Updating the Genomic Component of the UMLS Semantic Network

Barry Cohen, PhD[1], Yan Chen, MS[1, 2], Yehoshua Perl, PhD[1]
[1] New Jersey Institute of Technology, Newark, NJ ; [2] BMCC, CUNY, New York, NY

Abstract

The UMLS Metathesaurus and the Semantic Network (SN) were created in the absence of a comprehensive curated genomics terminology and before the recent quantitative and qualitative explosion of genomic knowledge. In this paper we evaluate the internal consistency of the SN's categories relevant to genomics and propose changes to improve its ability to express genomic knowledge. We evaluate the completeness of the SN with respect to genomic concepts by extracting genomics vocabulary from leading texts and databases of genomic information and comparing the extracted vocabulary to the SN. We propose corresponding extensions to the SN to fill identified gaps.

Introduction

Our knowledge of genes and genomes is one of the fastest expanding areas in biomedical research. Since the completion of the sequencing of the human genome, the volume of genomic sequence information has continued to expand at an exponential pace. Through techniques of comparative genomics, all of this information sheds light on the functioning of the human genomic system.

The UMLS Metathesaurus is a comprehensive biomedical resource that has been steadily extended by the incorporation of additional source vocabularies. However, as noted by McCray [1], "Some of the UMLS vocabularies contain terminology at the cellular and molecular level, but none has been created specifically for genetic resources." This has resulted in some gaps in its coverage. It is important that a user of genomic knowledge be able to connect it to the general body of biomedical knowledge. The UMLS enables this by integrating more than 100 biomedical source terminologies [2], e.g., the Gene Ontology (GO) [3] allows researchers to report results regarding genes and gene products.

Background

The UMLS Semantic Network [1] (SN) is an upper level terminology for biomedicine composed of broad categories called Semantic Types (STs). A step in the integration of a new source terminology into the UMLS is the assignment of STs of the SN to the concepts being added. In [4], Lomax and McCray discuss the successful mapping of all GO terms into the UMLS. The mapping of GO concepts by the STs, however, was less satisfactory. They note that the SN does not allow for some of the distinctions present in GO because there are a relatively small number of STs at the level of molecular phenomena. They point out that GO makes a distinction between a molecular function that is a "direct [molecular] activity" and a molecular function that consists of an ordered assembly of activities. GO categorizes the latter as a "biological process." They say that "No similar distinction, however, is made within the UMLS semantic network. Thus a large proportion of both molecular function and biological process terms were assigned the same ST, 'Molecular Function' (or its child, 'Genetic Function'), losing much of the resolution present in GO."

The Semantic Types to which GO's cellular component terms were assigned similarly show a lack of sufficient resolution. Most GO cellular component terms were assigned to the ST **Cell Component**[1], but additional children or siblings of **Cell Component** are needed to retain all the semantics captured in GO. Other GO categories not currently available as separate Semantic Types include developmental processes. As we see, a number of new Semantic Types are needed for proper coverage of genomic concepts. For example, the SN has no subhierarchy specifically concerned with the storage, replication and use of genetic information.

In [5], Yu *et al.* identify more than 30 existing UMLS STs as relevant to genomics and suggest extending the SN's coverage of genomics. They propose the addition of six Semantic Types: **Complex** and **Protein Structure**, and the latter's **1D**, **2D**, **3D** and **4D** children. Further, they identify relevant existing relationships and suggest more relationships to enhance genomics coverage in the SN. They deem 24 out of the 53 semantic relationships of the UMLS SN relevant, and add 16 new ones (*createbond, breakbond, follows, releases, signals, transports, activates, promotes, deactivates, similarity_related_to, physically similar to, 1D structure related to, 2D structure related to, 3D*

[1] Semantic types are indicated in bold.

structure related to, 4D structure related to, and *functionally similar to).*

Methods

Previous works have identified elements that would improve the genomic coverage of the SN. We use the work of Yu, *et al.* [5] identifying both STs of the SN that are relevant to genomics and new STs. We address the need expressed by Lomax and McCray [4] to provide finer granularity of molecular activity.

A second method we employ as a basis for identification of genomic semantic types (GSTs) is a manual review of terms used in comprehensive online genomic resources, including Entrez Gene [6] and OMIM (Online Mendelian Inheritance in Man) [7] and several biology, genetics and molecular evolution texts [8, 9, 10]. More than 200 relevant terms were extracted from these sources. We treat these genomics terms in one of two ways. If a term is of sufficiently high frequency, is a broad category and has become standardized in its usage, it is recruited as an ST. More specialized genomic terms are used to measure the inclusiveness of the broader semantic categories. The GSTs defined should be sufficiently broad so that every genomic term encountered is naturally susceptible to assignment of one or more GSTs, and every GST covers numerous concepts. Sometimes a broad category is also a concept in META as a result of being a concept in a source terminology. This is not a contradiction, since a META concept may have many descendents.

Third, we examined the existing genomically relevant categories within the SN for uniformity and internal consistency. For example, as noted in [4], a comparison of the SN's **Biologically Active Substance** subtree with its **Natural Phenomenon or Process** subtree finds that the former has **Immunological Factor** but the latter does not have a corresponding **Immunological Function** or **Immunological Process**. This reflects an inconsistency in the structure of the subhierarchies of the SN.

Understanding of genomic entities and processes often requires that they be viewed from multiple angles. Multiple parents are usually allowed in terminologies, including META, to permit this. We allow multiple parents for STs of the SN, as proposed in the Enhanced Semantic Network (ESN) [11], creating a directed acyclic graph (DAG) structure for the *isa* hierarchy. The proposed modifications of the SN were submitted to two domain experts for review.

Results

Figure 1 shows the modified genomic portions of the Semantic Network. The numbered items in bold italic in Figure 1 follow from the suggestions of the corresponding references. Our additions are in bold, including the concepts assigned multiple parents. A starred entry (*) has multiple parents and thus appears more than once in the hierarchy, in keeping with our suggestion that a directed acyclic graphic structure is required to encompass the multiple subsumption of concepts. For example, an **Enzyme** is both a **Biologically Active Substance** and an **Amino Acid, Peptide, or Protein**, corresponding to perspectives on its biological activity and its structure, respectively.

Genomic terms within the UMLS are currently assigned to the STs **Gene or Genome**; **Nucleic Acid, Nucleoside or Nucleotide**; **Biologically Active Substance**; **Idea or Concept**; **Cell Component**; **Amino Acid, Peptide or Nucleotide**; and **Genomic Function** and their subhierarchies.

The currently defined ST **Gene or Genome** is clearly a particularly important one for genomic concepts. Following our review of existing STs for internal consistency, we propose that this ST be divided into two separate STs, **Gene** and **Genome**. A genome is a collection of one or more macromolecules containing an organism's (or cell's) genetic complement. A gene is one of the functional elements of a genome (that is, it is a *part-of* a genome). The "or" conjunction of **Gene or Genome** implies a joining of types at an equal conceptual level and does not correctly express their relationship.

Molecular Sequence is likewise of fundamental significance for genomics. It is intended, according to its UMLS definition, to comprise the genetic sequences and gene product sequences "reported in the published literature and/or … deposited in … databanks such as GenBank, European Molecular Biology Laboratory (EMBL), National Biomedical Research Foundation (NBRF), or other sequence repositories." Its child STs include **Nucleotide Sequence** and **Amino Acid Sequence**. However, the assignment of **Molecular Sequence** to the **Conceptual Entity** hierarchy fails to capture the concrete character of the sequences in the sequence databases. We propose that **Molecular Sequence** be placed as a child of the ST **Anatomical Structure**. This places it as a grandchild of **Physical Object** and a sibling of **Protein Structure**. **Carbohydrate Sequence** (a child of **Molecular Sequence**) would

more accurately be called **Carbohydrate Structure**, since it is not necessarily linear.

Following our review of genomic terms in the literature, we add to the genomic repertoire of the SN two child STs of **Cell Component**. The first, **Organelle**, would include as child types the major categories of genomic entities that are the persistent repositories of an organism's hereditary information

and are the targets of sequencing (**Chromosome**, **Mitochondrion** and **Chloroplast**). The major genomic databanks (e.g., GenBank) are built on these entities. **Organelle** would also apply to a variety of on-genomic cell components.

This requires identification of the organism's genes and other functional elements, including regulatory and other non-coding elements; identification of the sequence and structure of the expressed products; and mapping of the regulatory networks of the various subsystems of the organism. The proposed child types of **DNA Element** include **Genome, Gene**, **Translated Region**, **Transcribed Region**, **Transcription Factor Binding Site**, **Exon**, **Intron**, **Control Region**, **Promoter** and **Chromosome Band**. A large number of specialized databases are devoted to these various types of elements (e.g., regulatory networks, functional RNAs). There are also specialized databases that contain elements, not necessarily naturally occurring, generated by genetic technologies (Expressed Sequence Tags, for example).

Each **Gene** is of one of two fundamentally different types, according to whether it is a template for a functional RNA or for a protein. Functional RNAs and proteins have many subdivisions according to the structure or function of the gene product. We include STs for these broadest divisions of genes, **Protein Coding Gene** and **RNA Coding Gene**.

In the **Event** hierarchy, we propose a **Biologic Process** ST to encompass ordered sequences of activities. This addresses the observation in [4] that it would be desirable to have a higher degree of granularity in processes at the molecular level, and to be able to discriminate composite processes from simple molecular activities. This is a semantic distinction that was lost in the incorporation of GO terms into the UMLS.

To provide a new ST **Biologic Process** with greater discrimination, three child STs are introduced: **Developmental Process**; **Regulation of Biologic Process** and **Reproduction**. These are major subhierarchies in GO. To improve the granularity of the SN's **Molecular Function** category related to genomics, three child STs are introduced: **Binding**; **Transcription Regulator Activity** and **Translation Regulator Activity.** These also are major subhierarchies in GO.

We requested two domain experts to review the ST hierarchy of Figure 1. Reviewer 1 suggested adding four STs: Multiorganism Process, Cellular Process, Organism Process and Catalytic Function. She also suggested moving two STs to different positions in the hierarchy, which is implemented in the present version. Reviewer 2 suggested removing four STs: Chromosome, Mitochondria and Chloroplast (for uniformity of granularity) and Chromosome Band. The feedback concerning the rest of the hierarchy was positive.

Discussion

We undertook to consider the SN as a conceptual framework for genomics from both internal and external perspectives. Our work shows that the consideration of internal consistency and the consideration of completeness with respect to the state of the field both yield notable improvements in the scope and organization of the SN's GSTs.

The mining of literature for genomics terms proved to be a useful method. However, the journal literature that should be surveyed for a thorough application of literature mining is vast and requires automated natural language processing techniques. We expect to extend our methods to enable a comprehensive literature review in the future.

To be included in the proposed additions to the SN, a GST had to meet rather strict criteria intended to guarantee that it is mature and of general significance. Each new GST must refer to a core biological phenomenon; have a standardized nomenclature; be backed by a very large data set; and have universal (or close to universal) applicability to living things. The result is a small and tightly grouped set of GSTs, closely reflecting the genome sequencing program that has driven advances in our genomic knowledge. We believe that adherence to these criteria provides a strong basis for agreement on their inclusion. However, the boundaries of inclusion are a matter of judgment.

The coherence of biological processes can be expected to be reflected in a high degree of symmetry between entities and events in the respective hierarchies of the SN. Thus, an Action Entity should be complemented by a corresponding Event. In genomic systems, a more complex triangular relationship often exists among the genomic information (encoded in a defined region of a

relatively passive macromolecule), an action agent (a smaller molecule) and an event. An example of such a triad is a transcription factor binding site (a sequence on a DNA molecule), a transcription factor (a molecular complex) and the initiation of transcription. While the categories we have introduced in the SN's **Entity** and **Event** hierarchies are somewhat complementary, such internal coherence does not yet fully exist among them. The STs representing encoded information are more complete than the action agents and events. Completeness and internal consistency should be goals of a more fully elaborated representation of genomic phenomena.

Some of the terms extracted by our literature search do not readily fall under either the **Entity** or **Event** hierarchies of the SN. For example, the UMLS concept Genetic Code is the relation of codons (nucleotide triplets) in a DNA coding region to the corresponding amino acids in the protein gene product. It is a function from a DNA strand to an amino acid sequence. Genetic Code is currently assigned the Semantic Type **Molecular Sequence** in the SN, though it is neither a nucleotide sequence nor an amino acid sequence, but the relation of the two. Measured by how strongly it is conserved and its universal nature, the genetic code must be considered among the basic elements of the system of hereditary information. The categorization of the concepts ancestor/descendant and paralogy and orthology suffer related problems. We will postpone our proposed resolution of these problems to a fuller elaboration of the representation of genomic concepts.

In the absence of a general genomics terminology as a UMLS source, many genomics terms found in the reviewed texts are absent from the Metathesaurus. These include, for example, tetrad, clade, synonymous codon, DNA strand, deme, RNA coding gene and reversion. Despite the incomplete nature of the UMLS as a genomics vocabulary, the SN, as an upper-level ontology, should be planned to accommodate a wide range of terms not yet present. Our understanding of the processes by which the elements of an organism's genetic program interact to replicate, differentiate and modulate the use of the information in the genome are still at an early stage. This is a rapidly developing area and many new concepts may be expected to be added.

We delay the introduction of associative semantic relationships pertinent to genomics until a stable set of GSTs has been established. The work of Yu *et al.* [5] in this regard will guide us in this future research.

We note that one of the key benefits of a representation in a semantic type hierarchy is that it permits the introduction and inheritance of correspondingly specific semantic relations, each with a defined and limited domain of values that participate in the relation.

This is a first version of our work on genomic semantic types. The limited expert review of this work was valuable. Methods for future work could include additional expert review and consensus, as well as automated analysis of the scientific literature to enhance reproducibility and completeness.

Conclusions

The SN was formulated before the quantitative and qualitative explosion of genomic knowledge precipitated by genome sequencing projects. It was designed to categorize the concepts contained in a large number of biomedical vocabularies, none of which focuses on genomics. An expansion of the set of STs of the SN to accommodate new knowledge and research directions is in order. We have proposed here some changes and additions to the set of genomically relevant STs that improve the SN's capacity to capture the current significance of genomics in the biomedical domain.

In the absence of a curated general ontology or terminology of genomics, an independent review of terms in the relevant literature forms an important foundation for the creation of a sound semantics of genomics. Given the significance of genomics to biology and biomedicine, it is important that the SN contain a consistent set of categories that systematically integrate genomic knowledge and link it to other biomedical domains.

Acknowledgments

This work was partially supported by the United States National Library of Medicine under grant R 01 LM008445-01A2. We thank Andrey Rzhetsky and James Cimino for helpful comments and Jane Lomax and Dennis Vitkup for their reviews of the genomic component.

References

1. McCray, A. T., An Upper Level Ontology for the Biomedical Domain, Comp Funct Genom. 2003;4:80–4.
2. http://www.nlm.nih.gov/research/umlssource_faq.html#what_vocab
3. GO Consortium. Creating the gene ontology resource: Design and implementation. Genome Res, 11:1424–1433, 2004.

4. Lomax J. and McCray A. T. Mapping the gene ontology into the unified medical language system. Comparative and Functional Genomics, 5(5):345–361, 2004.
5. Yu, H., Friedman C., Rzhetsky A., Kra P. Representing genomic knowledge in the UMLS semantic network. Proc AMIA. 1999:181–5.
6. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
7. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM.
8. Russell, P. J., Genetics, 1998.
9. Graur, D. and Li, W., Fundamentals of Molecular Evolution, 1999.
10. Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., Darnell, J., Molecular Cell Biology, 2000.
11. Zhang L., Perl Y., Halper M., Geller J., Cimino J.J. An Enriched Unified Medical Language System Semantic Network with a multiple subsumption hierarchy. JAMIA 2004 May-Jun; 11(3):195–206.

Entity
Physical Object
Anatomical Structure
Molecular Sequence
Nucleotide Sequence
Amino Acid Sequence
Carbohydrate Structure
Protein Structure [5]
Primary Structure [5]
Secondary Structure [5]
Tertiary Structure [5]
Quaternary Structure [5]
Fully Formed Anatomical Structure
Body Part, Organ, or Organ Component
Tissue
Cell
Cell Component
Organelle
Chromosome
Mitochondrion
Chloroplast
DNA Element
Genome
Gene
Protein Coding Gene
RNA Coding Gene
Translated Region
Transcribed Region
Transcription Factor Binding Site
Exon
Intron
Control Region
Promoter
Receptor [11]*
Membrane Component [5]
Substance
Chemical Viewed Functionally
Biologically Active Substance
Neuroreactive Substance or Biogenic Amine
Hormone
Enzyme *
Vitamin
Immunologic Factor
Receptor *
Chemical Viewed Structurally
Chemical Complex [4]
Organic Chemical
Nucleic Acid, Nucleoside, or Nucleotide
Amino Acid, Peptide, or Protein
Enzyme [11] *
Inorganic Chemical
Element, Ion, or Isotope
Event
Activity
Occupational Activity
Research Activity
Molecular Biology Research Technique
Phenomenon or Process
Natural Phenomenon or Process
Biologic Process
Developmental Process
Regulation of Biologic Process
Reproduction
Biologic Function
Physiologic Function
Organism Function
Immunologic Function [4]
Cell Function
Molecular Function
Genetic Function
Binding
Transcription Regulator Activity
Translation Regulator Activity
Pathologic Function
Disease or Syndrome
Neoplastic Process
Cell or Molecular Dysfunction
Experimental Model of Disease

Legend

Red Bold: New proposed ST

Green Italic: New proposed ST based on previous work [reference in square brackets]

Blue*: Existing ST with proposed second parent

Grey Background Bold*: Second position for existing ST

Brown: Existing ST in new position

Figure 1: Updated Genomic Component of the UMLS Semantic Network