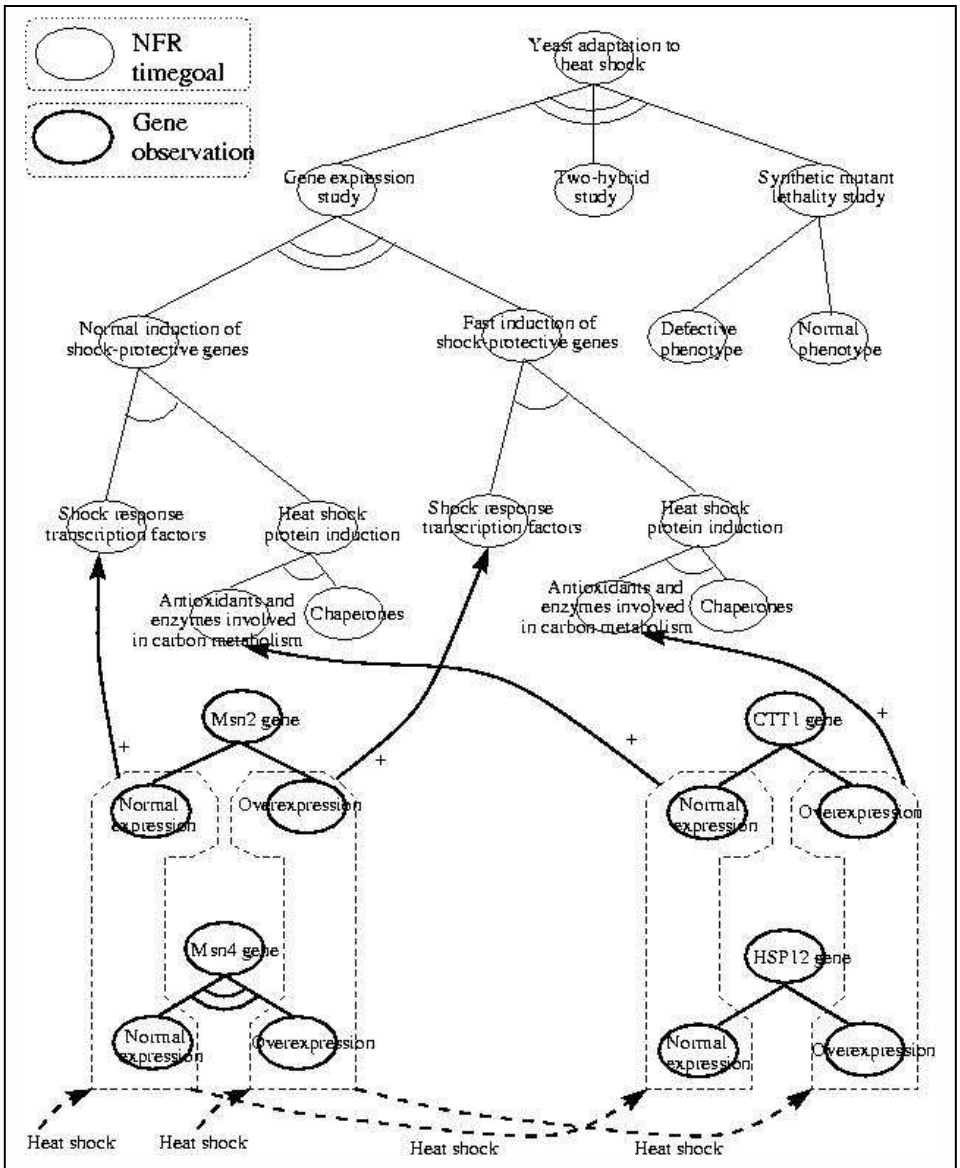

Figure 1: A typical Timegoal Interdependency Graph (TIG) for the yeast's adaptation to a heat shock.

2.2 *Transformations*

The IGIPI framework deals with time and the changes that may occur over time in a biological system. For this, it is necessary to represent processes that cause a change in the state of a biological system (both natural processes such as DNA transcription and experimental processes such as mixing[7]). The IGIPI framework refers to these processes as *transformations*. Transformations are represented as dotted lines connecting gene observation timegoals.

One of the major goals of representing transformations is to show their effects on the state of the participating genome components. For instance, a gene expressed at a high level at time *t* may be affected by a transformation, such that its expression at time *t+1* changes to a different level. A genome component's previous state may cease to exist and a new state may emerge as a result of the transformation. Figure 1 shows a "heat shock" transformation being applied to the overexpressed Msn2 and Msn4 genes, which causes the CTT1 and HSP12 genes to be overexpressed at the next time point.

The IGIPI framework represents significant events in a biological transformation as gene observation timegoals, including the starting and ending points. As shown in Figure 1, the structure of a transformation typically consists of the participating timegoals, the environmental conditions involved when the transformation was initiated (which may be preconditions for the transformation to occur) and the effects or changes engendered by the transformation on the participating timegoals[7].

Participating timegoals in a transformation are usually observations of genes' expression levels that contribute towards satisficing a high level biological function. Using this approach, it is possible to model the relationship between the output gene observations and the original input gene observations, by representing changes in the semantic categories of gene observations before and after a transformation. An example of this situation is shown in Figure 1; the Msn2 and Msn4 genes are labelled as "shock response transcription factors" and a "heat shock" transformation causes the transcription of the CTT1 and HSP12 "heat shock proteins" to be induced. Thus, our model of transformations goes beyond simply representing inputs (participant gene observations) and outputs.

2.3 *Complexes of Genome Components*

In a biological transformation, a major event may involve not one but several participating genes or proteins in specific states of expression[7]. The IGIPI framework attempts to build a complete picture of a transformation as it occurs over time, and thus it offers a structural abstraction for representing a group of participants in a transformation. This abstraction is called a *complex*.

A complex joins together several objects such as genes or proteins that are observed to participate in an experimental transformation simultaneously. Figure 1 illustrates several examples of gene complexes. For example, when a "normal expression" of Msn2 and a "normal expression" of Msn4 are joined in a complex, the resulting complex contributes towards satisficing the "shock response transcription factors" NFR timegoal, thus inducing the biological function of yeast adaptation to heat shock.

3. The IGIPI Framework as an Ontology-Based Clustering Approach

Classical mathematical-based clustering techniques pose many problems and are deemed to be insufficient for accurate prediction of gene function[3]. This section describes how the IGIPI framework can be used as an ontology-based clustering technique, to provide more effective prediction of gene function. For this purpose we present a similarity assessment algorithm complementing the IGIPI framework. The purpose of this algorithm is to measure how similar (or how different) two timegoals are, in the context of a timegoal interdependency graph[10].

Applying the IGIPI framework as an ontology-based clustering technique for function prediction could provide the following advantages over using solely mathematical-based clustering techniques[3]:

1) Background knowledge about protein interactions, including conditions under which they occur, will be incorporated into the function prediction process.
2) The correctness of the function predictions may be evaluated quantitatively, and thus questionable predictions could be rejected.
3) Function prediction will scale to a much larger number of genes.

3.1 Background in Assessing the Analogical Similarity of Objects

Our model uses a basic similarity algorithm (described in the next subsection) which measures the distance of the semantic models of two objects with respect to their *generalization*[10].

Generalization is a modeling abstraction used to build semantic models of objects, by extracting from two or more objects their common attributes and aggregating these into another more general object[10]. For example, *normal phenotype* and *lethal phenotype* can be generalized as the *outcome of an SML experiment*. This way, the common characteristics of *normal phenotype* and *lethal phenotype* are abstracted into the object *outcome of an SML experiment*.

In the IGIPI framework and the NFR framework generalization is represented by *decomposing* a timegoal into one or more low-level timegoals[1]. A decomposition between two timegoals expresses that the less general timegoal (i.e., the low-level timegoal) groups a subset of instances grouped by the more general timegoal (i.e., the high-level timegoal). For example, decomposing the *outcome of an SML experiment* into *lethal phenotype* expresses that the set of instances grouped by *lethal phenotype* is a subset of instances grouped by the *outcome of an SML experiment*.

According to the principle of *ontological uniformity*, similarity comparisons are only allowed between objects that have similar basic ontologies[10]. In our case this requirement is satisfied, because similarity comparisons always take place between timegoals in the same timegoal interdependency graph.

3.2 The Generalization Distance Algorithm for Assessing the Analogical Similarity of Timegoals

In order to measure how similar two timegoals are to each other, we have used the *Generalization Distance Algorithm*[10]. This algorithm evaluates the distance of two given timegoals with respect to generalization. We claim that the generalization distance algorithm provides a quantitative indication of the timegoals' semantic analogies and resemblancies.

The algorithm returns a real value between 0 and 1; a value close to 1 implies a large distance (and thus little similarity) between two timegoals. The basic idea of the algorithm is to determine the ancestors which the two timegoals do not share, and then to evaluate a function which weighs higher-level ancestors more than lower-level ancestors[10]. The algorithm uses the auxiliary *ClassDepth* algorithm, which computes the depth of a timegoal in a TIG[10].

4. XML and the IGIPI Software Tool

We provide an XML DTD schema (available on our web site) to materialize the abstract ideas of the IGIPI framework. We have used this schema to combine multiple sets of genomic and proteomic data in XML files, according to the rules of IGIPI. The proteomic data was provided to us by BIND[5], while the genomic data by Charlie Boone's yeast lab of the Banting and Best medical institute[12].

We also provide a graph visualization software tool (see Figure 2), employing a Java servlet and applet. This tool parses XML files stored on our server with experimental data, and dynamically generates graphs illustrating the data. The graphs illustrate the similarities

between the genes and proteins, by placing more similar objects spatially closer to each other. The tool also allows clustering of the data according to the Generalization Distance Algorithm (see Section 3.2).

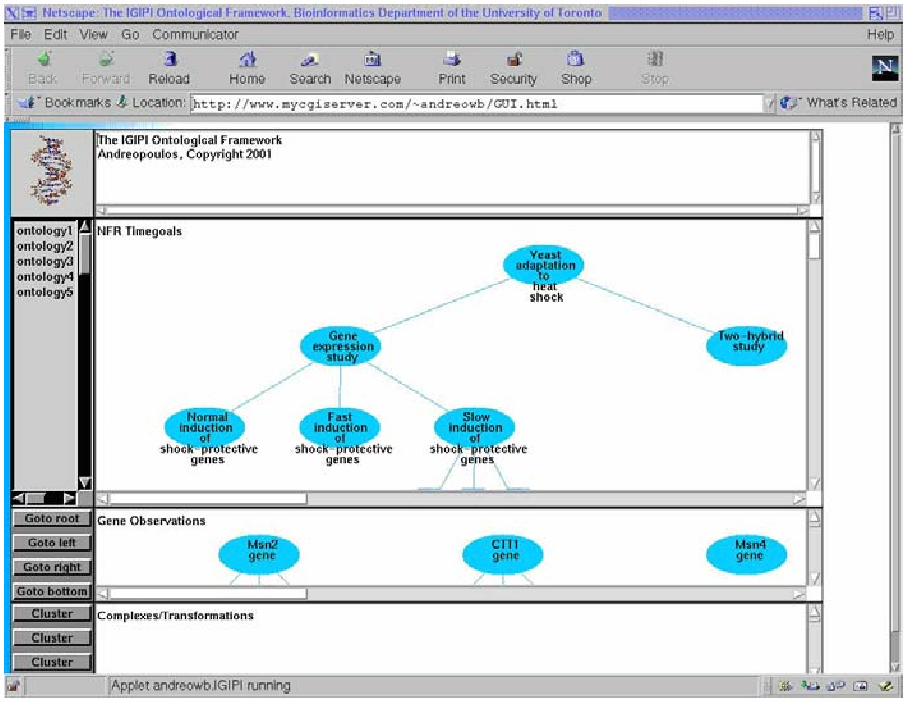

Figure 2: The IGIPI software tool for graph visualization and clustering.

5. Future Work

A major goal for the future is to apply the IGIPI framework to massive amounts of experimental data, primarily genomic and proteomic data from the yeast *Saccharomyces cerevisiae*. Recent developments in biotechnology tools have enabled *synthetic mutant lethality* (SML) studies to be applied to the entire yeast genome[11]. When the data from SML studies is combined with previously published genetic interactions from the yeast literature, the result is very large data sets containing thousands of genetic interactions.

Furthermore, the 6,200 proteins of yeast have been used extensively in yeast *two-hybrid* searches to detect interacting partners of proteins[11] (as opposed to genes).

A goal of our work is to integrate the protein interaction networks from yeast two-hybrid studies with the gene interaction networks from SML studies. The latter type of data is currently provided to us by the BIND database[5], while the former by the yeast lab of the Banting and Best medical institute[12]. Applications of the IGIPI framework for this purpose could provide a novel way of predicting gene function, given that the functions of many yeast proteins have already been elucidated[5,11].

The IGIPI framework may also be applied to more sophisticated modeling applications. Specifically, it could be used to model the typical circulation of a genetic network from state to state, around a cycle of states, until the network reaches a state it's been in before. The cycle of states in a genetic network typically fluctuates when a perturbation occurs. Furthermore, a very strong perturbation may cause a genetic network to abandon one cycle of states and enter a different cycle. Modeling these genetic phenomena presents challenges that the IGIPI framework may address successfully.

References

1. Kyungwha Lawrence Chung, *Representing and Using Non-Functional Requirements: A Process-Oriented Approach.* Ph.D. Thesis, Department of Computer Science, University of Toronto, June 1993.
2. Gene Ontology Consortium. http://www.geneontology.org/
3. Eisen, M.B. et al. Cluster analysis and display of genome-wide expression patterns. *Proc. of the National Acad. of Sciences USA* **95**, 14863-14868 (1998).
4. Gasch, A.P. et al. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell* **11**, 4241-4257 (2000).
5. Bader, G.D. and Hogue, C.W.V. BIND - a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **16(5).** 465-477 (2000).
6. J. Mylopoulos, E. Yu. Using Ontologies for Knowledge Management: A Computational Perspective. *Annual Conference of the American Society for Information Science*, Washington, DC, p. 482-496. (1999).
7. Hafner, C.D. and Fridman, N. Ontological Foundations for Biology Knowledge Models. *In the Proceedings of the 4th International Conference on Intelligent Systems for Molecular Biology (ISMB-96),* 78-87. AAAI Press (1996).

8. Stevens, R. et al. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. *Bioinformatics* **16(2),** 184-185 (2000).
9. Karp, Peter D. An ontology for biological function based on molecular interactions. *Bioinformatics* **16(3),** 269-285 (2000).
10. Spanoudakis G., Similarity Analyzer: An Implementation Overview, Working Paper #11, *Institute of Computer Science, Foundation for Research and Technology* – Hellas, Iraklion, Crete, Greece, September 1994.
11. Petra Ross-Macdonald. Functional analysis of the yeast genome. *Funct. Integr. Genomics* **1**, 99-113 (2000).
12. Roberts, C.J. et al. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**: 873-880 (2000).
13. S. Schulze-Kremer. Ontologies for Molecular Biology. *Proceedings of the Third Pacific Symposium on Biocomputing, Hawaii,* World Scientific Publishers, Singapore, pp.693-704. (1998).