Representing Genomic Knowledge in the UMLS Semantic Network

Hong Yu[1], Carol Friedman, Ph.D.[2,1], Andrey Rhzetsky, Ph.D[3,1], Pauline Kra, Ph.D[4].
[1]Department of Medical Informatics, Columbia University
[2]Department of Computer Science, Queens College CUNY
[3]Center for Genomics Research, Columbia University
[4]Yeshiva University

Genomics research has a significant impact on the understanding and treatment of human hereditary diseases, and biomedical literature concerning the genome project is becoming more and more important for clinicians. The Unified Medical Language System (UMLS) is designed to facilitate the retrieval and integration of information from multiple-readable biomedical information resources[1]. This paper describes our efforts to integrate concepts important to genomics research with the UMLS semantic network. We found that the UMLS contains over 30 semantic types and most of the semantic relations that are essential for representing the underlying genomic knowledge. In addition, we observed that the organization of the network was appropriate for representing the hierarchical organization of the concepts. Because some of the concepts critical to the genomic domain were found to be missing, we propose to extend the network by adding six new semantic types and sixteen new semantic relations.

Introduction

The Human Genome Project (HGP) is extracting information from the DNA strands that constitute our genetic inheritance[2]. The acquisition of a comprehensive human genome sequence will have unprecedented impact and value for basic biology, biomedical research, biotechnology, and medicine. Identification of a gene permits the development of diagnostic tests that can reveal aberrations prior to manifestation of clinical symptoms[3]. Knowledge of a patient's genetic makeup can allow physicians to minimize disease risk through preventive medicine, conventional drug therapies and gene therapy[4]. More than 100 human diseases are caused by alteration of a specific gene(s)[5,14]. Research in biomedical literature and retrieval of information produced by the genome project will be ultimately essential for the understanding of human hereditary disease genes, studies of carcinogenesis, design of antimicrobial drugs, and fundamental biomedical research.

The number of articles in biomedical research is growing exponentially and it is difficult if not impossible for researchers to manually keep track of information relevant to their areas of interest. It would be extremely valuable if information from the biomedical literature could be automatically extracted, organized and stored in a knowledge base. A possible way of accomplishing this is to process the documents using natural language (NLP) techniques to extract and structure the relevant information. The information could then be made accessible to researchers and other computerized processes.

As prelude to NLP, it is crucial to develop a domain model or ontology for the representation of pertinent information in the domain because information extracted using NLP will need to be mapped into a suitable representation. We developed an ontology concerning genomic concepts that specifies important concepts in the domain, their relationships, and also organizes the concepts into a hierarchy of classes.

The UMLS is a large-scale knowledge source designed to facilitate retrieval and integration of information from multiple-readable biomedical information resources[1]. The UMLS semantic network consists of classes of concepts and relations in biomedicine, and is an important knowledge source for biomedicine. Therefore, it is important that the UMLS network also contains concepts that are pertinent to genomics.

In this paper we analyze the UMLS semantic network with regard to genomics concepts and relations. For each concept or relation we identify a comparable node in the UMLS network or propose an extension. We also examine the organization or hierarchical relations of the network to determine if it is appropriate for the genomics information.

Background

Currently, there is no published knowledge base that completely represents information associated with the genome project. Hafner et al[7] adapted the hierarchical classification system of the Unified Medical Language System (UMLS) and developed a knowledge

1091—8280/99/$5.00 © 1999 AMIA, Inc.

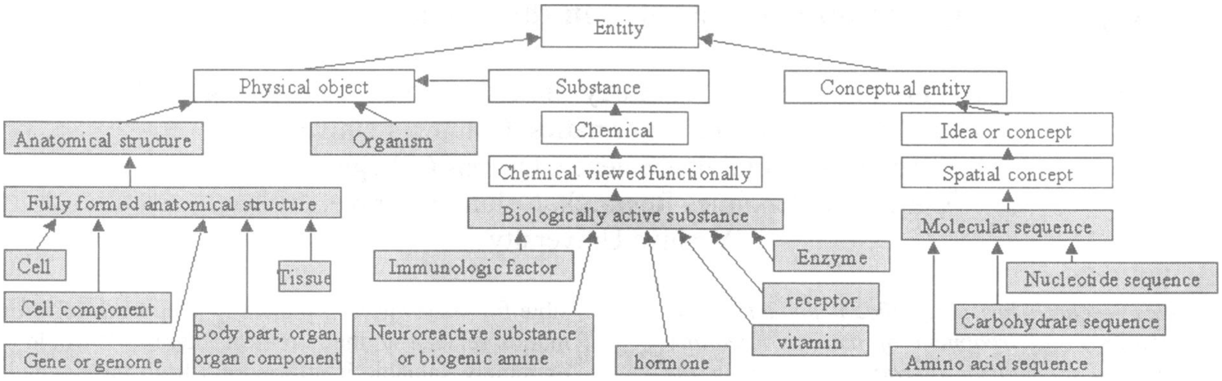

Fig 1. Notes under the simplified hierarchy of 'entity' of UMLS semantic network that are important for the genomic project are shown using gray rectangles. There are more than 30 nodes that are in the UMLS and relevant to the genomic project ('Event' taxonomy not shown).

representation associated with biomedical literature research papers. They proposed a new set of taxonomy for *entity, event* and *piece of information*. However, their knowledge base was focused on the material and methods sections of the literature, and important information associated with the hypotheses and results were not covered. Most of the other computerized resources in biomedical research, such as the Genbank[14] of the National Center for Biotechnology Information (NCBI), consist of databases that specify names and abbreviations for genes, proteins and gene related diseases. These databases contain crucial facts but do not contain a specification of many of the concepts and relations in the domain. Research developed for biomedical simulation was associated with an ontological design[8,9]. However, the knowledge domain was narrow and did not apply to the genome project.

The UMLS has incorporated diverse perspectives and approaches in constructing the terminology for biomedicine, and represents a large-scale cooperative and distributed effort[10]. The UMLS contains a Metathesaurus (MT) which specifies biomedical concepts in a format that integrates over 30 biomedical vocabularies. The UMLS also contains a semantic network (SN) that defines and organizes the semantic types. Each concept in the MT is associated with one or more semantic types. In this paper, we focus solely on the SN.

The Semantic Network (1999 version) has 134 semantic types and 54 relationships. The hierarchical relation is represented by means of an *isa* link. Three basic parts of the taxonomy are semantic types called *entity*, and *event*, and a semantic relation called *associate_with.* Figure 1 presents a simplified diagram of the hierarchy associated with *entity*. The part showing contains the types that are relevant to genomics information. The gray rectangles represent pertinent types that are already in the semantic network. Figure 3 contains a simplified diagram for the relation *associate_with.* The rectangles represent relevant relations that are in the network.

Methods

We created an ontology to represent the concepts and relations important for representing genomics[6]. Based on this work, we manually mapped the concepts of *substance, action*, and *relation* into the UMLS semantic network. We identified corresponding types in the UMLS by examining the network and looking at the definitions provided by the UMLS for each of the types. We also identified the genomic concepts that do not correspond to any type in the UMLS. We created new types and determined where in the UMLS network to attach the new types. This was also accomplished by manual analysis of the nodes in the network.

Results

A large number of the genomics concepts already existed in the UMLS semantic network. Figure 1 depicts an abbreviated version of the type hierarchy for *entity*, and represents the information we identified as being relevant to genomics using gray rectangles. For example, some of these types include *gene, nucleotide sequence, amino acid sequence, cell, cell component, tissue, and organ.*

Figure 2 shows the six new semantic types that we added to the UMLS network under *entity*. The new

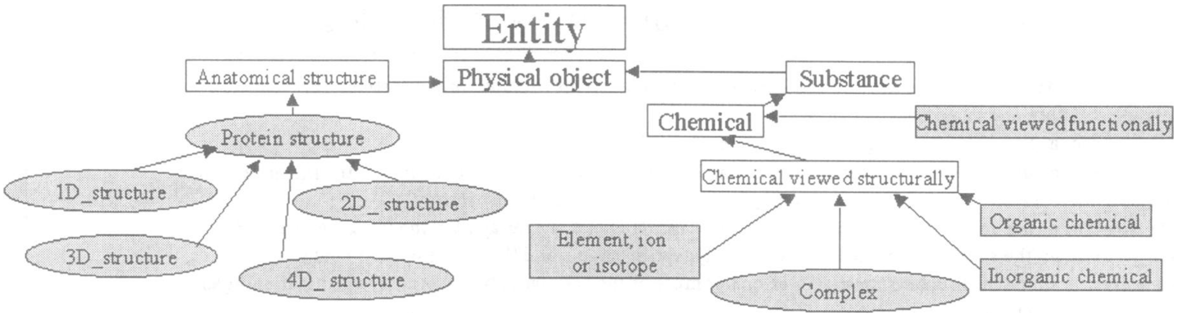

Fig 2. Classification of hierarchical relationship of 'entity'. Gray rectangular nodes are important for the genome project. Gray oval nodes are the integrated new types.

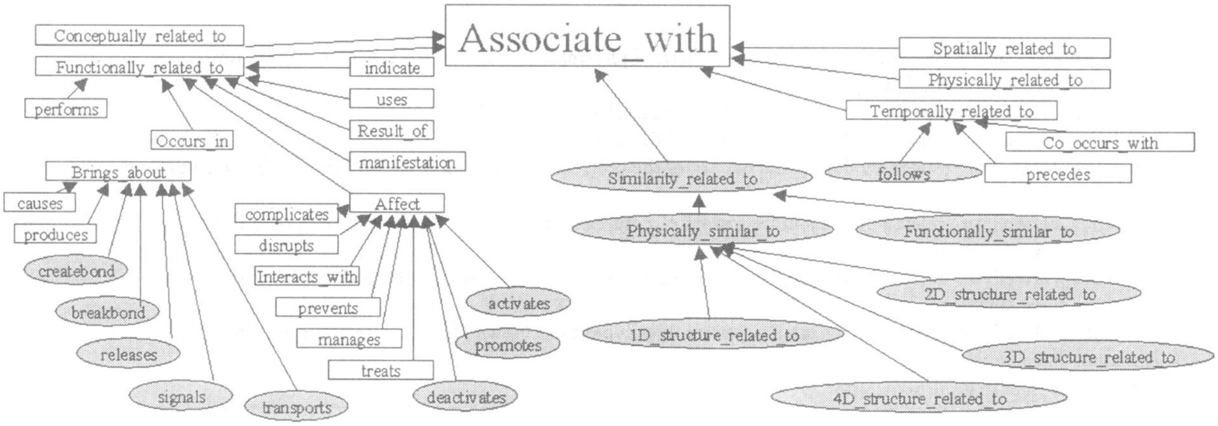

Fig 3. The simplified classification of 'associate_with' based on the '-isa-' relationship. The rectangular ones are the original types of UMLS. The gray ones are the additions.

types and definitions are shown by Table 1. *Protein structure* was introduced as a subclass of *anatomical structure* and *1D_structure*, *2D_structure*, *3D_structure* and *4D_structure* were introduced as subtypes. In another branch of the network, *complex* is introduced as a subclass of *chemical viewed structurally.*

The type *associate_with* representing non-hierarchical relations is shown in Figure 3. The relations include *physically_related_to, spatially_related_to, temporally_related_to, functionally_related_to and conceptually_related_to*, which are necessary in order to represent biomedical reactions and processes. We introduced 16 new semantic relations under *associates_with* as shown in Figure 3. The new types and definitions are described in Table 1. Eight types of chemical actions: *activates*, *deactivates*, *promotes*, *createbond*, *breakbond*, *releases*, *transports*, and *signals,* were added as subclasses of *functionally_related_to.* In addition, we integrated a sixth relation *similarity_related_to* as a subclass of *associate_with.* We further introduced *physically_similar_to* and *functionally_similar_to* as two subclasses of *similarity_related_to.* We added *1D_structure_related_to, 2D_structure_related_to, 3D_structure_related_to,* and *4D_structure_related_to* as four subclasses of *physically_similar_to.* Finally, we added *follows* as a subclass of *temporally_related_to.*

Discussion

Protein plays a crucial role in all the biomedical processes. An important characteristic of protein is the well-defined three-dimensional structure. The structure of the protein is the essential biophysical character determining the function of the protein. Four levels of the protein structure are the following: primary structure (*1D_structure)*, secondary structure

Table 1: The definitions of the new semantic types and relations related with genomic knowledge

Protein structure: The three dimensional structure of protein. This includes primary structure, secondary structure, tertiary structure and quaternary structure.
1D_structure: the amino acid sequence and the location of disulfides[13].
2D_structure: the spatial arrangement of amino acid residues that are near one another in the linear sequence[13]. Alpha helix, beta sheet, and collagen helix are elements of the secondary structure[13].
3D_structure: the spatial arrangement of amino acid residues that are far apart in the linear sequence[13].
4D_structure: the spatial arrangement of a few polypeptide chains and the nature of their contacts[13].
Complex: a chemical structure consisting of one or more of the following types: organic chemical, inorganic chemical, or an element, ion or isotope.
Activates: to make capable of reacting or of accelerating a chemical reaction.
Deactivates: to make inactive.
Breakbond: to break a covalent bond/connection within a molecule.
Createbond: to form a covalent bond between two molecules.
Promotes: to accelerate.
Releases: to set free.
Signals: to sign as the occasion for prearranged combined action.
Transports: to carry from one place to another.
Similarity_related_to: related either by the similar pattern of physical attribute or characteristic, or the similar pattern of carrying out of some function or activity.
Physically_similar_to: related by the similar pattern of physical attribute or characteristic.
Functionally_similar_to: related by the similar pattern of function.
1D_structure_related to: related by the similar pattern of the collected sequences of amino acids, carbohydrates, and nucleotide sequences.
2D_structure_related_to: related by the similar pattern of polypeptides viewed from the perspective of their 2D_structural characteristics.
3D_structure_related_to: related by the similar pattern of polypeptides viewed from the perspective of their 3D_structural characteristics.
4D_structure_related_to: related by the similar pattern of polypeptides viewed from the perspective of their 4D_structural characteristics.
Follows: occurs later in time, come after.

(*2D_structure*), tertiary structure (*3D_structure*), and quaternary structure (*4D_structure*), which we added as four subclasses of a new semantic type *protein structure* (Fig 2).

Complex and *dynamic* are the two essential properties of biomedical processes [9,11,12]. Biomedical processes involve numerous complexes. A structure such as an enzyme complex, ribosome, protein filament, and virus, is not a single, covalently linked molecule. Instead it is formed by the noncovalent assembly of many molecules, which are called the subunits of the final structure. The structure formed by these subunits is a *complex.* Each central process in a cell----such as DNA replication, RNA or protein synthesis, vesicle budding, transmembrane signalling, or apoptosis----is catalyzed by a complex of 10 or more proteins. Drug design normally targets a complex structure. Diseases caused by the disruption of normal cellular processes usually involve the disruption of a complex. Thus we introduce *complex* as a type to represent a structure which contains several proteins or other substances (Fig2). We added *complex* as a subclass of *chemical viewed structurally.* However, we are not satisfied with the taxonomy of *chemical viewed structurally.* We will do further research to find the best way to represent this type of knowledge.

Biomedical systems are dynamic. DNA transcription, RNA translation, protein modification, and all the biomedical processes are chains of biomedical reactions. The actions are related temporally and causally to other actions. For instance, RNA translation occurs after DNA transcription. The alteration of DNA transcription causes the alteration of the product of RNA translation—protein. Thus the biomedical systems can be interpreted and understood as complex, time dependent, interactive processes[9]. In addition to the actions like *prevents, causes, indicates, interacts_with, disrupts, and occurs_in*, which already exist in the network, we added actions important for the biomedical reactions. *Activates, deactivates, and promotes* represent the influence of substance(s) or process(es) on another substance(s) or process(es). *Breakbond, createbond, and releases* will represent biophysical and biochemical interaction and alteration of substances, such as phosphorylation and dephosphorylation. *Transports* is very important to identify the movement from one place to another. Most cellular processes involve transportation. For instance, protein synthesis involves a series of transportation: the transportation of RNA from the nucleus to ribosome, the transportation involved in post-translational modification processes, and the transportation of the protein to the finial target, such as the membrane. The time relations between biomedical reactions and processes are represented by *temporally_related_to*. UMLS has *precedes* and *co_occurs_with* as two subclasses of *temporally_related_to*. We added *follows* as a new subclass of *temporally_related_to* for the convenience and accuracy of knowledge representation. For example, the statement *B precedes A if C happens* is different from the statement *A follows B if C happens*. Actions such as *causes, promotes, activates, deactivates* and *signals* represent causal relations.

Similarity comparison is the alignment of gene and amino acid sequences. We introduce *similarity_related_to* as a new relationship. Similarity comparisons between genes and amino acid sequences are the important part of the genome project. The gene and protein family classifications are based on the similarities between the sequences. Genes and proteins resemble one another in DNA sequence and amino acid sequence, respectively, only if they have a common ancestor. Inherited genetic diseases can be caused by the alterations of the genes and amino acids sequences. For example, sickle-cell anemia can result from a change in a single amino acid in a single protein. We further introduced *structurally_related_to* as a subclass of *physically_similar_to*. Proteins can be classified according to the structural similarities. In addition, structural comparison is important to determine functional similarities. Similarity comparison at each of the four levels of the protein structure is equally important. Different level of comparison in protein structure reveals different aspect of knowledge. In our model, functional similarity is represented by *functional_similar_to*.

Conclusion

We identified the semantic types and relations of the genome project and integrated them with the UMLS ontology. We found over 30 semantic types and most of the semantic relations relevant to the genome project. We also found the organization of 1999 semantic network is suitable. We added twenty-two semantic types and relations to the categories of *entity* and *associate_with* to include concepts needed for the genome project. The successful mapping and extensions show the suitability and the adaptability of the UMLS semantic network for the representation of the growing domain of biomedical knowledge.

References

1. Lindberg DAB, Humphreys BL, McCray AT. The Unified Medical Language System. *Meth Inform Med* 1993;32:281-291.
2. Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R., Walters L, and the members of the DOE and NIH planning groups. New goals for the U.S. Human Genome Project: 1998-2003. *Science* 1998; 282: 682-689.
3. Lin JH. Divining and altering the future: Implications from the human genome project. *Science*, 1998. 282: 1532.
4. Karanjawala ZE and Collins FS. Genetics in the context of medical practice. *JAMA* 1998. 280 (17): 1533-1534.
5. McKuSICK, VA. Mendelian inheritance in man, a catalog of human genes and genetic disorders, 12th edition, 1998.
6. Rzhetsky A, Koike T, Kalachikov S. Kra P, Yu H, Friedman C. A knowledge model for analysis and simulation of regulatory networks in bioinformatics studies aiming at disease gene discovery. A poster to *AMIA 99 Fall Annual Symposium,* 1999.
7. Hafner CD, Baclawski K, Futrelle RP, Fridman N, Sampath S. Creating a knowledge base of biological research papers. In *2d Inter'l Conf. on Intelligent Systems for Molecular Biology.* AAAI Press, Stanford CA. 1994;
8. Fridman Noy, N. and Hafner, C. (1997). The state of the art in ontology design: a survey and comparative review. *AI Magazine*, Fall 1997; 53-74.
9. Sandblad B, Meinzer HP, Modelling and simulation of complex control structures in cell biology. *Meth Inform Med* 1992; 31(1): 36-43.
10. Campbell KE, Oliver DE, Shortliffe EH. The Unified Medical Language System: Toward a collaborative approach for solving terminologic probems. *JAMIA* 1998; 5: 12-16.
11. Jeuken M, A note on models and explanation in biology. *Acta Biotheoretica* 1968; 18(1) 284-90.
12. Hafner, C. and Fridman, N. Ontological foundations for biology knowledge models. *In 4th Inter'l Conf. on Intelligent Systems for Molecular Biology,* St. Louis, MO: AAAI Press 1996; 78-87.
13. Stryer L. Biochemistry. *W.H. Freeman and company/New York,* third edition, 1988, 31.
14. NCBI: *http://www.ncbi.nlm.nih.gov/*