Diploma Thesis Exposé

Ontology Construction from Phenotype Data

Christoph Böhm

February 4, 2008

Supervisors: M.Sc. Philip Groth and Prof. Dr. Ulf Leser
Knowledge Management in Bioinformatics

Department of Computer Science
Faculty of Mathematics and Natural Sciences II

1 Background

In the life sciences, many data are produced in a high-throughput manner. In the case of phenotypes, this is particularly achieved through RNA interference (RNAi), where the result of silencing one gene at a time is observed for a phenomic outcome (RNAi) [Tuschl and Borkhardt, 2002]. In large RNAi screens, the number of potential results is nearly unlimited [Wheeler et al., 2005]. A phenotype is the "manifestation of a set of traits in an individual that result from the combined action and interaction of genotype and environment"[1]. The phenotype data resulting from RNAi and other phenotype screening methods are often captured in journal articles, conference publications or specific databases. Many times, these descriptions are natural language using a very domain-specific vocabulary. As a consequence comparisons across datasets are performed either manually or use some sort of fuzzy text-mining. To overcome this shortcoming, the community is in need of a universal phenotype ontology. Such a structure should define phenotype-specific concepts in a more species-independent manner [Groth and Weiss, 2006].

1.1 Ontologies

For the purpose of this thesis, we define an ontology as a set of concepts that are, where applicable, connected by relations. There are two main types of relations: IS-A and PART-OF. The IS-A relation combines two concepts A and B where A is more general than B, i.e. B is a (subconcept of) A. In case of the PART-OF relation, a concept B consists of multiple concepts $A_1..A_n$, i.e. A_i is part of B $(i = 1..n)$. For further reading we refer the reader to [Staab and Studer, 2004].

Currently, there are a few ontologies assisting biologists in annotating their data. Two well-known examples are the Gene Ontology (GO)[2] [GO-Consortium, 2006] and the Mammalian Phenotype Ontology (MPO)[3] [Smith et al., 2005]. "The GO project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. [...] It has developed three structured controlled vocabularies that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner."[4] The MPO enables biologists to annotate documents at "different levels and richness of phenotypic knowledge. [...] It continues to develop dynamically via collaborative input from research groups, mutagenesis consortia, and biological domain experts."[5]

1.2 Phenotype Data

Whereas an ontology is a tool to provide information in a structured and standardized way, PhenomicDB[6] [Groth et al., 2007] is a database of phenotypes described through plain and heterogeneous text fragments. It is an integrated multi-species phenotype/genotype resource that contains phenotypes from several primary databases such as OMIM[7] or MGD[8] associated with their genotypes. PhenomicDB also includes GO annotations and orthology relationships from NCBIs HomoloGene database[9]. This project envisions "that integration of classical phenotypes

[1] quotation is taken from [Smith et al., 2005]
[2] GO is available at www.geneontology.org
[3] MPO is available at www.informatics.jax.org/searches/MP_form.shtml
[4] quotation is taken from www.geneontology.org/GO.doc.shtml
[5] quotation is taken from [Smith et al., 2005]
[6] PhenomicDB is available at www.phenomicdb.de
[7] OMIM is available at www.ncbi.nlm.nih.gov/omim, see [Hamosh et al., 2002]
[8] MGD is available at www.informatics.jax.org, see [Bult et al., 2007]
[9] HomoloGene is available at www.ncbi.nlm.nih.gov/sites/entrez?db=homologene, see [Wheeler et al., 2007]

with high-throughput data will bring new momentum and insights to our understanding. [...] It enables easy cross-species mining of phenotypes."[10] This integrated data will form the basis for this thesis.

1.3 Problem and Vision

The GO as well as the MPO were created by hand. This approach ensures high quality but requirers many expert groups that invest an enormous amount of time. For instance, currently 15 research groups are full members of the GO consortium committed to ontology utilization and development.

The GO covers concepts for multiple species. In contrast the MPO contains phenotype-specific information only for mammals. To the best of our knowledge, there is no ontology dealing with phenotype-specific concepts from a broad range of species.

If there was a multi-species ontology containing the phenotype concepts wich are described as text in PhenomicDB, one had a tool to support the exact formulation and machine readability of phenotypes. In case the life sciences adopted this ontology, i.e. started improvements and utilization, it could be the foundation of a high-quality knowledge base.

Since PhenomicDB contains thousands of phenotype-related concept, a manual process is not practicable. Therefore, we are targeting an automatic process to construct this ontology.

2 Preparatory Work

In a student research project [Böhm, 2007] we reviewed state-of-the-art methods for the automatic creation of term hierarchies. We described early works such as [Salton, 1971] and [Forsyth and Rada, 1986] through modern approaches, i.e. [Cimiano et al., 2005]. In experiments, we found that a simple method based on statistics of term cooccurrence [Sanderson and Croft, 1999] reaches a decent precision. The observed poor recall was most likely caused by the specific term matching that was used for concept search in documents. We expect a better recall by applying fuzzy concept search.

Other methods mentioned in the survey leverage linguistic properties of documents. A well-known approach uses so called Hearst patterns [Hearst, 1992] to extract hyponyms from text corpora. One such pattern is $NP(,NP)^*(or|and)otherNP'$. It implies that the NPs are sub-concepts of NP'.

3 Goal

The goal of this thesis is the automatic generation of a multi-species phenotype ontology. The result is not aimed to be a fully mature knowledge base as it would be defined by human experts, but a high-quality stating point for manual improvements. Concepts as well as relations will be extracted from PhenomicDB. Relations will be derived by a hybrid approach based on statistical properties, lexical patterns and domain-specific links between objects.

4 Approach

Our fundamental idea is to capture different sources of evidence for concept relations in a graph. The nodes are the concepts. The edges, or more precisely their weights, represent

[10] quotation is taken from [Groth et al., 2007]

evidences for relations between two concepts. We focus on IS-A relations, but will not tackle the disambiguation of IS-A, PART-OF and other relationships.

Given such a model the problem of identifying a concise yet broad set of superconcepts from a set of concepts can be understood as the dominating set problem (DSP): Given a graph $G = (V, E)$ and node weights w_v, find a subset $D \subset V$ (the superconcepts), such that for each node $v \in V - D$ a node $d \in D$ exists, where $\{v, d\} \in E$ and $\sum_{d \in D} w_d$ is maximal. In [Lawrie et al., 2001] this approach was applied to term cooccurrence as the only source of evidence.

Our process towards the phenotype ontology thus can be divided into four sub-problems: (1) the concept definition, (2) the localization of these concepts in the document corpus, (3) the evidence graph construction, and (4) the relation discovery.

4.1 Concept Definition

This initial step will extract terms and phrases representing phenotype concepts from PhenomicDB. This will use statistical analysis of single terms and n-grams of terms. A term that has a certain tf.idf value x in a background model (e.g. a sample of Medline) but a tf.idf value $y \gg x$ in the foreground model (i.e. PhenomicDB) is most likely a term specific to PhenomicDB, and therefore probably a phenotype concept. For n-grams of tokens, we will take fuzzyness into account. When determining the frequencies of phrases an n-gram p may be found in a window of size $w > n$. We shall also analyze the statistical significance of our results.

In addition, we consider to use existing sets of concepts, e.g. GO-, MPO-, UMLS[11]-, or MeSH[12] concepts. This decision considers the size and plausibility of the concept set extracted by the statistics above mentioned.

4.2 Concept Discovery

While in the first step the concepts are defined, this phase will locate them in the corpus. This is necessary for an evaluation of term/phrase cooccurrence and lexical patterns. We will implement a fuzzy search. When locating a phrase $p = t_1..t_i$ the algorithm will allow mismatches such that the phrase $p' = t_1..t'..t_i$ where $t' \neq t_1..t_i$ is also a match for p. This will be accomplished by the use of windows: find $p = \{t_1..t_i\}$ in a window of size $w > i$.

4.3 Evidence Discovery

This phase deals with the collection of evidences for relations between concepts. As multiple evidences are to be extracted, we will incorporate different approaches for inducing concept relations, e.g. syntactic properties or existing ontologies. This step's result is the graph capturing the evidences for relations between concepts. We plan to include the following sources of evidence for relations:

1. Concept cooccurrence in PhenomicDB
2. Hearst patterns
3. Links from phenotypes to genotypes and their annotation with GO concepts
4. Concept cooccurrence in a Medline sample for low frequency concepts in PhenomicDB (depending on the amount of low frequency terms)

[11] Unified Medical Language System: www.nlm.nih.gov/research/umls
[12] Medical Subject Headings: www.nlm.nih.gov/mesh

4.4 Relation Discovery

The exploration of concept relations is the main focus of this thesis. The relations are going to be discovered by the approach above mentioned: we will capture evidences for relations as edge weights in a graph and use the DSP to determine subsets of superconcepts. At first we might have to simplify the graph generated in the previous step, e.g. aggregate multiple edges between two concepts. Since the DSP is NP-hard, we will apply a heuristic. This algorithm has to take a weighting scheme for the different sources of evidence into account. The definition of this weighting scheme is part of the challenge to extract the relations.

5 Implementation

The data to be processed is mainly in PhenomicDB. Phenotypes are available in a flat file[13]. This file also includes the genotype link information. The genotype annotations can be downloaded from the NCBI[14].

The main programming language is Java 5[15]. To find Hearst patterns, we will need to do some linguistic preprocessing. Since it is not the focus of this thesis to (re)implement Natural Language Processing techniques, we will use public domain libraries for that purpose, e.g. openNLP[16] or LingPipe[17]. In these cases we will have to use a scripting language like Perl[18].

The output of the ontology generation process will be an XML file containing the concepts and their relations. Using XSLT it will be easy to transform this result to another XML Schema or a flat file format. Target formats could be OWL [McGuinness and van Harmelen, 2004] or OBO [Smith et al., 2007]. This decision depends on further use and is not part of the thesis.

One way to evaluate the relation discovery is to run the approach on GO- or MPO-concepts. For these two sets of concepts the precision and recall should be superior to the values from the student research project. Generally one will have to review the constructed relations with a biological background. We will use experts' evaluation of a subset of the relations, to estimate the overall accuracy.

[13] phenomicdb.de/downloads.html
[14] www.ncbi.nlm.nih.gov/Ftp
[15] java.sun.com/j2se/1.5.0
[16] opennlp.sourceforge.net
[17] www.alias-i.com/lingpipe
[18] www.perl.org, perl.com

References

[Böhm, 2007] Böhm, C. (2007). Methoden zur automatischen erstellung von term-hierarchien aus phänotypendaten (studienarbeit). Technical report, Knowledge Management in Bioinformatics, Humboldt University of Berlin.

[Bult et al., 2007] Bult, C., Eppig, J. T., Kadin, J. A., Richardson, J. E., Blake, J. A., and the Mouse Genome Database Groupy (2007). The mouse genome database (mgd): mouse biology and model systems. *Nucleic Acids Res.*, pages 1–5.

[Cimiano et al., 2005] Cimiano, P., Hotho, A., and Staab, S. (2005). Learning concept hierarchies from text corpora using formal concept analysis. *Journal on Artificial Intelligence Research*, 24:305–339.

[Forsyth and Rada, 1986] Forsyth, R. and Rada, R. (1986). *Adding an edge.*, pages 198–212. Chichester: Ellis Horwood: Halsted Press, New York.

[GO-Consortium, 2006] GO-Consortium (2006). The gene ontology (go) project in 2006. *Nucleic Acids Res.*, 34(Database-Issue):322–326.

[Groth et al., 2007] Groth, P., Pavlova, N., Kalev, I., Tonov, S., Georgiev, G., Pohlenz, H.-D., and Weiss, B. (2007). Phenomicdb: a new cross-species genotype/phenotype resource. *Nucleic Acids Res.*, 35(Database-Issue):696–699.

[Groth and Weiss, 2006] Groth, P. and Weiss, B. (2006). Phenotype data: A neglected resource in biomedical research? *Current Bioinformatics*, 1(3):347–358.

[Hamosh et al., 2002] Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., and Mckusick, V. A. (2002). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucl. Acids Res.*, 30(1):52–55.

[Hearst, 1992] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, pages 539–545, Morristown, NJ, USA. Association for Computational Linguistics.

[Lawrie et al., 2001] Lawrie, D., Croft, W. B., and Rosenberg, A. (2001). Finding topic words for hierarchical summarization. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 349–357, New York, NY, USA. ACM Press.

[McGuinness and van Harmelen, 2004] McGuinness, D. L. and van Harmelen, F. (2004). Owl web ontology language - overwiew, w3c recommendation. www.w3.org/TR/2004/REC-owl-features-20040210.

[Salton, 1971] Salton, G. (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[Sanderson and Croft, 1999] Sanderson, M. and Croft, B. (1999). Deriving concept hierarchies from text. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 206–213, New York, NY, USA. ACM Press.

[Smith et al., 2007] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech*, 25(11):1251–1255.

[Smith et al., 2005] Smith, C. L., Goldsmith, C. A., and Eppig, J. T. (2005). The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biology*, 6(1).

[Staab and Studer, 2004] Staab, S. and Studer, R. (2004). *Handbook on Ontologies (International Handbooks on Information Systems)*. International Handbooks on Information Systems. Springer.

[Tuschl and Borkhardt, 2002] Tuschl, T. and Borkhardt, A. (2002). Small interfering rnas: a revolutionary tool for the analysis of gene function and gene therapy. *Mol Interv*, 2(3):158–167.

[Wheeler et al., 2005] Wheeler, D. B., Carpenter, A. E., and Sabatini, D. M. (2005). Cell microarrays and rna interference chip away at gene function. *Nat Genet*, 37 Suppl:25–30.

[Wheeler et al., 2007] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L. Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Ostell, J., Miller, V., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2007). Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, 35(Database issue).