The Clinical Bioinformatics Ontology: A Curated Semantic Network Utilizing RefSeq Information

M. Hoffman, C. Arnoldi, and I. Chuang

Pacific Symposium on Biocomputing 10:139-150(2005)

THE CLINICAL BIOINFORMATICS ONTOLOGY: A CURATED SEMANTIC NETWORK UTILIZING REFSEQ INFORMATION

M. HOFFMAN , C. ARNOLDI, I. CHUANG

Cerner Corporation, 2800 Rockcreek Parkway
Kansas City, MO 64117, USA

{mhoffman1,carnoldi,ichuang}@cerner.com

Existing medical vocabularies lack rich terms to describe findings that are generated by modern molecular diagnostic procedures. Most bioinformatics resources were designed primarily to support the needs of the research community. We describe the development of a curated resource, the Clinical Bioinformatics Ontology (CBO), a semantic network appropriate for describing clinically significant genomics concepts. The CBO includes concepts appropriate for both molecular diagnostics and cytogenetics. A standardized methodology based on consistent application of RefSeq information is applied to the curation of the CBO in order to provide a reproducible and reliable tool. Challenges related to this curation process are discussed in this paper. At the time of submission the CBO included 4,069 concepts, associated by 8,463 relationships.

1. Introduction

The practice of medicine increasingly utilizes methods based on recent advances in genomics. Diagnostic tests based on the detection of single nucleotide polymorphisms, cytogenetic observations or gene expression patterns are utilized to confirm, classify or monitor inherited or pathological conditions. Clinical information systems will be expected to manage these results and to support the exchange of data based on genomics-related findings. Extension of clinical information systems to manage genomics finding should include the ability to standardize information using a controlled vocabulary and the ability to support genomics-based inference, ideally using ontology-based reasoning.

Current medical vocabularies lack expressions capable of describing the granular findings generated by molecular tests in sufficient detail. SNOMED Clinical Terms® (SNOMED-CT®)[a] is an ontology but lacks concepts appropriate for the detailed description of chromosomal structures below the level of arm and lacks terms for the accurate description of molecular findings. Additionally, there are errors in the hierarchical organization of those genomics-related concepts that are included in SNOMED-CT; for example, “Genome” is positioned as a child of “Gene”. The Logical Observation and Identification Codes (LOINC®) (1) resource has recently incorporated a limited number of

[a] www.snomed.org

molecular diagnostics-related concepts but is not structured as an ontology and thus was not designed to support conceptual-inferencing.

Of the many bioinformatics resources provided by the National Center for Biotechnology Information (NCBI)(2), the Online Mendelian Inheritance in Man™ (OMIM)(3), is the most clearly clinically oriented. However, OMIM is not provided in a machine readable format suitable for integration with relational databases. Furthermore, entries in OMIM apply inconsistent methods of describing positional information. For example, the description of the CFTR gene uses the first position of the DNA sequence in the reference sequence file, 132 bp before the beginning of the coding sequence, for referencing the position of mutations[a]. The HADHA gene content references mutations relative to the beginning of the coding sequence[b], the approach that is consistent with the standard naming conventions of the Human Genome Organization (HUGO)(4;5).

Other bioinformatics resources were not designed to support clinical applications or were not implemented in a machine readable format that is publicly available. The Gene Ontology (GO) (6) has become a valuable tool for the research community and includes concepts describing molecular processes, molecular functions and cellular components. However, GO lacks expressions related to the practice of medicine, for example terms appropriate for the description of chromosomal observations or expressions appropriate for describing human genetic variations of clinical significance. Another bioinformatics resource, the Human Gene Mutation Database (HGMD)(7), provides a curated set of uniquely identified expressions associated with human gene variations but is not available in a machine readable format to the general public. Likewise, the most widely utilized human gene mutation database, dbSNP(8), was designed primarily to support the needs of the research community and lacks the clarity needed to support the clinician.

There have been a variety of efforts to integrate existing bioinformatics resources and medical vocabularies (9-11). While useful for promoting the mining of research data for research purposes, these projects have not filled the gaps and inaccuracies in existing resources. To satisfy the need for a clinically oriented genomics vocabulary, we have developed a curated semantic network, the Clinical Bioinformatics Ontology™ (CBO). This resource combines the attributes of a medical vocabulary with the positional information offered by bioinformatics resources, especially RefSeq (12).

[a] http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=602421
[b] http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=600890

2. Materials

2.1. *RefSeq*

RefSeq is a part of the NCBI collection of bioinformatics resources (2). Characterized genes are assigned RefSeq identifiers which link to GenBank sequence files containing a DNA sequence generated by reverse transcription of the mRNA product of the gene, protein sequence and often a genomic contig file.

2.2. *SNOMED-CT*

SNOMED-CT (version 20030108) was evaluated using the CLUE Browser, version 5.5 (Clinical Information Consultancy, UK)[a].

2.3. *LExScape - Health Language Incorporated*

The CBO was managed using LExScape, a nomenclature management program produced by Health Language Incorporated (HLI) (Aurora, Colorado). LExScape controls the versioning, re-parenting and retirement of concepts. LExScape also assigns a global unique identifier (GUID) to each concept. In this paper we use the naming conventions associated with this system – concepts, facets and terms. Concepts are uniquely identified entities or ideas. Facets are associated with concepts and provide further descriptive information. Terms are alternative names (synonyms, abbreviations) for concepts.

1.4. *Protégé-2000*

Protégé-2000 (Stanford University)[b] is a publicly available frames-based ontology management tool.

1.5. *Controlled Medical Terminology (CMT)™*

The Cerner Controlled Medical Terminology (CMT)™ (Cerner Corporation, Kansas City, MO) is a meta-vocabulary associating most of the common medical vocabularies, including SNOMED and LOINC. CMT enables integration of concept-based capabilities into the Cerner Millennium® suite of clinical software. Cerner Pathnet Helix™ was used to demonstrate the capture of genetic test results in a laboratory information system (LIS) using the CBO.

[a] http://www.clininfo.co.uk/clue5/index.htm

[b] http://protege.stanford.edu

3. Methods

3.1. *Design*

The CBO was designed to reflect a structure similar to that of SNOMED-CT in order to favor compatibility with the phenotypic information that is codified in clinical information systems. Hierarchical "IS_A" relationships ("SUBSUMES" in HLI) were combined with lateral definitional relationships (Table 1) to create the semantic network.

Table 1 – Definitional relationships

Full name	Alias
Has Chromosomal Location	HAS_CHRM_LOC
Arm of	ARM_OF
Band of	BAND_OF
Has Arm Location	HAS_ARM_LOC
Has Band Location	HAS_BAND_LOC
Exon of	EXON_OF
Intron of	INTRON_OF
Allele of	ALLELE_OF
Has Constituent Variant	HAS_CONS_VAR
Nucleotide Variant of	NUC_VAR_OF
Has Effect	HAS_EFFECT
Transcript of	TRANSCRPT_OF
RT Product of	RT_PROD_OF
Amino Acid Variant of	AA_VAR_OF
Encodes	ENCODES
Has Phenotype	HAS_PHENOTYP
Has Location	HAS_LOC
Includes Band	INCLUDS_BAND
Mode of Inheritance	MODE_OF_INH
Had CH3 Status	HAS_CH3_STAT
Source Band	SOURCE_BAND
Target Band	TARGET_BAND

Some significant design decisions included:

- **Method of describing mutations and polymorphisms**. Associating a concept representing a mutation to the gene of origin would violate "IS_A" logic. Therefore, we applied definitional relationships to associate a mutation to its gene. Most of the mutations described in the CBO are documented relative to the cDNA sequence associated with the RefSeq. The model defines a series of relationships that must be traversed to associate a genomic gene concept to a mutation concept.

- **Consistent use of Reference Sequence data.** In order to create a reproducible methodology, we associate gene concepts with a facet that provides the RefSeq identifier for the mRNA derived DNA sequence (12). The sequence content of a Reference Sequence can change as corrections or revisions are submitted, therefore we include a facet with the "gi" identifier of the GenBank sequence in order to capture the version of the gene utilized at the time that content was created (13).
- **Allele naming**. We adopted a generalized allele naming approach that was based on the OMIM style of naming gene alleles {Gene.allele identifier}[a]. CBO allele concepts are associated with OMIM allele identifiers through facets storing the OMIM allele identifier.
- **Description of chromosomal structures.** We applied the standards of the International System for Human Cytogenetic Nomenclature (ISCN)(14) to the naming of chromosomal structures (arms, centromeres, telomeres and bands). To address the multiple resolutions applied to clinical observation of cytogenetic structures (400, 550 and 850), we prefixed chromosomal band concepts with the resolution at which the bands are observed. High resolution bands were associated to lower resolution bands using the BAND_OF definitional relationship. For example, the "550.7q31.2" concept is related to the "400.7q31" concept using the "BAND_OF" relationship.
- **Directionality of definitional relationships**. The nomenclature modeling tool that we utilized does not require explicit statement of inverse relationships because one attribute of a relationship (as modeled in LExScape) is whether or not that relationship automatically includes an implicit inverse.

3.2. *Curation*

Controlling the scope of the CBO was an important design parameter. The primary approach to the selection of concepts that are clinically significant was a detailed analysis of the test catalogues offered by molecular diagnostic laboratories and interviews with these laboratories. Additional clinically significant concepts were identified by reviewing GeneTests (15) and the primary literature. Concepts, as well as their associated terms and facets, were created by a content analyst. The curation process involved utilizing a documented methodology in order to ensure consistency. The white paper describing this methodology will be posted with the content on a publicly available web site (URL to be provided during final edit). After new content was created, a reviewer examined the content for accuracy, consistency with the methodology and completeness.

[a] http://www.ncbi.nlm.nih.gov/Omim/omimfaq.html#numbering_system

3.3. *HLI import and export*

All concepts, relations, facets and terms were imported into LExScape using a java program. After LExScape assigned GUIDs and formally defined the structure of the ontology, the content was exported from LExScape into CSV files. These files were imported into CMT and the Cerner Millennium system using a program written in the Cerner Command Language (CCL)® , a variant of the Structured Query Language (SQL) that is optimized for the clinical information systems developed by Cerner Corporation. The CSV files were also imported into Protégé (Stanford University), with an intermediate conversion to RDF, to further demonstrate the portability of the content.

4. Discussion

Clinical information systems will need an appropriate controlled vocabulary to manage information gathered during clinical genomics-based diagnostic procedures. In order to address the gaps identified in current medical vocabularies and bioinformatics resources, we developed the Clinical Bioinformatics Ontology (CBO). The CBO is structured to provide the benefits of a controlled vocabulary as well as the advantages of a semantic network. The scope of the CBO was limited to concepts currently applied in clinical practice, thus concepts related to microarray-based diagnostics, proteomics and other emerging technologies have not yet been incorporated into the CBO.

Table 2 – Summary of CBO elements (July 2004)

Element	Number
Concept	4069
Nucleotide variant	286
Facet	2110
Term	460
Relationships	8463

Central to the CBO is the use of a content curation process. Resources created through automated annotation can manage high volumes of information, however the resulting output is often heterogeneous in quality; in contrast a curated resource offers the opportunity to enforce a minimal level of annotation and apply a deliberate strategy to the growth of the resource (16). In order for the CBO to be useful in clinical settings, where there are high quality control expectations, we chose to apply a curation approach, despite the limited rate of growth associated with human curation.

Challenges

A variety of challenges were identified during the design of the model, examples of which are discussed below.

1. *Describing mutations using "IS_A" compliant methods.*

During the initial design of the CBO, models in which mutations would be children of the gene to which they are associated were evaluated. The inability to represent mutations in this way and maintain "IS_A" relationships led us to create a relatively flat model in which definitional relationships are used to describe clinically significant mutations. In our model all single nucleotide polymorphisms are children of the "Human Nucleotide Variant" concept, which is a child of "Human Nucleic Acid Variant" (of which the other children are "Human Allele" and "Human Haplotype").

2. *The lack of a consistent resource providing the number of exons in a gene.*

Many molecular diagnostic laboratories have begun to perform diagnostic sequencing. Often only a few of the exons for a particular gene are sequenced. The majority of the genes represented in the CBO are associated to concepts representing their constituent exons and introns. The efforts to generate these concepts were often hindered by the lack of a consistent resource providing the number of exons for a gene. Our approach was to utilize the RefSeq tables provided by the UCSC (17) combined with a review of literature relating to the gene being curated. The UCSC values are based on a computed number of exons generated by automated alignment of the genomic DNA sequence with the RefSeq mRNA sequence (18). Often the values determined by this method do not agree with published information (Table 3). When the UCSC value agreed with the published literature we added exon and intron concepts to the CBO. When they did not agree we deferred the creation of the exon and intron concepts.

Table 3 – Disparities in exon count

Gene	UCSC exons	Published value
CFTR	27	27 (19)
MTHFR	12	11 (20)
RHD	11	10 (21)

A further issue in naming exons is the use of suffixes to describe short exons (5a, 5b). We applied an absolute approach in which each exon, regardless of size, is assigned a unique integer value relative to the beginning of the gene. The convention used for naming exons in the CBO is "Gene name".e."exon number", with a similar approach for naming introns. Thus the 3rd exon for the CFTR gene is named "CFTR.e.3".

3. *Reconciling inaccuracies in existing resources and the literature.*

Another challenge that we identified were significant discrepancies between descriptions of mutations (or polymorphisms) in the literature and the reference information provided by RefSeq. One of our goals was to adopt a methodology which could be applied by any user yet result in the same concept naming conventions. On occasion the application of this approach generated concepts whose names disagree with values in the published literature.

For example, the human platelet antigen ITGA2 has polymorphisms that result in antigenic diversity that are factors in neonatal alloimmune thrombocytopenia. The paper describing the nomenclature of these antigens refers to the RefSeq NM_002203 and cites the polymorphism 1600G>A with the amino acid substitution result of E505K (22). Analysis of the DNA sequence for the coding region (CDS) associated with this RefSeq file (gi:6006008) clearly shows an "a" at this position:

```
                                               *
1561 aagaaaagagg aaggaagagt ctacctgttt actatcaaaa agggcatttt
```

In order to comply with our methodology we created the a concept to describe the mutation relative to the RefSeq, "ITGA2.c.1600A>G", and the associated amino acid substitution concept, "ITGA2.p.K534E". These concept names are accurate relative to the RefSeq but will require close communication with clinicians accustomed to the naming used in the literature. A log of such exceptions was created and is included with the CBO.

4. *The lack of a widely accepted system for describing alleles.*

Patient results are often reported to clinicians in terms that describe an aggregate of mutation or polymorphism findings for a particular gene. This required us to adopt an approach to allele naming that could be generalized. While many genetic systems, including the HLA locus (23)and the CYP2D6 gene[a] have locus or gene specific allele naming conventions, the only resource that offers a fairly generic approach to allele description was OMIM. However, many clinically significant alleles do not have OMIM allele identifiers; for example many of the CYP2D6 gene alleles are under-represented in OMIM. Therefore, we adopted an OMIM-like approach to allele naming in which the allele is named using the convention: [Gene name][.][unique allele number]. The allele concepts are related to their constituent nucleic acid variations through the definitional relationship type of "HAS_CONS_VAR". This approach can be extended to describe haplotypes, in which a haplotype is described by definitional relationships to the constituent alleles.

Applications of the CBO

Importing the CBO into a commercial healthcare information system allowed us to demonstrate that the general benefits of a controlled vocabulary

[a] www.imm.ki.se/CYPalleles/cyp2d6.htm

and a semantic network are applicable in the context of a clinical genomics software system. Discrete results captured in the PathNet Helix LIS system are associated to a CBO GUID. The benefits of this approach include:

- The comparison of results between patients is facilitated.
- The retrieval of historic results is enabled.
- The exchange of data is greatly simplified.

By associating discrete clinical genomics results to a CBO GUID, the user is free to utilize any display name for a genetic test result. For example, molecular diagnostics labs often perform allele-specific PCR to detect mutations in genes, such as the CFTR gene. Laboratories often report these DNA-based findings in terms of the amino acids substitutions that they cause. For example, the mutation most commonly associated with cystic fibrosis, the deletion of nucleotides 1522-1524 of the CFTR gene, results in the deletion of the phenylalanine at position 508 in the CFTR protein. Users can display their results using expressions such as: _508, del 508, 508 del, delF508 or any other name, but the underlying association between the discrete result and the GUID for the concept "CFTR.c.1522-1524del" remains.

Our implementation of the CBO within a commercial HIS also allowed us to demonstrate the benefits of an ontology applied within this context. Areas in which the semantic network aspects of the CBO have been demonstrated include:

- **Simplification of the database implementation process**. Using the series of definitional relationships connecting a gene concept to associated mutation concepts, users representing an orderable test procedure in the system can quickly access likely discrete results for that procedure and reduce the amount of effort needed to build the representations of these discrete results in the system.
- **Utilize the CBO as reference data**. Software applications designed to utilize the CBO can be implemented to access the CBO as a source of reference data. For example, the "MODE_OF_INHERITANCE" relationship can enable or disable display behaviors that are appropriate for X-linked conditions.
- **Flexible queries**. Most importantly, the semantic network created by the CBO allows research-oriented users to efficiently design queries that apply the CBO model for questions such as "Retrieve all results related to chromosome 7" or "retrieve all results related to arm p of chromosome 10" or "retrieve all results related to exon 28 of the von Willebrand factor gene".

A subset of the CBO is provided in Table 4 in order to demonstrate how the combination of "IS_A" (SUBSUMES) and definitional relations were used to associate concepts related to the CFTR gene, including the mutation discussed above.

Table 4 – Example relationships associating CFTR related concepts

Source Concept	Relationship Type	Target concept
Human Gene	SUBSUMES	(gDNA).CFTR
(gDNA).CFTR	HAS_CHRM_LOC	Chromosome 7
(gDNA).CFTR	HAS_ARM_LOC	7q
(gDNA).CFTR	HAS_BAND_LOC	850.7q31.2
850.7q31.2	BAND_OF	550.7q31.2
550.7q31.2	BAND_OF	400.7q31
Human mRNA	SUBSUMES	(mRNA).CFTR.0
(mRNA).CFTR.0	TRANSCRPT_OF	(gDNA).CFTR
Human Protein	SUBSUMES	(AA).CFTR.0
(mRNA).CFTR.0	ENCODES	(AA).CFTR.0
Human cDNA	SUBSUMES	(cDNA).CFTR.0
(cDNA).CFTR.0	RT_PROD_OF	(mRNA).CFTR.0
(gDNA).CFTR	MODE_OF_INH	Autosomal
Human Exon	SUBSUMES	CFTR.e.11
Human Amino Acid Variant	SUBSUMES	CFTR.p.508delF
CFTR.p.508delF	AA_VAR_OF	(AA).CFTR.0
Human Nucleotide Variant	SUBSUMES	CFTR.c.1522_1524del
CFTR.c.1522_1524del	NUC_VAR_OF	(cDNA).CFTR.0
CFTR.c.1522_1524del	HAS_EFFECT	CFTR.p.508delF
CFTR.c.1522_1524del	HAS_LOC	CFTR.e.11

Queries designed to use these relationships can thus support a wide variety of research applications using clinical results codified using the CBO.

While the primary focus of this work was the implementation of the CBO in a commercial HIS, we have also converted the ontology into formats that will support wider utilization, including RDF and Protégé.

We have developed a resource that extends the scope of existing medical vocabularies and benefits from the ability to standardize positional genomic information utilizing RefSeq information. The combined attributes of a controlled vocabulary and a semantic network allow the CBO to be useful for both the delivery of care and research applications. We continue to expand the scope and depth of the content provided in the CBO. Areas of current effort include representation of pathogen-related concepts, including those needed to describe HIV genotype findings, representation of cytogenetic abnormalities and adding further support for the representation of splice variants. The CBO content is available in CSV and RDF formats at www.cerner.com/cbo at no cost to non-commercial users.

Acknowledgments

We acknowledge the contributions of Siew-Tien Lai, Kevin Power, Ginger Kuhns, Chris Finn, Roll-Jean Cheng, Daniel Bellissimo, Barbara Zehnbauer, Arthur Hauck, David McCallie, Doug McNair and Daniel Peterson. This work was funded in its entirety by Cerner Corporation.

References

1. Huff SM, Rocha RA, McDonald CJ, et al. Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. J Am Med Inform Assoc. 1998;5:276-92.
2. Wheeler DL, Church DM, Edgar R, et al. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res. 2004;32 Database issue:D35-40.
3. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2002;30:52-5.
4. Antonarakis SE. Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group. Hum Mutat. 1998;11:1-3.
5. den Dunnen JT, Antonarakis SE. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat. 2000;15:7-12.
6. Harris MA, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 2004;32 Database issue:D258-61.
7. Stenson PD, Ball EV, Mort M, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003;21:577-81.
8. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29:308-11.
9. Lussier YA, Sarkar IN, Cantor M. An integrative model for in-silico clinical-genomics discovery science. Proc AMIA Symp. 2002;469-73.
10. Cantor MN, Lussier YA. A knowledge framework for computational molecular-disease relationships in cancer. Proc AMIA Symp. 2002;101-5.
11. Cantor MN, Lussier YA. Putting data integration into practice: using biomedical terminologies to add structure to existing data sources. Proc AMIA Symp. 2003;125-9.
12. Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res. 2001;29:137-40.
13. Andreas Baxevanis BFFO. Bioinformatics: A practical guide to the analysis of genes and proteins. ed. John Wiley & Sons, Inc, 2001:
14. AnonymousAn International System for Human Cytogenetic Nomenclature (1985) ISCN 1985. Report of the Standing Committee on Human Cytogenetic Nomenclature. Birth Defects Orig Artic Ser.

1985;21:1-117.

15. Pagon RA, Tarczy-Hornoch P, Baskin PK, et al. GeneTests-GeneClinics: genetic testing information for a growing audience. Hum Mutat. 2002;19:501-9.
16. Pruitt KD, Katz KS, Sicotte H, Maglott DR. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. Trends Genet. 2000;16:44-7.
17. University of California Santa Cruz. http://genome.ucsc.edu/cgi-bin/hgText.
18. Terry Furey, personal communication. (GENERIC)
19. Claustres M, Laussel M, Desgeorges M, et al. Analysis of the 27 exons and flanking regions of the cystic fibrosis gene: 40 different mutations account for 91.2% of the mutant alleles in southern France. Hum Mol Genet. 1993;2:1209-13.
20. Goyette P, Pai A, Milos R, et al. Gene structure of human and mouse methylenetetrahydrofolate reductase (MTHFR). Mamm Genome. 1998;9:652-6.
21. Wagner FF, Gassner C, Muller TH, Schonitzer D, Schunter F, Flegel WA. Molecular basis of weak D phenotypes. Blood. 1999;93:385-93.
22. Metcalfe P, Watkins NA, Ouwehand WH, et al. Nomenclature of human platelet antigens. Vox Sang. 2003;85:240-5.
23. Marsh SG. Nomenclature for factors of the HLA system, update March 2004. Tissue Antigens. 2004;64:108-9.