THE IGIPI ONTOLOGICAL FRAMEWORK: INTEGRATING GENE INTERACTIONS WITH PROTEIN INTERACTIONS

BILL ANDREOPOULOS[1], AIJUN AN[1], XIANGJI HUANG[2]
[1]*Department of Computer Science and Engineering, York University*
[2]*Department of Information Technology, York University*
Toronto, Ontario, Canada M3J 1P3
billa@cs.yorku.ca

Keywords: Integration, interaction, protein, gene, ontology, framework.

1. Introduction

Researchers in biological sciences often face the problem of combining the results of different types of genomic and proteomic studies. Integrating the results of different types of experimental studies using ontologies, could provide researchers with many benefits, such as predicting gene function more reliably and supporting the evolution of current knowledge by integrating it with new genomic data[5,10]. We address this important genomics problem by proposing the IGIPI ontological framework, standing for "Integrating Gene Interactions with Protein Interactions", for integrating the results of experimental studies. IGIPI is based on the concept of goals that need to be satisfied in an experiment. IGIPI views different experiments as pieces of a puzzle that if positioned properly will create a more complete model of the cell. Researchers can potentially use IGIPI for semantic markup of web sites that have biomedical content. For semantic markup of web sites, IGIPI-based ontologies in terms of the OWL Web Ontology Language can be used[1]. Section 1 introduces the problem domain and the motivation. Section 2 describes the IGIPI ontological framework and illustrates an example application on yeast.

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. Unfortunately, reality becomes complicated by what happens at the higher cellular level of proteins. For instance, protein interactions produced from two-hybrid studies often are not mapped directly to gene interactions from synthetic mutant lethality (SML) studies, adding fuzziness to predicting the gene functions[5]. The purpose of SML studies is to identify

[1] http://www.cs.yorku.ca/~billa/IGIPI/

interactions between genes in the genome, by knocking out pairs of genes until a cell dies[10.] 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.
- **Nonallelic noncomplementation**: Mutations in two genes may fail to complement, because the gene products are subunits of the same multi-protein complex.
- **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)[10].

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[10].

Thus, it is necessary to create a complete picture of the cell, by combining the results of different genomic and proteomic studies. Researchers need to be able to combine the protein interactions observed in two-hybrid studies with the gene interactions observed in SML studies[10]. For this, it is necessary to represent the experimental and environmental conditions under which any observation was made[5]. Integrating the events observed at the higher cellular level of protein interactions with the SML gene interaction data, allows assessing the meaning of the observed interactions with greater confidence[7,8,9]. Then one can draw more informed conclusions about the gene and protein functions.

We address the challenge of representing the conditions under which the protein and gene interactions were observed[2,6,7,8,9] with the IGIPI ontological framework. Representing information derived from different experimental studies requires solving the following knowledge representation problems:

1) Ability to represent the fact that some genes may repress or affect negatively a biological function, while simultaneously inducing other biological functions.
2) Ability to represent all experimental and environmental conditions under which biological functions are manifested.
3) Ability to represent the specific group (module) of genes involved in each manifestation of a biological function.
4) Ability to represent the 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]).
5) Ability to represent the relative time point at which an event in a process occurs[7].

2. Description of the IGIPI Framework

The IGIPI framework is an ontological framework used for combining data produced by multiple genomic experiments on a biological function. For the integration 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 *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 (i.e. gene 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].

The IGIPI framework is based on the semantic modeling abstraction of a "goal" for representing the different conditions and experimental techniques that lead to the results. This section describes the abstractions offered by the IGIPI framework that address the challenges presented when integrating data from different genomic experiments.

2.1 *Timegoals: NFRs and Observations*

The IGIPI framework is based on the concept of *timegoals*. A timegoal is a goal that needs to be satisfied at a specific time interval 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 representing goals in experiments that, if satisficed, will lead to observing a biological function. Each timegoal is represented as a node (cloud). The interdependencies between timegoals are represented as edges.

The IGIPI framework supports two types of timegoals: *NFRs* (high level goals) and *observations* (low level goals). The term NFR is derived from the software engineering term "non-functional requirement"[1]; in our context an NFR is a high level goal placed on a biological experiment, without stating anything about the precise means by which the goal will be satisficed in the experiment. A developer can construct an initial TIG by identifying the top-level function that is expected to be observed and sketching an NFR timegoal for it. Figure 1 shows observing the "yeast adaptation to a heat shock" in an experiment as a root NFR timegoal at the top of the TIG. All the different timegoals are arranged hierarchically; a general parent timegoal is decomposed into more specific offspring timegoals at lower levels. An offspring timegoal's time interval is included in the parent timegoal's time interval. To represent the timegoals that need to be satisficed for the "yeast 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 leads to observing the yeast's adaptation to a heat shock. The NFR timegoals do not represent knowledge about the genomic-level events that need to occur for the biological function to be observed; this is the purpose of observation timegoals.

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 only one of the timegoals for the three types of experimental studies needs to be satisficed, to satisfice the "yeast adaptation to a heat shock" timegoal. When no arc is shown it is an *OR* contribution by default.

The bottom of a TIG consists of the *observation timegoals* that represent goals concerning the events that need to occur at a low genomic level, to satisfice one or more high-level NFR timegoals. A observation represents an observation or manipulation of a gene or protein at a low genomic level. Since observations are considered timegoals they may be decomposed into more specific observations at a lower level. For example, Figure 1 shows a observation timegoal representing the general goal of observing the Msn2 gene; this timegoal gets decomposed into the timegoals of overexpressing the Msn2 gene and observing the Msn2 gene at its normal expression level.

Observation timegoals make a positive or negative contribution towards satisficing one or more high level NFR timegoals. Figure 1 shows how interdependency links are used to represent a observation timegoal's contribution towards satisficing an NFR timegoal; such a contribution can be positive ("+" or "++") or negative ("-"or "--"). Since an NFR timegoal can receive both positive and negative contributions from several other observation timegoals, it is hard to draw a line between whether an NFR timegoal is satisfied or not. Thus, we use the concept of satisficing an NFR timegoal, as described above, to indicate that an NFR timegoal receives enough positive contributions such that the person carrying out the experiment can consider the timegoal to be satisfied[1].

2.2 Transformations

The IGIPI framework deals with time and the changes that occur over time in a biological system. 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 broken lines connecting observation timegoals.

The IGIPI framework represents the starting and ending points of a biological transformation as observation timegoals. Timegoals participating in a transformation are observations of proteins or genes' expression levels that contribute towards satisficing a high level biological function. As shown in Figure 1, a transformation consists of the participating timegoals, the environmental conditions involved (which may be preconditions for the transformation to occur) and the effects or changes induced by the transformation on the participating timegoals[7].

One of the major goals of representing transformations is to show their effects on the states of the participating genome components. A genome component's previous state may cease to exist and a new state may emerge as a result of the transformation. For instance, a gene expressed at a certain level at time t may be affected by a transformation, such that its expression at time $t+1$ changes to a different level. 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.

It is also possible to model the relationship between the input and output timegoals in a transformation, by representing changes in the semantic categories of the timegoals after a transformation. An example of this situation is shown in Figure 1; the Msn2 and Msn4 genes are labeled as "shock response transcription factors" and a "heat shock" transformation induces the transcription of the CTT1 and HSP12 "heat shock proteins".

2.3 *Complexes of Genome Components*

In a transformation, an event at a time point may involve more than one participating genes or proteins in specific states of expression[7]. The IGIPI framework builds a complete picture of a transformation as it occurs over time, by offering a structural abstraction for representing a group of participants at a time point. This abstraction is called a *complex*.

A complex joins several objects such as genes or proteins that participate in a transformation simultaneously. Figure 1 illustrates several examples of gene complexes. When a “normal expression” of Msn2 and a “normal expression” of Msn4 are joined in a complex, together they contribute towards satisficing the “shock response transcription factors” NFR timegoal, thus inducing the function of “yeast adaptation to a heat shock”.

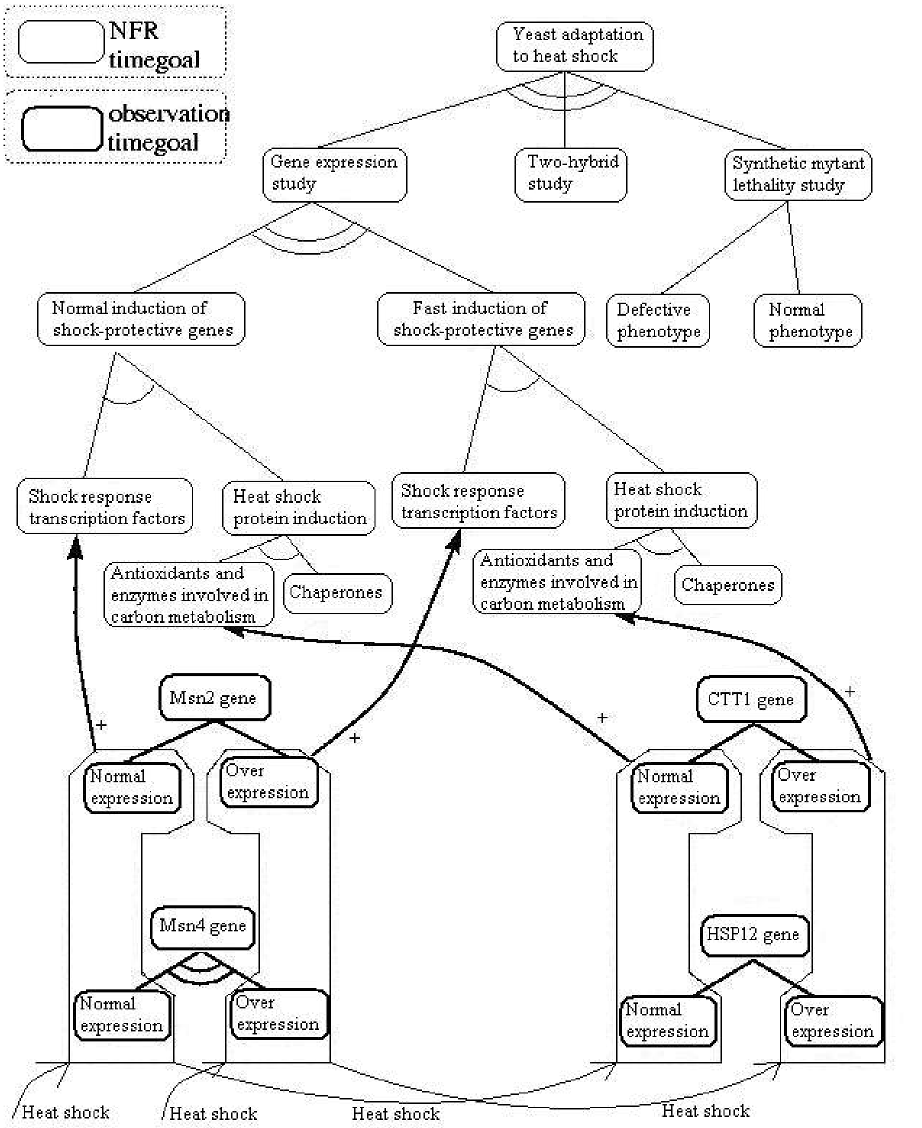
NFR timegoal
observation timegoal
Yeast adaptation to heat shock
Gene expression study
Two-hybrid study
Synthetic mytant lethality study
Normal induction of shock-protective genes
Fast induction of shock-protective genes
Defective phenotype
Normal phenotype
Shock response transcription factors
Heat shock protein induction
Shock response transcription factors
Heat shock protein induction
Antioxidants and enzymes involved in carbon metabolism
Chaperones
Antioxidants and enzymes involved in carbon metabolism
Chaperones
Msn2 gene
CTT1 gene
+
+
+
+
Normal expression
Over expression
Normal expression
Over expression
Msn4 gene
HSP12 gene
Normal expression
Over expression
Normal expression
Over expression
Heat shock
Heat shock
Heat shock
Heat shock

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

2.4 Contrib

We use the notion of a timegoal being satisficed, as opposed to satisfied. In figure 2, the symbol “V” on a timegoal means that it is satisficed, while a symbol “X” means that it is not satisficed – for example, the timegoal “Avastin” is satisficed meaning that this drug has been taken by a human. Figure 2 shows how contributions from lower timegoals are propagated upwards and contribute towards satisficing higher timegoals. The timegoal 'angiogenesis' contributes to timegoal 'lung cancer', but 'angiogenesis' receives a strong negative contribution from drug ‘Avastin’; thus timegoal 'lung cancer' is not satisficed.

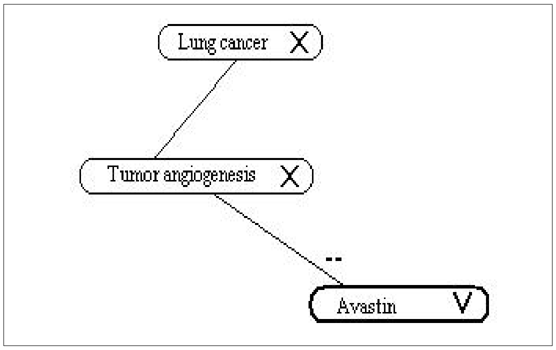

Figure 2: A negative contribution of an observation timegoal for the drug ‘Avastin’ contributes to not satisficing ‘tumor angiogenesis’ and ‘lung cancer’.

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. The Gene Ontology Consortium (2001). Creating the gene ontology resource: design and implementation. *Genome Research* 11: 1425-1433. 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. Petra Ross-Macdonald. Functional analysis of the yeast genome.*Funct. Integr. Genomics* **1**, 99-113 (2000).