Integration of Biological Data with Semantic Networks

Michael Hsing[*,1] and Artem Cherkasov[2]

[1]*CIHR/MSFHR Strategic Training Program in Bioinformatics, Genetics Graduate Program, Faculty of Graduate Studies, University of British Columbia, Vancouver, British Columbia, Canada*

[2]*Division of Infectious Diseases, Department of Medicine, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, Canada*

Abstract: In recent years, the broad utilization of high-throughput experimental techniques resulted in a vast amount of expression and interaction data, accompanied by information on metabolic, cell signaling and gene regulatory pathways accumulated in the literature and databases. Thus, one of the major goals of modern bioinformatics is to process and integrate heterogeneous biological data to provide an insight into the inner workings of a cell governed by complex interaction networks.

The paper reviews the current development of semantic network (SN) technologies and their applications to the integration of genomic and proteomic data. We also elaborate on our own work that applies a semantic network approach to modeling complex cell signaling pathways and simulating the cause-effect of molecular interactions in human macrophages.

The review is concluded with a discussion of the prospective use of semantic networks in bioinformatics practice as an efficient and general language for data integration, knowledge representation and inference.

Keywords: Semantic networks, biological data integration, protein interactions, knowledge representation, ontology, semantic web.

1. INTRODUCTION

The Current State of Biological Data

Recent progress in high-throughput genomics and proteomics has resulted in large volumes of data on protein expression, activation and interactions. One of the major challenges of the modern bioinformatics research is to not only store but also process and integrate biological data to understand the inner workings of cells governed by complex interaction networks.

The diverse biological data on intracellular processes can be conventionally classified into five major groups: sequences of genes and proteins, their expression levels, protein structures, molecular interactions and higher-level cellular functions. For instance, genomic sequences of over 400 organisms determined to date are stored in organism-specific databases such as FlyBase [1], the Saccharomyces Genome Database (SGD) [2] and the Mouse Genome Database (MGD) [3] and in centralized general resources such as Genbank [4]. In addition, protein databases like UniProt [5] archive protein-related information extracted from research articles.

Currently gene expression data are generated at an increasing speed, as can be assessed from the recent progress in the employment of microarray technology [6]. Furthermore the emerging protein-chip technologies [7-9] are expected to permit the large-scale measurement of protein expression levels. In addition, over 30,000 protein structures have been experimentally determined by X-ray diffraction and NMR. The corresponding structural data are stored in databases such as Protein Data Bank [10] and represent invaluable sources for understanding of protein structures, functions and interactions.

Importantly, the successful use of high-throughput protein interaction determination techniques such as yeast two-hybrid [11, 12], affinity purification followed by mass spectrometry [13, 14], and phage-display [15, 16] has shifted research focus from a single gene/ protein to more coherent network perspectives. Large-scale protein interaction data are currently available for a number of organisms including *Helicobacter pylori, Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*, and the data are stored in several interaction databases such as *BIND* [17], *DIP* [18], *IntAct* [19], *GRID* [20] and *MINT* [21] that are all equipped with basic bioinformatics tools for protein network analysis and visualization.

The content of these databases typically complements the experimentally determined protein interactions with ones that are predicted from gene proximity [22, 23], fusion [24, 25] and co-expression data [26-28] as well as those determined by using phylogenetic profiling [29], interologs identification [30] or interacting protein domains [31-34]. Some conventional bioinformatics approaches identify hypothetical interactions between proteins based on their three-dimensional structures [35, 36] or by applying text-mining techniques [37-42]. Currently, no single method is capable of predicting all possible protein interactions, and hence, such integrative resources as *SPRING* [43] and *Predictome* [44] combine multiple theoretical approaches to

*Address correspondence to this author at the *Address correspondence to this author at the Division of Infectious Diseases, Department of Medicine. D 452 HP, VGH. 2733 Heather Street. Vancouver, B.C. Canada V5Z 3J5; Tel: 604-451-9897; Fax: 604-875-4013;
E-mail: mhsing@interchange.ubc.ca

1574-8936/06 $50.00+.00

increase prediction accuracy and coverage. The higher level databases, such as *KEGG* [45], *MetaCyc* [46], Reactome [47], *STKE* [48] and *TRANSPATH* [49] among others, associate networks of interacting proteins with definite cellular processes including metabolism, signal transduction and gene regulation. These resources typically represent biological information in a form of individual pathway diagrams summarizing experimental results collected during years of research on particular cellular function(s).

Although various types of data can be obtained easily from bioinformatics resources, the majority of biological information is only present in unstructured formats such as literature that are difficult to process computationally.

The challenges of Biological Knowledge Representation

The recent developments in experimental methodologies, computational predictions, database technologies and methods of network analysis provide unique opportunities for systemic investigation of complex intracellular processes. At the same time, the rapid accumulation of genomic and proteomic data have made two major bioinformatics problems apparent. The first problem is the lack of communication between different bioinformatics resources whether they are databases or individual analysis programs. Biological data are hierarchical and highly-related (e.g. genes, transcripts, proteins, intracellular compartments, cells and organisms), but yet they are conventionally stored separately in individual databases and in different formats. Thus, even though these databases are often cross-referenced and possess modern searching capabilities, there is still no easy way to handle complex bioinformatics queries requiring answers from multiple sources. For example, a common strategy for answering questions such as "Does [protein A] in [organism A] have a [homologue] in [organism B]?" or "what is the [cellular response] when [protein A] is [inhibited] by [chemical B]" requires building an in-house database to collect data derived from all the relevant sources (e.g. genomic sequence, protein interaction and small molecule databases) and parsing the data into proper formats that are acceptable by different bioinformatics programs (e.g. sequence alignment, network analysis, visualization and simulation tools). Such an approach is very time-consuming and not scalable for other queries of similar nature.

The lack of communication is caused by the second and more fundamental problem in Bioinformatics, namely databases simply store and display biological data in static and often arbitrary tables or pathway diagrams. An automatic response to the above two questions requires computers to understand the *concepts* of each term in square brackets and the *relationships* between those concepts. To enable automatic communication, the specification of such concepts and relationships needs to be shared among the different bioinformatics resources, despite their different data formats or internal synonyms tables.

The problems with representation of biological knowledge can be illustrated by an analogy with the World Wide Web. The current plain-text representation of web pages in an HTML format allows information to be displayed and searched but, on the other hand, the lack of knowledge integration and inference mechanisms hinders the communication between different web services for automatic responses to complicated questions whose answers do no exist in a single web page [50].

The field of bioinformatics requires a common 'biological language' that is capable of representing not only biological data but also the knowledge and logics inherent in the data [51]. There is a wide range of data integration and knowledge representation technologies available to date [52-54], which include relational databases [55], object-oriented programming approaches [56], description logics [57], data warehousing tools [58, 59] as well as the creation of expert - [60], frame - [61, 62], and multiagent systems [63, 64].

A powerful method that is common to many knowledge representation approaches is that known as Semantic Networks (SN). In the next two sections, we present the theory and features of semantic network approaches and briefly review recent SN applications that are relevant to the bioinformatics challenges we faced. We anticipate that semantic networks can offer the required common 'biological language' for establishing communication and asserting the meaning inherent within bioinformatics data.

2. THEORY AND FEATURES OF SEMANTIC NETWORKS

A semantic network (SN) is a graph that expresses abstract knowledge in a form of nodes that designate individual concepts, and connecting edges that represent relationships between the concepts [65-67]. An identity and behavior of each node is defined by its relationships with other nodes, and the topology of a semantic network can represent any type of knowledge describable by natural languages [67].

As introduced by Charles Peirce [68] semantic networks are graphic representations of predicate calculus, which expresses logics in a linear format [52, 67]. A semantic network system is more computationally effective for knowledge retrieval and inference than a pure predicate calculus approach [69], and therefore, this methodology has influenced many areas of computer sciences including artificial intelligence, relational database technology and object-oriented programming, since its first computational implementation in early 1950's [65, 67]. The general properties of semantic networks have also been employed by many logic-based, frame-based or rule-based knowledge representation systems [70].

SN has several features that make it particularly useful for integrating biological data. These features include abilities to easily define inheritance hierarchy between concepts in a network format, allow economic information storage and deductive reasoning, represent assertions and cause-effect through abstract relationships, cluster related information for fast retrieval, and adapt to new information by dynamic modification of network structures [52, 67, 69, 70].

An SN example is illustrated on Fig. **1** where five nodes (referred to as semantic agents) are connected by four pairs of semantic relationships. Agent [Protein A] has a relationship {instance of} with [Protein] agent that, in turn, has the opposite relationship {prototype of} with agent [Protein A]. Thus, the semantic network expresses that the

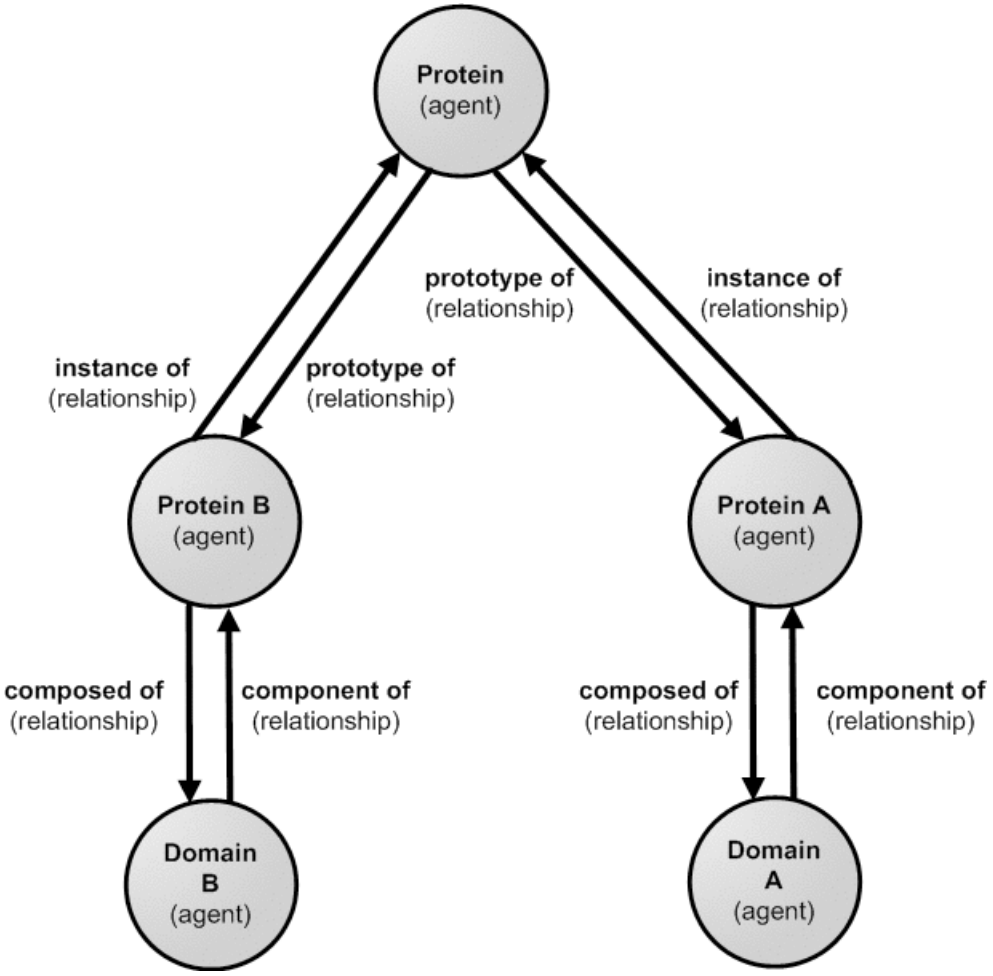

Fig. (1). An example of a semantic network. Characteristics and behaviors of a semantic agent are defined by its relationships with other agents. Semantic agents are represented as nodes, and relationships are depicted as edges. This semantic network expresses the information that Protein A and Protein B are instances of a Protein (a prototype), and they are composed of Domain A and Domain B respectively.

[Protein] agent is the '*prototype*' or '*parent*' of both [Protein A] and [Protein B].

The example in Fig. **1** illustrates important types of relationships in the SN: {prototype/instance} or often referred to as IS-A relationships, which define the hierarchy and inheritance transfer between semantic agents. The ability to specify hierarchy in SN allows common properties to be stored on a prototype, and the properties are inherited to all its members. For instance, a property such as "protein has a three-dimensional structure" only needs to be associated with the prototypical agent [Protein] in the above example, and both [Protein A] and [Protein B] will inherit the same property through the {instance of} relationships with [Protein]. Therefore, the inheritance hierarchy not only allows economic storage of biological information in semantic network systems, but also enables deductive reasoning from the existing information [67].

In addition to inheritance hierarchy (subsumption), other types of relationships can be created abstractly within semantic networks in order to represent assertions or statements about facts. Fig. **1** illustrates that the {composed/component} or PART-OF relationship defines the composition of semantic agents, and thus objects [Protein A] and [Protein B] are {composed of} their constituent components: [Domain A] and [Domain B]. Assertions on various social or physical interactions can be expressed through relationships such as {is a friend of}, {is a father of}, {binds} or {produces}. When more details are required to define the interactions, a high-level relationship such as {produces} is decomposed into a set of low-level agents and relationships. For instance, Fig. **2** illustrates that a [substrate A] agent is linked to a [product A] agent by a {produces} relationship in the first model. In the second model, the {produces} relationship is replaced by a [chemical reaction] agent, which connects the [substrate A] agent by a {from} relationship, and links [product A] *via* {to}. The [chemical reaction] agent allows additional properties such as enzymes and reaction rates to be properly included into the semantic network. Both models can co-exist in a single semantic network system since one model does not exclude the other. Assertional relationships are created among agents with the established hierarchy in semantic networks. For example, [substrate A], [product A] and [enzyme] agents in Fig. **2** all represent instances of [physical objects], while [chemical reaction] is an instance of [event] agents. Thus, the ability to create semantic agents and relationships is very useful to model complex interactions among different components in biological systems.

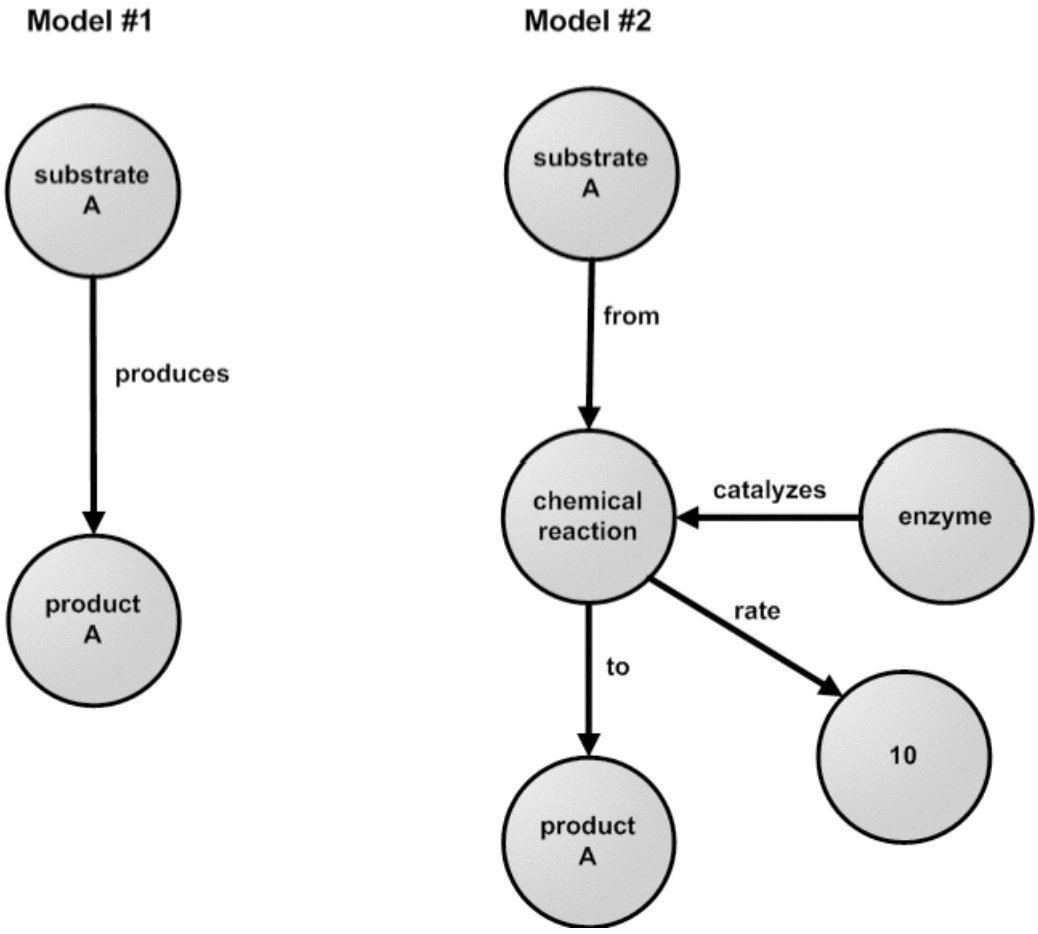

Fig. (2). Assertions are represented by abstract relationships and agents in a semantic network. The figure illustrates two semantic models for chemical reactions.

Another important feature of SN is the ease and speed to retrieve information concerning a particular concept. The use of semantic relationships ensures that related-concepts are clustered together in a network [70]. For example, protein synonyms, functional descriptions, coding sequences, interactions, experimental data or even relevant research articles can all be represented by semantic agents, each of which is directly linked to the corresponding protein agent. Thus, biological information can be retrieved effectively through simple relationship traversal starting from a query agent (such as a protein) in the semantic network [69].

The expressive power of semantic networks has been demonstrated in early artificial intelligence and machine translation projects as well as analyses in philosophy, psychology and linguistics [52]. For example, the *Nude* system, the first computationally implemented SN, is an interlingua or a conceptual language that serves as the medium for translation between two natural languages [71]. The *Nude* established fifty semantic primitives, representing a small set of core concepts where all other concepts were built upon. The KL-ONE (Knowledge Language One) system further established the inheritance hierarchy among the semantic primitives and their individual instances [72]. In addition, SNePS (Semantic Network Processing System) [73] and Conceptual Graphs [52, 74, 75] demonstrated that SN can effectively represent a wide range of natural language semantics, including propositions, if-then statements and logic operators such as conjunction, disjunction, negation and existential quantifiers.

The power of semantic network systems includes not only effective knowledge representation, but also knowledge inference procedures through relationship propagation [52, 67]. Agents that represent propositions can be linked through *implication* relationship, expressing the cause and effect between the agents [52]. For instance, Fig. **3** shows that a propositional agent, [gene A is expressed], implies a proposition, [the cell cycle is initiated], with a probability of 0.8. Semantic networks with implication relationships are considered as causal networks, which can be learned from observed data by applying Bayesian statistics [52, 76].

In dynamic semantic networks such as *Petri nets* [77], PSN (procedural semantic networks) [78] and *Visual Knowledge* [79], passive agents contain data while active agents contain procedures to manipulate the passive (data) nodes by creating/removing nodes or by modifying the existing semantic relationships. Therefore, the semantic networks can respond to new biological information by either modifying the network structures or changing the weights associated with the relationships as demonstrated by neural nets [52].

One of the most ambitious semantic network applications is the Cyc project, which aims to build a knowledge base comprised of all human common sense [80]. Currently, the

Cyc knowledge base contains 3.2 million assertions involving 280,000 concepts constructed by domain experts [81], and the knowledge is represented by the CycL language implementing the fundamental semantic network features [82]. The existing human knowledge in Cyc has been utilized for automatic acquisition of additional knowledge by natural language processing from unstructured texts, scientific literature and the World Wide Web [81].

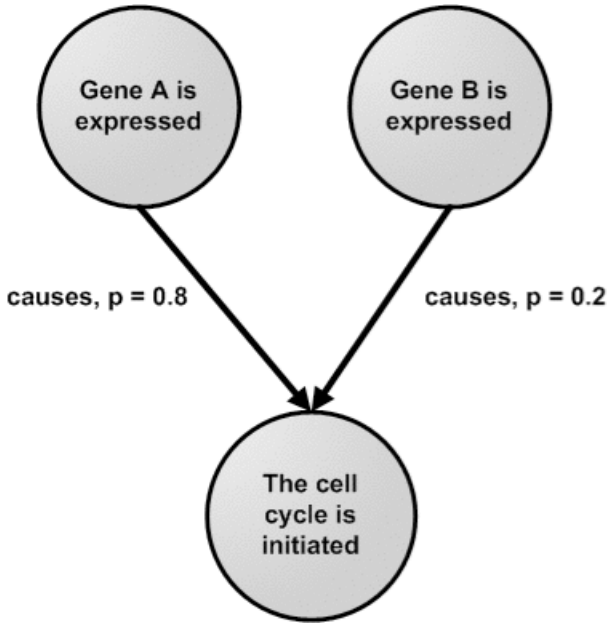

Fig. (3). Cause-effect relations in semantic networks.

3. RECENT APPLICATIONS OF SEMANTIC NETWORKS

Recently, several projects have attempted to organize and integrate biological knowledge using SN-like systems, although the formal methodology of semantic networks is not always recognized by these approaches. One example is biological ontologies, which have been actively researched in several biomedical knowledge domains.

Biological Ontology

An 'ontology', as defined in philosophy and information science, is a formal specification of a conceptualization [83]. In other words, ontologies represent the knowledge stored in a semantic network system, and the implied knowledge inherent within the ontologies can be retrieved by using automatic reasoning methods enabled by SN [52, 67]. Therefore, the first step of representing biological knowledge in SN is to establish an ontology specific to the domain of interest.

The Open Biomedical Ontologies (OBO) consortium [84] houses several well developed ontologies covering such topics as mammalian phenotypes [85], cell types [86] and mouse anatomy [87]. One of the most widely used ontologies is Gene Ontology (GO), a 'controlled vocabulary' for terms that describe functions of genes [88]. Gene Ontology organizes biological terms into three direct acyclic graphs for 'Cellular Component', 'Molecular Function' and 'Biological Process' respectively, where each term is linked to another through 'IS_A' or 'PART_OF' relationships. In its current form, GO provides a convenient and standardized list of vocabularies for gene annotation in databases, but suffers issues with circularity, inconsistence and incoherence [89]. For instance, the lack of distinction between 'types' and 'instances' in GO has created inconsistent interpretation of the relationships such that properties can not always be inherited correctly from a parent to a child node [89]. In addition, each of the three GO trees is entirely separated from each other as no relationship is allowed between terms of different trees. It has been argued that the focus of most biological ontologies to date is on rapid incorporation of new terms, but not on software implementation or correct logic expression of the terms and relationships [89]. Several efforts have been taken to transform biological ontologies into more formalized knowledge representation methods. For instance, in response to the lack of clearly defined and consistent relationships in current biomedical ontologies, Smith *et al.* [90] developed Relation Ontology, which formally defines 10 fundamental relationships, considering spatial and temporal aspects of biological phenomena, and useful for improving interoperability of the different ontologies. In addition, the Gene Ontology Next Generation project (GONG) has attempted to assist the automatic curation and delivery of Gene Ontology by translating GO into DAML+OIL, a description logic-based language with richer expression and reasoning capabilities [91].

The currently available biological ontologies have been established by different research groups and stored in individual databases, often complemented by ontology developing and editing programs such as COBrA [92], Protégé [93], Ontolingua Server [94] and Chimaera Ontology Environment [95]. It is feasible to speculate that the integration of the existing ontologies into a single semantic network system would be very useful for efficient queries in multiple biological knowledge domains. As discussed above, SN enables both hierarchical and assertional relationships between the concepts. Therefore domain ontologies can be simultaneously represented and interconnected with each other based on common concepts in an upper-level ontology, implemented in a single semantic network system. Such an approach has been taken by Unified Medical Language System (UMLS).

Unified Medical Language System

The Unified Medical Language System (UMLS) represents the most comprehensive biological ontology resource available to date that integrates more than 60 families of biomedical terminologies [96]. The controlled vocabularies in UMLS include NCBI taxonomy for model organisms, Medical Subject Headings (MeSH) for biomedical literature, OMIM for genetic knowledge bases, and Systematized Nomenclature of Medicine (SNOMED) as well as many established terminologies for anatomical and clinical domains. Notably, Gene Ontology has recently been integrated into UMLS to further expand the coverage on genomic concepts [97].

The Metathesaurus component of UMLS consolidated over 2.5 millions of synonyms from different terminology resources into 900,551 concepts, and the concepts are categorized into 135 upper-level semantic types in the UMLS semantic network [96]. Thus, the semantic network environment provides a universal framework in which

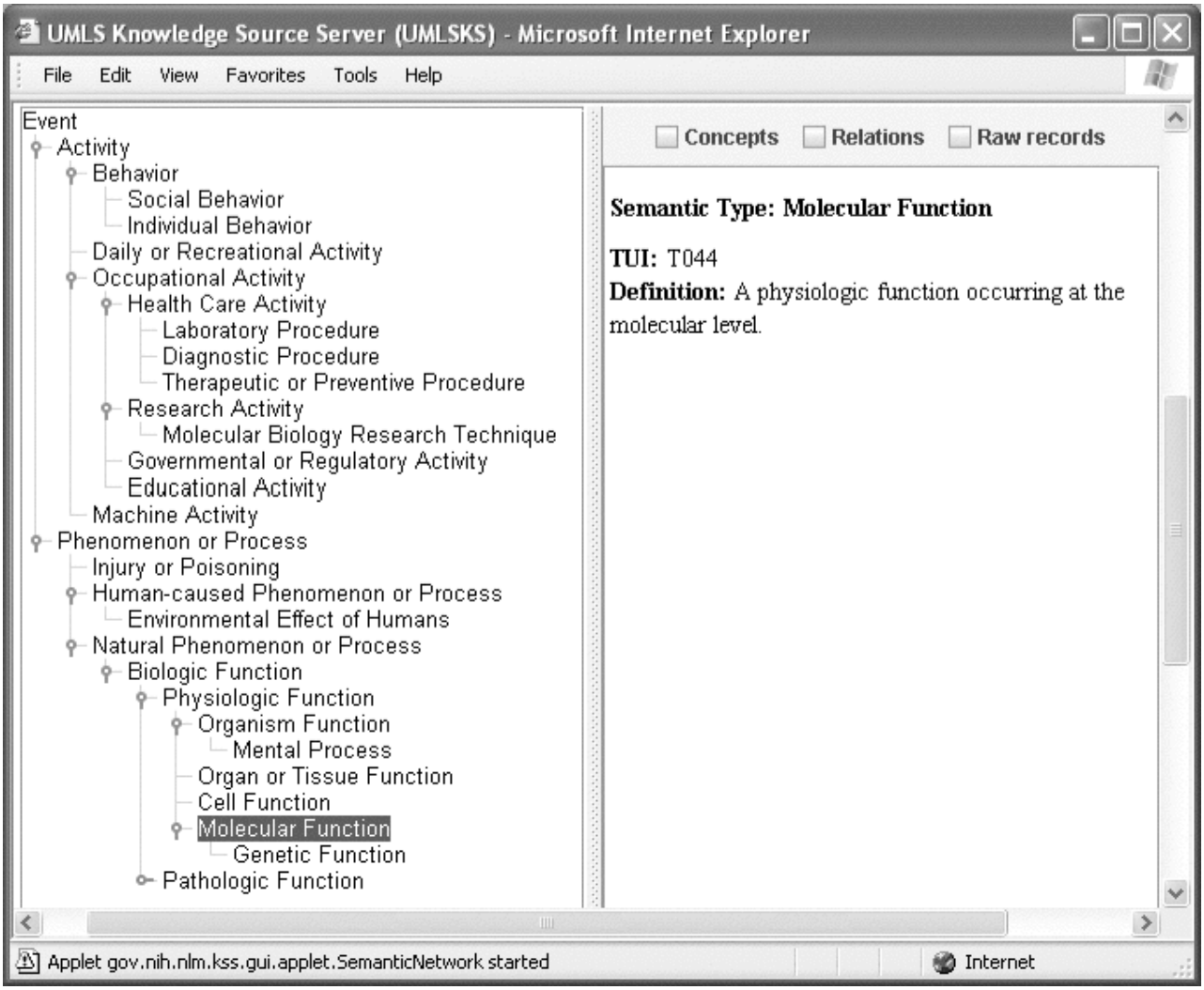

Fig. (4). The semantic types in the UMLS semantic network (release: 2005AB) were queried through the Knowledge Source Server (version 5.0). The left frame represents semantic types as words, connected by lines indicating {isa} relationship. The higher level semantic type is shown on the left of a lower type in the hierarchical browser. For instance [Behavior] {isa} [Activity], and [Activity] {isa} [Event]. The right frame shows the definition of the highlighted semantic type, [Molecular Function].

concepts from different biomedical domains are represented and related through semantic types and relationships [98, 99]. The semantic types are classified into two major hierarchical groups: [entities] and [events], while {is_a} and {associated_with} represent the two basic semantic relationships in UMLS [99]. Similar to other semantic network conventions, {is_a} relationships are used to define the inheritance hierarchy among semantic types, and assertions are established through {associated_with} relationships, which are further categorized into 54 lower-level relationships such as {part_of}, {location_of}, {affects} and {performs} [99]. The hierarchy for a subset of UMLS semantic types is illustrated on Fig. **4**.

Although the UMLS is not free of ontological problems such as inconsistence [100, 101] and incorrect relationship assignment [100], the integration of ontologies has been proven useful for extracting functional information of biological entities and mapping free-text to concepts from literature with natural language processing [102, 103]. Similarly, Gene Ontology has been utilized for functional annotation of proteins [104] and phenotypic associations of genes [105, 106] through literature text-mining and ontology mapping. With the increased awareness on issues related to ontological design [107] and the construction of reference ontologies such as the Foundational Model of Anatomy (FMA) [108] for correlating and validating different ontologies, we anticipate ontological development will be improved and continue to play an important role in integrating biological data.

Semantic Web

In addition to biological data integration and knowledge representation, the importance of semantic networks has recently been recognized for the World Wide Web [50, 51, 109, 110]. The 'Semantic Web' is proposed as a new technology to represent meanings in a web page and to enable communication of the meanings between web pages and services [50].

The web in its current form represents information as structured-texts in HTML (Hypertext Markup Language) or

in XML (Extensible Markup Language) with additional 'tags' for grouping information in a document. Both these formats store information in the form of natural languages, that are enabled to connect web pages through hyperlinks attached on certain key words. Such representation is sufficient to display information to humans, but limited for knowledge integration from multiple sources computationally [110].

Therefore, the RDF (Resource Description Framework) has recently been developed as an alternative language for the Semantic Web [50, 110]. The RDF represents information in sets of the triples, each of which contains a subject, a property and an object. Thus, a triple is a linear representation of an 'agent-relationship-agent' graph in a semantic network. The difference in RDF is that each of the subject, property and object can be represented by a URI (Universal Resource Identifier). The use of URI not only links web pages or resources in a meaningful way (as the way they are related in RDF), but also allows the definitions of concepts and relationships to be precisely described and shared on the web [50].

The advantage of the Semantic Web is that the information is now machine-readable and can be exchanged among web pages that rely on a common specification of concepts and relationships (i.e. ontologies). Thus, software agents can automatically collect relevant information from multiple web pages in response to a particular task (such as the question "Does [protein A] in [organism A] have a [homologue] in [organism B]") and pass the collected information to proper web services for further analyses.

Recently, the Semantic Web technology has been applied to bioinformatics by ^{my}Grid [111] and BioMOBY [112-114] to establish seamless and interoperable communication between bioinformatics service providers (e.g. genome sequence centers) and service consumers (e.g. individual labs or biologists). The communication is achieved through domain ontologies that not only standardize the biological concepts in service providers, but also allow semantic description of their service types. Thus, software agents can automatically discovery which service providers best satisfy the need of service consumers. Currently, the domain ontologies and service descriptions are written in RDF-based semantic languages such as OWL (Web Ontology Language), derived from DAML+OIL [115]. OWL represents information in a machine-friendly format and enables automatic reasoning through applications, such as Racer [116], Pellet [117] and KAON2 [118], that check consistency, completeness and redundancy in ontologies.

Using OWL, BioMOBY successfully integrated several on-line plant genome databases and analytical services in PlaNet, and provided a common user interface and web display for data access [114]. New services can be easily incorporated with the existing partners through the BioMOBY registry system without further coordination or reprogramming [114].

An essential advantage of the OWL is that it reinforces the expression of ontologies in terms of description logics, enabling efficient data communication among databases and analysis programs. For example, BioPAX [119] has recently utilized the OWL format in order to establish an ontology of biological pathway data and, on one hand, to enable exchange of pathway models between such databases as BioCyc [120] and Reactome [47] and, on another hand, to utilize inference engines for automatic check of models consistency. In addition, pathway data organized according to the BioPAX format can be easily browsed and edited using standard ontology editors such as Protégé [93].

Other contemporary Semantic Web applications to Bioinformatics include FungalWeb [121], YeastHub [122] and BioDASH [123]. Projects such as FungalWeb and YeastHub represent genomic and proteomic data using Semantic Web languages including OWL-DL and RDF; the data are stored in RDF-compatible database systems (Sesame [124], Kowari [125], Triplestore [126]), queried by the corresponding database query languages, and analyzed through reasoners such as Racer [116]. By contrast, BioDASH [123] uses Semantic Web technologies to integrate biochemical data on biological entities such as gene, proteins and compounds with drug development processes involving target identification, validation, drug testing and clinical trials. The BioDASH approach has been demonstrated on a drug target, Glycogen Synthase Kinase 3 beta, known to associate with diseases such as diabetes type 2 and Alzheimer's.

The wide variety of software that support different stages of ontology development including creating, populating, deploying, validating, evolving and maintaining have been reviewed by Pollock, J. T. and Hodgson, R. [53]. In addition, multi-agent systems have been applied for information gathering in bioinformatics; for instance, BioMAS automatically annotate genomic sequences from multiple bioinformatics sources by using software agents for information retrieval, integration, analysis and display [64].

All such semantic approaches aim to achieve 'semantic interoperability', a dynamic computational capability to integrate and communicate both the explicit and implicit meanings of digital content without human intervention [53].

Application Development and Advanced Biological Modeling

The power of semantic networks has recently been demonstrated in the fields of software development and modeling complex intracellular systems. For instance, the Visual Knowledge (VK) platform [79] utilizes the SN in combination with other contemporary computer-science approaches including the set theory, frame system, object-oriented modeling and multi-agent system to successfully create a declarative application development and executable semantic environment. Previous VK applications include corporate enterprise systems, flight scheduling and hardware maintenance, integrated currency exchange boards and various business applications [79]. Currently in the 5th generation of knowledge computing, VK operates by using several fundamental classes of semantic agents such as physical things, events, transformations and operations, some of which are presented in Fig. **5**.

Within the VK environment, each semantic class contains its own computer-codes and a unique set of constraints that define the intrinsic behaviors of all its instances. Therefore, each semantic agent is implemented as a reusable and active

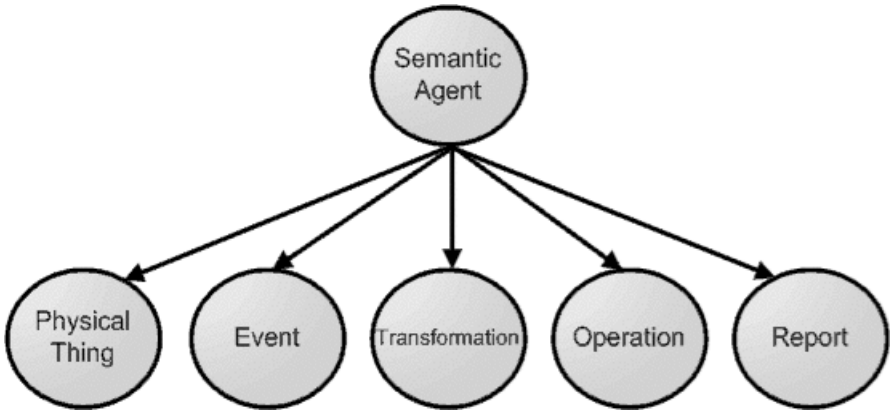

Fig. (5). The basic classes of semantic agents in Visual Knowledge. Semantic agents in the Visual Knowledge environment are classified according to their common properties and functions [79]. A semantic agent of the class 'physical thing' models a physical object that has a shape and occupies space. An agent of the class 'event' represents a phenomenon or a change that occurs on a physical object over a period of time. To enable application development such as a website, the VK contains application-specific agents including transformations, operations, and reports. A transformation agent creates other agents and modifies their relationships. An operation agent searches and collects other agents with certain properties, and a report agent displays the results in an application.

software agent with instructions to respond automatically when it is connected to other agents.

The overall infrastructure of the VK environment is consisted of three interdependent layers. In the bottom layer, data are organized in a form of a semantic database, and the basic interactions with the database are established through C^{++} and Smalltalk programs. To maximize the speed of data transformation among disk, memory and developer interface, the native format of data in the semantic database closely resemble their final form in a client program.

The middle layer of the VK represents a declarative application development interface that enables developers to conveniently create any semantic networks and computer applications without computer-code writing, but by simply dragging and dropping agents in-and-out of their relationships [79].

The top layer contains web-based applications that allow end-users to interact (create, query, edit, analyze, simulate) with the semantic models and data through a web user-interface, without the need for training in the VK-based knowledge representation language. In addition, Visual Knowledge provides automatic import tools that integrate data of different formats (plain-text, table, XML, RDF and OWL) into the semantic database. For instance, any ontology or data expressed in the OWL format can be parsed into the VK environment through creation and mapping of semantic agents and relationships.

One distinctive feature of the Visual Knowledge is unification of three essential semantic technology components: 'data', 'query language' and 'inference/ reasoning engine'. The majority of the contemporary data integration approaches separate (physically and concept-ually) these three components and implement each of them not only by different languages, but often by different software systems. For example, the data might be stored in the format of RDF and queried by SeRQL in Sesame, while the reasoning is done in an external program such as Racer that utilizes the data as inputs.

Such separation between data and the query languages can lead to the loss of important information, as queries themselves may contain the essential user-defined knowledge. Importantly, such separation between data and inference engine slows down the I/O flow.

Within the VK environment, data, query and inference engine are all implemented as software agents stored in a single, unified semantic database. This architecture allows very comprehensive capture of all domains of knowledge introduced by ontology builders, application developers and end-users. Such organization enables complex queries and reasoning with the data at runtime. Importantly, because the VK queries are composed of semantic agents and relationships, the queries can form 'derived' relationships that not only are reused in applications but also become part of the domain knowledge.

To address the current challenges in modeling biological systems, a specialized, biology-oriented VK-application package called BioCAD [127] has been developed by Visual Knowledge, Inc. and previously delegated to a bioinformatics company Upstream Biosciences, Inc. in collaboration with our group. The BioCAD environment utilizes the semantic network methodology to integrate the gene expression, protein expression and protein interactions data, and currently contains 30 millions agents and hundreds of millions of relationships (occupying about 1.7 GB of hard drive storage in a personal computer) representing biological entities such as genes, proteins and cells defined from various bioinformatics resources. Fig. **6** shows the graphical interface of a BioCAD client program, which enables creation and manipulation of biological agents stored at a central server.

The BioCAD currently contains over 80,000 prototypical gene or protein agents from *Homo sapiens*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Mus musculus, Rattus norvegicus* and others [127]. Within the BioCAD environment, a gene or a protein agent is connected to different annotation objects derived from the GenBank [4], RefSeq [128], and Gene Ontology [88] (an example of the

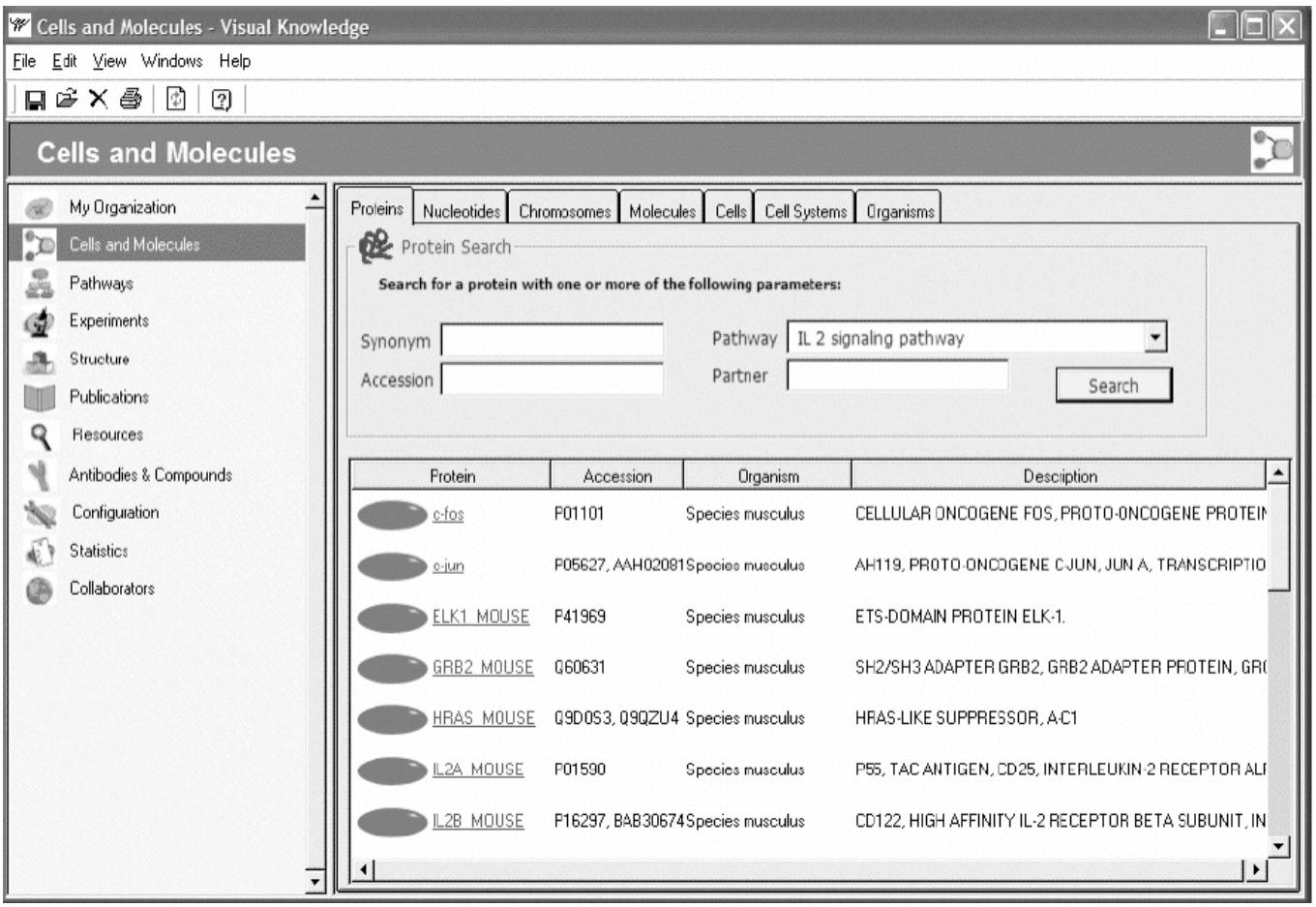

Fig. (6). BioCAD - integrative environment for biological data. A BioCAD client program allows users to create and manipulate a variety of biological agents including genes, proteins, chromosomes, cells, cell systems and organisms. In addition, BioCAD enables pathway modeling and experimental data management by semantic networks. The screenshot was obtained from the BioCAD program with permission from Visual Knowledge, Inc.

BioCAD annotation for the PIK3R1 protein agent is featured on Fig. **7**).

The experimental data on gene expression, protein expression or protein interactions obtained from other resources, as tables, XML records, RDF or OWL files, can be readily imported into the BioCAD database and converted into semantic agents and relationships. For instance, protein-protein interactions determined from the yeast two-hybrid experiments are represented in the BioCAD as 'event' agents connected to the protein prototypes.

In addition, we have assigned the protein agents implemented in the BioCAD to protein domains defined by the *Pfam* [129], *Prosite* [130] and *InterPro* [131] classifications and converted each domain into an individual SN agent. To enable the prediction of protein-protein interactions, the *InterDom* domain information [132] has also been imported into the BioCAD and the corresponding domain-domain interactions have been translated into the event agents.

Hundreds of conventional pathways maps on metabolism and signal transduction have also been integrated into the BioCAD in a form of semantic agents and relationships and can be readily queried by users. The BioCAD graphing capabilities allow visualizations of protein interaction networks enhanced with numerous layers of additional information – an example of a reconstruction of protein interaction networks can be found on Fig. **8**.

Visual Knowledge has successfully demonstrated the application of semantic technologies in application development and biological modeling in the BioCAD. However, there are several challenges that are faced by the VK technology. Visual Knowledge is a multi-user distributed application development environment where changes made by different developers in individual VK databases are packaged and exchanged through a federated change management system. However, certain issues arise when agents within a package from one developer overlap with the existing agents in the database of another developer. An advanced VK visualization tool is clearly required to represent the overlapping agents between a package and a database, and identify the potential impact of importing the package.

At the same time, an automatic ontology mapping tool needs to be created to efficiently integrate the various ontologies stored in the VK. The ontology mapping still represents a challenging open question for the Semantic Web community [133]. Despite the fact, that several standardized ontologies are currently available, it is inevitable that users

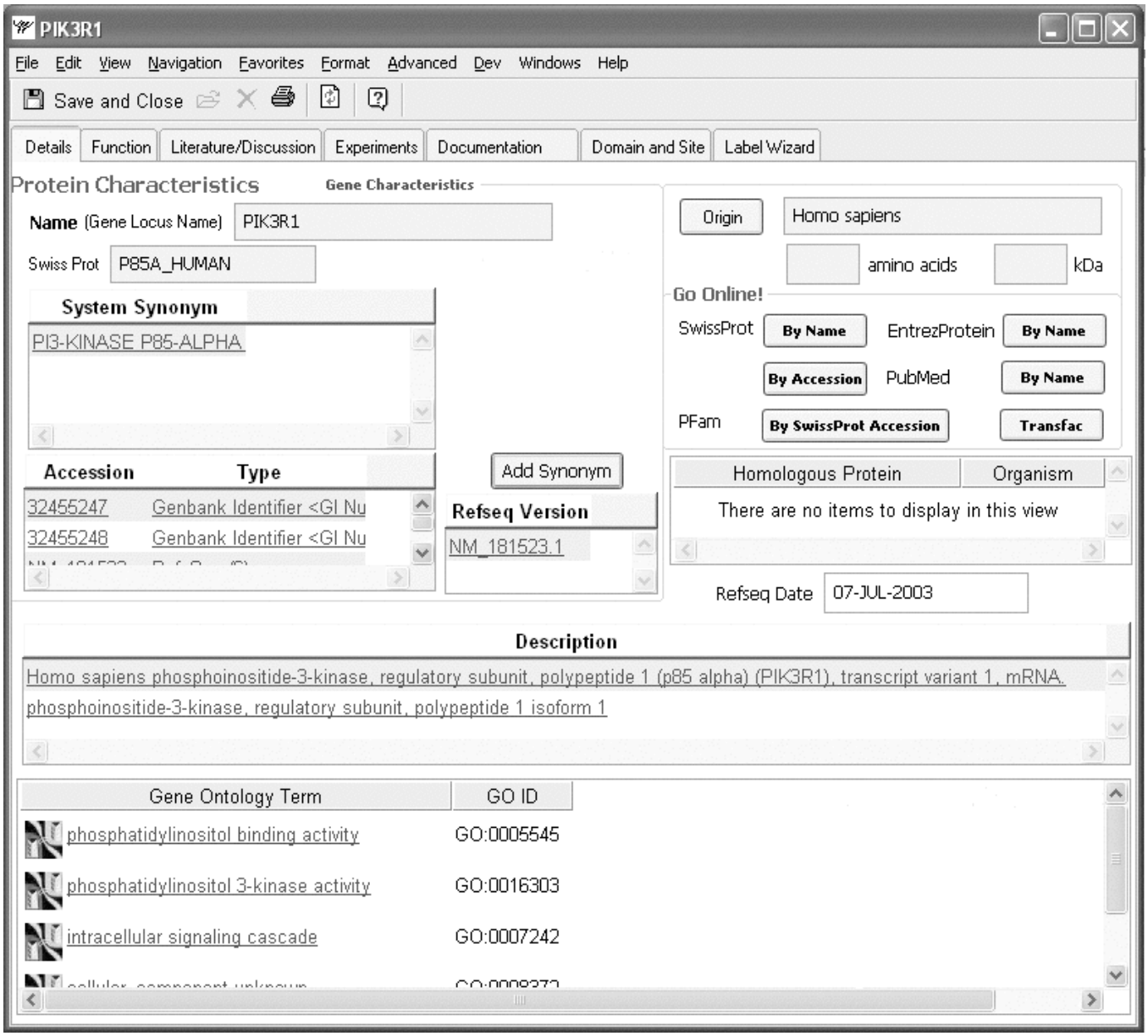

Fig. (7). Protein annotations in BioCAD. The protein detail page for PIK3R1 shows the annotation integrated from various resources such as Genbank and Gene Ontology. Each piece of information such as a gene name, a description or an accession number is represented by a semantic agent in the database. The page is generated by collecting and displaying the relevant semantic agents through dynamic operations. This screenshot was created from the BioCAD program with permission from Visual Knowledge, Inc.

will create personalized versions of ontologies to reflect their particular interests. Thus, the ontology mapping tool that can automatically connect similar concepts and relationships from different ontologies will greatly assist human-curated integration efforts.

A Case Study - Modeling And Simulating Molecular Interactions in Macrophages

The existing biological agents and resources in BioCAD provide an excellent environment for modeling and simulating the mechanistic details of intracellular interactions. Using the BioCAD environment, we have developed and implemented a semantic network capable of representing and simulating complex molecular mechanisms such as PI3K regulation of cell signaling in human macrophages.

Macrophages, which are essential components of human immune system, engulf and digest pathogens by phagocytosis and phagosome maturation – two of many other cellular processes that are regulated by the phosphoinositide 3-kinases (PI3Ks) [134-137]. Previous experimental studies demonstrated that *Mycobacterium tuberculosis* (MTB) can interfere with the PI3K signaling pathways in order to survive within macrophages [138, 139]. Based on the diverse roles of PI3Ks, we hypothesized that MTB's interference with the PI3K signaling could impact numerous other macrophage processes, exceeding those that are currently studied.

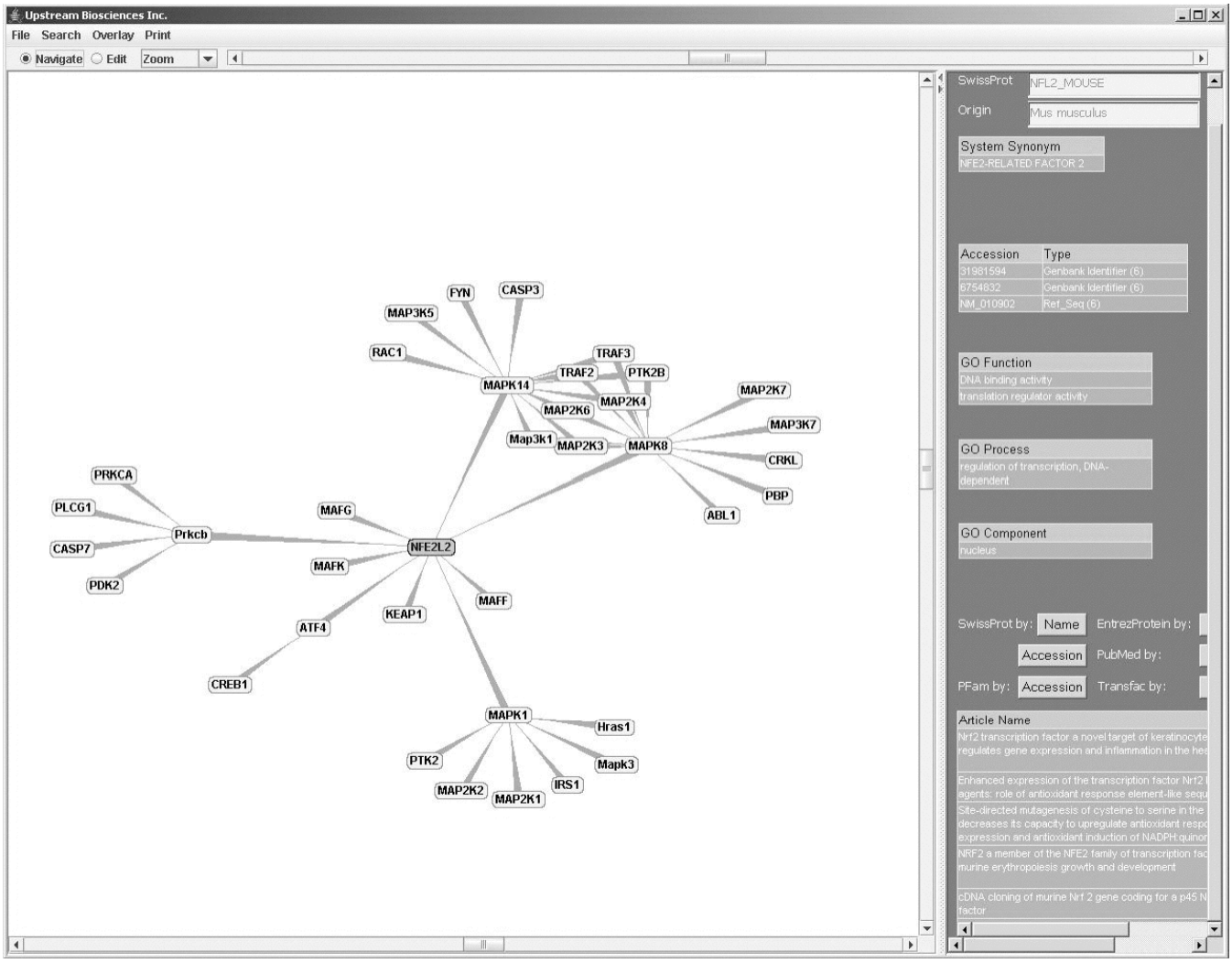

Fig. (8). Visualization of protein interaction networks in BioCAD. The auto-graphic tool in BioCAD visualizes the underlying protein interaction data in a two-dimensional graph, where a node represents a protein agent and a line depicts a physical interaction event, connecting two proteins. This picture shows proteins that directly interact with a NFE2L2 protein (grey node), and proteins that are one interaction away from NFE2L2. Clicking any node displays details about the protein in the right panel. This screenshot was created from the BioCAD's graphic tool with permission from Visual Knowledge, Inc.

We applied the semantic network approach to reconstruct the detailed picture of the PI3K signaling in human macrophages and used the resulting SN-model to predict additional cellular responses caused by the MTB interference [140, 141]. Thus, we constructed the SN of human macrophages where cellular components have been modeled by SN agents of the [physical thing] class, divided into six prototypes: [Cell], [Intracellular Compartment], [Macro-molecule], [Domain and Site], [Small Molecule and Molecular Fragment], and [Atom]. Fig. **9** illustrates that the spatial organization of the different types of structures is established through the {composition} relationships, linking the semantic agents.

To represent molecular interactions inside the macrophage, we created six event prototypes: [localization], [translocation], [non-covalent interaction], [covalent interaction], [allosteric regulation] and [cellular response]. A biological example for each type of events is given in Table **1**.

The mechanistic details and the cause-effect relationships in signal transduction are explicitly modeled through the above semantic events, each of which is composed of multiple semantic agents and relationships. For instance, a non-covalent event consists of two binding events, each connected to the corresponding binding domain, conformational states and binding states. Fig. **10** illustrates an example of a non-covalent interaction, representing the binding between proteins Ras and PI3K-p110, leading to a potential activation on the kinase domain.

An essential feature of signal transduction is the conformational and functional changes of proteins as a result of non-covalent binding with other molecules or chemical modification; however, such cause-effect information is poorly reflected by most interaction databases and pathway maps. In the semantic network, we represented this information by the [allosteric regulation] event agents, modeling proteins as logical, integrating and adaptive devices, where the functional responses of protein domains

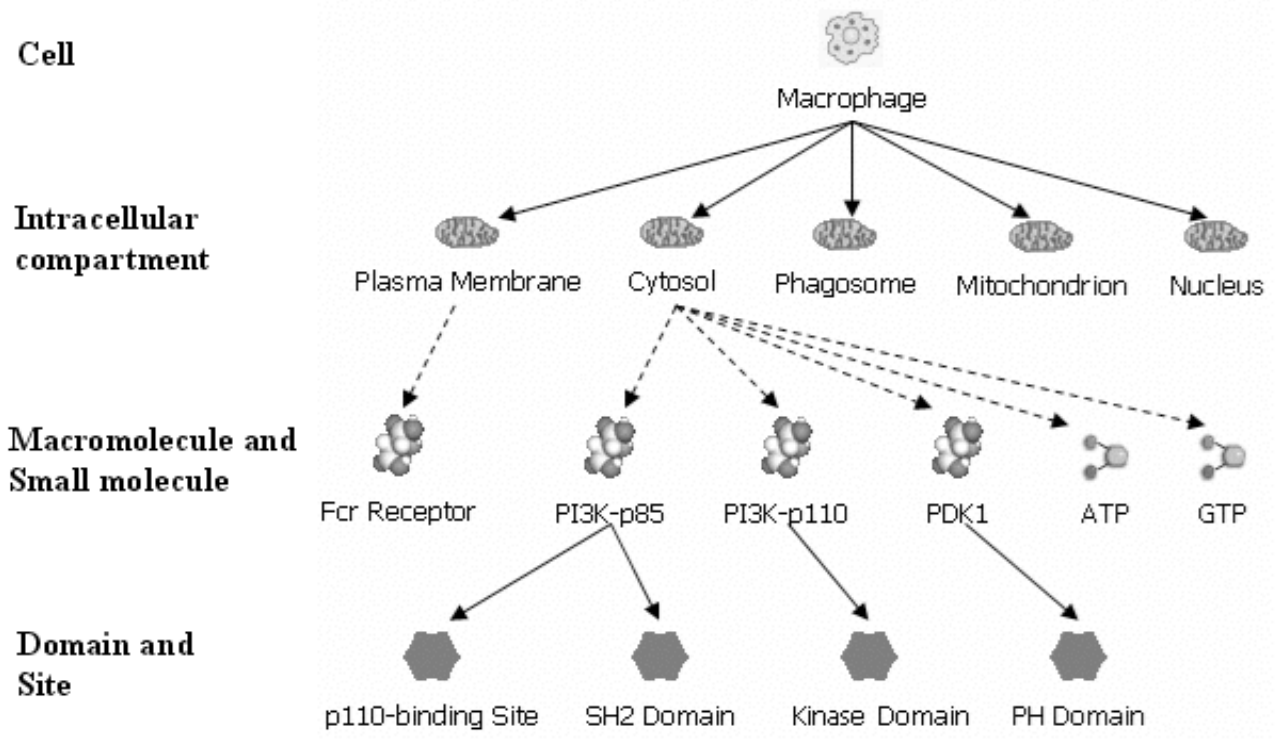

Fig. (9). Spatial organization of intracellular structures in the semantic network. Biological structures are modeled by semantic agents (visualized as icons), which are related to their components by the composition relationships (shown as solid arrows). A human macrophage has been modeled as a semantic agent of the [Cell] prototype, and it is composed of various [Intracellular Compartment] agents, including plasma membrane, cytosol, and nucleus. Each compartment such as cytosol has been linked to [Macromolecule] and [Small Molecule and Molecular Fragment] agents including proteins, ATP and GTP. A macromolecule such as a protein is further composed of [Domain and Site] agents. For simplicity, only one direction of the paired relationships is shown. We used dotted arrows to indicate that there are additional agents and relationships between two agents.

are determined by a set of conditions from other domains (Fig. **11**). The semantic models of the other interaction types are described in detail in [141].

Table 1. Six Major Event Prototypes Represent Interactions Among Biological Structures in the Semantic Network

Semantic Agent - Event	Biological Examples
Localization	A protein is located in the cytosol
Translocation	A protein moves from cytosol to plasma membrane.
Non-covalent Interaction	A ligand binds to a receptor.
Covalent Interaction	An enzyme catalyzes a chemical reaction where substrates are converted to products.
Allosteric Regulation	A ligand binding on site A of a protein causes a conformational change on site B of the protein.
Cellular Response	A qualitative cellular behavior such as cell survival, cell death, phagosome formation, and an increase of intracellular glucose level.

The first column contains the six prototypes, and the second column contains biological examples of the corresponding prototypes.

Thus, the BioCAD SN-framework allowed us to translate information on PI3K signaling in literature and pathway database into semantic agents and relationships. As a result, we collected and placed into the SN-context the bioinformatics data from major public databases and 27 PI3K-specialized research articles, covering 59 prototypical proteins and involving 46 non-covalent interactions, 17 covalent interactions, 27 allosteric regulations and 8 cellular responses. Fig. **12** illustrates one of the interaction maps that summarize the different MTB interference scenarios in macrophages.

To further analyze the dynamic behaviors of molecular interactions in macrophages, a semantic network cell-simulator has been developed by using the active agents such as transformations and operations in the BioCAD system. Molecules in a SN-simulation interact with each other according to their current *conformational states* and *locations*, and any successful interaction is recorded by an event agent connecting the participating entities including 'molecule', 'domain', 'location' and 'time' agents. Therefore, the SN-simulator provides a traceable 'history' of all the events that happened to every molecule and allows a detailed analysis of simulated molecular interactions [140, 141]. To illustrate this point, Fig. **13** features an example of one macrophage simulation run, involving two instances of IGHG3, FCGR1A, LYN and GAB2 molecules, each of which behaves according to its prototype in the semantic network. As can be seen on the figure, at a certain time frame ('time 6'), the SN-simulation identified the occurrence

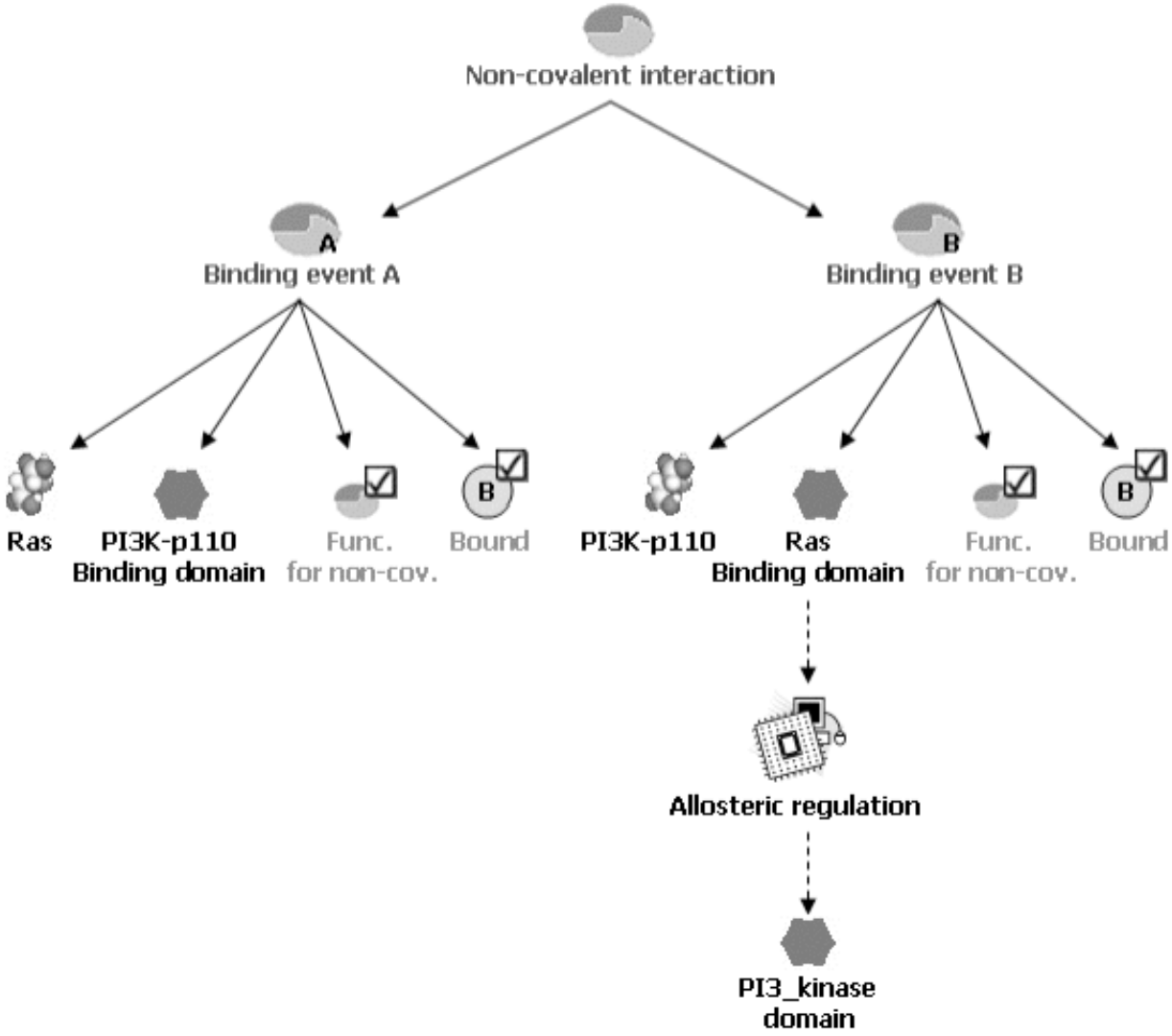

Fig. (10). A non-covalent interaction event models the binding of two molecules. This example features the interaction between proteins Ras and PI3K. The event links the binding domains and their corresponding states.

of 4 non-covalent interactions, 1 covalent interaction, 4 allosteric regulation and 11 translocation events.

Similarly, by running the simulation on the SN-model for human macrophages we were able to traverse along signaling events starting from the MTB surface molecules through the macrophage cell receptors into the downstream interactions and leading to induced cellular responses. Such pathway reconstruction within the SN environment enabled us to simulate four known macrophage responses and identify responses such as increased macrophage cell survival and intracellular glucose uptake that have not yet been appreciated in the literature (Table **2**).

The above application of the SN methodology to macrophage pathways has demonstrated that heterogeneous data for different biological entities such as small molecules, proteins, protein domains, intracellular compartments and cells can be successfully integrated by the use of semantic agents and relationships. Furthermore, the semantic integration enabled effective utilization of different analysis tools such as pathway walk and simulation, which generated testable hypotheses on the relationship between local perturbations (e.g. protein activation) and system responses (e.g. cellular behavior). Such insights on macrophage interactions, which can lead to new experiments, are often difficult to acquire from conventional integration approaches based on interaction tables and pathway diagrams.

We anticipate further semantic integration of other available data from pathway databases, such as BioCyc [120] and Reactome [47], will increase the size of the current macrophage interaction network and enhance our prediction on cellular responses. Future development of the SN methodology will enable simulating much larger numbers of interacting molecules in space- and time-dependent manner, leveraging on several public data sources for gene expression, protein activation profiles, subcellular localization and cause-effect interactions. In addition, a web-based collaborative pathway modeling environment, SNEC (Semantic NEtworks for Cell-modeling), which utilizes the VK semantic technologies for integrating and analyzing intracellular systems, will soon be available to the public.

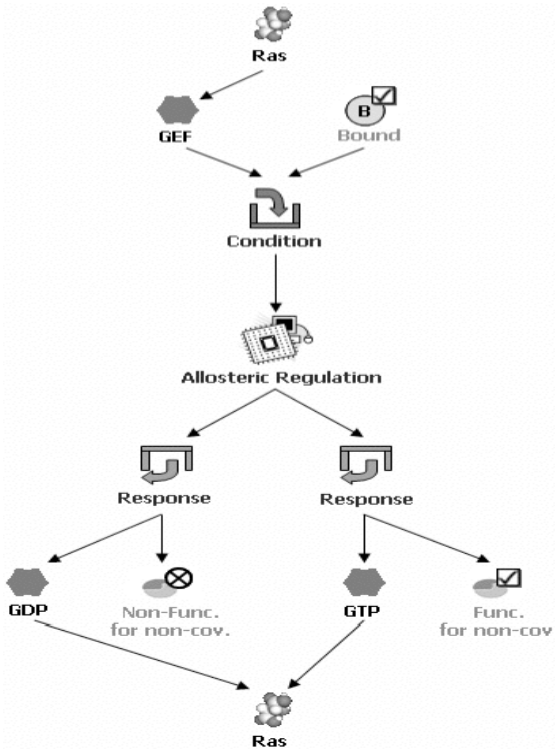

Fig. (11). Allosteric regulation event of the semantic model. An allosteric regulation event agent is composed of condition and response events. Each condition event considers a domain and its conditional state (binding or phosphorylation state). After the conditions are met, one or more response events would change the conformational states on other domains. The illustrated allosteric regulation indicates that the binding of GEF domain on Ras protein will 1) inhibit the function of GDP domain and 2) activate the function of GTP domain in Ras.

4. CONCLUSION

The rapid advance on genomic and proteomic experimental techniques and computational prediction methods has created a massive amount of biological data that are currently distributed in databases with different formats. The challenge of modern bioinformatics is to integrate heterogeneous data by utilizing formal knowledge representation techniques and establishing communication among various bioinformatics services.

The methodology of semantic networks allows active representation of any domain knowledge by abstract concepts and relationships. While the expressive power of semantic networks enables representation of inheritance hierarchies and assertions among different biological concepts, SN operations facilitate fast information retrieval, dynamic manipulation of network structures, and automatic knowledge inference and reasoning.

Semantic network systems have already shown their broad applicability in various fields of computer sciences including artificial intelligence, object-oriented programming, and database technologies. Recent applications have further demonstrated the significance of SN methodology in the construction of biological ontologies, integration of biomedical knowledge, communication among bioinformatics web services, application development and intracellular simulation. We anticipate further broader use of semantic networks in the field of bioinformatics as an efficient and general language for data integration, knowledge representation and computational prediction.

ACKNOWLEDGEMENTS

We acknowledge Visual Knowledge, Inc. and Upstream Biosciences, Inc. for the software and database support during the macrophage case study. We thank Conor Shankey

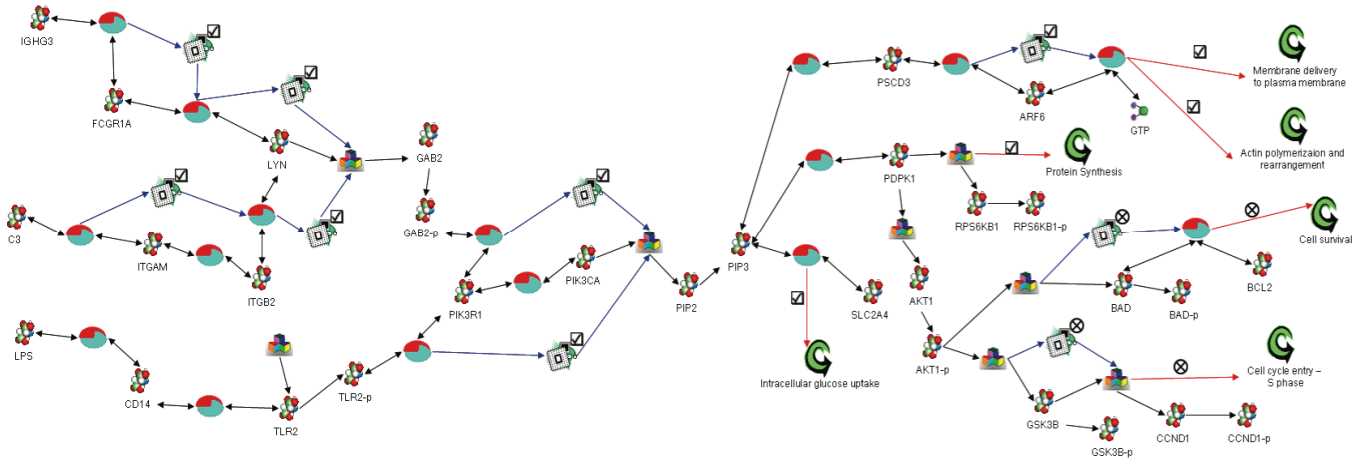

Fig. (12). The interaction map was generated by traversing among the different molecules in the macrophage model. The starting points of this graph are the three MTB surface molecules IGHG3, C3 and LPS. All the arrows represent 'derived relationships', which have consolidated several semantic agents and relationships. The black double-headed arrows are used to connect binding molecules to their non-covalent interactions (the double-headed arrow indicates the dual directionality in the interaction). The black single-headed arrows represent the connections among enzymes, substrates, and products. The blue arrows connect allosteric regulations to the molecular interactions. The map shows the cause-effect connections from the three surface molecules on MTB to the production of PIP3 in the macrophage. The map continues from the PIP3 downstream interactions to various cellular responses. The red arrow with a 'check' sign represents 'promoting' relationship (also a derived relationship). The red arrow with a 'cross' sign represents 'inhibitory' relationship. Icon definitions: Red/blue circle = non-covalent interaction event; coloured cubes = covalent interaction event; computer chip with '+' sign = positive allosteric regulation event; computer chip with '-' sign = negative allosteric regulation event; green arrow = cellular response event.

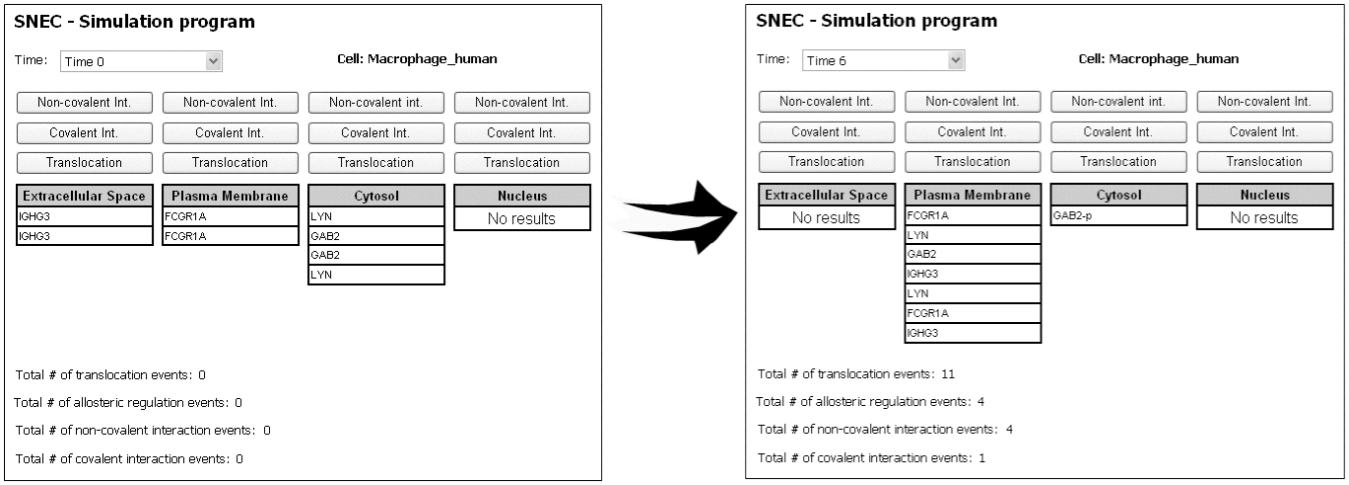

Fig. (13). A SN-based simulator. An instance of a human macrophage cell is composed of four compartments; extracellular space, plasma membrane, cytosol and nucleus. There are three operations for each compartment, and they are activated by the three action buttons respectively: 'Non-covalent int.', 'Covalent int.' and 'Translocation.' An operation such as 'Non-covalent int.' searches for a pair of molecules that have the ability to interact in the same location and creates an event agent accordingly. A translocation event moves a molecule to an adjacent compartment. The combo boxes located at the top is used to increment the time. Before the simulation run (the screenshot at left), there was no event occurred as shown in the reports at the bottom of the screen. The screenshot at right shows the simulation outcomes at time 6. The molecules have changed their original locations, and many events have accumulated. There were 11 translocation events, 4 allosteric regulation events, 4 non-covalent interactions events, and 1 covalent interaction events. In particular, the allosteric events switched the corresponding proteins to the active conformation that enabled different interactions in the following time steps. As a result, a GAB2 protein has been phosphorylated into GAB2-p by an activated LYN kinase. The phosphorylation event occurred at the plasma membrane, and GAB2-p moved from plasma membrane to cytosol through a subsequent translocation event.

for reviewing the manuscript section that describes Visual Knowledge and BioCAD. We thank Zakaria Hmama, Neil E. Reiner and Jimmy Lee (Division of Infectious Diseases, Department of Medicine, Faculty of Medicine, University of British Columbia) for their advice on modeling bacterial invasion processes. The research presented in the case study section of the paper was funded by the CIHR/MSFHR Strategic Training Program in Bioinformatics, Canadian Institutes of Health Research (CIHR), Michael Smith Foundation for Health Research (MSFHR), and Natural Sciences and Engineering Research Council of Canada (NSERC).

Table 2. Predicted Macrophage Responses Caused by MTB

Macrophage Responses Promoted by MTB	Supporting Evidence
Actin polymerization and rearrangement	[142]
Membrane delivery to plasma membrane	[142]
Cell survival	-
Cell cycle entry - S phase	-
Protein synthesis	-
Intracellular glucose uptake	-
Macrophage responses inhibited by MTB	
Recruitment of oxidase complex to phagosome	[143]
Phagosome-lysosome fusion	[144]

Column 2 indicates literature that supports the predicted responses.

REFERENCES

[1] Drysdale RA, Crosby MA. FlyBase: genes and gene models. *Nucleic Acids Res* **2005**; 33: D390-5.

[2] Christie KR, Weng S, Balakrishnan R, *et al*. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res* **2004**; 32: D311-4.

[3] Eppig JT, Bult CJ, Kadin JA, *et al*. The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. *Nucleic Acids Res* **2005**; 33: D471-5.

[4] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. *Nucleic Acids Res* **2005**; 33 Database Issue: D34-8.

[5] Bairoch A, Apweiler R, Wu CH, *et al.* The Universal Protein Resource (UniProt). *Nucleic Acids Res* **2005**; 33: D154-9.

[6] Chittur SV. DNA microarrays: tools for the 21st Century. *Comb Chem High Throughput Screen* **2004**; 7: 531-7.

[7] Zhou H, Roy S, Schulman H, Natan MJ. Solution and chip arrays in protein profiling. *Trends Biotechnol* **2001**; 19: S34-9.

[8] Kodadek T. Development of protein-detecting microarrays and related devices. *Trends Biochem Sci* **2002**; 27: 295-300.

[9] Lopez MF, Pluskal MG. Protein micro- and macroarrays: digitizing the proteome. *J Chromatogr B Analyt Technol Biomed Life Sci* **2003**; 787: 19-27.

[10] Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. *Nucleic Acids Res* **2000**; 28: 235-42.

[11] Uetz P, Giot L, Cagney G, *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **2000**; 403: 623-7.

[12] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci USA* **2001**; 98: 4569-74.

[13] Gavin AC, Bosche M, Krause R, *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **2002**; 415: 141-7.

[14] Ho Y, Gruhler A, Heilbut A, *et al.* Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **2002**; 415: 180-3.

[15] Crameri R, Kodzius R. The powerful combination of phage surface display of cDNA libraries and high throughput screening. *Comb Chem High Throughput Screen* **2001**; 4: 145-55.

[16] Walter G, Konthur Z, Lehrach H. High-throughput screening of surface displayed gene products. *Comb Chem High Throughput Screen* **2001**; 4: 193-205.

[17] Alfarano C, Andrade CE, Anthony K, *et al.* The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res* **2005**; 33: D418-24.

[18] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **2004**; 32: D449-51.

[19] Hermjakob H, Montecchi-Palazzi L, Lewington C, *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res* **2004**; 32 Database issue: D452-5.

[20] Breitkreutz BJ, Stark C, Tyers M. The GRID: the General Repository for Interaction Datasets. *Genome Biol* **2003**; 4: R23.

[21] Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G.. MINT: a Molecular INTeraction database. *FEBS Lett* **2002**; 513: 135-40.

[22] Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci* **1998**; 23: 324-8.

[23] Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* **1999**; 96: 2896-901.

[24] Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. *Science* **1999**; 285: 751-3.

[25] Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **1999**; 402: 86-90.

[26] Ge H, Liu Z, Church GM, Vidal M. Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* **2001**; 29: 482-6.

[27] Grigoriev A. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic Acids Res* **2001**; 29: 3513-9.

[28] Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein-protein interactions. *Genome Res* **2002**; 12: 37-46.

[29] Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* **1999**; 96: 4285-8.

[30] Matthews LR, Vaglio P, Reboul J, *et al.* Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Res* **2001**; 11: 2120-6.

[31] Gomez SM, Rzhetsky A. Towards the prediction of complete protein--protein interaction networks. *Pac Symp Biocomput* **2002**; 413-24.

[32] Ng SK, Zhang Z, Tan SH. Integrative approach for computationally inferring protein domain interactions. *Bioinformatics* **2003**; 19: 923-9.

[33] Obenauer JC, Yaffe MB. Computational prediction of protein-protein interactions. *Methods Mol Biol* **2004**; 261: 445-68.

[34] Reiss DJ, Schwikowski B. Predicting protein-peptide interactions *via* a network-based motif sampler. *Bioinformatics* **2004**; 20 Suppl 1: I274-82.

[35] Lu L, Lu H, Skolnick J. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins* **2002**; 49: 350-64.

[36] Aloy P, Russell RB. Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci USA* **2002**; 99: 5896-901.

[37] Marcotte EM, Xenarios I, Eisenberg D. Mining literature for protein-protein interactions. *Bioinformatics* **2001**; 17: 359-63.

[38] Temkin JM, Gilder MR. Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics* **2003**; 19: 2046-53.

[39] Albert S, Gaudan S, Knigge H, *et al.* Computer-assisted generation of a protein-interaction database for nuclear receptors. *Mol Endocrinol* **2003**; 17: 1555-67.

[40] Donaldson I, Martin J, de Bruijn B, *et al.* PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics* **2003**; 4: 11.

[41] Daraselia N, Yuryev A, Egorov S, Novichkova S, Nikitin A, Mazo I. Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* **2004**; 20: 604-11.

[42] Hoffmann R, Krallinger M, Andres E, Tamames J, Blaschke C, Valencia A. Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci STKE* **2005**; 2005: pe21.

[43] von Mering C, Jensen LJ, Snel B, *et al.* STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* **2005**; 33 Database Issue: D433-7.

[44] Mellor JC, Yanai I, Clodfelter KH, Mintseris J, DeLisi C. Predictome: a database of putative functional links between proteins. *Nucleic Acids Res* **2002**; 30: 306-9.

[45] Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* **2004**; 32: D277-80.

[46] Krieger CJ, Zhang P, Mueller LA, *et al.* MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* **2004**; 32 Database issue: D438-42.

[47] Joshi-Tope G, Gillespie M, Vastrik I, *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* **2005**; 33: D428-32.

[48] Gough NR. Science's signal transduction knowledge environment: the connections maps database. *Ann N Y Acad Sci* **2002**; 971: 585-7.

[49] Choi C, Krull M, Kel A, *et al*. TRANSPATH - a high quality database focused on signal transduction. *Comparative and Functional Genomics* **2004**; 5: 163-8.

[50] Berners-Lee T, Hendler J, Lassila O. The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American* **2001**; 284: 34-43.

[51] Baker CJO, Butler G, Haarslev V. Ontologies, Semantic web and Intelligent Systems for Genomics. Paper read at 1st Canadian Semantic Web Interest Group Meeting (SWIG-04), at Montreal, Quebec, Canada, 2004.

[52] Sowa JF. Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks/Cole Publishing Co., Pacific Grove, CA 2000.

[53] Pollock JT, Hodgson R. Adaptive information: improving business through semantic interoperability, grid computing, and enterprise integration, John Wiley & Sons., Hoboken, NJ 2004.

[54] Gardner SP. Ontologies and semantic data integration. *Drug Discov Today* **2005**; 10: 1001-1007.

[55] Date CJ. Database in depth: relational theory for practitioners, O'Reilly, Sebastopol, CA 2005.

[56] Budd T. An Introduction to Object-Oriented Programming, Addison Wesley, Boston, MA 2001.

[57] Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schneider P, Eds, The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press, Cambridge 2003.

[58] Shah SP, Huang Y, Xu T, Yuen MM, Ling J, Ouellette BF. Atlas - a data warehouse for integrative bioinformatics. *BMC Bioinformatics* **2005**; 6: 34.

[59] Yona G. Biozon. Date of access: Dec. 05, 2005. Available from: URL: http://biozon.org

[60] Giarratano JC, Riley GD, Expert Systems: Principles and Programming, Fourth Edition: Principles and Programming, Thomson Course Technology, Boston, MA 2004.

[61] Minsky M. A framework for representing knowledge, In: Winston, PH Ed, The Psychology of Computer Vision. McGraw-Hill, New York, NY 1975; 211-277.

[62] Massar JP, Travers M, Elhai J, Shrager J. BioLingua: a programmable knowledge environment for biologists. *Bioinformatics* **2005**; 21: 199-207.

[63] Wooldridge M, Introduction to MultiAgent Systems, John Wiley & Sons, Hoboken, NJ 2002.

[64] Decker K, Khan S, Schmidt C, Situ G, Makkena R, Michaud D. Biomas: A multi-agent system for genomic annotation. *Intl J Cooperat Inform Sys* **2002**; 11: 265-92.

[65] Griffith RL. Three Principles of Representation for Semantic Networks. *ACM Trans Database Syst* **1982**; 7: 417-42.

[66] Sowa JF, Ed, Principles of Semantic Networks: Explorations in the Representation of Knowledge. Morgan Kaufmann Publishers, San Mateo, CA 1991.
[67] Lehmann F. Semantic networks. *Comp Mathemat Applicat* **1992**; 23: 1-50.
[68] Roberts DD. The Existential Graphs of Charles S. Peirce, Mouton, The Hague 1973.
[69] Shastri L, Why Semantic Networks?, In: Sowa JF Ed, Principles of Semantic Networks: Explorations in the Representation of Knowledge. Morgan Kaufmann Publishers, San Mateo, CA 1991; 109-36.
[70] Schubert LK, Semantic Nets Are in the Eye of the Beholder, In: Sowa JF Ed, Principles of Semantic Networks: Explorations in the Representation of Knowledge. Morgan Kaufmann Publishers, San Mateo, CA 1991; 95-107.
[71] Richens RH. Preprogramming for mechanical translation. *Mechan Trans* **1956**; 3: 20-5.
[72] Woods WA, Schmolze JG. The KL-ONE family. *Comp Mathemat Applicat* **1992**; 23: 133-77.
[73] Shapiro SC, Rapaport WJ. The SNePS family. *Comp Mathemat Applicat* **1992**; 23: 243-75.
[74] Sowa JF. Conceptual graphs for a database interface. *IBM J Res Develop* **1976**; 20: 336-57.
[75] Sowa JF. Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, Reading, MA 1984.
[76] Pearl J, Russel S. Bayesian networks, In: Arbib, MA Ed, Handbook of Brain Theory and Neural Networks. MIT Press, Cambridge, MA 2003; 157-60.
[77] Petri CA. Nets, time and space. *Theoret Comp Sci* **1996**; 153: 3-48.
[78] Mylopoulos J. The PSN tribe. *Comp Mathemat Appl* **1992**; 23: 223-41.
[79] Visual Knowledge. Visual Knowledge. Date of access: Dec. 05, 2005. Available from: URL: http://www.visualknowledge.com
[80] Lenat DB. CYC: a large-scale investment in knowledge infrastructure. *Commun ACM* **1995**; 38: 33-8.
[81] Matuszek C, Witbrock M, Kahlert RC, Cabral J, Schneider D, Shah P, D. L. (2005). Searching for Common Sense: Populating Cyc from the Web. Paper read at In Proceedings of the Twentieth National Conference on Artificial Intelligence, at Pittsburgh, Pennsylvania.
[82] Lenat DB, Guha RV. Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project, Addison-Wesley, Reading, MA 1990.
[83] Gruber TR. A translation approach to portable ontology specifications. *Knowledge Acquisition* **1993**; 5: 199-220.
[84] OBO. Open Biomedical Ontologies. Date of access: Dec. 05, 2005. Available from: URL: http://obo.sourceforge.net/
[85] Smith CL, Goldsmith CA, Eppig JT. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol* **2005**; 6: R7.
[86] Bard J, Rhee SY, Ashburner M. An ontology for cell types. *Genome Biol* **2005**; 6: R21.
[87] Hayamizu TF, Mangan M, Corradi JP, Kadin JA, Ringwald M. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome Biol* **2005**; 6: R29.
[88] Harris MA, Clark J, Ireland A, *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **2004**; 32: D258-61.
[89] Smith B, Williams J, Schulze-Kremer S. The ontology of the gene ontology. *AMIA Annu Symp Proc* **2003**; 609-13.
[90] Smith B, Ceusters W, Klagges B, *et al.* Relations in biomedical ontologies. *Genome Biol* **2005**; 6: R46.
[91] Wroe CJ, Stevens R, Goble CA, Ashburner M. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. *Pac Symp Biocomput* **2003**; 624-35.
[92] Aitken S, Korf R, Webber B, Bard J. COBrA: a bio-ontology editor. *Bioinformatics* **2005**; 21: 825-6.
[93] Noy NF, Crubezy M, Fergerson RW, *et al.* Protege-2000: an open-source ontology-development and knowledge-acquisition environment. *AMIA Annu Symp Proc* **2003**; 953.
[94] Farquhar A, Fikes R, Rice J. The Ontolingua Server: A tool for collaborative ontology construction. *Intl J Human-comp Studies* **1997**; 46: 707-27.
[95] McGuinness DL, Fikes R, Rice J, Wilder S. The Chimaera Ontology Environment. Paper read at the Seventeenth National Conference on Artificial Intelligence (AAAI 2000), July 30 - August 3, at Austin, Texas, 2000.
[96] Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* **2004**; 32: D267-70.
[97] Lomax J, McCray AT. Mapping the Gene Ontology into the Unified Medical Language System. *Compar Funct Genom* **2004**; 5: 354-61.
[98] McCray AT, Nelson SJ. The representation of meaning in the UMLS. *Methods Inf Med* **1995**; 34: 193-201.
[99] McCray AT. An upper-level ontology for the biomedical domain. *Compar Funct Genom* **2003**; 4: 80-4.
[100] Cimino JJ, Min H, Perl Y. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. *J Biomed Inform* **2003**; 36: 450-61.
[101] Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. *J Biomed Inform* **2003**; 36: 414-32.
[102] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* **2001**; 17-21.
[103] Liu H, Johnson SB, Friedman C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *J Am Med Inform Assoc* **2002**; 9: 621-36.
[104] Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics* **2005**; 6 Suppl 1: S1.
[105] Perez-Iratxeta C, Wjst M, Bork P, Andrade MA. G2D: a tool for mining genes associated with disease. *BMC Genet* **2005**; 6: 45.
[106] Cantor MN, Sarkar IN, Bodenreider O, Lussier YA. Genestrace: phenomic knowledge discovery *via* structured terminology. *Pac Symp Biocomput* **2005**; 103-14.
[107] NCOR. National Center for Ontological Research. Date of access: Dec. 05, 2005. Available from: URL: http://ncor.us/
[108] Rosse C, Mejino JL, Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. *J Biomed Inform* **2003**; 36: 478-500.
[109] Jenssen TK, Hovig E. The semantic web and biology. *Drug Discov Today* **2002**; 7: 992.
[110] Hendler J. Communication. Science and the semantic web. *Science* **2003**; 299: 520-1.
[111] Stevens RD, Robinson AJ, Goble CA. myGrid: personalised bioinformatics on the information grid. *Bioinformatics* **2003**; 19 Suppl 1: i302-4.
[112] Wilkinson MD, Links M. BioMOBY: an open source biological web services proposal. *Brief Bioinform* **2002**; 3: 331-41.
[113] Lord P, Bechhofer S, Wilkinson MD, *et al.* Applying semantic web services to Bioinformatics: Experiences gained, lessons learnt. Paper read at ISWC 2004, at Springer-Verlag, Berlin Heidelberg. 350-64, 2004.
[114] Wilkinson M, Schoof H, Ernst R, Haase D. BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case. *Plant Physiol* **2005**; 138: 5-17.
[115] World Wide Web Consortium. OWL Web Ontology Language Overview. Date of access: Dec. 05, 2005. Available from: URL: http://www.w3.org/TR/owl-features/
[116] Haarslev V, Moller R. (2003). Racer: An OWL Reasoning Agent for the Semantic Web. Paper read at Proceedings of the International Workshop on Applications, Products and Services of Web-based Support Systems, in conjunction with the 2003 IEEE/WIC International Conference on Web Intelligence, October 13, at Halifax, Canada. 91-5.
[117] Mindswap. Pellet. Date of access: Dec. 05, 2005. Available from: URL: http://www.mindswap.org/2003/pellet/index.shtml
[118] KAON2. KAON2. Date of access: Dec. 05, 2005. Available from: URL: http://kaon2.semanticweb.org/
[119] BioPAX. BioPAX: Biological Pathways Exchange. Date of access: Dec. 05, 2005. Available from: URL: http://www.biopax.org/
[120] BioCyc. BioCyc. Date of access: Dec. 05, 2005. Available from: URL: http://biocyc.org/
[121] Shaban-Nejad A, Baker CJO, Butler G, Haarslev V. The FungalWeb Ontology: The core of a Semantic Web application for fungal genomics. Paper read at 1st Canadian Semantic Web Interest Group Meeting (SWIG' 04), at Montreal, Quebec, Canada, 2004.
[122] Cheung KH, Yip KY, Smith A, Deknikker R, Masiar A, Gerstein M. YeastHub: a semantic web use case for integrating data in the life sciences domain. *Bioinformatics* **2005**; 21 Suppl 1: i85-i96.

[123] BioDASH. BioDASH. Date of access: Dec. 05, 2005. Available from: URL: http://www.w3.org/2005/04/swls/BioDash/Demo/

[124] openRDF. Sesame. Date of access: Dec. 05, 2005. Available from: URL: http://www.openrdf.org/

[125] Kowari. Kowari Metastore. Date of access: Dec. 05, 2005. Available from: URL: http://www.kowari.org/

[126] Advanced Knowledge Technologies. AKT Triplestore. Date of access: Dec. 05, 2005. Available from: URL: http://triplestore.aktors.org/

[127] Visual Knowledge. BioCAD. Date of access: Dec. 05, 2005. Available from: URL: http://www.biocad.com

[128] Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **2005**; 33 Database Issue: D501-4.

[129] Bateman A, Coin L, Durbin R, *et al*. The Pfam protein families database. *Nucleic Acids Res* **2004**; 32: D138-41.

[130] Sigrist CJ, Cerutti L, Hulo N, *et al.* PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* **2002**; 3: 265-74.

[131] Mulder NJ, Apweiler R, Attwood TK, *et al.* InterPro, progress and status in 2005. *Nucleic Acids Res* **2005**; 33 Database Issue: D201-5.

[132] Ng SK, Zhang Z, Tan SH, Lin K. InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res* **2003**; 31: 251-4.

[133] Rousset MC. Small Can Be Beautiful in the Semantic Web. Paper read at Third International Semantic Web Conference, at Hiroshima, Japan, 2004.

[134] Tjelle TE, Lovdal T, Berg T. Phagosome dynamics and function. *Bioessays* **2000**; 22: 255-63.

[135] Stephens L, Ellson C, Hawkins P. Roles of PI3Ks in leukocyte chemotaxis and phagocytosis. *Curr Opin Cell Biol* **2002**; 14: 203-13.

[136] Cantley LC. The phosphoinositide 3-kinase pathway. *Science* **2002**; 296: 1655-7.

[137] Wymann MP, Zvelebil M, Laffargue M. Phosphoinositide 3-kinase signalling--which way to target? *Trends Pharmacol Sci* **2003**; 24: 366-76.

[138] Russell DG. Mycobacterium tuberculosis: here today, and here tomorrow. *Nat Rev Mol Cell Biol* **2001**; 2: 569-77.

[139] Fratti RA, Backer JM, Gruenberg J, Corvera S, Deretic V. Role of phosphatidylinositol 3-kinase and Rab5 effectors in phagosomal biogenesis and mycobacterial phagosome maturation arrest. *J Cell Biol* **2001**; 154: 631-44.

[140] Hsing M, Bellenson JL, Shankey C, Cherkasov A. Modeling of cell signaling pathways in macrophages by semantic networks. *BMC Bioinformatics* **2004**; 5: 156.

[141] Hsing M, Modeling of cell signaling pathways in macrophages by semantic networks. M.Sc. Thesis, University of British Columbia, Vancouver, BC 2005.

[142] Schlesinger LS, Bellinger-Kawahara CG, Payne NR, Horwitz MA. Phagocytosis of Mycobacterium tuberculosis is mediated by human monocyte complement receptors and complement component C3. *J Immunol* **1990**; 144: 2771-80.

[143] Moura AC, Modolell M, Mariano M. Down-regulatory effect of Mycobacterium leprae cell wall lipids on phagocytosis, oxidative respiratory burst and tumour cell killing by mouse bone marrow derived macrophages. *Scand J Immunol* **1997**; 46: 500-5.

[144] Xu S, Cooper A, Sturgill-Koszycki S, *et al*. Intracellular trafficking in Mycobacterium tuberculosis and Mycobacterium avium-infected macrophages. *J Immunol* **1994**; 153: 2568-78.

Received: October 5, 2005 Revised: December 6, 2005 Accepted: December 31, 2005