KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

| | |
|---|---|
| **Citation/Reference** | ElShal S., Simm J., Arany A., Zakeri P., Davis J., Moreau Y. (2016), **A comprehensive comparison of two MEDLINE annotators for disease and gene linkage: sometimes less is more** Bioinformatics and Biomedical Engineering, vol. 9656 of Lecture Notes in Computer Science, Springer International Publishing, 2016, pp. 765-778 |
| **Archived version** | Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher |
| **Published version** | http://link.springer.com/chapter/10.1007/978-3-319-31744-1_66 |
| **Journal homepage** | http://www.springer.com/gp/book/9783319317434?wt_mc=ThirdParty.SpringerLink.3.EPR653.About_eBook |
| **Author contact** | sarah.elshal@esat.kuleuven.be *+32 16 32 73 86* |
| **Abstract** | Text mining is popular in biomedical applications because it allows retrieving highly relevant information. Particularly for us, it is quite practical in linking diseases to the genes involved in them. However text mining involves multiple challenges, such as (1) recognizing named entities (e.g., diseases and genes) inside the text, (2) constructing specific vocabularies that efficiently represent the available text, and (3) applying the correct statistical criteria to link biomedical entities with each other. We have previously developed Beegle, a tool that allows prioritizing genes for any search query of interest. The method starts with a search phase, where relevant genes are identified via the literature. Once known genes are identified, a second phase allows prioritizing novel candidate genes through a data fusion strategy. Many aspects of our method could |

| | |
|---|---|
| | be potentially improved. Here we evaluate two MEDLINE annotators that recognize biomedical entities inside a given abstract using different dictionaries and annotation strategies. We compare the contribution of each of the two annotators in associating genes with diseases under different vocabulary settings. Somewhat surprisingly, with fewer recognized entities and a more compact vocabulary, we obtain better associations between genes and diseases. We also propose a novel but simple association criterion to link genes with diseases, which relies on recognizing only gene entities inside the biomedical text. These refinements significantly improve the performance of our method. |
| **IR** | url in Lirias |

*(article begins on next page)*

# A comprehensive comparison of two MEDLINE annotators for disease and gene linkage: sometimes less is more

Sarah ElShal [1, 2,*], Jaak Simm [1, 2], Adam Arany [1, 2], Pooya Zakeri [1, 2], Jesse Davis [3] and Yves Moreau [1, 2]

[1] Department of Electrical Engineering (ESAT) STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics Department, KU Leuven, Leuven, 3001, Belgium
[2] iMinds Future Health Department, KU Leuven, Leuven, 3001, Belgium
[3] Department of Computer Science (DTAI), KU Leuven, Leuven, 3001, Belgium

* sarah.elshal@esat.kuleuven.be

**Abstract.** Text mining is popular in biomedical applications because it allows retrieving highly relevant information. Particularly for us, it is quite practical in linking diseases to the genes involved in them. However text mining involves multiple challenges, such as (1) recognizing named entities (*e.g.*, diseases and genes) inside the text, (2) constructing specific vocabularies that efficiently represent the available text, and (3) applying the correct statistical criteria to link biomedical entities with each other. We have previously developed Beegle, a tool that allows prioritizing genes for any search query of interest. The method starts with a search phase, where relevant genes are identified via the literature. Once known genes are identified, a second phase allows prioritizing novel candidate genes through a data fusion strategy. Many aspects of our method could be potentially improved. Here we evaluate two MEDLINE annotators that recognize biomedical entities inside a given abstract using different dictionaries and annotation strategies. We compare the contribution of each of the two annotators in associating genes with diseases under different vocabulary settings. Somewhat surprisingly, with fewer recognized entities and a more compact vocabulary, we obtain better associations between genes and diseases. We also propose a novel but simple association criterion to link genes with diseases, which relies on recognizing only gene entities inside the biomedical text. These refinements significantly improve the performance of our method.

## 1 INTRODUCTION

MEDLINE is a very large biomedical corpus containing over 25 million abstracts on life science and biomedical research [1]. This huge amount of text makes it challenging for genetic researchers to extract the desired information in a reasonable amount of time [2]. Hence, text mining has become a popular tool to help researchers extract relevant information more easily. One application of text mining is to identify links between biomedical entities of interest, such as genes and diseases. Multiple approaches have been developed for this task, which rely on co-occurrence [3], concept profile similarity [4, 5, 6], or a combination of both [7]. These solutions address challenges including (1) recognizing the correct entity occurring in a given text, (2) selecting the correct set of concepts that defines a concept profile for a given entity, and (3) using the best criteria to link one entity with another. We introduce each challenge separately as follows.

Recognizing specific concepts (e.g., diseases and genes) within a given text is widely known as Named Entity Recognition (NER). NER is a basic step in text mining that involves (1) dividing the text into tokens that correspond to entities of interest, and (2) mapping the identified tokens to the correct entities [8, 9]. Different NER approaches exist to annotate a given text (e.g. MEDLINE abstracts) with biomedical entities [8-11]. Examples include MetaMap [8] and EXTRACT [9], which can be used to map MEDLINE abstracts to different sets of biomedical concepts. On the one hand, MetaMap maps the given text to the UMLS Metathesaurus [12]. On the other hand, EXTRACT maps the given text to a selection of biomedical ontologies (such as Gene Ontology [13] and Disease Ontology [14]). The resulting annotations can then be used to generate concept profiles for each

MEDLINE abstract, and consequently concept profiles for any desired biomedical entity that is linked to MEDLINE abstracts.

A crucial aspect to building concept profiles is selecting the set of concepts, often called the vocabulary that describes a given profile. When the annotations for all MEDLINE abstracts are available, one can simply choose the vocabulary as the set of all unique concepts that are annotated. However, this is not always optimal computationally. For example, MetaMap extracts more than 500,000 unique concepts from all MEDLINE abstracts. Hence choosing this as the concept vocabulary for describing human genes requires using a structure (e.g., a matrix) whose dimensions is around 20,000 x 500,000. Loading such a data structure requires a lot of memory and doing any computation (e.g., matrix multiplication) on this data is expensive. A more practical choice would be to narrow down this vocabulary to a smaller one that covers the most important concepts for the task at hand.

Deciding on whether a gene is linked to a disease or not can be approached from many directions. For example, a gene that frequently occurs in the abstracts that are linked to a given disease has a high chance of getting annotated with that disease. This is related to co-occurrence. Also, if a gene is linked with a set of concepts that is similar to that of the disease, the chances are high that both the gene and the disease are linked together. This is related to concept profile similarity. Both directions require taking into account a background set of abstracts and concepts such that we only keep the links with a given gene that are specific to one disease and not to every other disease. For example, we do not want a gene that frequently occurs in all abstracts to get highly annotated with a given disease. Also, we do not want a concept that frequently occurs in all profiles to highly influence a disease or gene profile such that it erroneously suggests a strong link between both profiles. Hence, selecting a criterion or measure to link a gene with a given disease is challenging.

In our previous work in Beegle [7], we applied a combination of co-occurrence and concept profile similarity to associate genes with diseases, such that we selected the best rank that results from each approach separately as the final rank by Beegle for a gene given a certain disease. We used the Jaccard Similarity to measure co-occurrence, and the Cosine Similarity to measure the similarity between concept profiles. Also, we employed MetaMap to extract the biomedical concepts from the MEDLINE abstracts. For more details about Beegle, we refer the reader to our previous publication [7].

In this work, we compare the concept profiles generated by MetaMap to their counterparts generated by EXTRACT. We evaluate the influence of each concept profile setting in finding links between genes and diseases. We investigate different choices of vocabulary that we generated either manually or automatically. Our manual choices were related to choosing the starting set of unique concepts (e.g., the unique set that comes out from considering only gene-related abstracts) and the set of sources that each concept belongs to (e.g., MeSH or Ensembl). Our automatic approaches were related to combining similar concepts with each other as one united concept (e.g., via Latent Semantic Indexing (LSI)) and hence reduce our vocabulary set without losing much information. Finally, we propose an association criterion to associate genes with diseases that simplifies the concept profile similarity measure and improves its performance. We evaluate this criterion in comparison to co-occurrence and concept profile similarity as two reference criteria.

## 2 MATERIAL AND METHODS

**Named Entity Recognition according to MetaMap and EXTRACT**

MetaMap is a tool that recognizes UMLS concepts inside a given text. It has been developed at the National Library of Medicine (NLM) to map biomedical text to the UMLS metathesaurus [8]. This corresponds to concepts recognized as MeSH terms, OMIM terms, Gene Ontology terms, SNOMED clinical terms, and many others. As of February 2014, MetaMap started to release its yearly-updated annotations for the MEDLINE baselines created November the year before. These baselines correspond to all the completed citations as of that date, which include the title and abstract texts for each included citation. MetaMap provides its annotations in the MetaMap Machine Output (MMO) format which is publicly available at their FTP website [15].

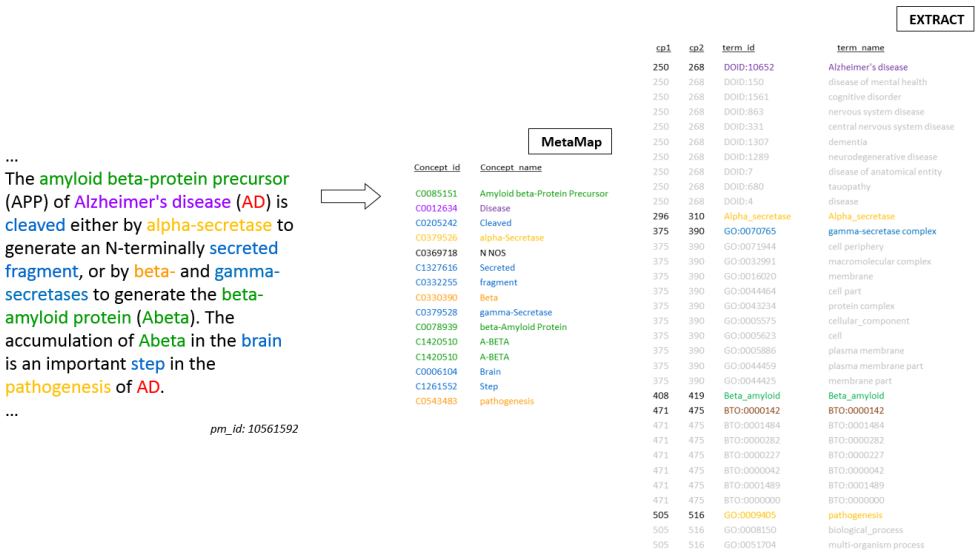
EXTRACT recognizes a collection of biomedical entities inside a given text, which corresponds to terms available in Gene Ontology (GO), Disease Ontology (DO), Ensembl, Brenda Tissue Ontology (BTO), NCBI Taxonomy, and others. It has been developed as a text mining pipeline at JensenLab [16] to serve many applications such as STRING [17]. It provides annotations for all MEDLINE titles and abstracts and it is updated every month. EXTRACT is available as a web service, and it can be downloaded as a tab separated file. The columns in this file correspond to information about the MEDLINE citation that is being annotated such as character positions and the annotated entities.

For more illustration, Table 1 provides a summary of the properties of each annotator. We also present the resulting annotations of MetaMap and EXTRACT given the same piece of text in Figure 1. We observe that MetaMap provides more annotations given that it relies on UMLS, which includes a large number of sources for biomedical concepts. We also observe that EXTRACT provides the whole hierarchy of terms (concepts) at a given character position, which is not the case for MetaMap that provides one concept at a given position. Note that we needed to parse the MMO of MetaMap to extract which concepts belong to which citation and construct the table as presented in Figure 1, which was not the case for EXTRACT where we directly received the annotations in the presented format. However we needed to integrate data from GO and DO for example to find out which terms correspond to the given term ids.

**Table 1.** A summary of the MEDLINE annotators

| | MetaMap | EXTRACT |
|---|---|---|
| Developed at | NLM | JensenLab |
| Annotations according to | GO, MeSH, OMIM, ... | GO, DO, BTO, Ensembl, … |
| Format | MMO | TSV |
| Frequently updated | Yearly | monthly |

**Fig. 1.** The annotations of MetaMap vs. EXTRACT given PubMed record 10561592




**Imposing manual and automatic vocabulary settings**

Given that we could obtain the concept annotations for all MEDLINE citations (in terms of titles and abstracts) either through MetaMap or EXTRACT, the question then was how to make use of these annotations to build concept profiles for diseases and genes to find links between such entities. This translates to choosing the sets of vocabulary used to build the concept profiles. The simplest choice would be to choose the unique set of concepts extracted from all the MEDLINE annotations; however, this was not optimal as we briefly introduced given the size of such vocabulary set. Hence, we tried different choices as follows:

1- Choose the vocabulary to be the unique set of concepts that we could extract from all MEDLINE citations that are linked with genes according to PubMed. We call this subset the PubMed vocabulary.

2- Choose the vocabulary set to be the unique set of concepts that we could extract from all the MEDLINE citations that are linked with gene functions according to GeneRIF [18]. We call this subset the GeneRIF vocabulary.

3- Choose the vocabulary set to be the unique set of concepts that only belong to a selection of biomedical sources inside a subset setting (e.g. GO, DO, and Ensembl concepts inside the GeneRIF vocabulary).

4- Apply automatic techniques such as LSI to reduce one subset setting (e.g. GeneRIF vocabulary) into a more representative set with fewer concepts.

We think that narrowing down the vocabulary corpus into the set of abstracts that talk about genes is a reasonable choice, given that it results in profiles that are focused on concepts which are gene-related and hence perform better in our problem of associating genes with diseases. For the PubMed vocabulary we used PubMed to download the ids of all the MEDLINE citations that were found to be linked with all human genes. This corresponds to a unique set of over than 2 million citations and 283,507 concepts (according to EXTRACT). For the GeneRIF vocabulary, we downloaded the ids from GeneRIF which corresponds to a unique set of 349,274 citations and 73,027 concepts (again according to EXTRACT). We applied different selections of sources inside the GeneRIF vocabulary (according to each annotator). Given the MetaMap annotations, we chose the following sources: GO, MeSH, OMIM, HUGO, and the Disease Database. This resulted in 72,822 concepts. Given

EXTRACT, we chose the following sources: GO, DO, and Ensembl. This resulted in 25,791 concepts. We selected these sources such that they are related to the two main entities in our text mining question (finding links between genes and diseases), and such that they are widely used within the annotation community [9, 19]. Finally we applied LSI via Singular Value Decomposition (SVD) to automatically reduce the GeneRIF vocabulary into a more representative subset where we could combine similar concepts together in one group. This group is called a dimension in an LSI context. We tried multiple dimension settings (starting from 2000 up to 10,000). We present a summary of the different vocabulary settings we just discussed in Table 2.

**Table 2.** A summary of the different vocabulary settings

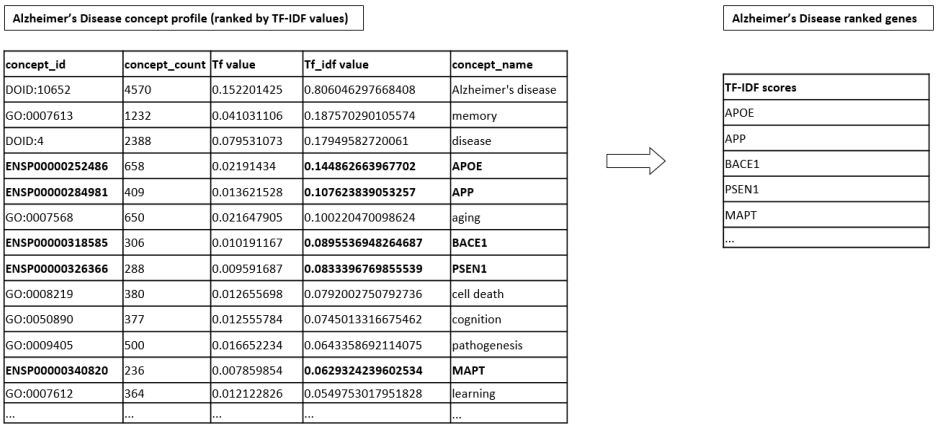
| | The PubMed vocabulary | The GeneRIF vocabulary | |
|---|---|---|---|
| # citations | 2,801,750 | 349,274 | |
| # concepts (complete set) | 283,507 (EXTRACT) | 73,027 (EXTRACT) | 119,336 (MetaMap) |
| # concepts (selecting sources) | n.a. | 25,791 (EXTRACT) | 72,822 (MetaMap) |
| # concepts (LSI) | n.a. | up to 10,000 | |

**Investigating multiple association measures**

Different measures exist to associate genes with diseases inside text mining. Co-occurrence and concept profile similarity are two examples. In co-occurrence, we rely on the disease being linked with a set of MEDLINE citations that is similar to the set of the gene. Hence the disease and the gene frequently co-occur, which either can happen in the full citation level, the abstract level, or even the sentence level. In concept profile similarity, we rely on the fact that a disease is found to share a similar concept profile to that of the gene. Hence the disease and the gene are described by the same biomedical concepts, from which we could infer that there is a strong link between the disease and the gene. Here we used the Jaccard Similarity to measure co-occurrence, and we used the Cosine Similarity to measure concept profile similarity. For more information about each measure, we refer the reader to our previous work [7].

In this work we propose a novel measure to associate genes with diseases, which can be seen as a mix between co-occurrence and concept profile similarity. In concept profile similarity we represent each concept inside the profile by its TF-IDF (Term Frequency – Inverse Document Frequency) value. This representation gives higher weights to concepts that frequently occur with the entity they describe but don't frequently occur in general, and it gives lower weights to concepts that frequently occur in general even though they frequently occur inside a given profile. Hence we decided to use the TF-IDF values for "gene" concepts inside a disease profile to be used as the score (or measure) that ranks how well a gene is linked with a given disease. We call this measure the TF-IDF scores. We show an example of this measure in Figure 2. On the left hand side we present the concept profile for Alzheimer's disease that is ranked by TF-IDF values in a descending order. We only show the top 13 concepts. We highlight the gene concepts in bold. On the right hand side we present the ranks of the genes against Alzheimer's disease according to their TF-IDF scores.

**Fig. 2.** An example for the TF-IDF scores



**Alzheimer's Disease concept profile (ranked by TF-IDF values)**

| concept_id | concept_count | Tf value | Tf_idf value | concept_name |
|---|---|---|---|---|
| DOID:10652 | 4570 | 0.152201425 | 0.806046297668408 | Alzheimer's disease |
| GO:0007613 | 1232 | 0.041031106 | 0.187570290105574 | memory |
| DOID:4 | 2388 | 0.079531073 | 0.17949582720061 | disease |
| **ENSP00000252486** | 658 | 0.02191434 | **0.144862663967702** | **APOE** |
| **ENSP00000284981** | 409 | 0.013621528 | **0.107623839053257** | **APP** |
| GO:0007568 | 650 | 0.021647905 | 0.100220470098624 | aging |
| **ENSP00000318585** | 306 | 0.010191167 | **0.0895536948264687** | **BACE1** |
| **ENSP00000326366** | 288 | 0.009591687 | **0.0833396769855539** | **PSEN1** |
| GO:0008219 | 380 | 0.012655698 | 0.0792002750792736 | cell death |
| GO:0050890 | 377 | 0.012555784 | 0.0745013316675462 | cognition |
| GO:0009405 | 500 | 0.016652234 | 0.0643358692114075 | pathogenesis |
| **ENSP00000340820** | 236 | 0.007859854 | **0.0629324239602534** | **MAPT** |
| GO:0007612 | 364 | 0.012122826 | 0.0549753017951828 | learning |
| ... | ... | ... | ... | ... |

**Alzheimer's Disease ranked genes**

| TF-IDF scores |
|---|
| APOE |
| APP |
| BACE1 |
| PSEN1 |
| MAPT |
| ... |

## The datasets

In our experiments, we used the 2014 release of MetaMap for the MEDLINE annotations. This corresponds to annotations for 22,076,054 MEDLINE citations. We used a version of EXTRACT that we downloaded in December 2014. This corresponds to annotations for 20,686,757 MEDLINE citations. As for the validation set, we used a benchmark of experimentally validated disease–gene annotations that we extracted from the OMIM morbidmap (downloaded in May 2015). This corresponds to 330 diseases, 2214 genes, and 2789 disease–gene pairs. We downloaded our gene data (ids and symbols) from the Ensembl database (in March 2013). This corresponds to 17,116 gene records. We only consider human genes in our experiments. In order to generate the gene concept profiles, we used GeneRIF to download the ids of the MEDLINE citations that are functionally linked with our Ensembl genes (downloaded in March 2015). This corresponds to a unique set of 349,274 citations, which we used to generate the GeneRIF vocabulary. Additionally we used PubMed to download the ids of the more general list of MEDLINE citations that were found to be linked with our Ensembl genes, which we used to generate the PubMed vocabulary. As for the disease concept profiles, we similarly used PubMed to download the corresponding list of MEDLINE ids. This corresponds to a set of 936,668 unique citations. Note that on PubMed, we restrict the maximum number of ids retrieved per entity to 6500. This is the maximum number of ids that we found linked to a gene in GeneRIF.

## Boltzmann-Enhanced Discrimination (BEDROC) evaluation

The Area Under the Receiver Operating Characteristic (ROC) curve (AUC) has been widely used to evaluate and compare prioritization tools. It can be interpreted as the probability of a disease-associated gene being ranked earlier than a gene selected at random by a uniform distribution. To estimate the AUC value of a prioritization model, we can simply take the average of the ranks of disease-associated genes considered as the test set. However, the AUC score often leads to a misinterpretation of the model's performance in early discovery of disease-associated genes [20, 21], especially from a researcher's perspective who is normally interested in the top results for a given disease. As a result, Boltzmann-Enhanced Discrimination of ROC (BEDROC) has been proposed [20] as a proper and robust evaluation measurement for the early discovery.

For n disease-associated genes ranked $< r_i >_{i=1}^{n}$ among N genes, the BEDROC score is calculated as follows:

$$\text{BEDROC} = \frac{\sum_{i=1}^{n} \exp(-\alpha p_i)}{\frac{n}{N}\frac{1-\exp(-\alpha)}{\exp(\frac{\alpha}{N}-1)}} + \frac{R_a \sinh(\frac{\alpha}{2})}{\cosh(\frac{\alpha}{2}) - \cosh(\frac{\alpha}{2} - \alpha R_a)} + \frac{1}{1-(\exp(\alpha(1-R_a)))} \quad (1)$$

where $p_i(\frac{r_i}{N})$ is the normalized rank of the i[th] disease-associated gene, $Ra = \frac{n}{N}$ is the ratio of the number of disease-associated genes to the total number of genes, and the parameter α tunes the importance given to early recognition. For example, when alpha equals to 275.5, 80% of BEDROC score is assigned to the top 100 ranked genes. The BEDROC value can be interpreted as the probability that a disease-associated gene being ranked better than a gene selected at random from an exponential probability distribution function of parameter α. In this study, we consider values of α equal to α = 160.9, α = 275.5 and α = 550.9, which correspond to 80% of the BEDROC being assigned to the top 1%, top 100 and top 50 ranked genes, respectively.

**Setting up the experiments**

In this work we had three objectives. First was to compare the contribution of MetaMap and EXTRACT as two MEDLINE annotators in generating concept profiles for diseases and genes; mainly in terms of how well each concept profile setting links the correct genes with their corresponding disease in our OMIM test set. Second was to check the impact of choosing the vocabulary on shaping the concept profiles and how that influences the disease–gene annotation process. Third was to compare the TF-IDF scores to concept profile similarity and co-occurrence as two traditional approaches. So we proceeded as follows:

1- Starting from the GeneRIF vocabulary, we used the complete annotations of MetaMap and EXTRACT to generate the concept profiles for our genes and diseases. Then we applied concept profile similarity and measured the cosine similarity on the TF-IDF representations of the profiles to score genes against diseases. We used these scores to rank the genes and calculate the BEDROC scores at the different α values. Note that for EXTRACT, we tried the complete annotations once using the whole hierarchy (including parent terms) and once using only the leaf terms at a given character position. We call this experiment the complete setup.

2- We used our manual selection of sources inside MetaMap and EXTRACT to generate a reduced version of the concept profiles we constructed in the first experiment. In parallel, we applied LSI. Again we measured the cosine similarity, computed the gene scores, and calculated the BEDROC score. Note that here for EXTRACT, we included only the leaf terms at a given position. We call this experiment the reduced setup.

3- We applied the TF-IDF scores measure on the (manually) reduced disease concept profiles of MetaMap and EXTRACT. We also tried a combination of TF-IDF scores and concept profile similarity by assigning the best rank that results from each approach as the gene's new score. Furthermore, we compared that to co-occurrence in which we applied the Jaccard-similarity to score a gene against a given disease.

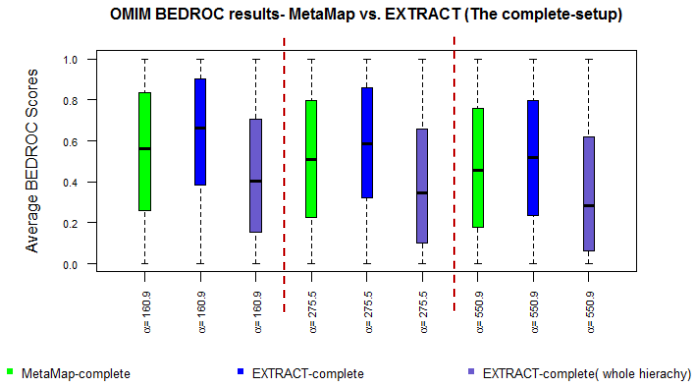
4- We additionally applied TF-IDF scores on the disease concept profiles resulting from the complete annotations of EXTRACT according to the PubMed vocabulary. Given the size of this vocabulary the TF-IDF scores measure was the most convenient computationally.

# 3 RESULTS

### MetaMap vs. EXTRACT (the complete setup)

We present the average BEDROC from the complete setup experiment in Figure 3. We observe that by applying concept profile similarity and including only the leaf terms of EXTRACT, we achieve the best average score of 62%, 57%, and 51% at $\alpha = 160.9$, $\alpha = 275.5$, and $\alpha = 550.9$ respectively. This compares to 54%, 51%, and 47% when employing MetaMap, and 44%, 40%, and 36% when considering the whole hierarchy of EXTRACT. Note that we highlight the black solid lines in the box plots correspond to the median value and not the average. This remark applies to the following box plots as well.

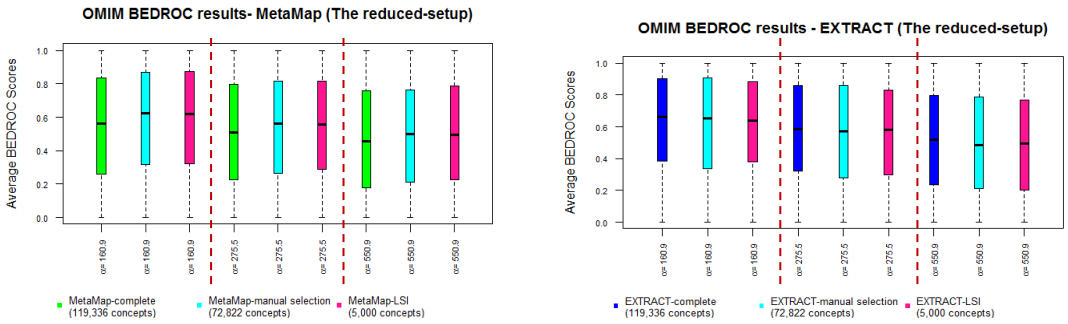
**Fig. 3.** MetaMap vs. EXTRACT (The complete setup)




### MetaMap vs. EXTRACT (the reduced setup)

We present the average BEDROC from the reduced setup when applying concept profile similarity and employing MetaMap on the left hand side of Figure 4. We observe that both the manual and the automatic reductions of the concept profiles result in an average score of 57%, 53%, and 49% at $\alpha = 160.9$, $\alpha = 275.5$, and $\alpha = 550.9$, which slightly improves the complete setup when employing MetaMap (especially at $\alpha = 160.9$). We also present the results from the reduced setup when employing EXTRACT on the right hand side of Figure 4. We observe that the reduced setup results in a comparable performance to the complete setup.

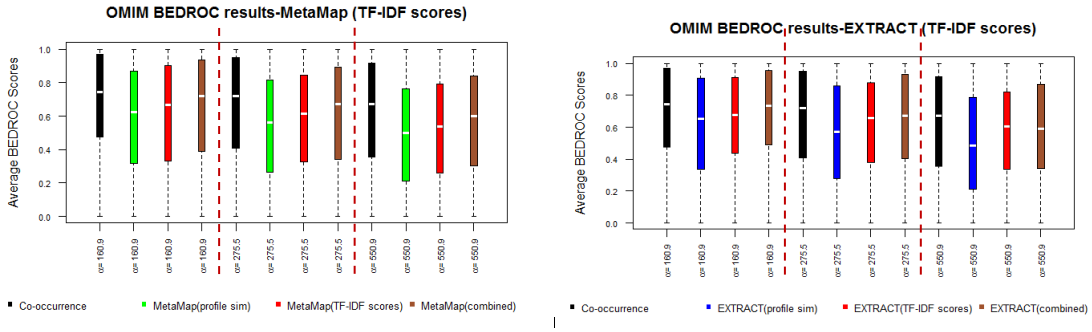
**Fig. 4.** MetaMap and EXTRACT (The reduced setup)



**TF-IDF scores**

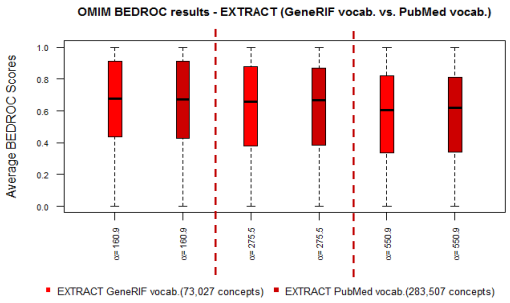
We present the average BEDROC when applying TF-IDF scores and employing MetaMap on the left hand side of Figure 5. We observe that the TF-IDF scores measure improves the BEDROC results of concept profile similarity such that it reaches an average of 59%, 56%, and 52% at $\alpha = 160.9$, $\alpha = 275.5$, and $\alpha = 550.9$. We also observe that when combining both TF-IDF scores and concept profile similarity, we achieve the best BEDROC results in this setting, which correspond to an average of 63%, 59%, and 55% at $\alpha = 160.9$, $\alpha = 275.5$, and $\alpha = 550.9$. We also present the performance of TF-IDF scores when employing EXTRACT on the right hand side of Figure 5. Again we observe that TF-IDF scores improve the results and when combined with concept profile similarity, we achieve the best results of 68%, 63%, and 58% at $\alpha = 160.9$, $\alpha = 275.5$, and $\alpha = 550.9$. We also observe that the improvement is more significant at the earlier discovery ($\alpha = 550.9$) in both models.

**Fig. 5.** MetaMap and EXTRACT (The TF-IDF scores)




We additionally present the results when applying TF-IDF scores and employing EXTRACT while including the PubMed vocabulary in Figure 6. We observe a comparable performance to the setting where we included GeneRIF as our vocabulary.

**Fig. 6.** EXTRACT (GeneRIF vocabulary vs. PubMed vocabulary)




# 4 DISCUSSION

In this work, we studied the contribution of MetaMap and EXTRACT as two different MEDLINE annotators in generating concept profiles for diseases and genes so that we could associate these entities with each other. We tried different vocabulary settings and compared different versions of the concept profiles generated by each annotator. We imposed these settings in manual and automatic fashions either by selecting the source vocabularies that generate the mapped concepts inside a given annotator or by applying LSI techniques. We also discussed TF-IDF scores as a criterion that we propose to associate genes with diseases.

We present a detailed summary of our results in Table 3. Our results show that EXTRACT outperforms MetaMap for disease-gene association in the complete setup experiment. This is achieved with more compact concept profiles and fewer concepts. We also show that when we further reduced the concept profiles generated from both annotators, either manually or automatically, we achieved at least as good performance with even fewer concepts. Furthermore, we showed that applying TF-IDF scores significantly improve the disease-gene associations especially when being combined with concept profile similarity. This combination approximates the performance of co-occurrence and it even improves it at the top 1% threshold. We additionally applied the t-test to assess the significance between our results (*e.g.* comparing EXTRACT_combined and MetaMap_combined at $\alpha = 160.9$, we achieved t=6.0629 and p-value=3.665e-09). The application of TF-IDF scores as an association criterion is interesting because it is simpler than concept profile similarity. In TF-IDF scores, we only need concept profiles for diseases, unlike concept profile similarity where we need concept profiles for disease and gene entities. Also in TF-IDF scores, we directly use the scores as the TF-IDF values of the gene concepts inside a disease profile, while in concept profile similarity we need to calculate the scores according to some similarity statistic (e.g., cosine similarity). Hence with TF-IDF scores we consume less space, do fewer computations, and arrive to better disease-gene associations.

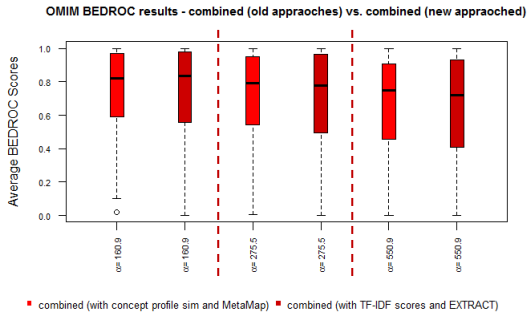
**Table 3.** A summary of the results

| **Methods** | **Average BEDROC score** | | |
|---|---|---|---|
| | $\alpha = 160.9$ | $\alpha = 275.5$ | $\alpha = 550.9$ |
| 80% of score given to the | **Top 1%** | **Top 100** | **Top 50** |
| TF-IDF scores + concept profile similarity<br>EXTRACT (manual selection) | **0.6800** | **0.6343** | **0.5756** |
| TF-IDF scores<br>EXTRACT | 0.6453 | 0.6110 | 0.5649 |
| Concept profile similarity<br>EXTRACT (LSI: 5000 dimensions) | 0.6037 | 0.5527 | 0.4891 |
| Concept profile similarity<br>EXTRACT (manual selection: 25,791 concepts) | 0.6048 | 0.5572 | 0.4980 |
| Concept profile similarity<br>EXTRACT (leaf terms: 73,027 concepts) | 0.6162 | 0.5704 | 0.5123 |
| Concept profile similarity<br>EXTRACT (whole hierarchy: 108,392 concepts) | 0.4408 | 0.4006 | 0.3575 |
| Concept profile similarity<br>MetaMap (119,336 concepts) | 0.5369 | 0.5069 | 0.4724 |
| Concept profile similarity<br>MetaMap (manual selection: 72,822 concepts) | 0.5661 | 0.5329 | 0.4920 |
| Concept profile similarity<br>MetaMap (LSI: 5000 dimensions) | 0.5752 | 0.5380 | 0.4907 |
| TF-IDF scores<br>MetaMap | 0.5906 | 0.5570 | 0.5187 |
| TF-IDF scores + concept profile similarity<br>MetaMap (manual selection) | 0.6313 | 0.5946 | 0.5464 |
| *Co-occurrence* | *0.6751* | *0.6504* | *0.6154* |

In comparison to our previous approaches in Beegle, we combined co-occurrence with TF-IDF scores on the disease profiles according to EXTRACT using best rank, and then computed the BEDROC scores against our previous OMIM benchmark. We compared this to our previous best approach where we combined co-

occurrence with concept profile similarity according to MetaMap. We present the results in Figure 7. We observe comparable BEDROC results.

**Fig. 7.** In comparison to old Beegle




We wanted to get an additional insight on the performance of each annotator and whether one works better on some disease queries that are different from the other or not. Hence we checked the diseases that achieved minimum recall (recall =0) in the top 100 ranked genes when applying TF-IDF scores given each annotator. We found out that the zero-recall set resulting from applying TF-IDF scores on the disease profiles according to EXTRACT is simply a subset of its counterpart according to MetaMap. It is also a subset of the zero-recall set when applying co-occurrence. We further investigated these disease queries and checked why they consistently lead to very poor recall. We present them in Table 4. We observe two things. First, most of the diseases are linked to very few citations, hence text mining cannot do much here and annotation with the correct genes fails. This is further confirmed when we checked the average number of citations for the one-recall set in the top 10 ranked genes, which is 2208.4 citations. Second, when enough text is available for the disease query, the corresponding top ranking genes are not random, however they share a fair number of citations with their corresponding disease query but they are not annotated in OMIM. Hence, text mining still returns some true biology here however the benchmark is probably not complete.

**Table 4.** Zero-recall diseases

| Disease name | #citations | Remarks |
|---|---|---|
| Barrett esophageal adenocarcinom | 1 | Very few text available |
| Cerebrooculofacioskeletal syndrome | 4 | Very few text available |
| Cirrhosis | 6500 | Enough text available however, top 3 genes are:<br>CFTR: 7844 common citations<br>CTGF: 1229 common citations<br>SMAD2: 861 common citations |
| Heinz body anemias | 5 | Very few text available |
| Microcephaly and chorioretinopathy | 27 | Very few text available |
| Coronary artery disease | 6500 | Enough text available however, top 3 genes are:<br>CRP: 1810 common citations<br>IL6: 138 common citations<br>MPO: 143 common citations |
| Lymphoma | 6500 | Enough text available however, top 3 genes are:<br>ALK: 2108 common citations<br>BCL6: 890 common citations<br>CD4: 5015 common citations |
| Major depressive disorder and accelerated response to antidepressant drug treatment | 21 | Very few text available |
| Renal plasia | 2 | Very few text available |

As for future work, we are currently integrating the annotations from EXTRACT to generate our concept profiles for genes and queries inside Beegle. We also plan to apply the TF-IDF scores measure in combination with the current approaches we apply there. Finally, we plan to study more automatic techniques (e.g., Latent Dirichlet Allocation and Logistic Regression) to sort out the most important concepts inside a concept profile and construct more relevant vocabularies.

## ACKNOWLEDGEMENTS

This work was supported by the Research Council KU Leuven [CoE PFV/10/016 SymBioSys, OT/11/051] to Y.M. and J.D.; the government agency for Innovation by Science and Technology to Y.M.; Industrial Research fund to Y.M.; Hercules Stichting to Y.M.; iMinds Medical Information Technologies [SBO 2015] to Y.M.; EU FP7 Marie Curie Career Integration Grant [\#294068] to J.D.; FWO-Vlaanderen [G.0356.12] to J.D.; and IMEC mandaat - Ph.D mandaat to A.A.. Funding for open access charge: Research Council KU Leuven.

## REFERENCES

1. United States National Library of Medicine (2002) PubMed: MEDLINE Retrieval on the World Wide Web. Fact Sheet.
2. Jensen, L. J., Saric, J., & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. Nature Reviews. Genetics, 7(2), 119–129.
3. Fleuren,W.W., Verhoeven,S., Frijters,R., Heupers,B., Polman,J., van Schaik,R., de Vlieg,J., Alkema,W. (2011) CoPub update: CoPub 5.0 a text mining system to answer biological questions, Nucleic Acids Res., 39
4. Jelier,R., et al. (2007) Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation. BMC Bioinformatics., 18, 8-14
5. Jelier,R., Schuemie,M.J., Roes,P.J., van Mulligen,E.M., Kors,J.A. (2008) Literature-based concept profiles for gene annotation: the issue of weighting. Int J Med Inform., 77, 354-362
6. Jelier,R., Schuemie,M.J., Veldhoven,A., Dorssers,L.C., Jenster,G., Kors,J.A. (2008): Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol., 9(6), R96
7. ElShal, S., Tranchevent, L.-C., Sifrim, A., Ardeshirdavani, A., Davis, J., & Moreau, Y. (2015). Beegle: from literature mining to disease-gene discovery. Nucleic Acids Research, 44 (2), e18.
8. Aronson, A. R., & Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3), 229–236.
9. Pafilis, E., et al. (2015). EXTRACT: Interactive extraction of environment metadata and term suggestion for metagenomics sample annotation. To appear in Database.
10. Netherlands Bioinformatics Centre. Peregrine literature indexing service.
11. United States National Library of Medicine. PubMed MeSH indexing.
12. Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res., 32, D267–270
13. Ashburner,M., Ball,C.A., Blake,J.A., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet., 25, 25–29
14. Kibbe,W.A., Arze,C., Felix,V., et al. (2015) Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res., 43, D1071–D1078.
15. United States National Library of Medicine. MetaMapped MEDLINE Baseline Results: http://ii.nlm.nih.gov/MMBaseline/index.shtml
16. Lars Juhl Jensen from the Novo Nordisk Foundation Center for Protein Research. JensenLab: Cellular Network Biology: http://jensenlab.org/
17. Szklarczyk, D., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Research, 43(Database issue), D447–452.
18. Mitchell, J. A., Aronson, A. R., Mork, J. G., Folk, L. C., Humphrey, S. M., & Ward, J. M. (2003). Gene indexing: Characterization and analysis of NLM's GeneRIFs. AMIA Annual Symposium Proceedings, 460–464.
19. Cheung, W. a, Ouellette, B. F., & Wasserman, W. W. (2012). Inferring novel gene-disease associations using medical subject heading over-representation profiles. Genome Medicine, 4(9), 75.
20. Truchon, J.F., Bayly C.I. (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. Journal of Chemical Information Modeling, 47:488–508
21. Zhao, W., Hevener, K., White, S., Lee, R., Boyett, J. (2009) A statistical framework to evaluate virtual screening. BMC Bioinformatics, 10, 225.