MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-1030

Aligning UniProt and MeSH – A Case Study on Human Protein Terms

Elena Beisswanger[a], Joachim Wermter[a], Udo Hahn[a]

[a]*Jena University Language and Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Jena, Germany*

Abstract

Terminologies which lack semantic connectivity hamper the effective search in biomedical fact databases and document retrieval systems. We here focus on the integration of two such isolated resources, the term lists from the protein fact database UNIPROT *and the indexing vocabulary* MESH *from the bibliographic database* MEDLINE*. The generated semantic ties result from string matching and term set inclusion. We investigated the implicit terminological overlap between both resources in the domain of human proteins and evaluated our approach on a sample of 550 randomly selected* UNIPROT *entries that were manually mapped to their corresponding* MESH *headings. We achieved 90% precision and 79% recall (applying taxonomy-sensitive metrics). Fortunately, those proteins we were able to map to the* MESH *are ten times as frequently discussed in the literature as those on which we failed.*

Keywords:

Terminological Alignment, Interoperability of Terminologies

Introduction

Over the past years, MEDLINE – with now over 19M entries – has gained world-wide reputation as the most authoritative and often used bibliographic resource for biomedical literature search via the PUBMED interface.[1] Much of its retrieval power can be attributed to the Medical Subject Headings (MESH),[2] a terminology from which index terms are manually derived as content descriptions for MEDLINE records. The MESH not only covers a controlled vocabulary (of about 25,000 descriptors spanning various domains such as anatomy, diseases, chemicals and drugs) but excels in the provision of a multi-hierarchical taxonomic thesaurus structure (plus synonyms).

For bioinformatics, a comparably authoritative resource for protein and gene fact search has emerged through the UNIPROT KNOWLEDGEBASE (UNIPROTKB) [1]. In particular, UNIPROTKB/SWISSPROT, the curated part of UNIPROTKB, is a comprehensive, high-quality protein database which contains over 400,000 manually annotated proteins from various species. Unlike the MESH, UNIPROTKB does not offer any taxonomic links between terms, but (just as the MESH) contains synonyms of the gene and protein names in a specific protein entry, which are all lined up with their associated unique database identifier. Generally, UNIPROTKB describes proteins on a more specific level than the coarser grained MESH, the latter dealing with proteins in branch D (Chemicals and Drugs).

PUBMED, the retrieval interface to MEDLINE, allows users to search for documents about protein families, groups, or complexes by entering suitable MESH terms. However, literature on a specific protein can only be retrieved by running a free-text search. Yet, this search mode is known to suffer from several shortcomings because protein names are notoriously complex and ambiguous and thus hard to nail down by free-text expressions. A new breed of semantics-based search engines such as SEMEDICO [2] or IHOP [3] aim to cope with these problems by incorporating named entity recognizers (cf., e.g., SEMEDICO's gene name normalizer GENO [4]) which enrich plain documents with semantic metadata, including links to UNIPROTKB identifiers. If a protein name, e.g., **"**Heat shock protein HSP 90-beta", is entered in such a search engine, not only a set of documents matching this term is retrieved, but also a link to the corresponding UNIPROTKB entry (in this case HS90B_HUMAN) is provided which holds additional factual information about the protein under scrutiny.

Even such advanced search engines are incapable of searching for a specific protein and (if requested by the user), at the same time, generalizing to proteins belonging to the same protein group, family or complex. For example, if a document search for **"**Heat shock protein HSP 90-beta" were conducted, one might also be interested in documents about other members of the HSP90 heat shock protein family, or even about heat shock proteins, in general. One way to enable users to retrieve those additional documents would be to establish explicit links from the specific protein name to the appropriate MESH term, in this case "HSP90 Heat Shock Proteins", from where even more generic terms could be reached, such as its parent term "Heat Shock Proteins". To realize such a semantics-rich search strategy we aligned knowledge available from the UNIPROTKB with the taxonomic structure of the MESH, thus enabling new 'taxonomic' search strategies (see Figure 1).

[1] http://www.ncbi.nlm.nih.gov/pubmed
[2] http://www.nlm.nih.gov/mesh

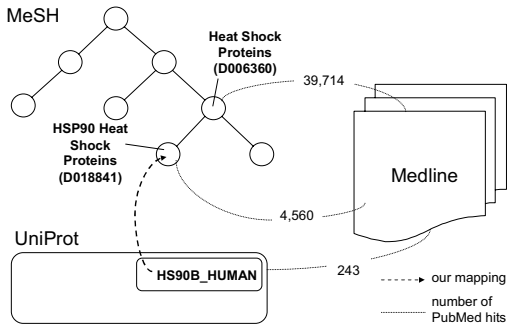

Figure 1- Linking UNIPROT and MESH for taxonomic search. The increasing number of hits (243 → 4,560 → 39,714) directly reflects the increasing conceptual generality of the search terms involved (HS90B_HUMAN → HSP90 Heat Shock Proteins → Heat Shock Proteins, respectively).

Materials and Methods

For our mapping study, we focused on the human subset of UNIPROTKB/SWISSPROT (RDF version from November 2008). Originally 20,328 entries were collected. In a subsequent cleansing step, we excluded those entries where the recommended name contained the phrase *"uncharacterized protein"* so that 19,052 human protein entries remained as input data for our mapping experiments. We also restricted the headings in the MESH thesaurus (MESH version 2008) such that permitted mapping targets were restricted to headings concerned with proteins and genes only, in the broadest sense though. This left us with all headings which belonged to one of the following MESH sub-hierarchies: D05 (Macromolecular Substances), D06 (Hormones, Hormone Substitutes, and Hormone Antagonists), D08 (Enzymes and Coenzymes), D09 (Carbohydrates) restricted to the Glycoproteins and Glycopeptides fraction, D12 (Amino Acids, Peptides, and Proteins), D13 (Nucleic Acids, Nucleotides, and Nucleosides), D23 (Biological Factors), and G14 (Genetic Structures).

We further utilized MESH's Supplementary Concept Records (SCR) as intermediate mapping target. SCR is a separate resource with rather specific, UNIPROT-like headings mainly concerned with chemicals and proteins. Each SCR heading comes with an explicit link to the closest possible MESH heading. Some SCR records are even linked to several MESH headings belonging to different classification axes. We exploited these links by mapping UNIPROT entries to SCR headings and then following the existing links to the associated MESH headings. Table 1 summarizes the sources of our mapping study.

Term Selection

For each UNIPROTKB entry we gathered all recommended and alternative protein names in their long form, as well as all gene names. In addition, for each entry we compiled a set of family and enzyme names based on three additional resources.

Table 1 - Quantitative data of terminological resources used for the mapping experiments

Source	Entries	Distinct Names
MESH protein	5,198	47,210
MESH SCR	182,890	462,673
UNIPROT human (cleansed)	19,052	90,920

We, first, extracted family names from the Similarity Annotation fields of UNIPROTKB entries, using simple regular expressions. A typical example for a Similarity Annotation is

"*Belongs to the protein kinase superfamily. TKL Ser/Thr protein kinase family. Pelle subfamily.*"

from the UNIPROTKB entry "IRAK3_HUMAN" ("Interleukin-1 receptor-associated kinase 3"). We here extracted the names "protein kinase", "TKL Ser/Thr protein kinase", and "Pelle".

Second, additional family names were harvested from INTERPRO,[3] a database of protein families and domains interlinked with UNIPROTKB. For the protein "KT81L_HUMAN" ("Keratin-81-like protein"), e.g., we followed the link to the protein family entry "IPR003054" from which we extracted the family name "Type II keratin".

Third, for entries coming with an Enzyme Commission (EC) number, this number was looked up in the Enzyme Nomenclature database,[4] to gather all attached enzyme names. For instance, UNIPROTKB entry "EYA2_HUMAN" ("Eyes absent homolog 2") is annotated with the EC number "EC 3.1.3.48". From the corresponding entry in the enzyme database we extracted the name "Protein-tyrosine-phosphatase".

While for MESH records, the heading itself and all associated entry terms were extracted, for MESH SCR records we considered the names of substances and all given synonyms.

Preprocessing of Terms

To cope with morphological term variations, we did not use a stemmer (that can be suspected to truncate many of the relevant domain-specific terms) but instead looked up each UNIPROTKB and MESH term in the UMLS Specialist Lexicon inflection file LRAGR.[5] If it is listed as a plural form of a noun, we extracted the corresponding singular form and added it to our term set. Then for all terms, punctuation marks were replaced by spaces, the task-specific stop words "gene", "protein", "family", "member", "domain", and "subunit" were removed from terms, and, finally, terms were lower-cased and tokenized, interpreting spaces as token boundaries.

A special preprocessing step was applied to MESH SCR terms. Many of these terms contain organism names, such as the substance name "IL2 protein, human" from record "C508594". These organism names were removed to make SCR terms compatible with UNIPROTKB terms that usually lack any kind of organism information (in UNIPROTKB, organism informa-

[3] http://www.ebi.ac.uk/interpro
[4] http://www.expasy.ch/enzyme/
[5] http://www.nlm.nih.gov/pubs/factsheets/umlslex.html

tion is kept in a separate field called 'Taxonomic Identifier'). We compiled a list of organism names from the NCBI taxonomy[6] and matched the names against all MESH / MESH SCR terms. If a name was found as substring of a MESH term, the organism-specific substring was removed from that term. However, we kept the NCBI taxonomy ID (TaxID), corresponding to the organism name that we removed, for later comparison with the TaxID associated with the UNIPROTKB entries. In our study on human proteins, this is only TaxID "9606" denoting "Homo sapiens (human)".

Term Mapping

In order to find for all human proteins in UNIPROTKB the closest related MESH heading we pursued a two-step approach. First, for each UNIPROTKB entry we matched all protein and gene names against all extracted MESH and MESH SCR terms. As matching criteria, we required, first, the MESH and the UNIPROTKB term to consist of the same tokens (their order was considered irrelevant) and, second, if a TaxID was associated with the MESH term, it had to match "9606" (human). Then for all these matches the corresponding MESH headings were selected as possible mapping targets for the UNIPROTKB entry. If no suitable MESH heading was found, in the second step, all enzyme and family names compiled for the UNIPROTKB entry were matched against all MESH / MESH SCR terms. Again the MESH headings corresponding to the successfully matched terms were selected as candidate mapping target for the corresponding UNIPROTKB entry. UNIPROTKB entries for which no target MESH heading was found after these steps were marked as *"not mapped"*.

In case several MESH / MESH SCR headings were found as possible mapping targets for a UNIPROTKB entry we determined the most suitable heading amongst the candidates with a LUCENE[7]-based ranking procedure basically relying on a fine-tuned TF-IDF weight (cf. Chapter 3.3 in [5]). In addition, our ranking mechanism took into account the type of terms that had matched. If, for instance, the recommended name of a UNIPROTKB entry matched a MESH term, we considered the associated MESH heading a better mapping candidate than a MESH heading of which a term matched an alternative name of a UNIPROTKB entry. If a UNIPROTKB entry was mapped to a MESH SCR heading, we followed the existing links to the corresponding MESH headings and selected them as mapping targets for the UNIPROTKB entry (only in this case several mapping targets were allowed per UNIPROTKB entry).

Mapping and Evaluation Results

In the first matching step (based on the comparison of protein and gene names from UNIPROTKB with MESH terms) our algorithm mapped 67% of all human protein entries to a MESH heading. In the second step (based on the comparison of family and enzyme names with MESH terms) mappings for additional 11% of the protein entries were found (see Table 2).

Table 2 - Results of the automatic mapping of human UNIPROTKB entries to MESH headings

Matching Step	Number of Matches (%)
Step1	12,691 (67%)
Step2	2,102 (11%)
Step1 + Step2	**14,793 (78%)**
Baseline	13,321 (70%)

As a baseline for comparison, we matched all protein and gene names of a UNIPROTKB entry to all MESH terms (terms were Porter-stemmed,[8] lower-cased, and punctuation marks were removed) and the MESH heading corresponding to the highest ranked match (cf. Section "Term Mapping") was selected as mapping target. Accordingly, mappings for 70% of all UNIPROTKB entries were found (see Table 2).

To assess the quality of automatically generated mapping results we compared them to a manually created gold standard. It consists of a random sample of 550 UNIPROTKB entries (drawn from the set of 19,052 entries) that were mapped by a domain expert to the closest related MESH heading(s). Since MESH is a multi-hierarchy, the expert was allowed to select more than one heading for each entry. A total of 58 entries (10.5%) were mapped to the general heading D011506 (Proteins). These entries were marked as *"not mapped"*.

As evaluation metric, we chose a relaxed precision and recall measure, *overlap proximity*, as introduced by Ehrig and Euzenat [6]. This metric pays tribute to the particularities of taxonomic hierarchies since, besides node-wise exact matches, it also incorporates the grounded intuition that even slightly more general or more specific matches within the taxonomic 'neighborhood' of a term are useful and reasonable matches, rather than treating them as absolute non-matches. Therefore, instead of taking the (strict) intersection of the set of automatically generated mappings (A) and the set of mappings in the manually created gold standard (G) as metrical criterion (as standard precision and recall metrics do), the relaxed variant takes the above considerations into account and measures the overlap proximity of the two sets with respect to a certain proximity function. Let M denote the matching between A and G ([6]) and σ be the proximity function between mappings a in A and g in G. Given the proximity ω

$$\omega(A,G) := \sum_{(a,g)\in M(A,G)} \sigma(a,g) \quad (1)$$

relaxed precision P_ω and recall R_ω are defined as

$$P_\omega(A,G) := \frac{\omega(A,G)}{|A|} \text{ and } R_\omega(A,G) := \frac{\omega(A,G)}{|G|} \quad (2)$$

We chose as proximity criterion for two mappings, $a = (u_a, m_a)$ and $g = (u_g, m_g)$, with u denoting a UNIPROTKB entry and m the MESH heading it was mapped to,

[6] http://www.ncbi.nlm.nih.gov/Taxonomy
[7] http://lucene.apache.org/
[8] http://tartarus.org/~martin/PorterStemmer/

$$\sigma(a,g) := \begin{cases} \dfrac{1}{p(m_a, m_g)} & \text{if } eq(u_a, u_g) \wedge (eq(m_a, m_g) \\ & \quad \vee s(m_a, m_g) \vee s(m_g, m_a)) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

with p denoting the length of the shortest path between two MESH headings (in terms of the number of nodes in the MESH taxonomy graph) plus one, such that $p(m, m) = 1$, eq denoting the equivalence and s the subclass relationship. As for standard precision and recall, a mapping in A that is also in G is scored '1'. Mappings in A where the predicted and the correct MESH heading are in a subclass relation to each other (hyponym relation, on the term level), are scored with the reciprocal value of the length of the shortest path between the correct and the predicted MESH heading. All remaining mappings are penalized with '0'.

Figure 2 illustrates the scoring logic for three automatically detected mappings. On the left, an exact mapping is shown, scored '1' (predicted and correct MESH heading are equal). In the middle, a too general mapping is shown. Since the predicted MESH heading is a direct parent of the correct heading, it is scored '0.5'. On the right, an incorrect mapping is shown where no subclass relationship holds between the predicted and the correct MESH heading, scored '0'.

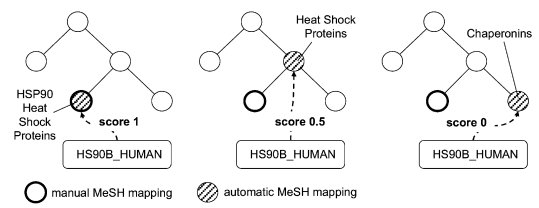

Figure 2 – Scoring of automatically computed mappings

As Table 3 reveals, when we apply the relaxed precision and recall measures we obtain 90% precision and 79% recall, resulting in 88% F-score, setting β to 0.5 in

$$F_\beta := ((1+\beta^2)PR)/(\beta^2 P + R) \quad (4)$$

in order to emphasize precision.

We also measured precision and recall for the constituent matching steps. Obviously, the first step based on matching protein and gene names to MESH terms results in more precise mappings (93% precision) than the second one based on looking up protein family names in MESH (73% precision). Still, adding the second step increases overall recall by 8 percentage points with a stable F-score of 88%. The mapping procedure outperforms the baseline (85% precision and 67% recall on the gold standard), in particular with respect to recall.

Table 3 - *Evaluation results in terms of relaxed precision P_ω, recall R_ω and F-score (values in %).*

Matching Step	P_ω	R_ω	F-score
Step1	93,0	70,7	87,5
Step2	72,7	8,0	27,7
Step1 + Step2	**90,2**	**78,5**	**87,6**
Baseline	85,3	66,6	80,8

This makes evident the value of the additional terms (e.g., from UNIPROTKB annotation fields and external databases like INTERPRO) for the mappings' outcome.

Discussion

For the evaluation of our approach we used a relaxed precision and recall measure and computed the F-score with emphasis on precision. Both decisions reflect requirements from the information retrieval scenario discussed in the beginning. The choice of the evaluation measure reflects our claim that even if the automatic procedure could not detect the fully correct MESH heading for a UNIPROTKB entry, detecting one of its parent or child headings would still enable the user to pass from UNIPROTKB to MESH and to correctly generalize / specialize the original search exploiting the MESH hierarchy. The emphasis on precision reflects our opinion that false mappings that would lead to the retrieval of irrelevant documents are worse than missing mappings due to the negligence of taxonomic relation that particular proteins share.

The terminological heterogeneity of protein and gene names might raise concerns about the size of our gold standard. To assess the plausibility of our evaluation based on this gold standard, let us assume that the random sample of 550 UNIPROTKB entries would have been drawn from an *infinite* set of entries, and the precision and recall estimates, resulting from the comparison of the automatically detected mappings with the gold standard, would be 0.5 (50%). Then the standard deviation of the estimates, ± 2.1, would be 'acceptable'. (In fact, the number is even an upper limit for the real standard deviation, since our sample was drawn from the finite set of 19,052 UNIPROTKB entries, and the determined precision and recall measures are 'far away' from 0.5.)

Still, about 20% of all human proteins in UNIPROTKB / SWISS-PROT could not be mapped properly to MESH headings by our algorithm, although based on the gold standard only 10% of non-matching entries should be expected. We tested two alternatives to increase our recall figures. First, we extended our procedure by a third matching step, again matching all UNIPROTKB terms (protein, gene, and family names) against all MESH terms. This time, we only required a partial token match and allowed for contradicting TaxIDs. Second, we considered additional UNIPROTKB name types for the mapping, *viz* short forms of gene and protein names, allergen names, CD antigen names, and the International non-proprietary names.

Although 1,267 additional UNIPROTKB entries (an increase of 7%) could be mapped, the overall effect (a decrease by 3 percentage points to 85% F-score) was negative due to decreasing precision (86%). The inspection of erroneously missed mappings revealed that many of the UNIPROTKB entries involved come with rather technical names such as "FAM75-like protein FLJ43859" ("YI020 HUMAN") that cannot easily be matched with MESH terms. Thus, we assume that only exploiting further annotations of UNIPROTKB entries (such as textual descriptions) might increase the number of correct mappings.

The good news is that our procedure deals satisfactorily with the most frequently occurring human protein names. We found that those proteins that we were able to map to MESH headings are mentioned, on average, in 10 times as many document as those on which we failed to map (111.3 documents, on the average, compared to 12.6).[9] In terms of precision, our procedure shows decent results. Still, we analyzed the false predictions and found that three-fourth of all incorrect mappings were due to the unresolved ambiguity of gene symbols.

Related Work

Biomedical ontologies and terminological resources are increasingly becoming important for knowledge management tasks in the life sciences [7]. This is witnessed by the rapid growth of single resources such as the GENE ONTOLOGY (GO)[10] which is massively used for the functional annotation of genes and gene products. Also large libraries of controlled vocabularies have emerged such as the UMLS[11] and OBO.[12] Efforts to foster interoperability have already been started, e.g., aligning GO with other OBO ontologies [8]. UNIPROTKB and the MESH have also been the target of deeper integration efforts. In [9], a procedure is described to link diseases mentioned in UNIPROTKB entries to the MESH disease terminology to make disease information in UNIPROTKB more easily accessible to clinical researchers. What has not been studied so far is the connection between protein entries in UNIPROTKB and MESH headings representing protein families, groups, or complexes, the goal of our investigation.

Conclusion

Despite the ever increasing number and size of single biomedical terminologies their usage for searching relevant facts and literature is currently hampered by a lack of semantic integration and interoperability. In this study, we proposed an automatic procedure to align human protein names in UNIPROTKB / SWISSPROT to suitable headings in the MESH.

The mappings we found were evaluated on a manually created gold standard of 550 match pairs resulting in 90% precision and 79% recall (with 88% F-score). Our approach outperformed a simple yet effective baseline by 7 percentage points F-score (5 and 12 percentage points in terms of precision and recall, respectively).

The mapping approach we propose can easily be applied to the whole of UNIPROTKB / SWISSPROT. A preliminary study on protein entries for a set of 29 important model organisms achieved promising results. For 78% of these entries mappings to MESH headings could be found. The research we have described is but a preparatory step for a more thorough evaluation that has to measure the effects of such alignments for the effectiveness of searches in real retrieval settings.

Acknowledgements

This work was funded within the STEMNET (No. 01DS001) and the JENAGE (No. 0315581D) projects by the Federal Ministry of Education and Research (BMBF), Germany.

References

[1] The UNIPROT Consortium. The Universal Protein Resource (UNIPROT). Nucleic Acids Res, 36 (Database issue): D190-D195, 2008.

[2] Schneider A, Landefeld R, Wermter J, and Hahn U. Do users appreciate novel interface features for literature search? A user study in the life sciences domain. Proceedings of the 2009 IEEE SMC 2009, 2009.

[3] Hoffmann R, and Valencia A. A gene network for navigating the literature. Nat Genet, 36(7):664, 2004.

[4] Wermter J, Tomanek T, and Hahn U. High-performance gene name normalization with GENO. Bioinformatics, 25(6):815-821, 2009.

[5] Gospodnetić O, and Hatcher E. Lucene in Action. Greenwich: Manning Publications, 2nd ed., 2005.

[6] Ehrig M, and Euzenat J. Relaxed precision and recall for ontology matching. Proceedings of the K-CAP'05 Workshop 'Integrating Ontologies', 2005, pp- 25-32.

[7] Bodenreider O, and Stevens R. Bio-ontologies: current trends and future directions. Brief Bioinform, 7(3):256-274, 2006.

[8] Bada M, and Hunter L. Identification of OBO nonalignments and its implications for OBO enrichment. Bioinformatics, 24(12):1448-1455, 2008.

[9] Mottaz A, Yip YL, Ruch P, and Veuthey AL. Mapping proteins to disease terminologies: from UNIPROT to MESH. BMC Bioinformatics, 9 (Suppl 5):S3, 2008.

Address for correspondence

Elena Beisswanger, elena.beisswanger@uni-jena.de

[9] The numbers are based on analyzing 4M abstracts from MEDLINE's Molecular Biology journals (1990-2008), which were annotated with genes/proteins by the gene name normalizer GENO [4].

[10] http://www.geneontology.org/

[11] http://www.nlm.nih.gov/research/umls/

[12] http://www.obofoundry.org/