Published online 17 September 2015

Nucleic Acids Research, 2016, Vol. 44, No. 2 ***e18***
doi: 10.1093/nar/gkv905

Beegle: from literature mining to disease-gene discovery

Sarah ElShal[1,2,*], Léon-Charles Tranchevent[1,2,3,4,5], Alejandro Sifrim[1,2,6], Amin Ardeshirdavani[1,2], Jesse Davis[7] 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]Inserm UMR-S1052, CNRS UMR5286, Cancer Research Centre of Lyon, Lyon, France, [4]Université de Lyon 1, Villeurbanne, France, [5]Centre Léon Bérard, Lyon, France, [6]Wellcome Trust Genome Campus, Hinxton, Wellcome Trust Sanger Institute, Cambridge CB10 1SA, UK and [7]Department of Computer Science (DTAI), KU Leuven, Leuven 3001, Belgium

Received March 03, 2015; Revised August 25, 2015; Accepted August 29, 2015

ABSTRACT

Disease-gene identification is a challenging process that has multiple applications within functional genomics and personalized medicine. Typically, this process involves both finding genes known to be associated with the disease (through literature search) and carrying out preliminary experiments or screens (e.g. linkage or association studies, copy number analyses, expression profiling) to determine a set of promising candidates for experimental validation. This requires extensive time and monetary resources. We describe *Beegle*, an online search and discovery engine that attempts to simplify this process by automating the typical approaches. It starts by mining the literature to quickly extract a set of genes known to be linked with a given query, then it integrates the learning methodology of *Endeavour* (a gene prioritization tool) to train a genomic model and rank a set of candidate genes to generate novel hypotheses. In a realistic evaluation setup, *Beegle* has an average recall of 84% in the top 100 returned genes as a search engine, which improves the discovery engine by 12.6% in the top 5% prioritized genes. *Beegle* is publicly available at http://beegle.esat.kuleuven.be/.

INTRODUCTION

Determining which genes cause which diseases is an important yet challenging problem (1). It has a variety of applications that range from DNA screening and early diagnosis, to gene sequence analysis and drug development (2). However, it is resource intensive both in terms of time investment and monetary cost. Traditionally, disease-gene identification is approached manually and is conducted in two phases. The first phase involves narrowing down a large set of candidate genes (e.g. the whole genome) into a significantly smaller set of genes that has a high probability of containing a disease causing gene. Different ways exist to tackle this phase, such as linkage analysis, genome sequencing and association studies (3–5). Then, in the second phase, experts experimentally evaluate the selected genes to confirm which of those candidates are truly disease causing. This involves wet lab experimentation for every selected gene. Consequently, an important advancement in this field has been the development of computational methods that can help the experts address the first phase of this process by automatically prioritizing a set of candidate genes for final experimental validation to maximize the yield of the second phase.

Many computational methods for human gene prioritization have been developed, and several review articles exist that describe their approaches, their differences, and how they can be used in practice (6–9). These methods differ in their expected inputs, their returned outputs and their prioritization strategies. A previous study compared the performance of eight of these methods that are publicly available as web-based tools (10). The evaluation setup used a realistic scenario where data prior to a certain date were used to generate the gene prioritizations and then the predictions were compared to disease-gene annotations discovered later. The results showed that *Endeavour* (11), *GeneDistiller* (12) and *ToppGene* (13) performed best when measuring the true-positive rates among the top returned genes. All three tools require a set of training genes (genes that are known to be linked to the disease of interest) or keywords (describing the disease under study) as input, which is then

[*]To whom correspondence should be addressed. Tel: +32 16 32 73 86; Fax: +32 16 32 19 70; Email: sarah.elshal@esat.kuleuven.be
Present address: Sarah ElShal, Department of Electrical Engineering (ESAT) STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics Department, KU Leuven, Leuven 3001, Belgium.

used to infer several models (according to different genomic sources) and rank a set of candidate genes based on the learned models. These approaches all require hand-selected input to compute the gene prioritizations. Normally, experts undertake this challenging and time-consuming process by collecting information from (i) generic or disease specific association databases (e.g. OMIM (14), GAD (15)), (ii) relevant literature or (iii) their own data and expertise (including for instance relevant patient records). Therefore, a tool that would support automatic identification of the training genes as an initial step to candidate gene prioritization would provide better usability to researchers interested in candidate gene prioritization (16).

Text mining is one popular strategy for automatically associating biomedical entities with each other (17–24). *MeSHOP* (22) and *Genie* (24) are two examples that associate genes with diseases. These tools can be used to rank a set of genes given a disease query, hence they could be used by the genetic experts to automatically search for the training input required by the gene prioritization tools. However these tools do not distinguish known gene associations from unknown ones. Hence, for a user who is interested in selecting a set of training genes prior to conducting a gene prioritization process, the existing tools are limited and require post processing to filter out the resulting gene associations.

This article presents *Beegle*, an online tool for disease-gene prioritization. First, *Beegle* mines the literature to automatically identify a list of genes known to be linked with a given query. Next, *Beegle* employs *Endeavour*, which integrates multiple genomic models to automatically rank a set of candidate genes (e.g. the human genome) according to a selected set of genes identified in the first step. We evaluated *Beegle* in two different ways. First, we evaluated its ability to identify known disease-gene associations from the literature. To do this, we have extracted a list of experimentally validated disease-gene associations from the OMIM database. Then, for each disease in this list, we compared the associations returned by *Beegle* to the known associations from OMIM. In addition, we compared *Beegle* to *MeSHOP*, a similar tool, on a subset of the OMIM list using the same experimental setup. Second, we evaluated the suitability of the returned genes to serve as input to train genomic models and generate novel hypothesis. Here, we employed an evaluation methodology that mimics real discovery by using rolled-back data to generate the gene prioritizations, and then by testing on disease-gene associations that were reported after the training data were collected. For this we have used two benchmarks: one based on literature that has already been described (10) and a new one that we have generated from the OMIM database. Our OMIM benchmark is a secondary contribution of this work and is made publicly available as supplementary data so that other researchers can use it to evaluate gene prioritization approaches.

MATERIALS AND METHODS

The pipeline

An overview of the current methodology of *Beegle* is shown in Figure 1. *Beegle* starts from a user query (e.g. a disease) and proceeds in two phases. First, it automatically analyses the literature to identify the genes that are potentially related with the given query. We call this phase the search phase. Second, it uses these genes (identified in the first step) as a seed set that is provided to *Endeavour*, which then analyses a number of genomic data sources to finally prioritize a set of candidate genes. We call this phase the discovery phase.

Annotating the literature

We use the biomedical database MEDLINE as our source of literature. Offline, we have indexed every MEDLINE abstract using MetaMap (25), which identifies the UMLS concepts (26) within a given abstract text. UMLS is a large, multi-purpose and multi-lingual thesaurus that brings together many health and biomedical vocabularies and standards (e.g. MeSH and SNOMED CT). With this strategy we have associated each MEDLINE abstract to a list of UMLS concepts. This corresponds to 12 308 151 abstract-concepts entries. We report the corresponding list of MEDLINE ids in Supplementary Data 1.

For every gene, we find the list of associated MEDLINE abstracts according to GeneRIF (downloaded in May 2012). Hence, we could generate a UMLS concept profile for all Entrez gene entries (16 493 genes in total, which we report in Supplementary Data 2). The gene profiles are described using 66 883 concepts, which we call the Genes-vocabulary. For every query, we find the list of associated MEDLINE abstracts according to PubMed, where we only consider the top 10 000 PubMed Ids to generate a corresponding UMLS profile. We restrict the query profiles to the concepts that already appear in the Genes-vocabulary.

The search phase

In the search phase, *Beegle* applies two text mining approaches to identify the genes most related to a given query. The first one is based on the number of abstracts in which the query and a given gene co-occur. We call this the explicit approach, since it relies on the explicit co-occurrence of a query and a given gene in the literature. We count three values: (i) the number of abstracts associated with the query, (ii) the number of abstracts associated with the gene and (iii) the number of abstracts associated with both the query and the gene. Then we use the Jaccard similarity to measure the strength of the association according to Equation (1):

$$\text{similarity\_explicit}(q, g) = \frac{X_{q,g}}{N_q + K_g - X_{q,g}} \tag{1}$$

where N is the number of abstracts associated with the query, K is the number of abstracts associated with the gene and X is the number of abstracts associated with both query and gene.

The higher the similarity score, the more confident we are that the association between the gene and the given query is real.

The second approach is based on the number of concepts shared between a gene and the given query profiles. We call this approach the implicit approach, since it goes one step further and tries to find hidden indirect associations between a gene and the given query. Given the UMLS

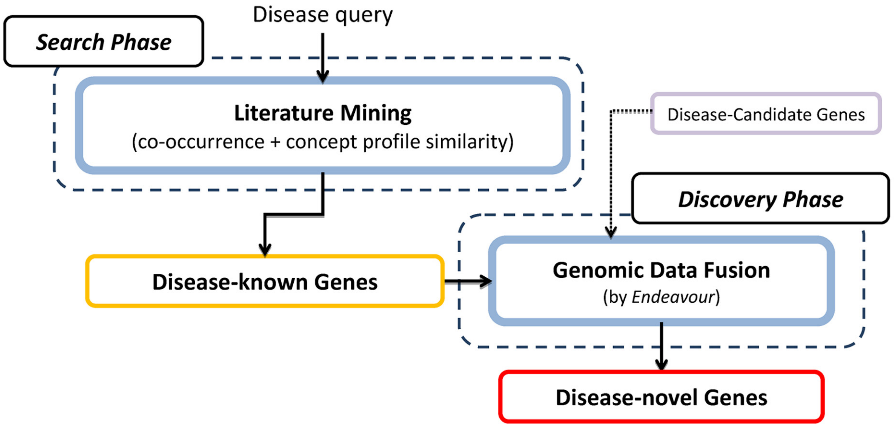


Figure 1. An illustration of *Beegle's* pipeline. The disease-gene annotation in *Beegle* follows two phases: search and discovery. The search phase involves two text mining techniques, while the discovery phase involves fusing different genomic models.

concept profiles that correspond to both the query and a gene, we apply the TF-IDF (Term Frequency-Inverse Document Frequency) transformation to each term in both the query and gene profiles. This transformation is commonly used in text mining and information retrieval (see (27) for more details), and it consists of two components: the term frequency (TF) and the inverse document frequency (IDF). The TF corresponds to the number of times the concept appears in all abstracts linked to a query or a gene. Intuitively, TF rewards concepts that are frequently associated with the gene. The IDF is calculated as the total number of documents divided by the number of documents that the concept appears in. The IDF gives higher weights to concepts that are commonly associated with a given gene but rare in general, and lower weights to concepts that are associated with many genes. Intuitively, IDF helps identify concepts that are more meaningful or discriminative for a given profile. Supplementary Table S1 illustrates how we compute TF-IDF scores. To measure how similar two concept profiles are, we calculate the cosine similarity between their TF-IDF mappings according to Equation (2):

$$\text{similarity\_implicit}(q, g) = \frac{(\text{tf} \times \log(\text{idf}))_q \bullet (\text{tf} \times \log(\text{idf}))_g}{\left\|(\text{tf} \times \log(\text{idf}))_q\right\| \left\|(\text{tf} \times \log(\text{idf}))_g\right\|} \quad (2)$$

where tf is the term frequency and idf is the inverse document frequency (note that we use the log for scaling purposes).

Beegle assigns a final gene-query score by using the best rank of both approaches. We call this approach, the combined approach. We show an example in Supplementary Table S2. Hence, *Beegle*'s output for this phase is an ordered list of the genes identified as being potentially related to the given query according to the literature.

The discovery phase

In the second phase, *Beegle* integrates *Endeavour* to generate the final gene prioritization for a given query. *Endeavour* relies on three inputs: (i) a set of training genes known to be linked to the query under study, (ii) a set of data sources that are used to build the query models using the training genes and (iii) a set of candidate genes to investigate (i.e. to prioritize). Per data source, *Endeavour* ranks the candidate genes according to how similar a gene is to the corresponding model, therefore providing one ranked list for each data source. To combine the lists, *Endeavour* applies the Order Statistics to produce a single ranking, which is the final prioritization list for the given query. For more details about *Endeavour*, we refer the reader to our previous work (11,28). Hence, in this phase, *Beegle* uses a training set (that the user selected from the known query-genes retrieved by *Beegle* in the search phase) to train the *Endeavour* models and rank a user-defined set of candidate genes. Note here that we restrict the data sources used by *Endeavour* in this phase to a predefined set that performed best in our experiments.

The web interface

Beegle is freely available online as a web interface that accepts any combination of one or many biomedical concepts (similar to *PubMed* queries such as 'Alzheimer's disease', 'Diabetes and Pregnancy' and 'Congenital Heart Defects or Eye Diseases'). Given the input query, *Beegle* retrieves the MEDLINE abstracts annotated with any of the query terms (according to PubMed) and generates the corresponding concept profile. Then for every human gene, it computes two scores according to our two text mining approaches (as discussed above). Finally, *Beegle* returns an ordered list of the genes most likely to be linked to the given query. Since in this phase we are only aiming to retrieve known annotations, we restrict the output list to genes that are co-mentioned with the query at least once in the literature. The response time for this phase ranges from few seconds to few minutes, depending on whether or not the query has been previously processed.

Next, the user can initiate the discovery phase by: (i) selecting a set of training genes (given the search output list), and (ii) defining a list of candidate genes that are of interest to prioritize. After the literature search, the interface allows users to review the evidence linked to the association by displaying relevant MEDLINE abstracts and concept profiles. This allows for the quick removal of spurious associations (false positives). Any association that might have been missed (false negative) can still be manually added to the list of genes to be used for the prioritization. Also, the user has the option to directly upload a set of training or candidate genes. Here the users have full control of fine tuning the selection process according to their expertise. Next, *Beegle* trains the genomic models and prioritizes the user-defined candidate set according to the selected training genes and the different genomic sources that are predefined (see list below). Finally, *Beegle* returns the prioritized list in a simple and user-friendly way. The response time for this phase ranges from few minutes (ranking a few candidate genes) up to 20 min (ranking the whole genome). Figure 2 presents three snapshots of the most important screens from the web interface. In addition, we invite the reader to visit the *Beegle* website and watch a simple tutorial on how to use system.

Data sources and background corpus

The public web interface of *Beegle* relies on the 2013 PubMed release 'downloaded on March 4th 2013'. Then it uses the 2013 (or the latest) version of the following data sources for the discovery phase: Gene Ontology (annotations for gene products) 'downloaded on May 15th 2013', Uniprot (protein sequence and functional information) 'downloaded on June 14th 2013', Text (MEDLINE literature) 'downloaded on March 4th 2013', STRING (genomic data integration) 'downloaded on June 10th 2013', Genetic Association Database 'downloaded on June 14th 2013', Rat Gene Database 'downloaded on June 20th 2013', gene predicted pathogenicity (29) and expression data (30).

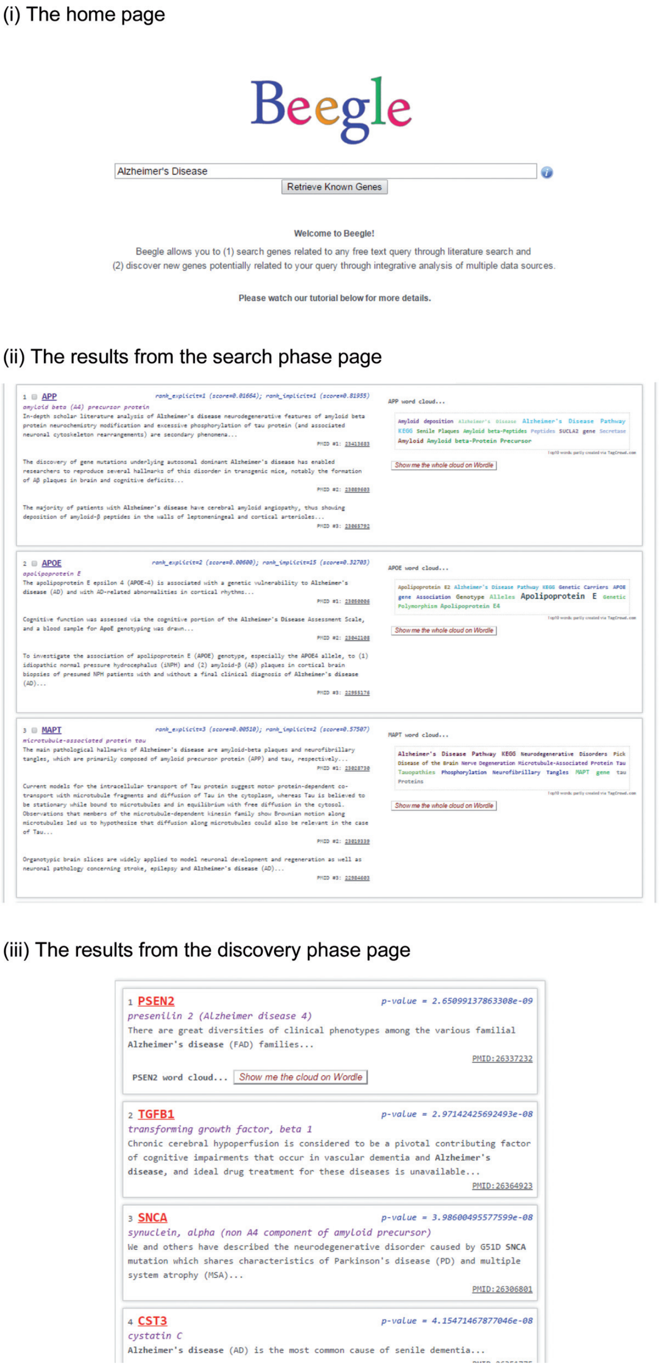
The literature-based benchmark

The first benchmark we apply to mimic novel discovery is an existing validation set (10), which was previously used to compare the predictive performance of several publicly available prioritization tools. Briefly, it was manually prepared by reviewing the scientific literature to gather novel disease-gene associations. For more details about the construction of this validation set, we refer the reader to our previous work (10). This set is composed of 34 queries and 42 annotations with at least one novel gene reported in 2010. In this benchmark, the training sets used for prioritization were manually extracted from the literature and dedicated databases.

The OMIM benchmark

OMIM provides a list of disease-gene annotations (6377 annotations in July 2013) based on experimental evidence. The annotation list is a combination of disease-gene entries that contains both confirmed and non-confirmed entries, as well as different mapping evidence. Note that each entry has a list of gene symbols, which includes both official symbols and aliases. Furthermore, many OMIM entries refer to the same disease concept. We refine this list in five steps:

(i) We remove non-confirmed entries.
(ii) We keep only the annotations whose evidence is based on mutations that are located within genes.
(iii) We keep only official gene symbols.
(iv) We combine disease entries that refer to the same disease concept.


(i) The home page

Alzheimer's Disease

Welcome to Beegle!

Beegle allows you to (1) search genes related to any free text query through literature search and
(2) discover new genes potentially related to your query through integrative analysis of multiple data sources.

Please watch our tutorial below for more details.

(ii) The results from the search phase page

1 □ APP rank_explicit=1 (score=0.01664); rank_implicit=1 (score=0.81955)
amyloid beta (A4) precursor protein
In-depth scholar literature analysis of Alzheimer's disease neurodegenerative features of amyloid beta protein neurochemistry modification and excessive phosphorylation of tau protein (and associated neuronal cytoskeleton rearrangements) are secondary phenomena...
PMID #1: 23413683

The discovery of gene mutations underlying autosomal dominant Alzheimer's disease has enabled researchers to reproduce several hallmarks of this disorder in transgenic mice, notably the formation of Aβ plaques in brain and cognitive deficits...
PMID #2: 23089603

The majority of patients with Alzheimer's disease have cerebral amyloid angiopathy, thus showing deposition of amyloid-β peptides in the walls of leptomeningeal and cortical arterioles...
PMID #3: 23065792

APP word cloud...

Amyloid deposition Alzheimer's Disease Alzheimer's Disease Pathway KEGG Senile Plaques Amyloid beta-Peptides Peptides SUCLA2 gene Secretase Amyloid Amyloid beta-Protein Precursor
Top10 words: partly created via TagCrowd.com
Show me the whole cloud on Wordle

2 □ APOE rank_explicit=2 (score=0.00600); rank_implicit=15 (score=0.32703)
apolipoprotein E
The apolipoprotein E epsilon 4 (APOE-4) is associated with a genetic vulnerability to Alzheimer's disease (AD) and with AD-related abnormalities in cortical rhythms...
PMID #1: 23050006

Cognitive function was assessed via the cognitive portion of the Alzheimer's Disease Assessment Scale, and a blood sample for ApoE genotyping was drawn...
PMID #2: 23042108

To investigate the association of apolipoprotein E (APOE) genotype, especially the APOE4 allele, to (1) idiopathic normal pressure hydrocephalus (iNPH) and (2) amyloid-β (Aβ) plaques in cortical brain biopsies of presumed NPH patients with and without a final clinical diagnosis of Alzheimer's disease (AD)...
PMID #3: 22955176

APOE word cloud...

Apolipoprotein E2 Alzheimer's Disease Pathway KEGG Genetic Carriers APOE gene Association Genotype Alleles Apolipoprotein E Genetic Polymorphism Apolipoprotein E4
Top10 words: partly created via TagCrowd.com
Show me the whole cloud on Wordle

3 □ MAPT rank_explicit=3 (score=0.00510); rank_implicit=2 (score=0.57507)
microtubule-associated protein tau
The main pathological hallmarks of Alzheimer's disease are amyloid-beta plaques and neurofibrillary tangles, which are primarily composed of amyloid precursor protein (APP) and tau, respectively...
PMID #1: 23028730

Current models for the intracellular transport of Tau protein suggest motor protein-dependent co-transport with microtubule fragments and diffusion of Tau in the cytoplasm, whereas Tau is believed to be stationary while bound to microtubules and in equilibrium with free diffusion in the cytosol. Observations that members of the microtubule-dependent kinesin family show Brownian motion along microtubules led us to hypothesize that diffusion along microtubules could also be relevant in the case of Tau...
PMID #2: 23019339

Organotypic brain slices are widely applied to model neuronal development and regeneration as well as neuronal pathology concerning stroke, epilepsy and Alzheimer's disease (AD)...
PMID #3: 22984603

MAPT word cloud...

Alzheimer's Disease Pathway KEGG Neurodegenerative Disorders Pick Disease of the Brain Nerve Degeneration Microtubule-Associated Protein Tau Tauopathies Phosphorylation Neurofibrillary Tangles MAPT gene tau Proteins
Top10 words: partly created via TagCrowd.com
Show me the whole cloud on Wordle

(iii) The results from the discovery phase page

1 PSEN2 p-value = 2.65099137863308e-09
presenilin 2 (Alzheimer disease 4)
There are great diversities of clinical phenotypes among the various familial Alzheimer's disease (FAD) families...
PMID:26337232

PSEN2 word cloud...

2 TGFB1 p-value = 2.97142425692493e-08
transforming growth factor, beta 1
Chronic cerebral hypoperfusion is considered to be a pivotal contributing factor of cognitive impairments that occur in vascular dementia and Alzheimer's disease, and ideal drug treatment for these diseases is unavailable...
PMID:26364923

3 SNCA p-value = 3.98600495577599e-08
synuclein, alpha (non A4 component of amyloid precursor)
We and others have described the neurodegenerative disorder caused by G51D SNCA mutation which shares characteristics of Parkinson's disease (PD) and multiple system atrophy (MSA)...
PMID:26306801

4 CST3 p-value = 4.15471467877046e-08
cystatin C
Alzheimer's disease (AD) is the most common cause of senile dementia...

Figure 2. The main three screens that *Beegle* returns on the web interface. These correspond to (i) the home page, (ii) the results from the search phase and (iii) the results from the discovery phase.

(v) We keep only disease entries that have at least three genes annotated.

This results in a refined list of disease-gene annotations (314 diseases and 2654 annotations) based on the OMIM database (version 2013). We call this list the OMIM-search validation set.

To generate a benchmark that allows a validation that mimics novel discovery, we used two versions of OMIM

(2010 and 2013) as follows: for both versions, we refined the list as described above. We then compared the output of Step 1 between both versions and discarded the diseases that did not appear in both lists. We also discarded diseases for which we could not find at least one 'novel' gene (i.e. reported in the 2013 version and not mentioned in the 2010 one). Finally, to avoid false positives, we verified the resulting entries by manually looking into the scientific literature. Our aim was to make sure the annotated genes were assigned to their correct disease concepts, and have indeed been reported in the correct period. This process resulted in a final OMIM list of 104 diseases that had annotations in both the 2010 and 2013 versions, and for which we had at least one novel gene. This corresponds to a total of 959 annotations reported in 2010, and a total of 277 annotations newly reported after 2010. We call these lists the OMIM-discovery validation set. The interest of this setup is that for a version of *Beegle* limited to using information obtained prior to 2010, the discoveries made in the 2010–2013 period will serve as a prospective validation of the tool.

Since we consider both the OMIM-search and the OMIM-discovery validation sets a secondary contribution of this work, we release the full lists in Supplementary Data 3–6. Hence, other researchers can use them for evaluating different gene prioritization methodologies. Table 1 provides a summary of the benchmark data sets.

The evaluation setup

The goal of the empirical evaluation is to (i) assess the quality of the text mining approaches in terms of retrieving known gene associations, and (ii) evaluate the ability of the genes retrieved to serve as training sets for building the Endeavour models and proposing novel hypothesis.

To address the first question, we employed the following methodology. First, we used the OMIM-search validation set to measure the percentage of recall in the top genes returned by *Beegle* in the search phase. Here, recall corresponds to the number of true positive genes retrieved at a certain rank threshold. For each query, we calculated the recall in the top 10, 25, 50 and 100 ranked genes. We then calculated the average recall over all disease queries. Note that our lists of ranked genes are based on the 2013 MEDLINE release. Second we compared *Beegle* to *MeSHOP*, which is a similar text mining tool. *MeSHOP* uses concept profile similarity to rank a list of human genes given an input MeSH term (22). We chose *MeSHOP* since their current results are also based on the 2013 PubMed release. We compared the two systems using a subset of the OMIM benchmark, such that each query returned a reasonable number of genes (ranging between 3 and 30). Since *MeSHOP* is restricted to using MeSH terms as queries, we further limited our subset to diseases where we could find equivalent MeSH terms. This resulted in 18 disease queries, which we provide in Supplementary Table S3. We call this the OMIM-comparison set. For every query in this set, we used each system to generate a gene ranking. Then, for each system and query, we computed the recall in the top 10, 25, 50 and 100 ranked genes (according to the corresponding OMIM genes reported in 2013). Finally, we computed both *Beegle* to *MeSHOP's* average recall over all queries for a specific rank threshold.

For the second evaluation, we used the literature-based validation set in one experiment, and then we used the OMIM-discovery validation set in a second one. For the literature-based set, we conducted the experiment as follows. We trained and compared three *Endeavour* models: one using the manually selected set of input genes, one using the top-10 genes retrieved by *Beegle,* and one using the top-*n* genes, where n is the number of genes in the query's manual training set. We chose the candidate set to be the whole genome. Given the results, we compared the recall, when considering the top 5%, 10% and 30% of the prioritized genes. We used these thresholds because they were previously used to compare *Endeavour* and other prioritization tools (10). On average, the thresholds correspond to the top 1000, 2000 and 6000 ranked genes. For the OMIM-discovery set, we ran *Endeavour* once with the set of OMIM genes reported until 2010, once with the top-10 genes retrieved by *Beegle*, and once with the top-*n* genes (where *n* is the number of OMIM genes for a given query). Again, we used the whole genome as our candidate set. Then we compared the recall, and the average AUC (Area Under the ROC Curve) results using each training set. For the ROC curves, we defined a number of rank thresholds (starting from 10 until 22000), then for each query we measured the TPR (True Positive Rate) and the FPR (False Positive Rate) at each threshold. Afterwards, we computed the average TPR and FPR results at each threshold given all the queries, and we used these average values to plot the ROC curves. Note that in both experiments, the genes retrieved by *Beegle* and the final prioritizations generated by *Endeavour* were based on literature and genomic data sources before 2010. The two validation sets are based on disease-gene associations reported from 2010 onwards. Thus our prioritizations were not contaminated by novel information.

RESULTS

Evaluating the search phase

We present the results of the OMIM-search set in Table 2. We observe that using co-occurrence alone results in an average recall of 54% in the top 10 versus 73% in the top 100. Similarly, using concept profile similarity alone results in an average recall of 48% in the top 10 versus 68% in the top 100. However, we observe that using the best rank achieves the best recall of 56% in our top 10 and 78% in our top 100 ranked genes. The same table also reports the number of genes that are confirmed by both approaches (separately and on average) in the top 10, 25, 50 and 100 ranked genes. We observe a 41% intersection in the top 10 ranked genes versus 21.7% at the top 100. In addition, we report the average recall at different *P*-value levels using co-occurrence in Table 3. We observe similar results to measuring the recall at different rank levels as reported in Table 2.

Table 4 compares *Beegle* and *MeSHOP* on the OMIM-comparison set. We observe that *Beegle* results in an average recall of 69% in the top 10 and 84% in the top 100. Similarly, *MeSHOP* results in an average recall of 51% in the top 10 and 63% in the top 100. Supplementary Data 7 and 8 contain the full lists of returned genes by each tool.

Table 1. A summary of the OMIM benchmarks

	No. of diseases	No. of genes	No. of disease-gene pairs
OMIM-search set	314	2055	2654
OMIM-discovery set 2010	104	859	959
OMIM-discovery set 2013	104	1107	1236
OMIM-discovery set 2013–2010	104	265	277

A summary of the different OMIM benchmarks that we used in our evaluation mentioning the counts of covered diseases, genes and disease-gene pairs. We used two versions (2010 and 2013).

Table 2. The results of the OMIM-search set

	Top 10	Top 25	Top 50	Top 100
Co-occurrence	54%	64%	69%	73%
Concept profile similarity	48%	57%	62%	68%
Best rank	**56%**	**67%**	**74%**	**78%**
The average number of confirmed genes by both approaches	*4.1*	*8.1*	*13.3*	*21.7*

A comparison of the average recall in the top 10, 25, 50 and 100 returned genes using co-occurrence, concept profile similarity and best rank. The best rank returns the best recall results. We also present the average number of confirmed genes by both co-occurrence and concept profile similarity. We observe higher intersection in the top10 versus top100.

Table 3. The results of the OMIM-search set at different *P*-value levels using co-occurrence

The *P*-value level	0.0000001	0.001	0.005	0.01	0.1	0.5	1
The average rank level	24	44	55	60	104	158	16493
The average recall	0.6316	0.6944	0.7091	0.7139	0.7451	0.7678	1.0000

The average recall at different *P*-value levels using co-occurrence. We observe similar results to measuring the recall at different rank levels.

Table 4. The results on the OMIM-comparison set

	Top 10	Top 25	Top 50	Top 100
Beegle	**69%**	**80%**	**83%**	**84%**
MeSHOP	51%	60%	62%	63%

A comparison of the average recall in the top 10, 25, 50 and 100 returned genes using *Beegle* and *MeSHOP*. *Beegle* obtains better recall results, where it improves *MeSHOP* by 18% in the top 10 returned genes.

Evaluating the discovery phase

Table 5 presents the results of the literature-based set. Using the top-10 genes retrieved by *Beegle* as an input set to train the *Endeavour* models results in an average recall of 41.2%, 48.5% and 77.5% in the top 5%, 10% and 30% prioritized genes, and using the top-*n* genes similarly results in an average recall of 33.3%, 46.6% and 73.0%. In comparison, using manually extracted input sets result in recalls of 28.6%, 38.1% and 71.4%. Hence *Beegle*'s automatic input set can improve the recall by up to 12.6%, 10.4% and 6.1% in at the top 5%, 10% and 30% prioritized genes, respectively.

Table 6 presents the recall results for the OMIM-discovery set. The results are comparable when using *Beegle*'s top retrieved genes and OMIM-reported genes as input sets. This corresponds to an average recall of 34.3%, 43.3% and 65.7% in the top 5%, 10% and 30% prioritized genes. Figure 3 shows the ROC curves comparing all input sets. On average, the AUC is 0.73.

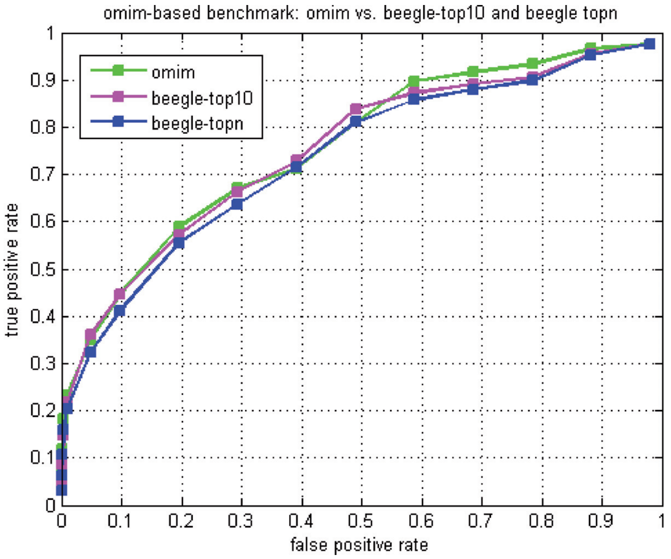


Figure 3. The ROC curves for the final gene prioritizations on the OMIM-discovery set. The green curve corresponds to OMIM 2010 derived training sets, the magenta curve corresponds to *Beegle's* top-10 training sets, and the blue curve corresponds to *Beegle's* top-*n* training sets. The results are comparable.

DISCUSSION

Our results show the potential of *Beegle* to annotate genes to diseases starting from the literature. On the one hand, the results from the search phase show that counting on ei-

Table 5. The results on the literature-based set

Input set	Recall in top 5%	Recall in top 10%	Recall in top 30%
Manually-extracted	28.6%	38.1%	71.4%
Beegle's top-10	**41.2%**	**48.5%**	**77.5%**
Beegle's top-n	33.3%	46.6%	73.0%

A comparison of the average recall results for the final gene prioritizations (using the literature-based set) in the top 5%, 10% and 30% prioritized genes using the manually-extracted, *Beegle*-extracted top-10, and *Beegle*-extracted top-*n* input sets. The automatic input set from *Beegle* improves the recall results (with slight improvement using top-10).

Table 6. The results on the OMIM-discovery set

Input set	Recall in top 5%	Recall in top 10%	Recall in top 30%
OMIM-reported	35%	45%	67%
Beegle's top-10	**36%**	**44%**	**66%**
Beegle's top-*n*	32%	41%	64%

A comparison of the average recall results for the final gene prioritizations on the OMIM-discovery set for the top 5%, 10% and 30% of the prioritized genes using the manually-extracted, *Beegle*-extracted top-10, and *Beegle*-extracted top-n input sets. The automatic input set from *Beegle* shows comparable results to the one extracted from OMIM (with slight improvement using top-10).

ther an explicit signal (that comes directly from the text), or an implicit one (that is interpreted from the fraction of shared concepts), to annotate genes to diseases tend to work successfully. They also show that combining both the explicit and the implicit signals results in a stronger retrieval as demonstrated by the increased number of experimentally-validated genes that appear in the top ranked genes. In addition, we found that *Beegle* improves the recall by an average of 18% and 21% at the top 10 and 100 returned genes when compared to *MeSHOP*. On the other hand, the results from the discovery phase demonstrated that *Beegle* can automatically generate an interesting training set to build models for predicting novel genes. They also show that using the top-10 returned genes as a training set slightly improves the performance relative to using the top-*n* genes. While we believe that expert input remains invaluable, we were surprised to observe that the results of automatic retrieval are at least as good as those for manually curated gene sets. We do however believe that additional review of potential training genes identified by *Beegle* (and the addition of any important gene missed during the retrieval phase) will further enhance the performance of the approach significantly (although this is difficult to quantify in a benchmark).

Beegle allows experts to find existing gene associations and predict new ones in an easy and straightforward manner. First, through the free-text query support, users can try any combination of biomedical concepts of interest, for which they can explore gene associations. Second, in the search phase, for every gene returned, *Beegle* presents two additional outputs: (i) at least one piece of literature in which the gene and the query are reported together, and (ii) a word cloud that views the concepts that most describe the gene in comparison to those of the query. These extra outputs provide an additional level of insight through which users can further assess the query-gene association and decide whether to add a given gene to the training set or not. Finally, in the discovery phase, in addition to the global rank, *Beegle* presents a detailed rank of the candidate genes according to the different genomic sources employed. Hence users are provided with additional insights that help them to assess the viability of the candidate gene to further decide whether or not to take it to the next step (e.g. for wetlab experiments).

Methodologically, *Beegle* mines the abstracts on MEDLINE to retrieve a ranked list of known genes using two text mining techniques. The first measures co-occurrence and the second measures concept profile similarity between genes and biomedical concepts. This is novel compared to previous research that has focused on using one of the two techniques in isolation to generate biomedical associations (17–22). *CoPub* (21) and *MeSHOP* (22) are two examples of such methodology. *Beegle* separates the search for known genes from the search for possible candidate genes. This is different from existing work that merge known and unknown gene associations in the same ranking (21–24). *BITOLA* (23) and *Genie* (24) are two examples of such methodology. Furthermore, *Beegle* supports the search using any free text query, which among existing systems is only possible in *Genie*. The rest of the existing tools are limited to specific vocabularies (e.g. MeSH terms). Supplementary Table S4 conceptually compares *Beegle* and four of the most closely related systems: *CoPub, MeSHOP, BITOLA* and *Genie*.

Nevertheless, *Beegle* is limited in the following ways. On the one hand, the user of *Beegle* is not allowed to choose the genomic sources which are used in the discovery phase and has to follow our preselection of sources (which proved to work best in our experiments). Also the user is expected to manually add training genes only in the form of gene symbols (that have a corresponding Entrez id). On the other hand, the response time of *Endeavour* is relatively slow and it can take up to 20 min to prioritize the whole genome. This is not optimal when our users expect an instantaneous response time, based on their experience with other search engines, such as Google for example.

For future work, we plan to enhance *Beegle* as follows. One way would be by improving the identification of known genes, which would be possible through applying enhanced text mining approaches. For instance, one approach could be to use a refined vocabulary set to generate the concept profiles for our queries and genes. This could be achieved by selecting high-quality concepts and discarding confusing ones. Another way would be to develop even better validation sets to measure the quality of the gene prioritizations. In this work, most of our control diseases are linked to just one future gene. We thus believe that a more extensive set with better gene coverage will give us a better insight into the performance of our tool. We also plan to integrate *Beegle* with variant prioritization tools (that are complementary to gene prioritization tools), such as *eXtasy* (31). We also plan to enhance the web interface, which is possible through (i) adding user accounts support (for managing personal queries, gene lists, etc.) and (ii) improving the response time by using a compact version of our data sets (e.g. compacting the vocabulary).

AVAILABILITY

Beegle is publicly available at:
http://beegle.esat.kuleuven.be/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Research Council KU Leuven [CoE PFV/10/016 SymBioSys to Y.M. and OT/11/051 to J.D.]; 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.]; IMEC mandaat [Ph.D. mandaat to A.A.]. Funding for open access charge: Research Council KU Leuven.

Conflict of interest statement. None declared.

REFERENCES

1. Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, **31**, 316–319.
2. Moody,S.E., Boehm,J.S., Barbie,D.A. and Hahn,W.C. (2010) Functional genomics and cancer drug target discovery. *Curr. Opin. Mol. Ther.*, **12**, 284–293.
3. He,M., Xu,M., Zhang,B., Liang,J., Chen,P., Lee,J.Y., Johnson,T.A., Li,H., Yang,X., Dai,J. *et al.* (2014) Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. Hum Mol Genet , **24**, 1791.
4. Qiu,Y.H., Deng,F.Y., Li,M.J. and Lei,S.F. (2014) Identification of novel risk genes associated with type 1 diabetes mellitus using a genome-wide gene-based association analysis. *J. Diabetes Investig.*, **5**, 649–656.
5. Penttilä,S., Jokela,M., Bouquin,H., Saukkonen,A.M., Toivanen,J. and Udd,B. (2014) Late-Onset spinal motor neuronopathy is caused by mutation in CHCHD10. *Ann. Neurol.*, **77**, 163–172.
6. Tranchevent,L.C., Capdevila,F.B., Nitsch,D., De Moor,B., De Causmaecker,P. and Moreau,Y. (2011) A guide to web tools to prioritize candidate genes. *Brief. Bioinform.*, **11**, 1–11.
7. Doncheva,N.T., Kacprowski,T. and Albrecht,M. (2012) Recent approaches to the prioritization of candidate disease genes. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **4**, 429–442.
8. Piro,R.M. and Di Cunto,F. (2012) Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J.*, **279**, 678–696.
9. Moreau,Y. and Tranchevent,L.C. (2012) Computational tools for prioritizing candidate genes: boosting disease-gene discovery. *Nat. Rev. Genet.*, **13**, 523–536.
10. Börnigen,D., Tranchevent,L.C., Bonachela-Capdevil,F., Devriendt,K., De Moor,B., De Causmaecker,P. and Moreau,Y. (2012) An unbiased evaluation of gene prioritization tools. *Bioinformatics.*, **28**, 3081–3088.
11. Tranchevent,L.C., Barriot,R., Yu,S., Van Vooren,S., Van Loo,P., Coessens,B., De Moor,B., Aerts,S. and Moreau,Y. (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res.*, **36**: W377–W384.
12. Seelow,D., Schwarz,J.M. and Schuelke,M. (2008) GeneDistiller–distilling candidate genes from linkage intervals. *PLoS One*, **3**, e3874.
13. Chen,J., Bardes,E.E., Aronow,B.J. and Jegga,A.G. (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. , **37**, W305.
14. Amberger,J., Bocchini,C. and Hamosh,A. (2011) A new face and new challenges for Online Mendelian Inheritance in Man (OMIM®). *Hum. Mutat.*, **32**, 564–567.
15. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The Genetic Association Database. *Nat. Genet.*, **36**, 431–432.
16. Jensen,J.L., Saric,J. and Bork,P. (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, **7**, 119–129.
17. Jelier,R., Jenster,G., Dorssers,L.C., Wouters,B.J., Hendriksen,P.J., Mons,B., Delwel,R. and Kors,J.A. (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.
18. Jelier,R., Schuemie,M.J., Roes,P.J., van Mulligen,E.M. and Kors,J.A. (2008) Literature-based concept profiles for gene annotation: the issue of weighting. *Int. J. Med. Inform.*, **77**, 354–362.
19. Jelier,R., Schuemie,M.J., Veldhoven,A., Dorssers,L.C., Jenster,G. and Kors,J.A. (2008) Anni 2.0: a multipurpose text-mining tool for the life sciences. *Genome Biol.*, **9**, R96.
20. Glenisson,P., Coessens,B., Van Vooren,S., Mathys,J., Moreau,Y. and De Moor,B. (2004) TXTGate: profiling gene groups with text-based information. *Genome Biol.*, **5**, R43.
21. Fleuren,W.W., Verhoeven,S., Frijters,R., Heupers,B., Polman,J., van Schaik,R., de Vlieg,J. and Alkema,W. (2011) CoPub update: CoPub 5.0 a text mining system to answer biological questions. *Nucleic Acids Res.*, **39**, 450.
22. Cheung,W.A., Ouellette,B.F. and Wasserman,W.W. (2012) Inferring novel gene-disease associations using medical subject heading over-representation profiles. *Genome Med.*, **4**, 75.
23. Hristovski,D., Friedman,C., Rindflesch,T.C. and Peterlin,B. (2006) Exploiting semantic relations for literature-based discovery. AMIA Annu. Symp. Proc. , 349–353.
24. Fontaine,J.F., Priller,F., Barbosa-Silva,A. and Andrade-Navarro,M.A. (2011) Genie: literature-based gene prioritization at multi genomic scale. *Nucl. Acids Res.*, **39**, W455–W461
25. Aronson,A.R. and Lang,F.M. (2010) An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.*, **17**, 229–236.
26. Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.
27. Manning,C.D., Raghavan,P. and Schutze,H. (2008) *Introduction to Information Retrieval*. Cambridge University Press, 100–123.
28. Aerts,S., Lambrechts,D., Maity,S., Van Loo,P., Coessens,B., De Smet,F., Tranchevent,L.C., De Moor,B., Marynen,P., Hassan,B. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.
29. Lopez-Bigas,N. and Ouzounis,C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.*, **32**, 3108–3114.
30. Lukk,M., Kapushesky,M., Nikkilä,J., Parkinson,H., Goncalves,A., Huber,W., Ukkonen,E. and Brazma,A. (2010) A global map of human gene expression. *Nat. Biotechnol.*, **28**, 322–324.
31. Sifrim,A., Popovic,D., Tranchevent,L.C., Ardeshirdavani,A., Sakai,R., Konings,P., Vermeesch,J.R., Aerts,J., De Moor,B. and Moreau,Y. (2013) eXtasy: variant prioritization by genomic data fusion. *Nat. Methods.*, **10**, 1083–1084.