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

Bill Andreopoulos[ab†], Aijun An[b], Xiangji Huang[b], and Dirk Labudde[a]

[a]Biotechnological Centre, Technische Universität Dresden, Germany
[b]Dept. of Computer Science and Engineering, York University, Toronto, Canada
† williama@biotec.tu-dresden.de

Abstract. Researchers are faced with the challenge of integrating, on the basis of a common semantic web framework, the information on biological processes resulting from genomic and proteomic experimental studies. Researchers would like to integrate the biological processes' roles in larger medical conditions, by taking account of the time dimension. This integration 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 process or medical condition over time. By representing the relative time points of events, this framework represents a biological process based on *how* it might be observed in an experiment over time. IGIPI involves integrating different ontologies and vocabularies, such as GO and UMLS/MeSH. We applied IGIPI to yeast and cancer examples. **Availability:** `http://www.cse.yorku.ca/~billa/IGIPI/`

1 Introduction and Motivation

Biomedical ontologies are often developed in an uncoordinated manner, reflecting hierarchical relations between domain concepts for the purposes of online database annotation and information retrieval. Individual ontologies often are not built for semantic web integration of information from different sources, such as that produced by labs employing different research methods. Ontology development often overlooks the time dimension in biological processes, how to reuse existing ontologies, and consequently ontologies are not interoperable. Ontology interoperability and information integration on the semantic web will support automated analysis, machine processing and reasoning about dispersed online literature, as well as markup of research results online [3]. It is necessary to create a common bioinformatics framework for representing experimental results on biological processes, while integrating the time dimension and previously developed ontologies. Combining these resources is a step towards utilizing their full potential for functional knowledge discovery on the semantic web [2, 9].

We present the IGIPI framework for integrating the time-relative information on a biological process, resulting from genomic and proteomic experimental studies. A *biological process* is a network of gene or protein interactions. Integration involves representing the experimental and environmental conditions associated with different studies, under which a biological process 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. The contributions of biological processes to a medical condition, such as cancer, can also be represented with IGIPI. A *medical 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 conditions that need to be satisfied to observe a biological process or medical 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 conditions as goals contributing to the overall outcome. A separate representation is developed for each biological process and medical condition, based on an OWL Web Ontology Language specification of IGIPI. Researchers can refine and reuse existing representations for semantic markup of websites.

Researchers need to be able to create a complete picture of the cell by integrating the information resulting from different genomic and proteomic studies [15, 13]. One needs to combine the protein interactions observed in two-hybrid studies with the gene interactions observed in synthetic mutant lethality (SML) studies. For this, it is necessary to be able to represent the conditions at different time points under which the protein and gene interactions were observed. Integrating the events observed at the higher cellular level helps to draw more informed conclusions about gene and protein functions and their process involvement, and supports finding knowledge through automated reasoning.

Our objectives in this paper are to provide the ability to represent:

1. a gene/protein inducing a biological process (i.e., a network of gene or protein interactions), while repressing other biological processes.
2. all experimental and environmental conditions under which a biological process was observed.
3. a module of genes/proteins inducing or repressing a biological process.
4. a process consisting of events that changes the module of genes/proteins inducing a biological process, e.g., by attracting more genes to join the module or repelling other genes from the module.
5. the *relative time points* of active modules of genes/proteins and other events in a process.

This paper is organized as follows. Section 2 describes the IGIPI abstractions for biological processes with application on yeast [5]. Section 3 describes extensions of IGIPI for biomedical information with application on cancer. Section 4 discusses analyzing and reasoning with online information. Section 5 puts in the context of related work. Section 6 concludes the paper.

2 Integrating Biological Process Information: 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 process 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 $satisfied$ or not. In the IGIPI framework, $satisfying$ 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).* Figure 1 shows an example of a TIG. A TIG records all timegoals representing goals in experiments that, if satisfied, will lead to observing the root biological process. A timegoal is represented as an oval shape, and 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; 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 process, 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 process 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 processes. This domain is the *GO Gene Ontology biological process.* The root NFR timegoal is decomposed into timegoals that represent more specific information about how the biological process may be observed.

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 satisfied to satisfy 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 satisfied to satisfy the parent. Figure 1 shows that only one of the timegoals for the three types of experimental studies needs to be satisfied, to satisfy 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 satisfy one or more high level NFR timegoals. An observation timegoal is shown as a thick oval shape and represents specific information about a manipulation or an expression of a gene or protein. Observations may be decomposed into more specific observations at a lower level. Observation timegoals make a positive or negative contribution towards satisfying one or more high level NFR timegoals. Figure 1 shows how interdependency links are used to represent an observation

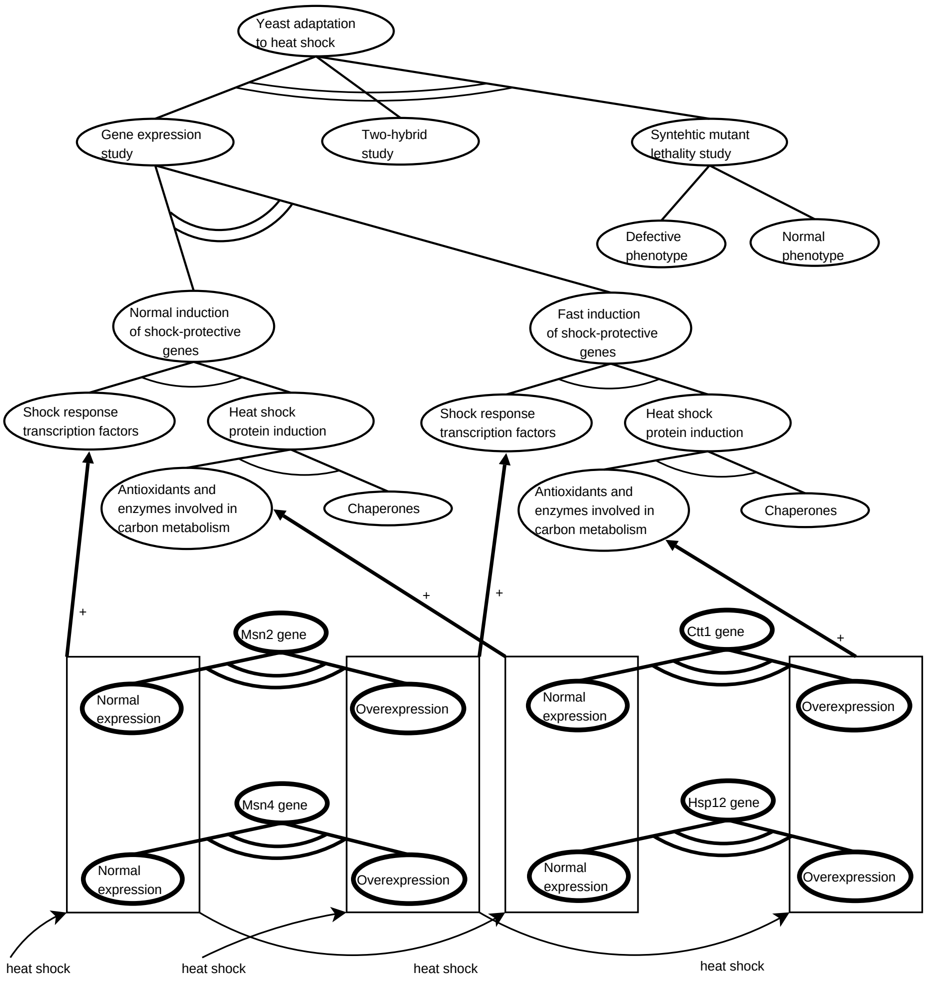
Yeast adaptation to heat shock
Gene expression study
Two-hybrid study
Syntehtic mutant lethality study
Defective phenotype
Normal phenotype
Normal induction of shock-protective genes
Fast induction of shock-protective genes
Shock response transcription factors
Heat shock protein induction
Antioxidants and enzymes involved in carbon metabolism
Chaperones
Shock response transcription factors
Heat shock protein induction
Antioxidants and enzymes involved in carbon metabolism
Chaperones
+
+
+
+
Msn2 gene
Normal expression
Overexpression
Msn4 gene
Normal expression
Overexpression
Ctt1 gene
Normal expression
Overexpression
Hsp12 gene
Normal expression
Overexpression
heat shock
heat shock
heat shock
heat shock

Fig. 1. The biological process Timegoal Graph (TIG) for "yeast adaptation to heat shock" [5]. Thick ovals are observation timegoals and thin ovals are NFR timegoals. This figure 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 satisfied 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 process to be observed; this is the purpose of observation timegoals, shown as thick oval shapes. At the bottom is 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. This figure 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.

timegoal's contribution towards satisfying an NFR timegoal; such a contribution can be positive ('+' or '++') or negative ('-' or '–').

2.1 Transformations

The IGIPI framework deals with 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 [3, 15, 13]. 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 satisfying a high level biological process. Figure 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.

2.2 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 observation timegoals, such as genes or proteins that participate in a transformation simultaneously. Figure 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 satisfying the "shock response transcription factors" NFR timegoal, thus inducing the process "yeast adaptation to a heat shock".

3 Integrating Medical Condition Information

There are web sites with bits and pieces of dispersed information on medical conditions, their symptoms and common side-effects of drugs. Besides building Timegoal Graphs (TIGs) for biological processes, the IGIPI framework can also be used to build TIGs representing information about how medical conditions are manifested. These TIGs can help to integrate biomedical information on the semantic web. We use the term *medical condition timegoals* to distinguish the timegoals of medical condition TIGs from NFR and observation timegoals of biological process TIGs. The root timegoal of a medical condition TIG has a value taken from a domain of medical conditions, such as "ischemic stroke", "haemorrhagic stroke", "lung cancer" etc. This domain is the *UMLS Unified Medical Language System* that integrates 100 biomedical vocabularies. Medical condition timegoals can be decomposed into offspring timegoals, which make an AND/OR contribution to a parent timegoal.Medical condition timegoals may also receive positive or negative contributions from the NFR and observation timegoals of biological process TIGs. An NFR or observation timegoal may contribute positively or negatively towards several medical condition timegoals. The contributions are propagated upwards.

Figure 2 shows an example for "lung cancer". The symptoms of a medical condition and the side-effects of drugs are represented as subtrees of the root medical condition timegoal. An observation timegoal is decomposed to represent how it may be observed under the influence of drugs. Figure 3 shows the propagations of contributions for satisfying timegoals.

4 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 with traditional online databases. Integrating information on the semantic web means that autonomous agents can scan websites and return to a physician any semantically annotated information which s/he is not aware of. This involves considering all of the contributions (positive/negative, AND/OR) that are propagated between timegoals in Timegoal Graphs (TIGs). In this section we discuss a case of finding which medical condition is the most probable cause of a symptom observed in a patient.

As an example, suppose a patient is observed with the symptom *hemoptysis*, the act of coughing up blood (Figure 4). Hemoptysis is an important symptom since it frequently reflects serious underlying lung disease. Hemoptysis is often a sign of lung cancer, but it may be caused by different underlying events in lung cancer patients. Hemoptysis also occurs in patients with acute or chronic bronchitis, as well as tuberculosis and pneumonia. Determining the cause(s) of hemoptysis is often not a trivial matter due to the numerous possible causations and their complexities. A physician could use the semantic web to find if the cause of hemoptysis in a patient is likely to be bronchitis or lung cancer or something else.

5 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. Moreover, they do not usually allow the representation of time in modelling. In this section, we outline relevant work of the past two years. In [1], the GeneOntology is recast in OWL to support reasoning. In [4], a top-level ontology of functions is provided. In [6], ontological relations are proposed for enhancing Knowledge Representation and Reasoning. In [7], integrating protein interaction data using the semantic web is

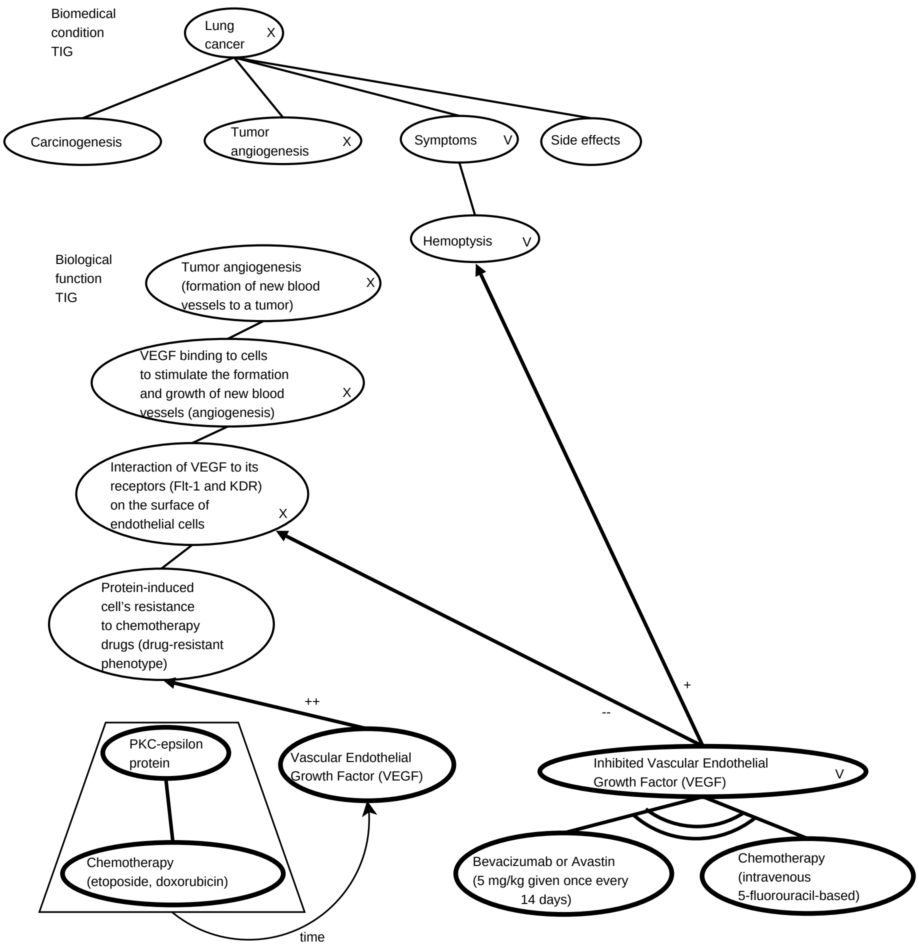
Biomedical
condition
TIG
Lung cancer X
Carcinogenesis
Tumor angiogenesis X
Symptoms V
Side effects
Hemoptysis V
Biological
function
TIG
Tumor angiogenesis (formation of new blood vessels to a tumor) X
VEGF binding to cells to stimulate the formation and growth of new blood vessels (angiogenesis) X
Interaction of VEGF to its receptors (Flt-1 and KDR) on the surface of endothelial cells X
Protein-induced cell's resistance to chemotherapy drugs (drug-resistant phenotype)
++
--
+
PKC-epsilon protein
Chemotherapy (etoposide, doxorubicin)
Vascular Endothelial Growth Factor (VEGF)
time
Inhibited Vascular Endothelial Growth Factor (VEGF) V
Bevacizumab or Avastin (5 mg/kg given once every 14 days)
Chemotherapy (intravenous 5-fluorouracil-based)

Fig. 2. The medical condition TIG for “Lung cancer”. This figure 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. This figure 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. The symptoms of root timegoal “lung cancer” are grouped under an offspring timegoal named “symptoms”. The 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. This figure shows the decomposition of the Vascular Endothelial Growth Factor (VEGF) into the timegoals “VEGF under Bevacizumab” and “VEGF under Chemotherapy”, which represents that the protein is in different states under the influence of Bevacizumab and Chemotherapy. 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 satisfied. This contributes to the root timegoal of the biological process TIG “Tumor angiogenesis”. The biological process “Tumor angiogenesis” contributes to the medical condition TIG that represents information about “Lung cancer”. Since “Interaction of VEGF to its receptors” is not satisfied, this contribution is propagated upwards to timegoals “Tumor angiogenesis” and “Lung cancer”, neither of which is satisfied either.

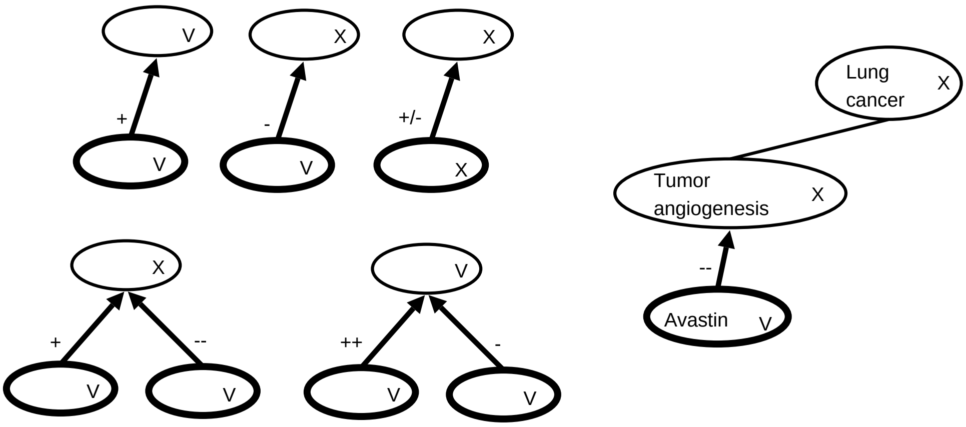

Fig. 3. A negative contribution of the "Avastin" drug observation timegoal negatively affects satisfying "Tumor angiogenesis" and "Lung cancer". The symbol 'V' on a timegoal means that it is satisfied, while the symbol 'X' means that it is not satisfied. "Avastin" is satisfied meaning that this drug is taken by a patient. Contributions from lower timegoals are propagated upwards and contribute towards satisfying 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 "Lung cancer" is not satisfied.

discussed. In [8], it is proposed to combine the semantic web with multi-agent systems for integrated access to biomedical information. In [10], integration of neurodegeneration data using RDF is described. In [12], a method exploiting syntactic patterns and the UMLS semantics for aligning biomedical ontologies is proposed. In [16], a Semantic Web system that facilitates cross-database queries and information retrieval in proteomics is proposed. In [17], leveraging the structure of the Semantic Web to enhance information retrieval for proteomics is proposed. In [11], using web ontology language to integrate heterogeneous databases in the neurosciences is discussed. In [14], advancing translational research with the Semantic Web is described.

6 Conclusion

We have presented the IGIPI knowledge representation framework, which provides a basis for a powerful use of semantic web technologies (OWL and RDF). Our contributions include the ability to represent the relative time points of events. This approach addresses the problem of biological information integration on the semantic web, through Timegoal Graphs (TIGs) that are built dynamically by the biological community on the basis of the IGIPI framework. This approach supports interoperability between ontologies and vocabularies on the semantic web, including the Gene Ontology, MGED Ontology and UMLS

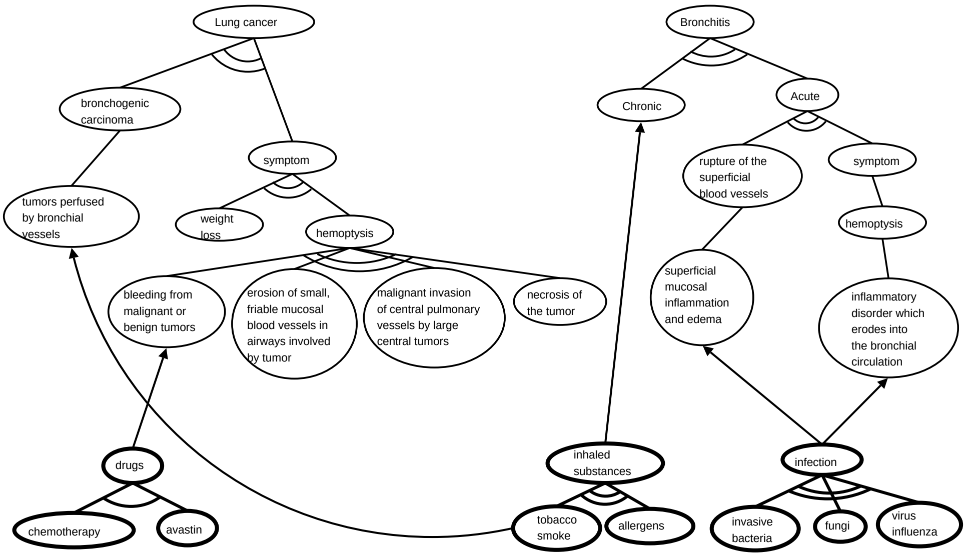
Lung cancer
bronchogenic carcinoma
symptom
tumors perfused by bronchial vessels
weight loss
hemoptysis
bleeding from malignant or benign tumors
erosion of small, friable mucosal blood vessels in airways involved by tumor
malignant invasion of central pulmonary vessels by large central tumors
necrosis of the tumor
drugs
chemotherapy
avastin
Bronchitis
Chronic
Acute
rupture of the superficial blood vessels
symptom
hemoptysis
superficial mucosal inflammation and edema
inflammatory disorder which erodes into the bronchial circulation
inhaled substances
tobacco smoke
allergens
infection
invasive bacteria
fungi
virus influenza

Fig. 4. The TIGs for "Lung cancer" and "Bronchitis" both involve the "Hemoptysis" symptom. hemoptysis could be a symptom of lung cancer or bronchitis and in the former case there could be several causes. Serious hemoptysis often occurs in patients with lung cancer when treated with chemotherapy and Avastin. In this case, the incidence of hemoptysis is relatively high in patients receiving chemotherapy and Avastin, as compared to no cases in patients treated with chemotherapy alone. This figure shows that hemoptysis could also be a symptom of bronchitis; in this case it is mild and self-limited. Bronchitis is often a *viral* or *bacterial* disease which follows a cold or infection. A physician who diagnoses hemoptysis as a symptom of bronchitis may be wrong, since the patient could suffer from lung cancer instead. In fact, physicians who work long hours frequently make such errors. How can a physician tell which of all possible conditions holds for a patient with hemoptysis? With a unifying framework to integrate hemoptysis information online for fast lookup and analysis, a physician could make more informed decisions concerning the underlying cause of hemoptysis in a patient.

Unified Medical Language System. The Gene Ontology gives values to the root timegoals of biological process Timegoal Graphs (TIGs). The UMLS Unified Medical Language System gives values to the root timegoals of medical condition TIGs. Moreover, this approach supports easy incorporation of new research results.This framework can support automated reasoning on the semantic web, which may allow a physician to relate observed symptoms to a known medical condition, or find likely side-effects of a drug.

References

1. M.E. Aranguren, S. Bechhofer, P. Lord, U. Sattler, R. Stevens. Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. BMC Bioinformatics. 2007 Feb 20;8:57.
2. 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, Springer LNCS 3208 2004.
3. O Brazhnik, JF Jones. Anatomy of data integration. J Biomed Inform. 2007 Jun;40(3):252-69. Epub 2006 Sep 24.
4. P Burek, et al. A top-level ontology of functions and its application in the Open Biomedical Ontologies. Bioinformatics. 2006 Jul 15;22(14):e66-73.
5. KH Cheung, et al. YeastHub: a semantic web use case for integrating data in the life sciences domain. Bioinformatics. 2005 Jun;21 Suppl 1:i85-96.
6. K Denecke. Enhancing knowledge representations by ontological relations. Stud Health Technol Inform. 2008;136:791-6.
7. L Dhanapalan, JY Chen. A case study of integrating protein interaction data using semantic web technology. Int J Bioinform Res Appl. 2007;3(3):286-302.
8. F Garcia-Sanchez, et al. Combining Semantic Web technologies with Multi-Agent Systems for integrated access to biological resources. J Biomed Inform. 2008.
9. J Koehler, S Philippi, M Lange. SEMEDA: ontology based semantic integration of biological databases. Bioinformatics. 2003 Dec 12;19(18):2420-7.
10. HY Lam et al. AlzPharm: integration of neurodegeneration data using RDF. BMC Bioinformatics. 2007 May 9;8 Suppl 3:S4.
11. HY Lam HY, et al. Using web ontology language to integrate heterogeneous databases in the neurosciences. AMIA Annu Symp Proc. 2006:464-8.
12. G Marquet, J Mosser, A Burgun. A method exploiting syntactic patterns and the UMLS semantics for aligning biomedical ontologies: the case of OBO disease ontologies. Int J Med Inform. 2007 Dec;76 Suppl 3:S353-61. Epub 2007 May 22.
13. C Pasquier. Biological data integration using Semantic Web technologies. Biochimie. 2008 Apr;90(4):584-94.
14. A Ruttenberg, et al. Advancing translational research with the Semantic Web. BMC Bioinformatics. 2007 May 9;8 Suppl 3:S2. Review.
15. DL Rubin, NH Shah, NF Noy. Biomedical ontologies: a functional perspective. Brief Bioinform. 2008 Jan;9(1):75-90.
16. AK Smith, KH Cheung, KY Yip, M Schultz, MK Gerstein. LinkHub: a Semantic Web system that facilitates cross-database queries and information retrieval in proteomics. BMC Bioinformatics. 2007 May 9;8 Suppl 3:S5. Review.
17. A Smith, K Cheung, M Krauthammer, M Schultz, M Gerstein. Leveraging the structure of the Semantic Web to enhance information retrieval for proteomics. Bioinformatics. 2007 Nov 15;23(22):3073-9. Epub 2007 Oct 7.