Published in Proceedings of the 4th BioCreative Challenge Evaluation Workshop vol. 1, p. 139-145
which should be cited to refer to this work

BiTeM/SIBtex group proceedings for BioCreative IV, Track 4: Gene Ontology curation

Gobeill Julien[1,2,*], Pasche Emilie[3] , Vishnyakova Dina[3] and Ruch Patrick[1]

[1] University of Applied Sciences - HEG, Library and Information Sciences Geneva, Switzerland
[2] SIBtex, SIB Swiss Institute of Bioinformatics, Geneva, Switzerland.
[3] University and Hospitals of Geneva, Division of Medical Information Sciences
Geneva, Switzerland

* Corresponding author: Tel: 0041 22 388 17 86, E-mail: julien.gobeill@hesge.ch

Abstract

For the BioCreative IV Track 4, we exploited the power of our machine learning Gene Ontology classifier, GOCat. GOCat computes similarities between an input text and already curated instances in order to infer GO terms. GO Annotations (GOA) and MEDLINE are used for populating the knowledge base (almost 100000 curated abstracts). For the subtask A, we designed a state-of-the-art statistical approach, using a naïve Bayes classifier and the official training set. We also investigated exploiting GeneRIFs for an alternative forty times bigger training set, but the results were disappointing, probably because of the lack of correct negative instances. For the subtask B, we applied GOCat to the first subtask output and reached promising results, up to 0.65 for Recall at 20 with hierarchical metrics. Thanks to BioCreative IV, we were able to design a complete workflow for curation. Given a gene name and a full text, this system is able to deliver highly relevant GO terms along with a set of evidence sentences; observed performances are sufficient for being used in a real semi-automatic curation workflow.

Introduction

The problem of data deluge in proteomics is well known: the available curated data lag behind current biological knowledge contained in the literature (1–3), and professional curators needs assistance from text mining in order to keep up with the literature (4–6). One particularly time-consuming and labor-intensive task is gene function curation of a full text with Gene Ontology (GO) terms. Such curation from literature is a highly complex task, because it needs expertise in genomics but also in the ontology itself. For that matter, this task was studied since the first BioCreative challenge in 2005 (7) and is still considered as both unachieved, and long-awaited by the community (8).

Our group participated in the first BioCreative. At this time, we extracted GO terms from full texts with EAGL, a locally developed Dictionary-Based classifier (9). Dictionary-Based approaches tend to exploit lexical similarities between the information about GO terms (descriptions and synonyms) and the input text. Such approaches are limited by the complex nature of the GO terms; identifying GO terms in text is highly challenging, as they often do not appear literally or approximately in text. Another smaller part of systems evaluated in BioCreative I relied on machine learning approaches. Such algorithms empirically learn behaviours from a knowledge base that contains training instances, i.e. instances of already curated publications. At that time, machine learning approaches produced lower results; the lack of a standard training set was notably pointed out.

We recently report on GOCat (10, 11), our new machine learning GO classifier. GOCat exploits similarities between an input text and already curated instances contained in a knowledge base to infer a functional profile. GO Annotations (GOA) and MEDLINE make now possible to exploit a growing amount of almost 100000 curated abstracts for populating this knowledge base. Evaluated on the first BioCreative benchmark, GOCat achieved performances close to human curators, with 0.65 for Recall at 20, against 0.26 for our dictionary-based system. Moreover, we showed in (11) that the quality of the GO terms predicted by GOCat continues to improve across the time, thanks to the growing number of high-quality GO terms assignments available in GOA: since 2006, GOCat performances have improved by 50%.

The BioCreative IV Track 4 was the occasion to exploit the GOCat power in a reference challenge. The subtask A aimed at evaluating system for filtering relevant sentences for GO curation, given a gene name and a full text. For this subtask, we designed a robust state-of-the-art approach, using a naïve Bayes classifier and the official training set (1346 positive sentences). We also investigated exploiting GeneRIFs for an alternative training set (76000 positive sentences). Then, the goal of the subtask B was to use these relevant sentences for assigning GO terms to the given gene. For this subtask, we submitted results computed with GOCat with different numbers of proposed GO terms.

Material and Methods

Subtask A

The goal of the subtask A was to determine, given a training set of curated sentences, whether new sentences are relevant for curation or not, and if possible to support the decision with a confidence score. Some state-of-the-art methods suitable for such supervised binary classification task include naïve Bayes classifiers and Support Vector Machines (SVM) (12,13). For implementation reasons, we chose a naïve Bayes classifier first, and finally did not investigate SVM due to a lack of time.

As we mentioned above with the GOCat description, we are used to work with statistical GO classification at the abstract/paragraph level, but we rarely apply our system at the sentence level.

Thus, for this subtask A, we further analysed the data in order to design a training set, and finally made some strong assumptions about them. First of all, we studied the length of evidence texts: as mentioned in the guidelines (14), the evidence texts for GO annotations may be derived from a single sentence, or multiple continuous, or discontinuous, sentences. In the training data, 66% of evidence texts contained only one sentence, 20% contained two sentences, 14% three and more. Hence, our first assumption was to consider only sentences: for example, a block of three positive sentences was considered as three independent positive sentences. Then, we compared, given a full text and a gene name, the set of the positive sentences, and the set of sentences where we were able to identify the gene name. For retrieving a given gene name in sentences, we relied on mapping patterns. With a simple case-insensitive mapping, we found the given gene name in 65% of the positive sentences. Then, we searched hyphens in gene names and generated a couple of variants (e.g. for "rft-1" we also tried to map "rft1"). With this rule, we reached 80%. We then investigated how to exploit the gene id in order to find supplementary synonyms and variants in reference databases, but we quickly concluded that this strategy would have brought too much noise. A further look to the data revealed that for most sentences in the 20% missed, the gene name was not explicit but often mentioned via pronouns, or such grammatical expressions that require a syntactic analysis and that is beyond statistical approaches. Hence, we accepted this limit, and our second assumption was to only consider sentences that contained the gene name. So, 80% of positive sentences contain the gene name. On the other hand, 20% of sentences that contain a given gene name are positive, 80% are negative (i.e. not positive). This was our third assumption: the training data should contain this 4:1 ratio, four negative sentences for one positive sentence. Finally, for the design of training data, we replaced all the gene names we identified by the word "genemention".

We thus were able to design training sets for our naïve Bayes classifier. For the gotaska_bitemteam_run1, we built the training set from the official training set that contained 100 curated articles. With our assumptions, we finally obtained a set of 9251 sentences containing gene names: 1346 positives and 7905 negatives. The ratio is slightly different (85% of negatives), possibly because positive sentences can apply for several enumerated genes. For the gotaska_bitemteam_run2, we added the development set (50 curated articles) to the previous training set, and thus obtained 683 supplementary positive sentences and 3912 supplementary negative sentences.

Finally, we investigated a second way for designing our training set, based on GeneRIFs. GeneRIFs are concise phrases identified in journal papers and describing a protein function, recorded in the reference databases by a curator. GeneRIFs are not GO annotations, but potentially provide positive sentences for our task. We first downloaded all available GeneRIFs (http://www.ncbi.nlm.nih.gov/gene/about-generif). In July 2013, there were approximately 826000 entries in the database. Each entry is provided with the gene ID, the GeneRIF text, and the PMID that was used. As GeneRIFs are taken in full texts, we only considered papers whose full text was available in PubMed Central (http://www.ncbi.nlm.nih.gov/pmc/). We were able to locate 76000 GeneRIFs in 48000 full texts. Thus, these 76000 GeneRIFs were considered as positive sentences. For negative sentences, we first retrieved all sentences containing the given gene names, and considerer that all non-positive sentences were negative, which is a strong assumption. We finally sampled this negative set in order to keep the 4:1 ratio between positive and negative instances. As for the first training sets, we replaced all identified gene names by "genemention". This GeneRIFs training set was used for the gotaska_bitemteam_run3.

Hence, these three training sets were used to train our naïve Bayes classifier. For each sentence, each word was considered as a feature. We also add several meta-features, such as the type of section (paragraph, title, caption…), the relative position of the sentence in the full-text (an integer between 1 and 20), the percentage of common words with the abstract, and the sentence length. Once the classifier was trained, we parsed the test set. For each article and each gene, we extracted the sentences containing the gene name. Then, each sentence was sent to the classifier and obtained a class (positive or negative) and a confidence score. As only 20% of sentences containing a given gene name were positive in the training set, we chose to return only the first 20% best ranked sentences.

Subtask B

The goal of the subtask B was to predict GO terms for a given gene in a given article. For this purpose, we used our GO classifier GOCat. GOCat relies on a *k*-Nearest Neighbors (*k*-NN), a remarkably simple algorithm which assigns to a new text the categories that are the most prevalent among the *k* most similar instances contained in the knowledge base. The GOCat knowledge base contains the nearly 100000 MEDLINE abstracts that were used for manual GO curation in the GOA database. GOCat is comprehensively described in (11).
Obviously, we discarded all the test set PMIDs from the knowledge base. Then, we started from the gotaska_bitemteam_run1. For each article and each gene name, we built a paragraph with the submitted sentences, then we sent the paragraph to GOCat. GOCat was used with *k*=100. As the *k*-NN usually outputs all possible GO terms along with a confidence score, we only kept the five most confident GO terms for gotaskb_bitemteam_run1, the ten most confident for gotaskb_bitemteam_run2, and the twenty most confident for gotaskb_bitemteam_run3.

Results and Discussion

Subtask A

Table 1 presents our results for the subtask A, computed with the official evaluation script, with two values used for the parameter (0 for partial match and 1 for exact match).

Run	Parameter	Precision	Recall	F1	Training set for Naive Bayes
gotaska_bitemteam_run1	0	0.344	0.213	0.263	Official training set
	1	0.206	0.128	0.158	
gotaska_bitemteam_run2	0	0.354	0.22	0.271	Official training and development set
	1	0.217	0.134	0.166	
gotaska_bitemteam_run3	0	0.204	0.127	0.156	GeneRIFs training set
	1	0.107	0.066	0.082	

Table 1. Official results of BiTeM SIBtex for subtask A.

The best results were obtained by the first two runs, computed with the official training and development set. The contribution of the development set in regards to performances is manifest but light: +3% for F1. These two runs were computed with a state-of-the-art statistical approach, relying on simple and strong – thus robust – assumptions, and the use of a simple binary classifier. At this stage, we don't know the others participants' results, so it is difficult to situate our performance. But we can compare the first two runs and the third one, which used GeneRIFs as training data. This third run was significantly weaker (appr. -50% for F1) while the used training set was forty times bigger. There is obviously a quality problem in the GeneRIFs training set. Its positive instances are built on the assumption that GeneRIFs are relevant sentences for GO annotation; this assumption seems *a priori* true, but maybe curators would make some distinctions between these two roles. But the weaker point seems to be the construction of the negative set. For the GeneRIFs training set, we considered that all sentences that mentioned the gene and were not positive were negative. Yet, GeneRIFs do not aim to produce an exhaustive set of evidence sentences in a paper, but only keep one sentence as evidence, while the annotation was exhaustive in the official BioCreative training set. Thus, there were 13 positive sentences per article in the BioCreative training set, against 1.6 in our GeneRIFs training set. The probability of false negatives sentences in the GeneRIFs training set is high and could mainly explain this counter-performance.

Subtask B

Table 2 presents our results for the subtask B, computed with the official evaluation script, with two values used for last parameter (0 for standard metrics and 1 for hierarchical metrics).

Once again, at this stage we do not know the other participants' results, but we can compare the GOCat performances with the performances we observed in previous studies. In (11), GOCat was evaluated on its ability to retrieve GO terms that was associated to a given PMID, without taking account of the gene. For Recall at 20 (R20), GOCat achieved performances ranging from 0.56 for new published articles to 0.65 for BioCreative I test set. These performances were obtained by using the abstract for the input text. In this subtask B, the observed R20 is 0.306. But this performance was obtained by taking account of the gene, as the input was a set of sentences dealing with a given gene, and the output was GO terms relevant for this gene. Anyway, these performances are beyond the maximum performances observed in (11) with Dictionary-Based approaches, which exploit similarities between the input text and GO terms themselves. Thanks to its knowledge base designed from real curated articles, GOCat is able to propose GO terms that do not appear literally or even approximately in text.

Run	Last parameter	Precision	Recall	F1	# GO terms returned
gotaskb_bitemteam_run1	0	0.117	0.157	0.134	5
	1	0.323	0.356	0.339	
gotaskb_bitemteam_run2	0	0.092	0.245	0.134	10
	1	0.248	0.513	0.334	
gotaskb_bitemteam_run3	0	0.057	0.306	0.096	20
	1	0.179	0.647	0.280	

Table 2. Official results of BiTeM SIBtex for subtask B.

Regarding hierarchical metrics, it is quite surprising to observe such a difference (R20 0.647, +111%), while GOCat aims at returning the GO terms that were most used by curators in GOA. Yet, this performance is remarkable, and is promising in a workflow where the curators would give the gene name and the PMID, then screen and check the proposed GO terms. In a fully automatic workflow, the best setting would be to return five GO terms. In this case, the observed F1 (0.134) still is far from human standards for strict curation, but the hierarchical F1 (0.339) seems sufficient for producing added value data. In this perspective, GOCat was used to profile PubChem bioassays (15), or within the COMBREX project to normalize functions described in free text format (16).

Conclusion

The main limit of GOCat, both observed by reviewers and mentioned in our papers, was the difficulty to integrate it in a curation workflow: it is stated that GOCat proposes more accurate GO terms, but these terms are inferred from the whole abstract, then the curator still has to locate the function in the publication and to link the correct GO term with a gene product. Thanks to BioCreative IV, we were able to design a complete workflow for curation and to evaluate it. Given a gene name and a full text, this system is able to deliver relevant GO terms along with a set of evidence sentences; observed performances are sufficient for being used in a real semi-automatic curation workflow.

Funding

This work was supported by the Swiss National Fund for Scientific Research (BiND project 3252B0-105755) and the FP7 program (Khresmoi project FP7 – 257528).

References

1. Blake,J.A. and Bult,C.J. (2006) Beyond the data deluge: data integration and bio-ontologies. J. Biomed. Inform., 39, 314–320.

2. Howe,D., Costanzo,M., Fey,P. et al. (2008) Big data: the future of biocuration. Nature, 455, 47–50.

3. Baumgartner,W., Bretonnel Cohen,K., Fox,L., Acquaah-Mensah,G. and Hunter L (2007) Manual curation is not sufficient for annotation of genomic databases. Bioinformatics 2007, 23:i41-i48.

4. Bodenreider,O. (2008) Ontologies and data integration in biomedicine: success stories and challenging issues. Data Integr. Life Sci.,5109, 1–4. doi:10.1007/978-3-540-69828-9_1.

5. Spasic,I., Ananiadou,S., McNaught,J. et al. (2005) Text mining and ontologies in biomedicine: making sense of raw text. Brief. Bioinformatics, 6, 239–251.

6. Hirschman,L., Gully,A., Krallinger,M. et al. (2008) Text mining for the biocuration workflow. Database, 2012, bas020.

7. Blaschke, C., Leon, E.A., Krallinger, M., Valencia, A. (2005) Evaluation of BioCreAtIvE assessment of task 2. BMC Bioinformatics, 6, S16.

8. Lu, Z., Hirschman, L. (2012) Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. Database (Oxford), 2012, bas043.

9. Ruch,P. (2006) Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics, 22, 658–664.

10. Gobeill,J., Pasche,E., Teodoro,D. et al. (2012) Answering Gene Ontology terms to proteomics questions by supervised macro reading in Medline. In: Proceedings of NETTAB Conference, EMBnet.journal, North America 18, Nov. 2012.

11. Gobeill,J., Pasche,E., Vishnyakova,D. and Ruch,P. (2013) Managing the data deluge: data-driven GO category assignment improves while complexity of functional annotation increases. Database (Oxford). 2013 Jul 9;2013:bat041. doi: 10.1093/database/bat041.

12. Huang,J., Lu,J. and Ling,C. (2003) Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy. In Proceedings of the Third IEEE International Conference on Data Mining (ICDM '03). IEEE Computer Society, Washington, DC, USA.

13. Colas,F. and Brazdil,P. (2006) Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks Artificial Intelligence in Theory and Practice, 169-178.

14. Van Auken,K., Schaeffer,M., McQuilton,P. et al. (2013) Corpus Construction for the BioCreative IV GO Task. BioCreative IV Proceedings.

15. Guha,R., Gobeill,J. and Ruch,P. (2009) GOAssay: from Gene Ontology to Assays IDentifiers – Towards Automatic Functional Annotation of PubChem BioAssays, Available from Nature Precedings http://dx.doi.org/10.1038/npre.2009.3176.1.

16. Anton,B.,Chang,Y.,Brown,P. et al. (2013) The COMBREX Project: Design, Methodology, and Initial Results. PLoS Biol 11(8): e1001638. doi:10.1371/journal.pbio.1001638.