Terminological Mapping for High Throughput Comparative Biology of Phenotypes

Y.A. Lussier and J. Li

Pacific Symposium on Biocomputing 9:202-213(2004)

TERMINOLOGICAL MAPPING FOR HIGH THROUGHPUT COMPARATIVE BIOLOGY OF PHENOTYPES

Y.A. LUSSIER[1,2] AND J. LI[1]
1- Department of Biomedical Informatics,
2- Department of Medicine,
Columbia University College of Physicians and Surgeons,
New York, NY 10032 USA
E-mail: lussier@dbmi.columbia.edu

Comparative biological studies have led to remarkable biomedical discoveries. While genomic science and technologies are advancing rapidly, our ability to precisely specify a phenotype and compare it to related phenotypes of other organisms remains challenging. This study has examined the systematic use of terminology and knowledge based technologies to enable high-throughput comparative phenomics. More specifically, we measured the accuracy of a multi-strategy automated classification method to bridge the *phenotype gap* between a phenotypic terminology (MGD: Phenoslim) and a broad-coverage clinical terminology (SNOMED CT). Furthermore, we qualitatively evaluate the additional emerging properties of the combined terminological network for comparative biology and discovery science. According to the gold standard (n=100), the accuracies (precision | recall) of the composite automated methods were 67% | 97% (mapping for identical concepts) and 85% | 98% (classification). Quantitatively, only 2% of the phenotypic concepts were missing from the clinical terminology, however, qualitatively the gap was larger: conceptual scope, granularity and subtle, yet significant, homonymy problems were observed. These results suggest that, as observed in other domains, additional strategies are required for combining terminologies.

1 Introduction

Comparative biological studies have led to remarkable biomedical discoveries such as evolutionarily conserved signal transduction pathways (*C. elegans*) and homeobox genes (*D. melanogaster*). Recently, comparative genomic studies to elucidate conserved gene functions have made significant advances principally via complementary integrative strategies such as functional genomics and standard notations for gene or gene function (e.g., Gene Ontology[1]). However, there is a pressing demand of technologies for greater integration of phenotypic data and phenotype-centric discovery tools to facilitate biomedical research[2,3,4,5,6,7,8,9,10]. While automated technologies permit increasingly efficient genotyping of organisms' cohorts across distinct species or individuals with distinct phenotype, our ability to precisely specify an observed phenotype and compare it to related phenotypes of other organisms remains challenging[11] and does not match the throughput capabilities of genotypic studies. Further, phenotypic "qualifiers" span biological structures and

functions extending from the nanometer to populations[12]: proteins, organelles, cell lines, tissue, Model Organism, clinical, genetic and epidemiologic databases. This diversity of scales, disciplines and database usage[13] has lead to an extensive variety of uncoordinated phenotypic notations including 1) differences in the definition of a phenotype[14] (e.g. trait, quantitative traits, syndromes), 2) differences in the terminological granularity and composition[15,16,17,18] and 3) distinct usage of identical terms according to the context (e.g. organism, genotype, experimental design, etc.). For example, there are multiple phenotypic terms that illustrate various granularities related to the eye: Iris dysplasia (goniodysgenesis)[19] [OMIM] , MP:0002092 eye: dysmorphology [Phenoslim] [52], uveitis severity [RGD][20], 368808003 Aberrant retinal artery [SNOMED CT], 81745001 Entire eye [SNOMED CT]. Moreover, the lack of timely and accurately access to relevant phenotypes across databases is another limiting factor that hinders the progress of phenotypic research.

The heterogeneity of phenotype notation can be found in both the clinical and biological databases. While each Model Organism Database Systems has standardized the phenotypic notation for its own research community, bridging the gap of phenotypic data across species remains a work in progress. In this regard, the Phenotype Attribute Ontology (PAtO) is an initiative stemming from the Gene Ontology Consortium[21] to derive a common standard for various existing phenotypic databases. In addition, the standardization of the database schema emerging from the PAtO collaboration will considerably increase the interoperability of phenotypic databases and may also clarify problems related to the terminological representation. In contrast, while heterogeneous database systems have been shown to unify disparate representational database *schema*[22,23], to our knowledge, the semantic modeling of the notation representation remains manually edited (e.g., structural naming differences, semantic differences and content differences).[24] In addition, these general-purpose heterogeneous database systems have not been specifically adapted to the complexity of phenotypic data reuse for comparative biology and genomics. The most prominent barrier to the integration of heterogeneous phenotypic databases is associated with the *notational (terminological) representation*. While terminologies can be manually or semi-automatically integrated, as illustrated by the meta-terminologies (e.g. Unified Medical Language System), such a process is both time consuming and labor expensive[25,26]. An alternative approach employing ontology[27,28] and lexicon-based mapping utilizes knowledge-based and semantic-based terminological mapping[29,30,31,32,33,34]. While single-strategy mapping systems have demonstrated limited success (only capable of mapping 13 - 60% of terms[35,36,37,38]), systems using a methodical combination of multiple mapping methods and semantic approaches have demonstrated significantly improved accuracy[39,40,41,42].

In our current study, we have developed an automated multi-strategy mapping method for high throughput combination and analysis of phenotypic data deriving from heterogeneous databases with high accuracy. Further, this mapping strategy also

allowed us to assess the qualitative discrepancies of phenotypic information between a clinical terminology and a phenotypic terminology.

2 Materials

*2.1 Phenoslim terminology (**PS**)*

Phenoslim is a particular subset of the phenotype vocabularies developed by Mouse Genome Database[52] (MGD) that is used by the allele and phenotype interface of MGD as a phenotypic query mechanism over the indexed genetic, genomic and biological data of the mouse. We used the 2003 version of PS containing 100 distinct concepts in our study. MGD is also currently developing comprehensive mammalian phenotype ontology and the Phenotype Attribute Ontology via collaboration with the Gene Ontology Consortium.

*2.2 Systematized Nomenclature of Human and Veterinary Medicine—Clinical Term® (**SNOMED CT**)*

The SNOMED CT terminology[53] (version 2003) is a comprehensive clinical ontology that contains about 344,549 distinct concepts, 913,697 descriptions (test string variants for a concept). SNOMED-CT satisfies the criteria of controlled computable terminologies and, in addition, provides an extensive semantic network between concepts, supporting polyhiearchy and partonomy as directed acyclic graphs (DAGs) and twenty additional types of relationships. It also contains a formal description of "roles" (valid semantic relationships in the network) for certain semantic classes. SNOMED CT has been licensed by the National Library of Medicine for perpetual public use as of 2004 and will likely be integrated to UMLS.

*2.3 The Unified Medical Language System® (**UMLS**) and Norm*

UMLS[54] is created and maintained by the National Library of Medicine. The 2003–version of the UMLS consisting of about 800,000 unique concepts and relationships taken from over 60 diverse terminologies were used in our studies. In addition, UMLS includes a curated semantic network of about 120 semantic types overlying the terminological network. Moreover, it contains an older version of SNOMED (SNOMED 3.5, 1998) that houses about half the number of concepts and descriptions of the SNOMED –CT. By design, the relationships found in the source terminologies in UMLS are not curated. Thus transformations over the unconstrained UMLS network are required to obtain a DAG and to control convoluted terminological cycles.[55]

Norm is a lexical tool available from the UMLS.[56] As its name implies, *Norm* converts text strings into a normalized form, removing punctuation, capitalization, stop words, and genitive markers. Following the normalization process, the remaining words are sorted in alphabetical order.

2.4 Applications and Scripts

All the applications and scripts pertaining to implementation of the methods discussed in this paper were written in Perl and SQL. The Database used was IBM BD2 for workgroup, version 7. Additionally, the Norm component of the UMLS Lexical Tools was obtained from the National Library of Medicine in 2003. Applications were run on a Dual-processor SUN UltraSparc III V880 under the SunOS 5.8 operating system.

3 Methods

3.1 Mapping of the Phenotypic Terminology to SNOMED CT

Phenoslim was mapped to SNOMED CT using the Molecular Medical Matrix (M^3) tools that we have developed[57,39,40,41], an architecture that integrates lexical, terminological/conceptual and semantic approaches to methodically take advantage of pre-coordination and post-coordination mechanisms. The specific methods used sequentially were a) decomposition of Phenoslim concepts in components, b) normalization of Phenoslim and SNOMED CT, c) mapping of PS components to SNOMED CT, d) conceptual processing, and e) semantic processing. Steps a), b) and c) are "term processing" steps that have been separated for clarity. Retired concepts and descriptions of SNOMED were not used in the study, though they are present in the SNOMED files.

a. ***Decomposition*** *of Phenoslim concepts in components*. Each Phenoslim concept is represented by one unique text string consisting of several words. Every combination of word was generated for each unique text string (including the full string) and mapped back to the original concept. A *terminological component* (TC) is a string of text consisting of one of these combinations.
b. ***Normalization*** *of Phenoslim and SNOMED CT*. Each terminological component of Phenoslim and each term associated with a SNOMED CT concept (SNOMED descriptions) was normalized using *Norm* (ref. material section).
c. ***Mapping*** *of PS components to SNOMED CT*. Subsequently, each normalized TC was mapped against each normalized SNOMED description using the DB2 database.

Table 1 Included Semantic Classes of SNOMED CT

Concept Identifier	SNOMED CT Concept Name
257728006	Anatomical Concepts
118956008	Morphologic Abnormality
64572001	Disease (disorder)
363788007	Clinical history/examination
246188002	Finding
246464006	Functions
105590001	Substance
243796009	Context-dependent categories
246061005	Attribute
254291000	Staging and scales
71388002	Procedure
362981000	Qualifier value

d. ***Conceptual Processing***. This process simplifies the output of the mapping methods. The Conceptual Processor is a database method that identifies all distinct pairs of conceptual identifiers of Phenoslim and SNOMED CT (PS-CT Pairs) that have been mapped by the previous terminological processes.

e. ***Semantic Processing***. The semantic processing consists of two successive subprocesses: **(i)** ***semantic inclusion criteria***, and **(ii)** ***Subsumption***. For Inclusion criteria, mapped SNOMED CT concepts were sorted according to the criteria "that they must be a descendant of at least one semantic class shown in table 1". This process eliminates erroneous pairs arising from homonymy of terms due to the presence of a variety of semantic classes in SNOMED that are irrelevant to phenotypes. An inclusion criteria was chosen since valid concepts may inherit multiple semantic classes. The list of SNOMED codes related PS concept was further reduced by subsumption with the relationships found in the relationship table of SNOMED as follow: two ancestor-descendant tables (one from the "is-a" relationship of the relationship table of SNOMED CT and another one from the partonomy relationships "is part of") were constructed. Each network of SNOMED CT concepts paired to a unique PS concept was then recursively simplified by removing "is-a" ancestors that subsume other concepts of the network concept, based on the hypothesis that most specific match is also the most relevant. The same procedure was repeated for the "is part of" relationship. Further, additional relationships of the disease and finding categories were explored in the relationship table and the concept related to a disease or finding was considered subsumed and then removed (within the scope of SNOMED concepts paired to the same PS concept). The remaining set of PS-CT pairs were considered valid for the evaluation.

3.2 Quantitative Evaluation of the Mapping Methods

The mapping methods previously described produces from none to multiple putative SNOMED concepts for every Phenoslim concept. Every group of distinct SNOMED concepts related to a unique PS concept was further assessed according to the following criteria: (i) classification - the SNOMED CT concepts are valid classifier or descriptor of part of the Phenoslim concept (Good/Poor), (ii) identity - the meaning of the SNOMED CT concept is exactly the same as that of the Phenoslim concept, (iii) completeness of representation of the meaning by SNOMED concepts, (iv) redundancy of representation of SNOMED concepts, (v) presence of erroneous matches. In addition, SNOMED CT was looked up to find an identical identifier or a class that could represent every PS concept that was not paired using the automated method. The problem of organizing the post-coordinated set of SNOMED concept was not addressed. We measured the efficacy of the mapping method using precision and recall. 3.1 Mapping of the Phenotypic Terminology to the Clinical one

3.3 Qualitative Evaluation of Mapping Problems between the Clinical and Phenotypic Terminologies

The qualitative evaluation and discussions focus on the description of types of mapping problems encountered, their methodological cause and proposed avenues of further research.

4 Results and Discussion

Using the mapping methods of M^3, every combination of words contained in each term associated with the 100 concepts of Phenoslim were computed yielding 4,016 terminological components. These components were processed in Norm by every possible mapping with a SNOMED –CT description calculated in DB2 in less than 2 minutes (about 3,5 billion possible pairs). 4,842 distinct terminological pairs were found. The conceptual processing reduced this number to 1,387 pairs between Phenoslim and SNOMED CT concepts. As shown in table 2, the final semantic processing provided the final set consisting of 740 distinct pairs (426 pairs did not meet the semantic inclusion criteria and 221 pairs were removed by subsumption). Three Phenoslim concepts were not mapped, one of which could not be mapped or classified in SNOMED CT (the only true negative map). 79 PS concepts were fully mapped to a valid composition of SNOMED concepts, 15 of which also contained one erroneous and superfluous SNOMED code. 18 PS concepts were incompletely mapped, two of which also contained an erroneous and superfluous concept. Overall, 18 concepts were also redundantly mapped (not shown in

Table 2. Evaluation of the Quality of the Mapping between each Group of SNOMED Concepts associated to each Concept of Phenoslim

		Validity of the Mapping to a Cluster of SNOMED Concepts	
		Valid	False
Phenoslim's Concepts Mapped by M3	**Complete Map (identity and classification)**	64	15
	Incomplete Map (classification)	18	2

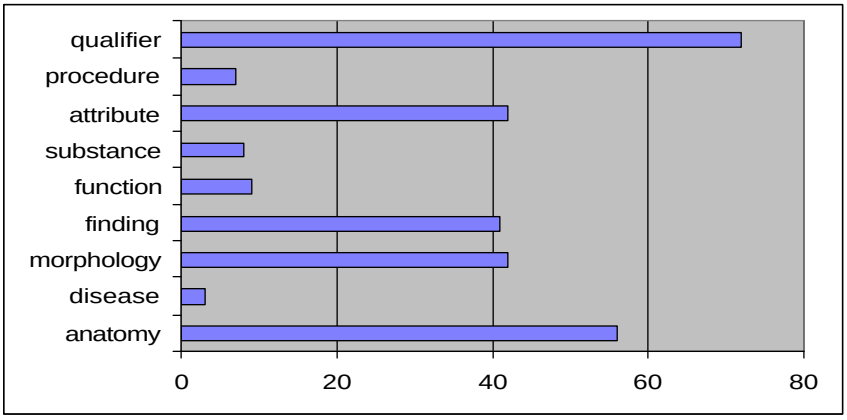

figure 1. Proportion of Phenoslim Concepts that can be mapped to the Semantic Types of SNOMED CT described in Table 1 (%)

the table) – having more than one representation of the same concept or an overlapping group of concepts. Figure 1 shows the proportion of Phenoslim concepts that can be classified to the semantic types of SNOMED, on average each concept is mapped to 2.9 semantic classes.

4.1 Quantitative Evaluation

Norm and the conceptual processing performed together at a precision of 11% (TP=64+18, FP=15+426+221). The precision of M^3's terminological classification accuracy is 98% (TP=725, FP=15). The precision and recall of M^3 to *classify* Phenoslim concepts in SNOMED CT are 85% and 98%, respectively (TP= 64+18, FP=15, FN=2); while the accuracy scores are 67%(precision) and 97%(recall) for M^3 used to map the *full meaning* in SNOMED (TP= 64, FP= 15+18, FN=2).

Table 3. Examples of Problematic Mappings

Mapping Problem	Examples	
	Phenoslim	**SNOMED**
(i) erroneous mapping	"…premature death"	"*immature*" + "death"
(ii) partial mapping	"Hematology…"	Partially mapped missing "hematological system"
(iii) relevant mappings omitted by M[3]	"…postnatal lethality""	"post*neo*natal death"
(iv) redundancy	"coat: hair texture defects"	"hair texture (body structure)", "Texture of hair (observable entity), Hair texture, function (observable entity)
(v) ambiguity	"renal system…",	Including the bladder, the urogenital?
(vi) inconsistency	"neurological/behavioral: … movement anoma*lies*" "neurological/behavioral: … nociception abnorma*lities*"	
(vii) Not in SNOMED	"Coat…", "Vibrissae…"	-
(viii) Context / Representation Scope	"Embryonic…"	"Fetal…" + "Embryonic…"

4.2 Qualitative Evaluation and Discussion

Table 3 illustrates examples of mapping problems. Erroneous mapping occurred for primarily due to slightly different meanings of related concepts with taken out of their context. For example, the conepts "human fetus" (>8wks gestation) and "human embryo" (<8wks) are subsumed by the concept "mammalian embryo" (vertebrate at any stage of development prior to birth). In SNOMED, the parent of fetus and embryo is "developmental body structure" which is the one desired for mapping this mammalian concept. In addition, SNOMED is used for human and veterinary purposes, thus the representation of "embryo" probably requires reengineering as well. The absence of "unaccompanied" adjectival forms of anatomical locations and systems contributed to the majority of the partial mapping problems. In contrast to SNOMED CT, SNOMED 98 in the current UMLS version contains adjectives mapped to the anatomical structure for corneal, skeletal, cellular, etc. In SNOMED CT, these adjectival forms are "accompanied" of the qualifier "structure" or "system structure" or "entire" as in "skeletal system", "skeletal system structure" or "entire skeleton". With additional semantic information in the phenotype terminology (e.g., anatomical location, or system), one could easily pre-process and extend terms with this contextual information before submitting them to norm. Some redundancy can be solved by enriching SNOMED CT with a complete network of relationship: "the entire central nervous system" does not have a partonomy relationship with the "entire nervous system" which led to an overlap of mapping. More specifically for

phenotypes of model organisms and genetics, the following concepts are incompletely conceptualized in SNOMED: "*normal embryogenesis*", "*tumor resistance*", "*tumor sensitivity*", or "*maternal effect*".

While significant efforts have been put forward to address the problems arising from *context, scale and granularity* in mediated schema, interoperability of databases and integration of ontologies, these three issues afflict the manual mapping of terminologies and, as demonstrated in this study, become daunting in presence of automated mapping methods due to rapid amplification. A careful modeling of semantic criteria could further improve the accuracy but may require machine learning approaches to avoid overtraining. For example, a phenotype must necessarily have an anatomical local coded or explicitly mapped from the relationships of its coded concept, to help discriminate between completely and incompletely mapped concepts. Context and scale from the source terminology can be processed as additional semantic criteria: phenotypes from the yeast should map to cellular and smaller SNOMED concepts, etc.

Finally, once coded in SNOMED, additional classification properties emerge from the associated anatomical locations: regional anatomy, tissular anatomy, cellular, subcellular anatomies, functional anatomy, organ/system anatomy. IN addition, the whole network can be considered as a semantic filter as it is generally consistent due to the rigorous representation language underlying the development of SNOMED CT.

6 Caveats and Implications for Future Work

It is important to point out that the manual curation used in the present evaluation was carried by one expert and employed a relatively small, domain-specific subset of the mammalian phenotypes. Mapping the phenotypes of yeast, worm or Drosophila may not yield as good accuracies and are currently investigated. The redundancy of terminological representation has not been addressed and remains necessary for automated processing. Knowledge engineering and additional studies are required to understand how phenotypes can be automatically integrated across species. Nonetheless, venues such as semantic constraints on the scale of the mapping appear promising: mapping yeast to structures and morphologies smaller than a cell, etc. Finally, more comprehensive approaches than lexical ones are required to interoperate the intricate combinations of implicit and explicit semantics nested in the database schema of complex biomedical databases.

7 Conclusions

Phenotypic analyses are critical to unlock the gene-disease relationships of complex diseases. The requirements for high throughput phenotypic genomics in which very large numbers of phenotype variants are related to a wide range of genes or gene patterns further motivate our research and development of the proposed methods. In addition, while manual mapping and the methathesaurus approaches remain the gold standards for accuracy, they are rate limiting. M^3 will require additional improvements to provide accurate solutions to the obstacles of phenotypic research, yet in its present condition it can automatically keep pace with new representations of phenotypes as they appear in databases. We are concurrently addressing the limitations of M^3 with additional semantic and language understanding tools.

Acknowledgments

Partial Support for this work came from a New York State Office of Science, Technology, and Academic Research (NYSTAR)-sponsored Center for Advanced Technology at Columbia University (Grant C020054).

References

[1] Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. *Gen Res.* 11(8):1425-1433. (2001)

[2] Freimer N, Sabatti C. The human phenome project. *Nat Genet.* **34**(1):15-21(2003)

[3] Gerlai R. Phenomics: fiction or the future? *Trends Neurosci.* **25**(10):506-9(2002).

[4] Bogue CW. Genetic Models in Applied Physiology: Invited Review: Functional genomics in the mouse: powerful techniques for unraveling the basis of human development and disease. *J Appl Physiol.* 2003 Jun;94(6):2502-9.

[5] Pool R, Esnayra. Bioinformatics – Converging Data to Knowledge Workshop Summary. Borad on Biology, Commission on Life Sciences. National Research Council. *National Academy Press* 41p (2000)

[6] Altman RB & Klein TE. Challenges for Biomedical Informatics and Pharmacogenomics. *Ann Rev Pharmaco & Toxicol*. **42**:113-133. (2002)

[7] Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet.* **33** Suppl:228-37(2003)

[8] Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. *Science*. **300**(5617):286-90(2003)

[9] Balmain A, Gray J, Ponder B. The genetics and genomics of cancer. *Nat Genet.* **33** Suppl:238-44(2003)

[10] Peltonen L, McKusick VA. Genomics and medicine. Dissecting human disease in the postgenomic era. *Science.* **291**(5507):1224-9 (2001)

[11] Navarro JD, Niranjan V, Peri S, Jonnalagadda CK, Pandey A. From biological databases to platforms for biomedical discovery. *Trends Biotechnol.* **21**(6):263-8(2003)
[12] Blois MS. Information in Medicine: The Nature of Medical Descriptions. Berkeley, California: University of California Press(1984)
[13] Rector AL, Rogers J, Roberts A, Wroe C. Scale and context: issues in ontologies to link health- and bio-informatics. *Proc AMIA Symp*:642-6(2002)
[14] Mahner M. Kary M. What exactly are genomes, genotypes and phenotypes? And what about phenomes?. J Theoret Biol **186**(1):55-63(1997).
[15] Elkin PL, Tuttle MS, Keck K, Campbell K, Atkin G, Chute C. The role of compositionality in standardized problem list generation. *Proceedings MEDINFO*, 660-664(1998)
[16] Elkin PL, Bailey KR, Chute C. A randomized controlled trial of automated term composition. In Chute CG, ed. Proceedings *AMIA Ann. Symp*, 765-774(1998)
[17] Mays E, Weida R, Dionne R, et al.. Scalable and expressive medical terminologies. In Cimino JJ, ed. Proceedings *AMIA Ann Symp*, 259-263(1998)
[18] Stuart NJ, Nels OE, Lloyd F, Tuttle MS, William CG, Sherertz DD. Identifying concepts in medical knowledge. *MEDINFO Proc*, 33-36(1995)
[19] Online Mendelian Inheritance in Man, OMIM (TM). Johns Hopkins University, Baltimore, MD. MIM #:270240:July 12, 2003: http://www.ncbi.nlm.nih.gov/omim/
[20] Steen, R.G., et al, 1999. A high density integrated genetic linkage and radiation hybrid map of the laboratory rat. Research Genetics, Rat Genome Database, ftp://rgd.mcw.edu/pub/publications/1999/steen_genome_research/
[21] Ashburner M, Ball CA, Blake JA, Botstein D, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**(1):25-9(2000)
[22] Hucka M, Finney A, Sauro HM, Bolouri H, Doyle J, Kitano H. The ERATO Systems Biology Workbench: enabling interaction and exchange between software tools for computational biology. *Pac Symp Biocomput.* 450-61(2002)
[23] Mork P, Shaker R, Halevy A, Tarczy-Hornoch P. PQL: a declarative query language over dynamic biological schemata. *Proc AMIA Symp*.533-7(2002)
[24] Sujansky W. Heterogeneous database integration in biomedicine. *J Biomed Inform.* **34**(4):285-98(2001)
[25] Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology.*JAMIA* **1**(1):35-50(1994)
[26] Burgun A, Bodenreider O. Mapping the UMLS Semantic Network into general ontologies. *Proc AMIA Symp* 81-5(2001)
[27] Lambrix P, Edberg A. Evaluation of ontology merging tools in bioinformatics.*Pac Symp Biocomput*. 589-600(2003).
[28] Li Q, Shilane P, Noy NF, Musen MA. Ontology acquisition from on-line knowledge sources. *Proc AMIA Symp* 497-501 (2000)
[29] Hill DP, Blake JA, Richardson JE, Ringwald M. Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. Genome Res. 2002 Dec;12(12):1982-91.
[30] Bodenreider O, Mitchell JA, McCray AT. Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. Proc AMIA Symp. 2002;:61-5.

[31] Burgun A, Bodenreider O, Le Duff F, Moussouni F, Loreal O. Representation of roles in biomedical ontologies: a case study in functional genomics. *Proc AMIA Symp* 86-90 (2002)
[32] Lussier YA, Shagina L, Friedman C. Automating SNOMED Coding Using Medical Language Understanding: A Feasibility Study. *Proc AMIA*: 418-422. (2001)
[33] Tuttle MS, Sherertz DD, Erlbaum MS, et al. Adding Your Terms and Relationships to the UMLS Metathesaurus. 1991 *Proc AMIA:*219-223(1991)
[34] Tuttle MS, Suarez-Munist ON, Olsen NE, et al. Merging Terminologies. 1995 *MEDINFO*. **8**(Pt 1):162-166. (1995)
[35] Lussier YA, Shagina L, Friedman C. Automating SNOMED Coding Using Medical Language Understanding: A Feasibility Study. *Proc AMIA:* 418-422. (2001).
[36] McCray AT, Srinivasan S, Browne AC. Lexical Methods for managing variation in biomedical terminologies. In Ozbolt JG, ed. *Proceedings of the Eighteenth Annual Symposium in Computer Applications in Medical Care*. Philadelphia: Hanley & Belfus, 235-239(1994)
[37] Rocha R, Rocha B, Huff SM. Automated translation between medical vocabularies using a frame-based interlingua. In Ozbolt JG, ed. *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*. 690-694(1994)
[38] Zeng, Q & Cimino JJ. Mapping Medical Vocabularies to the Unified Medical Language System. *Proc AMIA* 105-109. (1996)
[39] Cantor MN, N. SI, Hartel F, Bodenreider O, Lussier YA. An evaluation of hybrid methods for matching biomedical terminologies: Mapping the Gene Ontology to the UMLS. *Stud Health Technol Inform* 62-67(2003).
[40] Sarkar IN, Cantor MN, Gelman R, Hartel F, Lussier YA. Linking biomedical language information and knowledge resources: GO and UMLS. *Pac Symp Biocomput.* 439-50(2003)
[41] Cantor MN, Lussier YA. Putting Data Integration into pratice: using biomedical terminologies to add structure to existing data scources. *AMIA Symposium* (2003) *Accepted*.
[42] Zeng Q, Cimino JJ. Mapping medical vocabularies to the Unified Medical Language System. Proc AMIA Annu Fall Symp. 1996;:105-9.
[52] Blake JA, Richardson JE, Bult CJ, Kadin JA, Eppig JT; Mouse Genome Database Group. MGD: the Mouse Genome Database. *Nucleic Acids Res.* **1**;31(1):193-5(2003)
[53] Spackman KA & Campbell KE. Compositional Concept Representation using SNOMED: Towards Further Convergence of Clinical Terminologies. *Proc AMIA*: 875-879(1998)
[54] Lindberg DA, Humphries BL, McCray AT. The Unified Medical Language System. Methods Inf Med. **32**(4):281-291. (1993)
[55] Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. *Proc AMIA Symp.* 57-61(2001)
[56] National Library of Medicine. UMLS Lexical Tools. Application and Documentation available at http://umlsks.nlm.nih.gov.
[57] Lussier YA, Sarkar IN, Cantor M. An integrative model for in-silico clinical-genomics discovery science. *Proc AMIA Symp* 469-73(2002)