Integration of Genomic, Proteomic and Biomedical Information on the Semantic Web

Bill Andreopoulos

Aijun An

Xiangji Huang

Technical Report CS-2005-14

October 2005

Department of Computer Science and Engineering

4700 Keele Street North York, Ontario M3J 1P3 Canada

Integration of Genomic, Proteomic and Biomedical Information on the Semantic Web

Bill Andreopoulos

Department of Computer Science and Engineering, York University, Toronto, Ontario, Canada, M3J1P3, billa@cs.yorku.ca

Aijun An

Department of Computer Science and Engineering, York University, Toronto, Ontario, Canada, M3J1P3, aan@cs.yorku.ca

Xiangji Huang

Department of Information Technology, York University, Toronto, Ontario, Canada, M3J1P3, jhuang@yorku.ca

Abstract: Researchers are faced with the challenge of integrating on the basis of a common Semantic Web framework the information on biological functions resulting from genomic and proteomic experimental studies. Researchers would also like to integrate the biological functions' roles in larger biomedical conditions, which will support automated analysis and reasoning on the Semantic Web. We address these challenges by proposing the IGIPI framework, standing for "Integrating Gene Interactions and Protein Interactions". IGIPI views different experimental studies as pieces of a puzzle that if positioned properly will contribute to a more complete representation of a biological function or biomedical condition. This framework allows representing the relative time points of events. The IGIPI framework involves integrating different ontologies and vocabularies, including the Gene Ontology, MGED Ontology and UMLS Unified Medical Language System. Researchers can semantically markup their websites through reusing and refining IGIPI representations in the OWL Web Ontology Language. We applied IGIPI to yeast and cancer information.

Keywords: integration, gene, protein, interaction, ontology, biomedical, information.

1 Introduction and Motivation

Biomedical ontologies are often developed in an uncoordinated manner, sometimes merely reflecting hierarchical relations between concepts in a domain and serving the purposes of annotating online databases for information retrieval. Individual ontologies do not allow integration on the Semantic Web of information derived from different sources,

such as that produced by labs employing different research methods. Ontology development often overlooks reusing existing ontologies and often ontologies are not interoperable. Ontology interoperability and information integration on the Semantic Web will support automated analysis and reasoning about the machine processable dispersed online literature, as well as quick online markup of the latest research results [2,4,10,23,33,34,39,47,49,52]. It is necessary to create a common bioinformatics framework for representing knowledge about biological networks and pathways, while integrating gene expression data and previously developed ontologies [3]. A combined use of these resources will enable functional knowledge discovery, thus utilizing all of these resources to their full potential.

We present the IGIPI framework for integrating on the Semantic Web the information resulting from different types of genomic and proteomic experimental studies on a biological function. A *biological function* is a network of gene or protein interactions. Integration involves representing the experimental and environmental conditions associated with different studies, under which a biological function may be observed. Moreover, genes' and proteins' contributions under different conditions should be unambiguously represented. Different studies and conditions often suggest conflicting results on a network of gene or protein interactions, highlighting the non-triviality of integration [21,23,41,42]. Information on the contributions of biological functions to a high level biomedical condition, such as cancer, can also be represented with IGIPI. A *biomedical condition* is a condition observed in an organism that is of interest to the biomedical community. IGIPI is based on the notion of 'goals' representing the specific conditions that need to be satisfied to observe a biological function or biomedical condition outcome. If an outcome can be observed by two or more different types of experimental studies, such as gene expression studies and two-hybrid studies, then a researcher's aim is to represent the different conditions as goals contributing to the overall outcome [18,23,34,47-49]. A separate representation is developed for each biological function and biomedical condition, based on an OWL Web Ontology Language specification of IGIPI [56]. Researchers can refine and reuse existing representations for semantic markup of websites [40].

This approach benefits the biomedical community. It supports interoperability of concepts from separate ontologies [11,36]. It supports evolution of information by allowing its easy integration with the latest research results [24,40]. It allows physicians who have no time to search the latest research results to quickly retrieve from the Semantic Web the latest results on a biomedical condition, such as cancer [46]. Finally, it supports finding knowledge through automated reasoning, such as predicting the side effects of drugs, interpreting and diagnosing medical symptoms and predicting genes' and proteins' functions [3,4,24].

We often consider 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 interactions between genes in the genome, by knocking out pairs of genes until a cell dies [41]. 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) [42].

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 [42].

Thus, researchers need to be able to create a complete picture of the cell by integrating the information resulting from different genomic and proteomic studies [47-49]. To combine the protein interactions observed in two-hybrid studies with the gene interactions observed in SML studies it is necessary to be able to represent the experimental and environmental conditions under which the protein and gene interactions were observed. 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 [18,23,48]. Then one can draw more informed conclusions about the gene and protein functions. Addressing this challenge on the Semantic Web requires:

1) Ability to represent the fact that a gene/protein may induce a biological function (i.e. a network of gene or protein interactions) while repressing other biological functions.

2) Ability to represent all experimental and environmental conditions under which a biological function may be observed.

3) Ability to represent a module of genes/proteins inducing or repressing a biological function.

4) Ability to represent a process consisting of events that changes the module of genes/proteins inducing a biological function, e.g., by attracting more genes to join the module or repelling other genes from the module.

5) Ability to represent the relative time points of active modules of genes/proteins and other events in a process [18,49].

The outline of the paper is as follows: Section 2 describes related work. Section 3 describes the IGIPI abstractions for biological functions with application on yeast. Section 4 describes extensions of IGIPI for biomedical information with application on cancer. Section 5 discusses using IGIPI on the Semantic Web with the OWL Web Ontology Language. Section 6 discusses analyzing and reasoning with online information. Section 7 concludes and discusses future work. The Appendix gives the OWL specification of IGIPI.

2 Related Work

Individually developed ontologies often support the annotation of online databases for information retrieval purposes. However, they are often not interoperable and do not always allow integration of information derived from different sources and automated reasoning and analysis on the Semantic Web. Our approach differs from other approaches, since it is designed specifically for integrating genomic, proteomic and biomedical information on the Semantic Web on the common basis of 'goals'. Our approach supports automated reasoning about dispersed online literature and analyzing it for knowledge of interest to researchers.

2.1 Information Integration and Ontology Mappings

To support ontology-based information integration, the ontologies have to be connected to the contents of an information system. Many of the existing information integration systems such as [31] or [38] use two or more ontologies to describe the information. If several ontologies are used in an integration system, mapping between the ontologies is important. A mapping between two terms from two different ontologies implies that the terms have the same or similar meanings. We give an overview of some general approaches to mapping different ontologies in knowledge engineering [10,35,52].

Defined Mappings: Mappings can be defined between different ontologies manually or semi-automatically. This approach is taken in *KRAFT* [38]. Different kinds of mappings are distinguished in this approach, starting from simple one-to-one mappings between classes and values up to mappings between compound expressions. In this approach the user is free to define conflicting mappings that might not make sense. Semi-automated mapping methods such as *CAIMAN* [28], *OntoMapper* [37] and *GLUE* [12-14] assign a set of

relevant documents to each term capturing the meaning of the term, measure similarity between terms and search for mappings based on the similarity matrix obtained.

Probabilistic Mappings: Since semantic similarities between concepts can be represented probabilistically, Bayesian Network approaches to ontology mapping that take the degree of uncertainty in the Semantic Web into consideration have been proposed. Such an approach is defined by Ding et al. [11,36]. The source and target ontologies are translated into Bayesian networks. Then, the concept mappings between the two ontologies are treated as evidential reasoning between the two translated Bayesian networks.

Lexical Relations: An attempt to provide at least intuitive semantics for mappings between concepts in different ontologies (primarily linguistic or lexical ontologies) is made in the *OBSERVER* system [31], *WordNet*, *Cyc* and *SENSUS*. OBSERVER defines inter-ontology relations as *synonym, hypernym, hyponym, overlap, covering* and *disjoint*. Though these relations are similar to description logic constructs they do not have formal semantics.

Top-Level Ontology: To define mappings between different ontologies on the basis of the same semantics, all ontologies can be related to a single top-level ontology. This is done by inheriting concepts from the same top-level ontology. Concepts from different ontologies are connected in terms of common superclasses. This approach can help to resolve conflicts and ambiguities [8,19-20]. However, this approach does not establish a direct correspondence between different ontologies, which can make it hard to find exact matches.

Semantic Correspondences: An approach that tries to overcome the indirect mapping of concepts via a top-level ontology is to identify direct semantic correspondences between concepts from different ontologies. To avoid conflicting mappings between concepts, a common vocabulary defines a common concept lattice across different ontologies. Wache [51] uses semantic labels in order to compute correspondences between database fields.

2.2 Identifiers, Ontologies, Databases and Other Tools

Shared ontologies help describe biological concepts, but do not help the community agree on how to name them. For this purpose, several *identifier standards* of important biomedical and biological terms have been developed. *UMLS* is the National Library of Medicine's (NLM's) Unified Medical Language System project that develops and distributes multi-purpose, electronic "Knowledge Sources" and associated lexical programs [44]. *MeSH* is Medical Subject Headings at NLM. *CBIL* is the Controlled Vocabulary Terms for human anatomy. *PROW* is NLM's Proteine Review on the Web. *Enzyme Nomenclature* comes from the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. The *Mouse Anatomical Dictionary Browser* exists on the Mouse Genome Informatics site. The *mmCIF* dictionary stands for macromolecular

Crystallographic Information File. The *HUGO* gene nomenclature contains names and synonyms denoting known genes in various organisms. The *Life Science Identifier* (LSID) standard assigns unique database-dependent LSIDs to biological objects, standardizing the naming conventions for RDF-encoded biological information [39]. LSID combines the internet domain name of the source database with the local database object identifier.

Many *ontologies* have been developed for the biomedical community. The *Gene Ontology (GO) Consortium* consists of three ontologies of terms used for molecular functions, biological processes and cellular locations and the relations between the terms [1]. *IMGT*, the international *ImMunoGeneTics* information system, is a high-quality integrated knowledge resource specializing in immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC) and related proteins of the immune system of human and other vertebrate species [29]. It contains the *IMGT-ONTOLOGY*, an ontology which allows the management of the immunogenetics knowledge for all vertebrate species. The *TAMBIS project* aims to aid researchers in biological science by providing a single access point for biological information sources round the world [48]. The access point is a single interface (via the World Wide Web) which acts as a single information source. It will find appropriate sources of information for user queries and phrase the user questions for each source, returning the results in a consistent manner which will include details of the information source. *RiboWeb* is a prototype for new structural information resources that tightly link models and their coordinates with experimental (and other) data sources [54]. The project initially focused on the structure of the prokaryotic 30S ribosomal subunit, which initiates the translation of mRNA into protein and is the site of action of numerous antibiotics; the project has since been expanded to include structural data pertaining to the entire ribosome of prokaryotes (but primarily E. coli).

Previous ontologies have often focused on gene function, by modeling how different gene functions relate to each other [23-24]. Some projects involving developing ontologies for functional classification purposes are *Gene Ontology* (GO) [1], *EcoCyc* [25,53], *MIPS* [32] and *KEGG* [22]. Previous functional classification ontologies assist in annotation of gene functions, a practice that usually involves semantically annotating genes in databases while publishing the experimental methods and results. Other ontologies describe general concepts in biology, such as ‘gene’ and ‘protein’ that might possibly be used with different meanings across databases [34,43]. Examples of ontologies for this purpose are the *Sequence Ontology Project* and the *OMB* [43,55]. We were given many ideas for our work by the publications of Hafner and Fridman [18], who examined problems concerning representing information on complex biochemical substances and transformations of such substances into different forms. Our method for representing transformations of biochemical substances (as we describe in Sections 3.2-3.4) addresses the knowledge

representation problems described by Hafner and Fridman by modeling relationships between transformation inputs and outputs, as well as how the semantic category of the inputs changes after the transformation occurs.

Many *databases* of biomedical content have been developed. The *MIPS Munich Information center for Protein Sequences* is a database of information on proteins in various organisms, particularly yeast, including complex and sequence information [32]. *Swiss-Prot* is a curated protein sequence database which strives to provide a high level of annotations (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases [6]. The *BioCyc* Knowledge Library is a collection of Pathway/Genome Databases [23-24]. Each database in the BioCyc collection describes the genome and metabolic pathways of a single organism, with the exception of the *MetaCyc* database, which is a reference source on metabolic pathways from many organisms. *EcoCyc*, a part of the BioCyc library, is a scientific database for the bacterium Escherichia coli [25,53]. The EcoCyc project performs literature-based curation of the entire E. coli genome, and of E. coli transcriptional regulation, transporters, and metabolic pathways. *EMBL* at the European BioInformatics Institute is a Nucleotide Sequence Database (also known as EMBL-Bank) that constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications. *EpoDB* (Erythropoiesis database) is a database of genes that relate to vertebrate red blood cells. It includes DNA sequence, structural features, protein information, gene expression information and transcription factor binding sites. *FlyBase* is a Database of the Drosophila Genome. Six database volumes of biological information about proteins comprise Incyte's *Proteome BioKnowledge Library*: *HumanPSD*, *GPCR-PD*, *YPD*, *PombePD*, *WormPD* and *MycoPathPD*. Each volume focuses on a different organism important in pharmaceutical research. *PharmGKB* is an integrated resource about how variation in human genes leads to variation in our response to drugs. *InterPro* at the European BioInformatics Institute: InterPro is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to unknown protein sequences. *BIND* is a database designed to store and search for protein interactions from various organisms [5].

Various interesting bioinformatics integration and analysis tools have been developed. The *Integrated Genome Database* combined more than a dozen source databases including GenBank and the Genome Database (GDB), but collapsed because each source database changed its data model too frequently. The cross-database query languages *Kleisli* and *K2* can be used to access several databases, but query processing is too slow [47]. Troyanskaya et al. propose a Bayesian method for predicting gene functions that uses different data

types. Expert knowledge is incorporated as a prior by questioning several experts about the relative accuracies of using data types as evidence [50]. Li et al. apply genetic algorithms for predicting gene function by integrating gene expression and metabolic data [30]. The *BioGrid* platform aims to integrate gene expression and protein interaction data, through assigning domains and superfamilies to gene products using the SUPERFAMILY tool and the Structural Classification of Proteins (SCOP) database [9]. The *PROVA* tool focuses on integrating the rules specifying bioinformatics workflows on the Semantic Web [26].

2.3 Text Mining and Information Retrieval

The volume of biomedical research literature has increased so rapidly that *text mining* and *information retrieval (IR)* methods are essential for locating papers. *Text mining* can be viewed as a part of *IR* and largely involves pre-processing a document collection using the technique of *text categorization*, i.e., tagging unlabelled documents by categories. Ontologies provide the framework for the semantic representation of textual information. Terms in the text of an abstract are linked to ontology terms. Several approaches have been proposed for ontology-based text mining of literature in a database such as *PubMed* [27,33,46]. *GoPubMed* allows users to search PubMed on the basis of the Gene Ontology [15]. GoPubMed categorizes the abstracts according to GO, allowing users to navigate through the abstracts by category of molecular biology. It also shows ontology terms related to the user's query, which do not appear in the abstract and the user might be unaware of.

Information retrieval (IR) methods aim at finding the documents that best satisfy a user's information need. The query paradigm is used by the PubMed database [15]. IR methods have also been used for finding functional relationships among genes. This is done by collecting a large set of PubMed abstracts covering the literature relevant to an organism. Each gene is mapped to one *kernel* abstract in the collection which represents the gene through discussing its biological function. Each kernel abstract is mapped to the set of the most related abstracts and a set of terms is produced summarizing the abstract set, thus producing for each gene a body of related abstracts and list of terms. Several clustering methods have been used for grouping proteins based on their annotations and PubMed abstracts based on their keywords. This gives insight into common protein functions and common paper subjects [45].

3 Integrating Biological Function Information

This section describes the modeling abstractions offered by the IGIPI framework that are used for integrating biological function information on the Semantic Web.

3.1 Timegoals: NFR timegoals and Observation timegoals

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 Graph* (TIG). Fig. 1 shows an example of a TIG. A TIG records all timegoals representing goals in experiments that, if satisficed, will lead to observing the root 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: *NFR* timegoals (high level) and *observation* timegoals (low level). The term NFR is derived from the term *non-functional requirement* used in software engineering [7]; in our context an NFR timegoal is a high level goal in an experiment, such as an experimental or environmental condition that needs to be satisfied for observing a biological function, without stating anything about the low level genomic or proteomic events that need to occur. A developer starts constructing a TIG by identifying the top level biological function that is expected to be observed and sketching a root NFR timegoal for it. The root NFR timegoal of a TIG has a value taken from a domain of biological functions. This domain is the *GO Gene Ontology* [1]. The root NFR timegoal is decomposed into timegoals that represent more specific information about how the biological function may be observed.

Fig. 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 information about the low level genomic events that need to occur for the biological function to be observed; this is the purpose of observation timegoals described later.

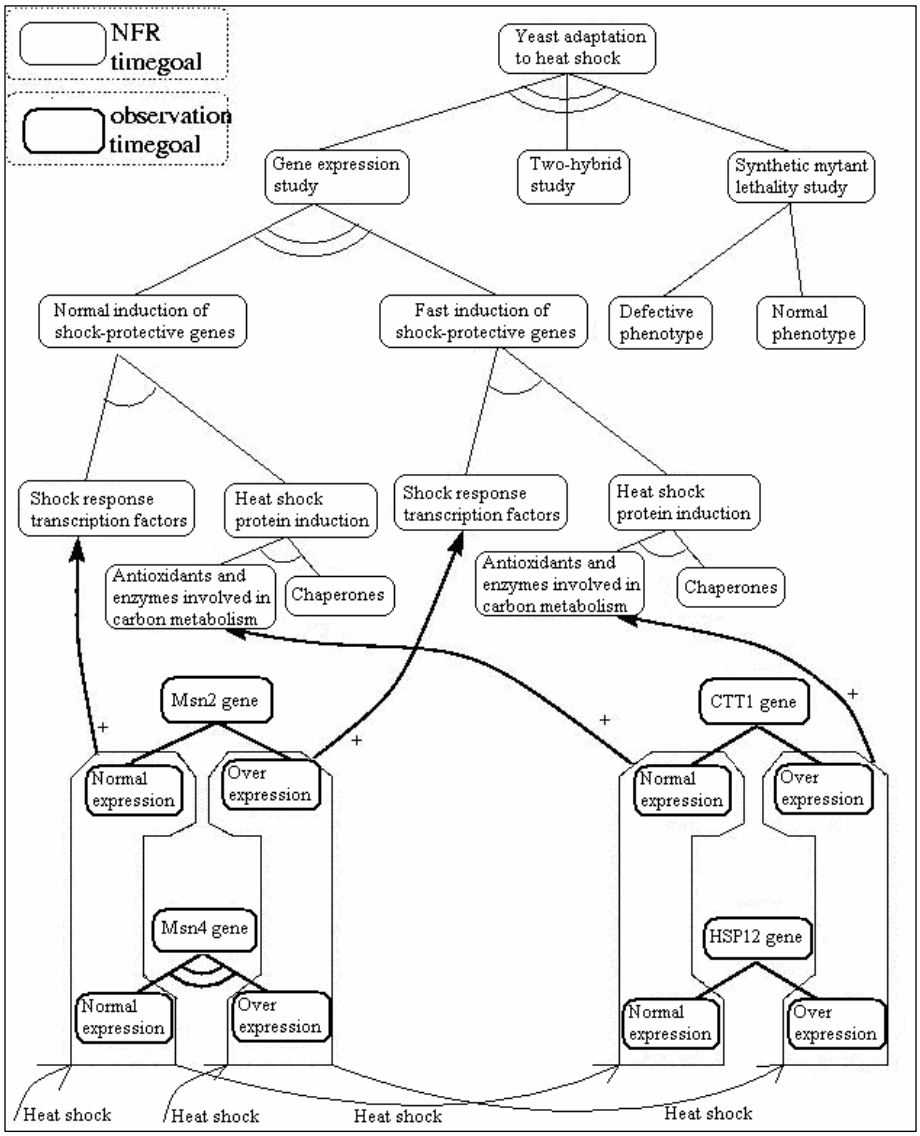

Fig. 1. The biological function Timegoal Graph (TIG) for "yeast adaptation to heat shock".

Timegoals are connected by interdependency links, which show *decompositions* of parent timegoals downwards into more specific offspring timegoals. In some cases the interdependency 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 interdependency 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. Fig. 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* representing goals concerning events that need to occur at a low genomic or proteomic level, to satisfice one or more high level NFR timegoals. An observation timegoal is drawn as a 'dark node' and represents specific information about a manipulation or an expression of a gene or protein. Since observations are considered timegoals they may be decomposed into more specific observations at a lower level. Fig. 1 shows an 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. Fig. 1 shows how interdependency links are used to represent an 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 [7].

3.2 Transformations

The IGIPI framework deals 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 [18]. 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. Fig. 1 shows that 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.

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. Fig. 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. Fig. 1 shows an example of this situation; 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".

3.3 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. 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. Fig. 1 shows 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".

3.4 Prerequisites for a contribution to occur

This framework allows using transformations to model that an event is a prerequisite for a timegoal to make a contribution to another timegoal. When a transformation precedes a timegoal or complex's contribution to a high level timegoal, it means that the transformation and anything before it are prerequisites for the contribution to occur.

4 Integrating Biomedical Condition Information

There exist uncountable biomedical web sites containing bits and pieces of information. Beyond building Timegoal Graphs (TIGs) for biological functions, the IGIPI framework can also be used to build TIGs representing information about how biomedical conditions are manifested. These TIGs can help to integrate the biomedical information on the Semantic Web. The root timegoal of a biomedical condition TIG has a value taken from a domain of biomedical conditions, such as "ischemic stroke", "haemorrhagic stroke", "lung cancer" etc. This domain is the *UMLS Unified Medical Language* System that integrates 100 biomedical vocabularies [44]. The root timegoal is decomposed into timegoals that represent information about the biomedical condition.

To distinguish the timegoals of biomedical condition TIGs from the NFR and observation timegoals of a biological function TIG, we use the name *biomedical condition timegoals*. Fig. 2 shows the biomedical condition TIG for "lung cancer". Like NFR timegoals, biomedical condition timegoals are decomposed downwards into more specific offspring timegoals. The offspring biomedical condition timegoals make an AND/OR contribution to the parent timegoal.

Biomedical condition timegoals may also receive contributions from the NFR and observation timegoals of a biological function TIG. An NFR or observation timegoal may contribute positively or negatively towards a biomedical condition timegoal. The contributions are propagated upwards and the root timegoal may or may not be satisficed, as described next.

4.1 Observation Timegoals Under the Influence of Drugs

An observation timegoal is decomposed to represent how it may be observed under the influence of drugs. Fig. 2 shows the decomposition of the Vascular Endothelial Growth Factor (VEGF) into the timegoals "VEGF under Bevacizumab" and "VEGF under Chemotherapy". This represents that the protein is in different states under the influence of Bevacizumab and Chemotherapy.

Fig. 2 shows that in some cases it is possible for cells to become drug resistant after chemotherapy. Although it is still not certain how this mechanism works, patients with a turned on gene PKC-epsilon seem to develop drug resistance. Fig. 2 shows that for a "protein-induced cell's resistance to chemotherapy drugs (drug-resistant phenotype)" to occur, it is first necessary for "chemotherapy" to occur, combined with the special protein "PKC-epsilon" that is not found in all humans.

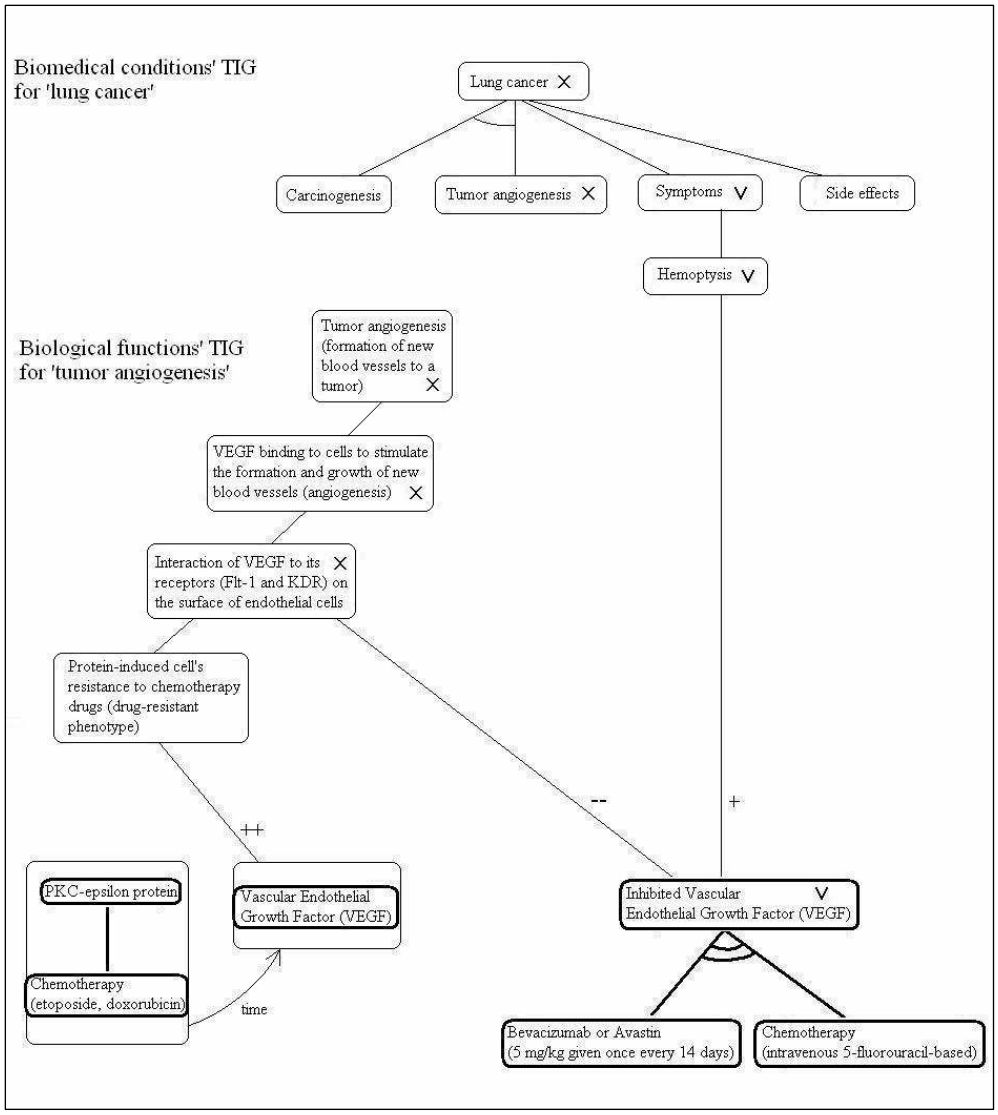

Fig. 2. The biomedical condition TIG for "Lung cancer".

4.2 Representation of symptoms of medical conditions and side-effects

The symptoms of a biomedical condition and the side effects of drugs are represented as subtrees of offspring timegoals of the root biomedical condition timegoal. Fig. 2 shows an example for "lung cancer". All symptoms of root timegoal "lung cancer" are grouped under

an offspring timegoal named "symptoms". All side-effects of drugs are grouped under an offspring timegoal named "side effects". Observation timegoals make positive or negative contributions to symptoms and side effects timegoals that are propagated upwards.

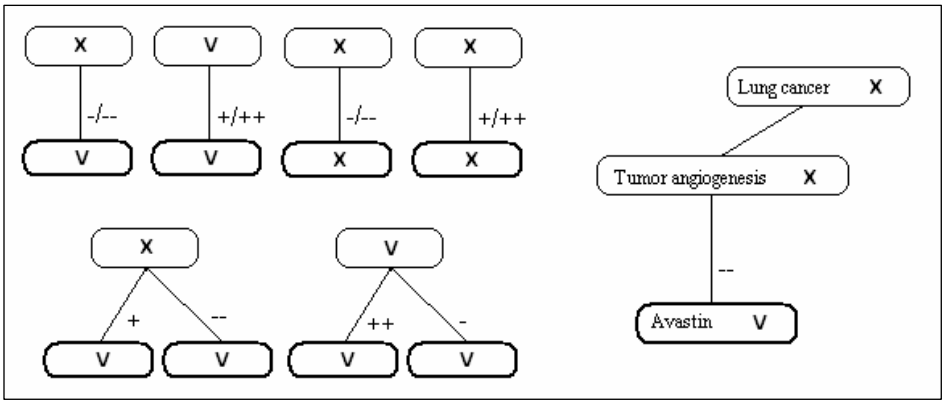

Fig. 3. A negative contribution of the "Avastin" drug observation timegoal negatively affects satisficing "Tumor angiogenesis" and "Lung cancer".

4.3 Propagations of Contributions for Satisficing Timegoals

We use the notion of a timegoal being *satisficed*. The symbol '*V*' on a timegoal means that it is satisficed, while the symbol '*X*' means that it is not satisficed. In Fig. 3 the timegoal "Avastin" is satisficed meaning that this drug is taken by a patient. Fig. 3 shows how contributions from lower timegoals are propagated upwards and contribute towards satisficing higher timegoals. The timegoal "Tumor angiogenesis" contributes to timegoal "Lung cancer", but "Tumor angiogenesis" receives a strong negative contribution from the drug "Avastin" that is taken by a patient; thus timegoal "Lung cancer" is not satisficed.

Fig. 2 shows the TIG for "Lung cancer". The drug "Avastin" inhibits the VEGF protein. In turn, this contributes negatively to the NFR timegoal "Interaction of VEGF to its receptors" which is getting a negative contribution and thus it is not satisficed. This contributes to the root timegoal of the biological function TIG "Tumor angiogenesis". The biological function "Tumor angiogenesis" contributes to the biomedical condition TIG that represents information about "Lung cancer". Since the function "Interaction of VEGF to its receptors" is not satisficed, this contribution is propagated upwards to timegoals "Tumor angiogenesis" and "Lung cancer", neither of which is satisficed either.

5 Practical Utility of the IGIPI Framework

For the IGIPI framework to be usable in practice, it must allow researchers to easily markup biological and biomedical websites with semantic information.

5.1 Semantic Markup of Websites

The Timegoal Graphs (TIGs) are represented in terms of the OWL Web Ontology Language [56]. Our goal is to eventually possess a library of TIGs for integrating all of the web-based information including all biological functions and all biomedical conditions, through semantic markup of websites. Genomic and phenotypic information are mapped onto TIGs that serve as the point of entry on the Semantic Web for all biological functions and biomedical conditions. By mapping biological and biomedical information to TIGs, Semantic Web applications are given direct access to the information through the TIGs. For building these TIGs, we are following an approach similar to the wikipedia online library which allows readers to update each article with new information. Researchers can visit an online library of the current TIGs and select the ones to use for semantic markup of biomedical web sites. A researcher's goal is to annotate his/her biomedical website with annotations taken from a TIG that contains sufficient semantic information. If the current state of a TIG is not refined enough for a researcher, then he/she can propose extensions or refinements for the TIG through a special web form, until the TIG is granular enough to annotate his/her website with precise semantic information [40]. Our current library contains the root timegoals for all biological function TIGs derived from the GO Ontology, the root timegoals for all biomedical condition TIGs derived from the UMLS Unified Medical Language System (that integrates 100 biomedical vocabularies) and names of known genes and proteins in human, yeast, fly, worm.

Several existing ontologies are mapped and integrated using OWL, like pieces of a puzzle. The root NFR timegoal of a biological function TIG is taken from the domain of the Gene Ontology (GO) [1]. The following block specifies that a root NFR timegoal is taken from the GO. It also specifies that a root NFR timegoal is OR decomposed into three NFR timegoals representing different experiments.

```
<owl:Class rdf:ID="#root_NFR_timegoal">
<rdfs:subClassOf rdf:resource="#NFR_timegoal"/>
<rdfs:subClassOf>
        <owl:Restriction>
                <owl:onProperty rdf:resource="#is_a"/>
                <owl:allValuesFrom rdf:resource="#GO_molecular_function"/>
        </owl:Restriction>
</rdfs:subClassOf>
```

```
<rdfs:subClassOf>
        <owl:Restriction>
                <owl:onProperty rdf:resource="#gets_OR_contribution_by"/>
                <owl:hasValue>"#gene_expression_study"</owl:hasValue>
                <owl:hasValue>"#two_hybrid_study"</owl:hasValue>
        <owl:hasValue>"#synthetic_mutant_lethality_study"</owl:hasValue>
        </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
```

The ontologies are linked together using OWL object properties. We represent the positive and negative contributions of the low-level timegoals to the high-level timegoals using relationships. As shown in the OWL schema below, each observation timegoal has a relationship called "contributes_positively_to" to zero or more NFR timegoals. This OWL code defines that an observation timegoal may contribute positively to zero or more NFR timegoals. Similar types of relationships exist for negative contributions, as well as between a biomedical condition timegoal and an NFR timegoal:

```
<owl:Class rdf:ID="#observation_timegoal">
<rdfs:subClassOf rdf:resource="#biological_function_timegoal"/>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource="#contributes_positively_to"/>
         <owl:allValuesFrom rdf:resource="#NFR_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">  0
</owl:minCardinality>
     </owl:Restriction>
. . . . . .
</rdfs:subClassOf>
```

The following represent the AND and OR contributions of NFR timegoals:

```
<owl:Class rdf:ID="#NFR_timegoal">
<rdfs:subClassOf rdf:resource="#biological_function_timegoal"/>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_OR_contribution_by"/>
         <owl:allValuesFrom rdf:resource="#NFR_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_AND_contribution_by"/>
         <owl:allValuesFrom rdf:resource="#NFR_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
```

Under the root NFR timegoal of a biological function TIG, experimental and environmental conditions are represented for which the biological function (protein and gene interactions) may be observed. The MGED Ontology for microarray experiment annotation gives values to the subtree under the NFR timegoal "gene expression study". Ontologies to give values to the subtrees under the NFR timegoals "two-hybrid study" and "SML study" are under development. The following specifies that the NFR timegoal for gene expression study is decomposed into NFR timegoals taken from the MGED Ontology:

```
<owl:Class rdf:ID="#gene_expression_study">
<rdfs:subClassOf rdf:resource="#NFR_timegoal"/>
<rdfs:subClassOf>
        <owl:Restriction>
          <owl:onProperty rdf:resource="#gets_OR_contribution_by"/>
          <owl:allValuesFrom rdf:resource="#MGED_timegoal"/>
          <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
      </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
        <owl:Restriction>
          <owl:onProperty rdf:resource="#gets_AND_contribution_by"/>
          <owl:allValuesFrom rdf:resource="#MGED_timegoal"/>
          <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
      </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="#MGED_timegoal">
<rdfs:subClassOf rdf:resource="#NFR_timegoal"/>
<rdfs:subClassOf>
        <owl:Restriction>
                <owl:onProperty rdf:resource="#is_a"/>
                <owl:allValuesFrom rdf:resource="#MGED_ontology"/>
        </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
        <owl:Restriction>
          <owl:onProperty rdf:resource="#gets_OR_contribution_by"/>
          <owl:allValuesFrom rdf:resource="#MGED_timegoal"/>
          <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
      </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
        <owl:Restriction>
          <owl:onProperty rdf:resource="#gets_AND_contribution_by"/>
          <owl:allValuesFrom rdf:resource="#MGED_timegoal"/>
          <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
      </owl:Restriction>
```

```
</rdfs:subClassOf>
</owl:Class>
```

An instance of the representation of NFR timegoals in an OWL TIG is:

```
<root_NFR_timegoal rdf:ID="#tumor angiogenesis">
<is_a rdf:resource="#tumor angiogenesis"/>
</root_NFR_timegoal>

<NFR_timegoal rdf:ID="#Interaction of VEGF to its receptors">
<contributes_OR_to rdf:resource="#tumor angiogenesis"/>
</NFR_timegoal>
```

A website is annotated with a root biological function timegoal *x*, such as "Tumor angiogenesis" [1]. A website is also annotated with one or more genome components (genes or proteins) $a_1..a_N$ that may be decomposed into drug contributions. The website can specify that the components $a_1..a_N$ make a positive or negative contribution to certain NFR timegoals of the OWL TIG for function *x*, thus contributing to the function *x*. If the website's content involves the biological function timegoal *x* contributing to some biomedical condition, then the researcher can annotate the website with the root biomedical condition timegoal *b* and specify that timegoal *x* contributes to timegoal *b* or to some of its offspring timegoals. For instance, failing to satisfice the biological function "Tumor angiogenesis" may contribute negatively to the biomedical condition "Lung cancer". In the case of a website that describes the negative contribution of drug "Avastin" to biomedical condition timegoal "Lung cancer", the website would have the following annotations describing the drug's negative contribution to the timegoal "Interaction of VEGF to its receptors" which eventually contributes negatively to the timegoals "Tumor angiogenesis" and "Lung cancer":

```
<root_biomedical_condition_timegoal rdf:ID="#lung cancer">
    <is_a rdf:resource="#lung cancer"/>
</root_biomedical_condition_timegoal>

<root_NFR_timegoal rdf:ID="#tumor angiogenesis">
    <is_a rdf:resource="#tumor angiogenesis"/>
</root_NFR_timegoal>

<observation_timegoal rdf:ID="#Avastin"/>

<NFR_timegoal rdf:ID="#Interaction of VEGF to its receptors">
    <gets_negative_contribution_by rdf:resource="#Avastin"/>
</NFR_timegoal>
```

It should be pointed out that the timegoals and contributions mentioned above require the corresponding OWL TIGs to be refined enough to allow representing everything. If the TIGs are not specific enough, a researcher can extend them using the online IGIPI tool.

5.2 IGIPI Online

We present an online tool[1] with a library of existing Timegoal Graphs (TIGs) in OWL that allows researchers to look up existing TIGs, refine and reuse them for annotating their websites. This website also contains the OWL schema for the IGIPI framework. Researchers can propose extensions to an OWL TIG that is not refined enough, thus actively participating in the TIGs' evolution [40]. Fig. 4 shows snapshots of browsing the website and proposing TIG extensions.

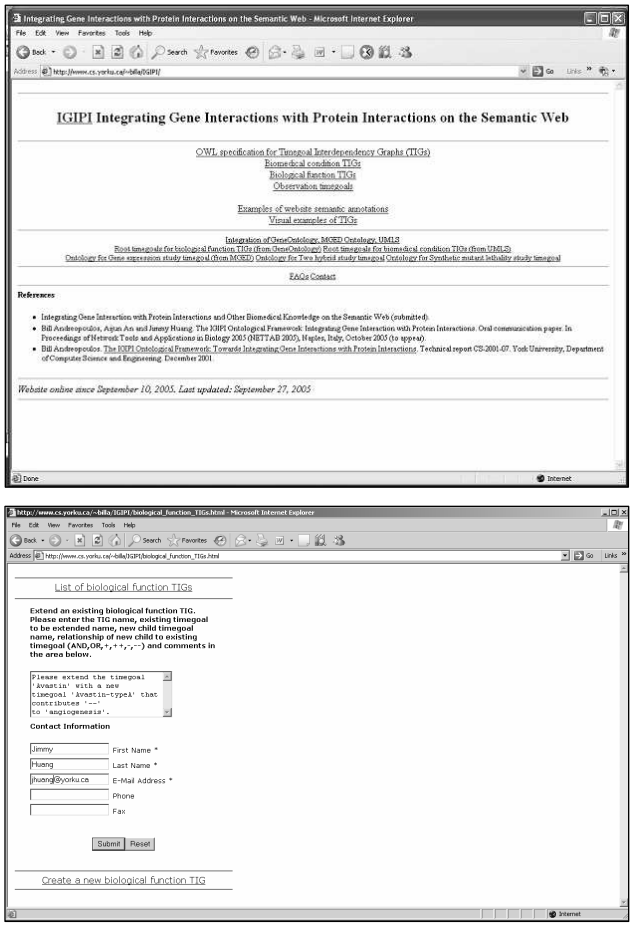

Fig. 4. The website of IGIPI TIGs in OWL allows looking up, refining and reusing the TIGs.

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

We also provide an online visualization tool shown in Fig. 5, employing a Java servlet and applet. This tool parses IGIPI-based XML files and dynamically generates graphs illustrating the file content.

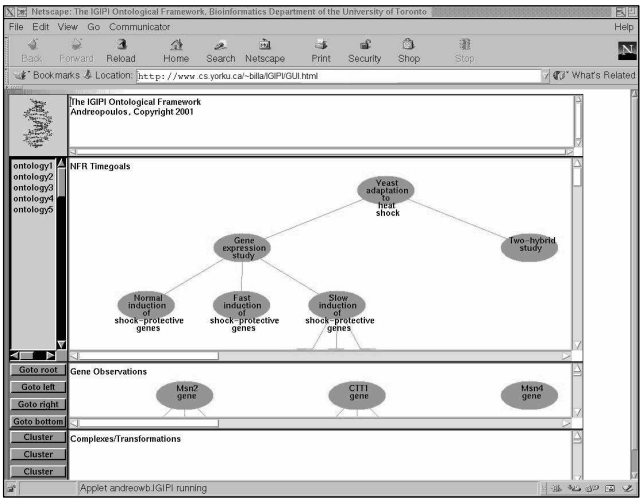

Fig. 5. The IGIPI online tool for TIG graph visualization.

6 Reasoning with Information on the Semantic Web

The ultimate purpose of semantically marking up websites on the basis of IGIPI is to reason on information in a unified manner that could not have been done using traditional online databases. Besides integrating information on the Semantic Web, autonomous agents can potentially use semantically annotated websites for web mining [3,4,17,33,45]. This involves considering all of the contributions (positive/negative, AND/OR) that are propagated between timegoals in Timegoal Graphs (TIGs). This section gives cases of mining the Semantic Web for information that is interesting to researchers.

6.1 Side Effects of a Drug

A physician might be debating whether to prescribe the "Avastin" drug to a patient with a biomedical condition (such as a type of cancer). The physician is considering the potential side effects of the drug on the patient and s/he wants to know about all of the possible side effects of the drug. In Section 4 we described representing side effects as timegoals in biomedical condition TIGs. Recently labs around the world have done various studies on the side effects of "Avastin" and posted the conclusions on annotated web pages,

but the physician has not had time to read all this material and be updated. "Avastin" has different side effects that depend on a patient's genetic makeup, other drugs s/he is taking and the specific type of biomedical condition.

The physician can find on the Semantic Web all known side effects of "Avastin" for similar biomedical conditions. Our approach allows the physician to also consider the genetic makeup of other patients on whom the side effects have been observed and other drugs they were taking. Thus, a physician can derive hints on whether "Avastin" is appropriate for the patient in question. This involves a knowledge reasoning technique of estimating the likelihood of each possible side effect. An agent can find on the Semantic Web all biomedical condition TIGs to which the drug "Avastin" timegoal makes a contribution (via semantically annotated websites). Then an agent analyzes: *a.* the similarity of each TIG to the biomedical condition information provided about the patient, *b.* the overlap between the genes and proteins contributing to each TIG and the genetic makeup of the patient, *c.* the overlap between drugs contributing to each TIG and other drugs the patient is taking.

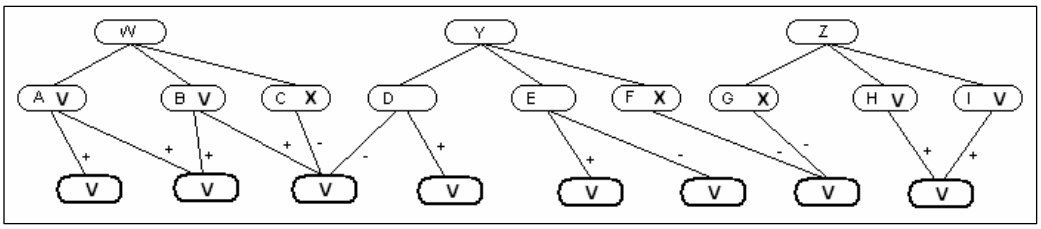

Fig. 6. The TIGs W, Y and Z are identified that are similar to a patient's biomedical condition. The contributions of genes, proteins and drugs are propagated upwards to the side effect timegoals.

To mine the Semantic Web for side effects that "Avastin" might have on the patient in question, the contributions (positive/negative, AND/OR) of genes, proteins and other drugs to side effect timegoals in the TIGs identified are propagated upwards, as we described in Section 4.3. An agent identifies for each of the TIGs identified the timegoals representing side effects that receive the strongest positive contributions and thus are most likely to be satisficed. Fig. 6 shows three TIGs, *W*, *Y* and *Z*, that receive a contribution from "Avastin" and are similar enough to the biomedical condition information provided about the patient. The 'dark' timegoals across the bottom are observation timegoals representing genes, proteins and drugs that overlap with the genetic makeup or drugs taken by the patient. As shown, through propagation of contributions upwards the side effect timegoals *A* and *B* receive the strongest positive contributions and are most likely to be satisficed. The side

effect timegoals *H* and *I* are also satisficed, though not as strongly. The side effect timegoals *D* and *E* are inconclusive. Thus, the physician can consider in his/her diagnosis the genetic makeup of the patient in question and any drugs already taken by the patient.

6.2 Symptoms of a Biomedical Condition

If a physician is considering what biomedical condition a set of symptoms observed in a patient are most likely to correspond to, an erroneous diagnosis might be made, especially if several biomedical conditions involve overlapping symptoms. In Section 4 we described representing symptoms as timegoals in biomedical condition TIGs. An agent can find on the Semantic Web all biomedical condition TIGs with similar sets of symptom timegoals. The agent then analyzes: *a.* the overlap between the genes and proteins contributing to each TIG and the genetic makeup of the patient, *b.* the overlap between drugs contributing to each TIG and other drugs the patient is taking. Through propagating the contributions upwards in the identified TIGs (see Fig. 6 and Section 4.3) the physician can identify which symptom timegoal is most likely to be satisficed. Then, a conclusion can be drawn about which biomedical condition the patient's symptoms are most likely to correspond to. Thus, the physician can consider in his/her diagnosis the genetic makeup of the patient in question and any drugs already taken by the patient.

6.3 Time Point of a Gene or Protein's Contribution

In a situation where a physician is considering what the sequence of contributions of a gene or protein g_0 is likely to be in a TIG where a set of genes or proteins $g_1 \ldots g_N$ are known to be involved, the researcher can find on the Semantic Web other TIGs in which the same or similar set of genes or proteins $g_0 \ldots g_N$ are known to be involved. Using the transformations and complexes as we described in Section 3, the researcher can derive conclusions about the most likely time points of g_0's involvement in the TIG.

7 Conclusions and Future Work

We have presented a novel framework for integrating biological and biomedical information on the Semantic Web. This framework supports any researcher's goal of being able to integrate his/her latest research results with existing information on the Semantic Web, through annotating the research results with semantics. This framework supports automated reasoning upon information on the Semantic Web which provides many benefits, such as allowing a physician to find likely side effects of a drug, or relate observed

symptoms to a known biomedical condition, or derive hints on genes' and proteins' roles in biological functions. The practical utility of this tool is obvious from the fact that it took us several minutes to read abstracts containing yeast and cancer information and represent the information as Timegoal Graphs (TIGs encoded in OWL). We allow users to reuse existing TIGs representing information on biological functions and biomedical conditions. Moreover, we allow users to actively participate in expanding and refining the existing TIGs, through an online website that gives a user the ability to easily send us feedback [40]. Our contributions include proposing a novel framework supporting the interoperability of different ontologies and vocabularies on the Semantic Web, including the Gene Ontology, MGED Ontology and UMLS Unified Medical Language System. The Gene Ontology gives values to the root timegoals of biological function TIGs. The UMLS Unified Medical Language System gives values to the root timegoals of biomedical condition TIGs. The MGED Ontology gives values to the subtree under the "Gene Expression Study" timegoal of a biological function TIG. Our contributions also include the ability to represent the relative time points of events.

One important research direction for the future is to design and implement a technology for ranking the hits returned by a Semantic Web search, similar to the Swoogle search engine [16]. Developing, refining and applying IGIPI-based OWL Timegoal Graphs (TIGs) to biological and biomedical information are ongoing tasks. We are developing and applying IGIPI-based OWL TIGs to large 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 [42]. Combining the data from SML studies with previously published genetic interactions from the yeast literature results in very large data sets with 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, as opposed to genes [42]. Our goal is to integrate the protein interactions from yeast two-hybrid studies with the gene interactions from SML studies and DNA microarray gene expression studies. The latter type of data is provided by the BIND database [5], while the former by the yeast lab of the Banting and Best Medical Institute [41].

Acknowledgements

This research was supported in part by a research grant from the Natural Science and Engineering Research Council (NSERC) of Canada and by an Ontario Graduate Scholarship.

References

[1] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9. http://www.geneontology.org/

[2] R. Backofen, M. Badea, P. Barahona, L. Badea, F. Bry, G. Dawelbait, A. Doms, F. Fages, C. Goble, A. Henschel, A. Hotaran, B. Huang, L. Krippahl, P. Lambrix, W. Nutt, M. Schroeder, S. Soliman and S. Will. Towards a Semantic Web for bioinformatics. In: Proceedings of "Bioinformatics 2004", Linköping, Sweden (3rd - 6th June 2004), SocBIN - Society for Bioinformatics in the Nordic countries.

[3] L. Badea and D. Tilivea. Integrating biological process modelling with gene expression data and ontologies for functional genomics (position paper), Proc. of the International Workshop on Computational Methods in Systems Biology University of Trento, 24-26 February 2003 -- Rovereto, Italy. Springer Verlag.

[4] L. Badea, D. Tilivea and A. Hotaran. Semantic Web Reasoning for Ontology-Based Integration of Resources. Principles and Practice of Semantic Web Reasoning, PPSWR 2004: 61-75, Lecture Notes in Computer Science 3208 Springer 2004.

[5] G.D. Bader and C.W.V. Hogue. BIND - a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* 16(5). 465-477 (2000).

[6] A. Bairoch, R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, L.S. Yeh. The Universal Protein Resource (UniProt). *Nucleic Acids Res., 33: D154-159, 2005.* (Protein Information Resource PIR, http://pir.georgetown.edu/)

[7] K. L. Chung, *Representing and Using Non-Functional Requirements: A Process-Oriented Approach*. Ph.D. Thesis, Department of Computer Science, University of Toronto, June 1993.

[8] D. Calvanese, G.D. Giacomo, and M. Lenzerini. Description logics for information integration. In *Computational Logic: From Logic Programming into the Future (In honour of Bob Kowalski)*, Lecture Notes in Computer Science. Springer-Verlag, 2001.

[9] P. Dafas, A. Kozlenkov, A. Robinson, and M. Schroeder. Integrating gene expression data, protein interaction data, and ontology-based literature search. In Werner Dubitzky and Francisco Azuaje, editors, *Artificial Intelligence and Systems Biology*, pages 107--126. Springer, 2004.

[10] L. Ding, P. Kolari, Z. Ding, S. Avancha, T. Finin, A. Joshi. Using Ontologies in the Semantic Web: A Survey.

[11] Z. Ding, Y. Peng, R. Pan and Y. Yu. A Bayesian Methodology Towards Automatic Ontology Mapping, AAAI-05 Workshop on Contexts and Ontologies: Theory, Practice and Applications (C&O-2005), Pittsburgh, PA, July 9, 2005.

[12] A.H. Doan, J. Madhavan, P. Domingos and A. Halevy. Learning to Map between Ontologies on the Semantic Web. In: WWW 2002.

[13] A.H. Doan, J. Madhavan, R. Dhamankar, P. Domingos, A. Halevy. Learning to match ontologies on the Semantic Web. The VLDB Journal 2003;12(4):303–319.

[14] A.H. Doan, J. Madhavan, P. Domingos, A. Halevy. Ontology Matching: A Machine Learning Approach. Springer-Verlag; 2004. p. 397–416.

[15] A. Doms and M. Schroeder. GoPubMed: Exploring PubMed with the GeneOntology. *Nucleic Acid Research*, 33(Web Server Issue):W783--W786, 2005.

[16] T. Finin, Y. Peng, R.S. Cost, J. Sachs, A. Joshi, P. Reddivari, R. Pan, V. Doshi, D. Li. Swoogle – a search and metadata engine for the Semantic Web, in *Proceedings of Conference on Information and Knowledge Management (CIKM04)*, 2004.

[17] Y. Gao, Z. Pan and J. Heflin. LUBM: A benchmark for OWL knowledge base systems. Journal of Web Semantics. 3(2005) 158-182.

[18] C.D. Hafner and N. Fridman. 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).

[19] J. Heflin and J. Hendler. Dynamic ontologies on the web. In *Proceedings of American Association for Artificial Intelligence Conference (AAAI-2000)*, Menlo Park, CA, 2000. AAAI Press.

[20] J. Heflin and J. Hendler. Semantic interoperability on the web. In *Extreme Markup Languages 2000*, 2000.

[21] R. Jansen, D. Greenbaum and M. Gerstein. Relating whole-genome expression data with protein-protein interactions. Genome Res. 2002 Jan;12(1):37-46.

[22] M. Kanehisa and S. Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).

[23] P.D. Karp. An ontology for biological function based on molecular interactions. *Bioinformatics* 16(3), 269-285 (2000).

[24] P.D. Karp, M. Krummenacker, S. Paley and J. Wagg. Integrated pathway/genome databases and their role in drug discovery, Trends in Biotechnology 17:275, 1999. http://biocyc.org/

[25] I.M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I.T. Paulsen, M. Peralta-Gil and P.D. Karp. EcoCyc: A comprehensive database resource for *Escherichia coli*, *Nucleic Acids Research* 33:D334-7 2005.

[26] A. Kozlenkov and M. Schroeder. PROVA: Rule-based Java-Scripting for a Bioinformatics Semantic Web. In E. Rahm, editor, *International Workshop on Data Integration in the Life Sciences DILS*, Leipzig, Germany, 2004. Springer.

[27] M. Krallinger and A. Valencia. Text-Mining and Information-Retrieval Services for Molecular Biology. Genome Biology 2005, 6:224, 2005.

[28] M.S. Lacher and Groh. Facilitating the Exchange of Explicit Knowledge through Ontology Mappings. In: Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference. AAAI Press; 2001. p. 305–309.

[29] M.-P. Lefranc. IMGT, the international ImMunoGeneTics database. Nucleic Acids Res., 31, 307-310 (2003) PMID:12520009.

[30] Z. Li and C. Chan. Integrating gene expression and metabolic profiles. J. Biol. Chem, 10.1074 2004.

[31] E. Mena, V. Kashyap, A. Sheth and A. Illarramendi, OBSERVER: An approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies, Proceedings of the 1 st IFCIS International Conference on Cooperative Information Systems (CoopIS '96), Brussels, Belgium, June 1996.

[32] H.W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. Stumpflen, J. Warfsmann and A. Ruepp. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* 32 Database issue:D41-4, 2004. PMID: 14681354.

[33] S. Mukherjea. Information Retrieval and Knowledge Discovery Utilising a Biomedical Semantic Web. Briefings in Bioinformatics. 6(3) 252-262, September 2005.

[34] J. Mylopoulos and 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).

[35] N.F. Noy. Semantic Integration: A Survey of Ontology-Based Approaches. SIGMOD Record, Special Issue on Semantic Integration 2004;33(4).

[36] R. Pan, Z. Ding, Y. Yu, Y. Peng. A Bayesian Network Approach to Ontology Mapping. In Proceedings of ISWC 2005. Nov. 6 - Nov. 10, 2005. Galway, Ireland.

[37] S. Prasad, Y. Peng and T. Finin. A Tool For Mapping Between Two Ontologies Using Explicit Information. In: AAMAS-02 Workshop on Ontologies and Agent Systems, Italy; 2002.

[38] A.D. Preece, K.-J. Hui, W.A. Gray, P. Marti, T.J.M. Bench-Capon, D.M. Jones, and Z. Cui. The kraft architecture for knowledge fusion and transformation. In *Proceedings of the 19th SGES International Conference on Knowledge-Based Systems and Applied Artificial Intelligence (ES'99)*. Springer, 1999.

[39] D. Quan, S. Martin and D. Grossman. Applying Semantic Web Techniques to Bioinformatics. Technical report, MIT.

[40] D. Quan and D. Karger. Capturing Ontological Information from Users. Technical report, MIT.

[41] C.J. Roberts, B. Nelson, M.J. Marton, R. Stoughton, M.R. Meyer, H.A. Bennett, Y.D. He, H. Dai, W.L. Walker, T.R. Hughes, M. Tyers, C. Boone, S.H. Friend. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287: 873-880 2000.

[42] P. Ross-Macdonald. Functional analysis of the yeast genome. *Funct. Integr. Genomics* 1, 99-113 (2000).

[43] S. Schulze-Kremer. Ontologies for Molecular Biology. *Proceedings of the Third Pacific Symposium on Biocomputing, Hawaii,* World Scientific Publishers, Singapore, pp.693-704. (1998).

[44] P.L. Schuyler, W.T. Hole, M.S. Tuttle, D.D. Sherertz. The UMLS Metathesaurus: representing different views of biomedical concepts. Bull Med Libr Assoc. 1993 Apr;81(2):217-22.

[45] H. Shatkay and R. Feldman. Mining the Biomedical Literature in the Genomic Era: An Overview. Journal of Computational Biology. 10(6) 2003. 821-855.

[46] I. Spasic, S. Ananiadou, J. McNaught and A. Kumar. Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text. Briefings in Bioinformatics. 6(3) 239-251, September 2005.

[47] L. Stein. Integrating Biological Databases. Nature Genetics Reviews. Vol. 4, No. 5, May 2003.

[48] R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N.W. Paton, C.A. Goble and A. Brass. TAMBIS: Transparent access to multiple bioinformatics information sources. *Bioinformatics*, 16(2):184–186, 2000. The TAMBIS Ontology (TaO) http://img.cs.man.ac.uk/tambis.

[49] R. Stevens, C.A. Goble and S. Bechhofer. Ontology-based knowledge representation for bioinformatics. Brief. Bioinform. 1, 398-414 (2000).

[50] O.G. Troyanskaya, K. Dolinski, A.B. Owen, R.B. Altman and D. Botstein. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in S. cerevisiae). Proc Natl Acad Sci USA 100(14): 8348-53, 2003.

[51] H. Wache. Towards rule-based context transformation in mediators. In S. Conrad, W. Hasselbring, and G. Saake, editors, *International Workshop on Engineering Federated Information Systems (EFIS 99)*, Kuhlungsborn, Germany, 1999. Infix-Verlag.

[52] H. Wache, T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hubner. Ontology-Based Integration of Information — A Survey of Existing Approaches.

[53] The EcoCyc ontology http://ecocyc.PangeaSystems.com/ecocyc/ecocyc.html

[54] The RiboWeb ontology http://smi-web.stanford.edu/projects/helix/riboweb.html

[55] The Schulze-Kremer ontology for molecular biology (MBO) http://igd.rz-berlin.mpg.de/~www/oe/mbo.html

[56]The OWL Specification http://www.w3.org/2004/OWL/

Appendix A. OWL Specification of the IGIPI Framework

This Appendix is organized as follows. The first part describes the timegoal class and its subclasses, including biomedical condition timegoal and biological function timegoal. The second part describes the subclasses of biological function timegoal, including NFR timegoal and observation timegoal. The third part describes the timegoals for gene expression study, two-hybrid study and synthetic mutant lethality study. The fourth part describes the specifications of relations that connect the timegoals, such as AND/OR and positive/negative contributions. The fifth part describes the transformations and complexes.

A.1Biological Function Timegoals and Biomedical Condition Timegoals

A timegoal is the parent class of all classes in the OWL Specification.

```
<owl:Class rdf:ID="#timegoal">
<rdfs:comment>A timegoal is a goal that needs to be
              satisficed at a point of time.</rdfs:comment>
</owl:Class>
```

A biomedical condition timegoal extends the timegoal class to represent biomedical condition Timegoal Graphs. It gets AND/OR decomposed into other biomedical condition timegoals and it receives positive/negative contributions from biological function timeoals.

```
<owl:Class rdf:ID="#biomedical_condition_timegoal">
<rdfs:subClassOf rdf:resource="#timegoal"/>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_positive_contribution_by"/>
         <owl:allValuesFrom rdf:resource="#biological_function_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">  0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_negative_contribution_by"/>
         <owl:allValuesFrom rdf:resource="#biological_function_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">  0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_OR_contribution_by"/>
         <owl:allValuesFrom rdf:resource="#biomedical_condition_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_AND_contribution_by"/>
```

```
        <owl:allValuesFrom rdf:resource="#biomedical_condition_timegoal"/>
        <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
```

A root biomedical condition timegoal is the root of a biomedical condition Timegoal Graph. It takes its values from the domain of the UMLS Unified Medical Language System.

```
<owl:Class rdf:ID="#root_biomedical_condition_timegoal">
<rdfs:subClassOf rdf:resource="#biomedical_condition_timegoal"/>
<rdfs:subClassOf>
    <owl:Restriction>
        <owl:onProperty rdf:resource="#is_a"/>
        <owl:allValuesFrom rdf:resource="#UMLS_library"/>
    </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
```

A biological function timegoal extends the timegoal class to represent the timegoals of biological function Timegoal Graphs.

```
<owl:Class rdf:ID="#biological_function_timegoal">
<rdfs:subClassOf rdf:resource="#timegoal"/>
</owl:Class>
```

A.2 NFR timegoals and observation timegoals

An NFR timegoal extends the biological function timegoal to represent the high level goals (experimental/environmental conditions) on a biological function Timegoal Graph. It gets AND/OR decomposed into other NFR timegoals and it receives positive/negative contributions from observation timeoals.

```
<owl:Class rdf:ID="#NFR_timegoal">
<rdfs:subClassOf rdf:resource="#biological_function_timegoal"/>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_OR_contribution_by"/>
         <owl:allValuesFrom rdf:resource="#NFR_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_AND_contribution_by"/>
         <owl:allValuesFrom rdf:resource="#NFR_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
```

An observation timegoal extends the biological function timegoal to represent the low level goals (genomic/proteomic events that need to occur) on a biological function experiment. It gets AND/OR decomposed into other observation timegoals and it contributes positively/negatively to NFR timeoals.

```
<owl:Class rdf:ID="#observation_timegoal">
<rdfs:subClassOf rdf:resource="#biological_function_timegoal"/>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource="#contributes_positively_to"/>
         <owl:allValuesFrom rdf:resource="#NFR_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">  0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource="#contributes_negatively_to"/>
         <owl:allValuesFrom rdf:resource="#NFR_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">  0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_OR_contribution_by"/>
         <owl:allValuesFrom rdf:resource="#observation_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_AND_contribution_by"/>
         <owl:allValuesFrom rdf:resource="#observation_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
```

A root NFR timegoal is the root of a biological function Timegoal Graph. It takes its values from the domain of the Gene Ontology.

```
<owl:Class rdf:ID="#root_NFR_timegoal">
<rdfs:subClassOf rdf:resource="#NFR_timegoal"/>
<rdfs:subClassOf>
    <owl:Restriction>
            <owl:onProperty rdf:resource="#is_a"/>
            <owl:allValuesFrom rdf:resource="#GO_molecular_function"/>
    </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
    <owl:Restriction>
            <owl:onProperty rdf:resource="#gets_OR_contribution_by"/>
```

```
            <owl:hasValue>"#gene_expression_study"</owl:hasValue>
            <owl:hasValue>"#two_hybrid_study"</owl:hasValue>
            <owl:hasValue>"#synthetic_mutant_lethality_study"</owl:hasValue>
    </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
```

A.3 Gene expression study, two-hybrid study and synthetic mutant lethality study

A gene expression study timegoal extends an NFR timegoal to represent the gene expression experimental conditions by which the root NFR timegoal may be observed. It gets OR decomposed into timegoals that get their values from the domain of the MGED Ontology.

```
<owl:Class rdf:ID="#gene_expression_study">
<rdfs:subClassOf rdf:resource="#NFR_timegoal"/>
<rdfs:subClassOf>
    <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_OR_contribution_by"/>
         <owl:allValuesFrom rdf:resource="#MGED_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
    <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_AND_contribution_by"/>
         <owl:allValuesFrom rdf:resource="#MGED_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="#MGED_timegoal">
<rdfs:subClassOf rdf:resource="#NFR_timegoal"/>
<rdfs:subClassOf>
    <owl:Restriction>
           <owl:onProperty rdf:resource="#is_a"/>
           <owl:allValuesFrom rdf:resource="#MGED_ontology"/>
    </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
    <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_OR_contribution_by"/>
         <owl:allValuesFrom rdf:resource="#MGED_timegoal"/>
         <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
<rdfs:subClassOf>
    <owl:Restriction>
         <owl:onProperty rdf:resource="#gets_AND_contribution_by"/>
         <owl:allValuesFrom rdf:resource="#MGED_timegoal"/>
```

```
        <owl:minCardinality rdf:datatype="&xsd;nonNegativeInteger">   0
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
```

A two hybrid study timegoal and a synthetic mutant lethality study timegoal extends an NFR timegoal to represent the two hybrid or SML experiments by which the root NFR timegoal may be observed.

```
<owl:Class rdf:ID="#two_hybrid_study">
<rdfs:subClassOf rdf:resource="#NFR_timegoal"/>
</owl:Class>

<owl:Class rdf:ID="#synthetic_mutant_lethality_study">
<rdfs:subClassOf rdf:resource="#NFR_timegoal"/>
</owl:Class>
```

A.4 Relations

The is_a relation involves domains and ranges of timegoals.

```
<owl:ObjectProperty rdf:ID="#is_a">
  <rdfs:domain rdf:resource="#timegoal"/>
</owl:ObjectProperty>

<owl:TransitiveProperty rdf:ID="#is_a"/>
```

The relation defining AND/OR decompositions involves domains and ranges of timegoals.

```
<owl:ObjectProperty rdf:ID="#gets_AND_contribution_by">
     <rdfs:domain rdf:resource="#timegoal"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="#contributes_AND_to"/>
     <owl:inverseOf rdf:resource="#gets_AND_contribution_by"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="#gets_OR_contribution_by">
     <rdfs:domain rdf:resource="#timegoal"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="#contributes_OR_to"/>
     <owl:inverseOf rdf:resource="#gets_OR_contribution_by"/>
</owl:ObjectProperty>
```

The relation defining positive/negative contributions involves domains and ranges of timegoals.

```
<owl:ObjectProperty rdf:ID="#contributes_positively_to">
     <rdfs:domain rdf:resource="#observation_timegoal"/>
</owl:ObjectProperty>
```

```
<owl:ObjectProperty rdf:ID="#gets_positive_contribution_by">
     <rdfs:domain rdf:resource=”#biomedical_condition_timegoal”/>
     <owl:inverseOf rdf:resource=”#contributes_positively_to”/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID=”#contributes_negatively_to”>
     <rdfs:domain rdf:resource=”#observation_timegoal”/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:ID="#gets_negative_contribution_by">
     <rdfs:domain rdf:resource=”#biomedical_condition_timegoal”/>
     <owl:inverseOf rdf:resource=”#contributes_negatively_to”/>
</owl:ObjectProperty>
```

A.5 Transformations

Transformations representing events across time (as we described in Sections 3.2-3.4) are represented as sequences of complexes, where each complex consists of a set of observation timegoals.

```
<owl:Class rdf:ID="#transformation">
<rdfs:subClassOf rdf:resource=”#observation_timegoal”/>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource=”#contains”/>
         <owl:allValuesFrom rdf:resource="#complex"/>
         <owl:minCardinality rdf:datatype=”&xsd;nonNegativeInteger”>  1
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>

<owl:Class rdf:ID="#complex">
<rdfs:subClassOf rdf:resource=”#observation_timegoal”/>
<rdfs:subClassOf>
     <owl:Restriction>
         <owl:onProperty rdf:resource=”#contains”/>
         <owl:allValuesFrom rdf:resource="#observation_timegoal"/>
         <owl:minCardinality rdf:datatype=”&xsd;nonNegativeInteger”>  1
</owl:minCardinality>
     </owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
```