Assisting Document Triage for Human Kinome Curation via Machine Learning

Yi-Yu Hsu, Chih-Hsuan Wei and Zhiyong Lu*
Affiliation: National Center for Biotechnology Information, Bethesda, MD 20894, USA
*Correspondence: zhiyong.lu@nih.gov

***Abstract*—Manual curation efforts are often tedious and have largely been limited by the need to process and integrate large numbers of biomedical literature in finite amounts of time. Document triage with automatic methods can compensate for the weakness of manual curation efforts and help provide more efficient and effective curation workflows. Machine-assisted document triage involves automatic identification of entities and relation extraction with natural language processing techniques. Here, we developed a system that can automatically predict which articles have a higher relevance for curation. In the triage task for Human Kinome Curation in BioCreative (BC) VI, we applied several machine learning methods for exploring articles with curatable knowledge, in particular the bio-concept relations among kinases, diseases, and biological processes. We used rich co-occurrence and linguistic features to assess the influence of human kinome articles from the neXtProt database. We expect this method can provide neXtProt biocurators with ranking lists for specific queries, thus facilitating the process of reviewing relevant information in the literature.**

Keywords—document triage, machine learning, relation extraction

I. Introduction

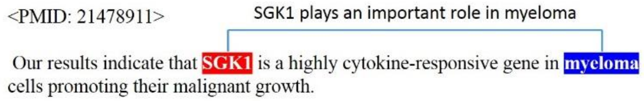
Assisting biocurators in the retrieval of relevant articles and passages for the curation of protein kinases proves to be an ongoing challenge. The BC VI Human Kinome Curation Track addresses this problem and is designed as an information retrieval task (literature triage) aimed at retrieving relevant articles for specific curation efforts (i.e. biological process (BP)/disease (DIS)). To help develop and evaluate approaches for this task, the neXtProt data (cite 25593349) is used and includes 300 proteins protein kinases. The articles in this database contain comprehensive manual annotations including gene ontologies, biological processes, and the National Cancer Institute (NCI) diseases.

This paper describes our submission to the human kinome curation track at the BC VI. This track includes three subtasks: (1) abstract triage, (2) full text triage, and (3) snippet selection. We participated in the first subtask, which is about retrieving curatable articles based on abstracts. As biocurators spend significant amounts of time surveying and reviewing articles using specific queries, a precise document triage classification could be helpful in reducing the workload of biocurators and allowing them to customize their own curation patterns (1, 2).

II. System Description

A. Data preprocessing

For the training set, the BC VI organizers provided two datasets, each including 1,615 and 1,844 pairs (<kinase, PMID>) with its associated axis, which can either be a biological process or disease. However, the datasets do not annotate which biological processes or diseases correspond to the annotated kinase. In this study, we combined the two sets (1,615 + 1,844 = 3,459) and generated triples <kinase, axis, PMID>. For instance, there is a relationship between "SGK1" and "myeloma" in Fig. 1., which would be noted as <SGK1, myeloma, 21478911> in the triple.



<PMID: 21478911>

SGK1 plays an important role in myeloma

Our results indicate that SGK1 is a highly cytokine-responsive gene in myeloma cells promoting their malignant growth.

Fig. 1 An example of positives in the training set

First, we used our Named Entity Recognizer (NER) taggers (3, 4) to recognize all kinase, disease, and biological process mentions. We filtered out the articles without kinase mentions, and narrowed our results to 2,775 triples. In order to evaluate our method, we kept 225 triples as a development set, leaving 2,550 triples for training (Table 1).

We also selected articles with 100 target proteins from 5.3 million citations, therefore creating 894,312 triples. Although there are no negative training instances in the datasets provided, we generated pseudo-negatives by using the following process. First, we used a Support Vector Machine (SVM)-one class classifier to train on 2,550 triples and test on 894,312 triples, and then selected the lowest 2,500 scores as our negative training instances. Note that we now have a positive set (2,550 triples) and a negative set (2,500 triples). We then trained our models on different classifiers described in the methods section using both the 2,550 positive set and 2,500 negative set. After the models were built, we added the 225 triples to the 894,312 triples to evaluate the ranking scores of the development set. Fig. 2 shows the workflow of our proposed document classification system for human kinome curation using machine learning.

Table 1. Statistics of the training set

	Official Training set	Utilized	
		Training set	Development set
# Triple	3,459	2,550	255
# PMID	3,018	2,282	221

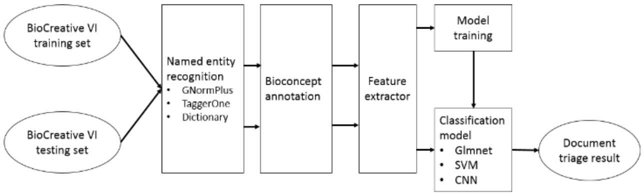


Fig. 2 The workflow of our human kinome curation system

B. *Methods*

Our submission to the abstract triage task utilizes several machine learning methods including lasso (least absolute shrinkage and selection operator) and Elastic-Net Regularized Generalized Linear Models (Glmnet) (5), SVM (6), and Convolutional Neural Network (CNN) (7). As mentioned, we applied our bio-concept taggers (TaggerOne (3) and GNormPlus (4)) to recognize all disease and kinase mentions, and then built a dictionary-based tagger to annotate biological process mentions. The performance of bio-concept recognition on the training set is shown in Table 2. Overall, the performance of the target kinase recognition is lower than the performance of the bio-concept recognition. When the recognition rate of DIS and BP mentions achieves over 98%, the recognition rate of the target kinase achieves less than 80% in recall. In our observation, 60% of missed target kinases are not mentioned in the abstract. The input data for the machine learning based models includes only the title, abstract, and bio-concept annotations of the taggers. Additionally, our methods did not distinguish between the data for disease and biological process mentions. The texts of the two types were trained together by using the same features.

Table 2. The NER results of target kinase, DIS, and BP

Mentions	Set	Method	Recall	Found	All
DIS	DIS_train.qrel	TaggerOne	0.9807	1375	1402
BP	BP_train.qrel	Dictionary	0.9975	1612	1616
Target Kinase	DIS_train.qrel	GNormPlus	0.7879	1304	1655
	BP_train.qrel		0.6302	1162	1844

- *Glmnet*

The features of large datasets would suffer from a curse of dimensionality, and they usually generate large sparse data matrices. To reduce the high dimensional features, Glmnet is a widely used algorithm for fitting various probability distributions in statistical computing and machine learning. When analyzing the high dimensional data, Glmnet uses the lasso or the elastic net to interpret and fetch important features with efficient computation. Therefore, Glmnet increased in stability and made predictions with a path of penalty parameters.

- *SVM*

A Support Vector Machine (SVM) is a robust machine learning algorithm for classification analysis. The SVM has been applied to many classification problems related to supervised learning with multidimensional data. After the SVM classifier is built, the model can correctly determine the hyperplane, which separates the data into different classes.

- One-class classification: this model aims to find the support vectors of the one-class training set, and allows for outlier/novelty detection (8). The goal is to distinguish new data as either similar or different from the normal training set.

- Binary classification: The original SVM is designed for determining the optimal separating hyperplane between the two groups. In practice, the SVM project samples on a higher dimensional space to approach the optimal hyperplane with less empirical classification errors (6).

- *CNN*

A Convolutional Neural Network (CNN) is derived from deep artificial neural networks that consist of receptive fields, local connectivity, and shared weights (7). The CNN has been well known for its excellent performance on image recognition. In this work, we trained a simple CNN with two layers of convolution on top of word vectors obtained from an unsupervised learning algorithm (9). We then designed and aligned the CNN with different parameters including an input layer, convolution layer, pooling layer, fully connected layer, and output layer.

- *Features*

The following features are applied to all the methods, as shown in Table 3. The features can be grouped into three categories: A) frequency features (feature 1-2): calculated the number of kinase and axis mentions in each abstract. B) location features (feature 3-7): the location of kinase and axis is detected. C) natural language processing (NLP) features (feature 8-11): kinase key words include a list of keyword groups, which is shown in Table 4. Each group includes manually generated key words of the genetic disease field. Furthermore, we applied tmVar (10, 11) to recognize variation mentions in the text as an additional variation key word group. The bag of words feature includes the lemma form of words around kinase, disease, and biological process mentions in the abstracts. Parsing tree path features use the dependency relation of dependency grammars to record the syntactic structure of kinase, disease, and biological process mentions (12). All features are transformed to document-term matrices.

Table 3. Statistical and linguistic features

Feature		Type
1	Number of target kinase	Numeric
2	Number of target axis	Numeric
3	Target kinase in 1st sentence	Boolean
4	Target axis in 1st sentence	Boolean
5	Target kinase in last sentence	Boolean
6	Target axis in last sentence	Boolean
7	The same sentence	Boolean
8	Kinase key words	String
9	Bag of words	String
10	Parsing tree path	String
11	Parsing tree path w/o ancestors	String

Table 4. The keyword groups

Group	Key words
Verb	involve, enhance, inhibit, regulate, increase, associate, phosphorylate
Patient	patient, men, women
genetic	detectable, survival, genetic, tumorigenesis, overexpression, mutation, translate, transcript, change, lymphangiogenic, neurotrophic
scale	mg, kg
period	day(s), during
examine	examine, experiment, screen, role, risk, significant
Variation	recognized by tmVar

EVALUATION

Before submitting official runs, we used the following evaluation metrics to assess each of our models:

- MAP (Mean Average Precision) is the mean of the precision scores for various queries.
- We also defined an estimated score ($Escore$) for measuring the ranking result of a triple (t) including a kinase, an axis, and a PMID. Note that γ is the rank of a triple after we combined the triples of the training and development set ($|D|$). We then summarized the score $\frac{\gamma}{|D|}$ of all 225 triples from the development set. For example, if we assume there are 10 PMIDs mentioned for a target kinase and the rank of one specific PMID is the top one among all 10 PMIDs, then $\frac{\gamma}{|D|}$ is 0.1. Therefore, the lower $Escore$ represents a better performance.

$$Escore = \sum_{|t|} \frac{\gamma}{|D|}$$

The following evaluation metrics are used by organizers for official results.

- P10 is the precision at rank 10: we calculated the number of documents that are relevant among the top ten documents returned by the system. If the system returns ten documents and only four documents are relevant, the P10 is 0.4. Also, P30 and P100 are the precision at rank 30 and 100 respectively.
- R30 is the recall at rank 30: we calculated the number of relevant documents retrieved in the top 30 documents returned by the system. We assume that within each query there are only 20 relevant documents within the collection. If the system returns ten of these relevant documents, then the R30 is 0.5. Also, R100 is the recall at rank 100.
- P at R0 is the maximum precision observed (for any rank value).
- R-Prec is the precision observed at rank r, where r is the number of relevant documents in the collection. If a given query contains twenty relevant documents, R-Prec is the precision at rank 20.

As shown in Table 5, we trained different models with features described in Table 3. For the Glmnet classifiers, both BP and DIS triples are included in the training set. An SVM (binary) is the model that we applied both positives and negatives as a binary classifier, while an SVM (one class) uses only the positives to train a one-class classifier. For the CNN classifiers, we constructed multiple hidden layers between the input and output layers, and modeled complex non-linear relationships. The evaluations of different methods on training sets (including disease and biological process sets) are reported in Table 6. In the last stage, we used the entire positives in the training set and features of better performance in our evaluation for method 9 and 10. In this case, we did not have the testing triples to evaluate both models. Overall, the performance of the Glmnet classifiers is superior compared to the other two classifiers. After reviewing and optimizing the parameters in the training set, we then used the following methods and features in Table 5 as our ten submitted runs.

Table 5. Combinations of different methods and features

Method	Features	Positives of Training set	Classifier
1	9	2500	Glmnet
2	1-9	2500	Glmnet
3	1-10	2500	Glmnet
4	1-9, 11	2500	Glmnet
5	1-10	2500	SVM (Binary)
6	1-9, 11	2500	SVM (Binary)
7	1-9, 11	2500	SVM (One class)
8	1-9, 11	2500	CNN
9	1-9, 11	2775	Glmnet
10	1-9, 11	2775	CNN

Table 6. Evaluation of training set

Method	$Escore$	MAP
1	58.09	0.0401
2	49.84	0.0535
3	46.46	0.0593
4	49.85	0.0598

5	52.88	0.0460
6	57.32	0.0470
7	82.20	0.0227
8	83.68	0.0204

Table 7 and 8 demonstrate the official results of the abstracts triage (provided by the task organizers). Both tables show that Glmnet classifiers have a better performance of MAP than that of SVM and CNN classifiers, which is consistent with our observation in the training & development phases. When considering the *Escore* metrics, Glmnet classifiers are also consistent with the best performance compared to all classifiers. As for CNN classifiers, the approach sketched here fails to compete with other classifiers because when running on a small dataset an over-fitting problem develops. CNN classifiers might be able to identify more relations using larger datasets.

Table 7. Official results of kinases/diseases for the ten submitted runs.

Method	MAP	R-Prec	P at R0	P10	P30	P100	R30	R100
3	0.109	0.147	0.458	0.152	0.098	0.052	0.222	0.327
9	0.109	0.145	0.453	0.148	0.097	0.052	0.223	0.327
4	0.108	0.142	0.455	0.151	0.098	0.052	0.225	0.326
5	0.088	0.125	0.351	0.119	0.081	0.044	0.203	0.304
6	0.088	0.125	0.351	0.117	0.081	0.044	0.201	0.304
1	0.081	0.098	0.370	0.103	0.075	0.042	0.184	0.286
7	0.079	0.099	0.338	0.103	0.075	0.042	0.182	0.288
2	0.073	0.084	0.338	0.094	0.064	0.038	0.166	0.269
8	0.062	0.079	0.224	0.075	0.054	0.036	0.150	0.265
10	0.060	0.079	0.227	0.065	0.057	0.034	0.154	0.259

Table 8. Official results of kinases/biological processes for the ten submitted runs.

Method	MAP	R-Prec	P at R0	P10	P30	P100	R30	R100
3	0.195	0.182	0.450	0.176	0.121	0.065	0.399	0.563
9	0.192	0.184	0.430	0.171	0.122	0.064	0.397	0.563
4	0.191	0.178	0.437	0.171	0.122	0.064	0.396	0.561
5	0.172	0.168	0.379	0.143	0.107	0.057	0.361	0.526
6	0.170	0.169	0.378	0.140	0.107	0.057	0.361	0.524
1	0.159	0.150	0.379	0.138	0.105	0.057	0.362	0.535
2	0.155	0.141	0.373	0.137	0.104	0.056	0.346	0.529
8	0.127	0.109	0.251	0.086	0.074	0.045	0.292	0.468
7	0.119	0.109	0.242	0.101	0.077	0.046	0.285	0.457
10	0.109	0.078	0.219	0.075	0.064	0.044	0.266	0.455

- *Error analysis*

We applied various statistical and linguistic features to prioritize the abstracts with relationships between target kinase and DIS/BP mentions. However, this task is very challenging by nature. First, there are no “non-curatable” or negative documents provided in the training set. Without such negative cases, most classification methods are difficult to distinguish the curatable versus non-curatable documents. Second, the relationships between DIS/BP mentions and the target kinases are not clearly curated in the training set. For example, one abstract may have various DIS/BP mentions. Therefore, it is difficult to find the correct triples for feature extraction in our method. Third, the low recognition rate of GNormPlus missed about 20~40% of the kinases (60% of the missed kinases are not in the abstracts). Thus, about 20~40% of the relevant documents would not be found in our results. Finally, according to the official results, our best performance acquired a P100 = 0.052, R100 = 0.327 in kinases/diseases and P100 = 0.065, R100 = 0.563 in kinases/biological processes. The average number of curatable articles are 15.90 and 11.55, respectively. For each kinase with an axis, the average number of curatable articles among the entire PubMed database is extremely low. Thus, we suspect that human curators may only include articles with evidence within the experimental results section. Therefore, those excluded articles cannot be filtered by our methods which use only titles and abstracts. According to the above reasons, this task is much difficult than many other traditional document classification tasks.

CONCLUSION

In summary, we used several machine learning methods with frequency, location, and NLP features for the neXtProt triage task, which aims to specifically retrieve PubMed articles with biomedical relations among kinases, diseases, and biological processes. The average numbers of curatable articles in the testing set are rare (15.90 of target kinases/diseases articles and 11.55 of target kinases/biological processes articles). Thus, the biocurators using our methods can retrieve 32.7% (5.2 articles) and 56.3% (6.5 articles) of curatable articles among all PubMed articles with only reviewing 100 articles returned by our best method. Therefore, we believe our method can effectively accelerate the manual curation efforts used today. In our future work, we plan to examine the triage task based on full texts, and investigate a robust machine learning approach which is capable of using curatable labeled data only.

ACKNOWLEDGMENT

The authors would like to thank the organizers of the BioCreative VI Track 2. We also thank Shazia Dharssi for proofreading the manuscript and the authors of the individual open source systems used in this work. This research is funded by the National Institutes of Health, National Library of Medicine.

REFERENCES

1. Mottin, L., et al., *neXtA5: accelerating annotation of articles via automated approaches in neXtProt.* Database (Oxford), 2016. **2016**.
2. Poux, S., et al., *On expert curation and scalability:UniProtKB/Swiss-Prot as a case study.* Bioinformatics, 2017.
3. Leaman, R. and Z. Lu, *TaggerOne: joint named entity recognition and normalization with semi-Markov Models.* Bioinformatics, 2016. **32**(18): p. 2839-46.
4. Wei, C.-H., H.-Y. Kao, and Z. lu, *GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains*. Vol. 2015. 2015. 918710.
5. Friedman, J.H., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* 2010, 2010. **33**(1): p. 22.

6. Smith, L.H., W. Kim, and W.J. Wilbur, *PROBE: Periodic Random Orbiter Algorithm for Machine Learning*. 2012.

7. Zhang, Y. and B. Wallace, *A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification*. 2015.

8. Meyer, D., *Support Vector Machines The Interface to libsvm in package e1071*. Vol. 1. 2001.

9. Pennington, J., R. Socher, and C. Manning. *Glove: Global Vectors for Word Representation*. in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014. Association for Computational Linguistics.

10. Wei, C.H., et al., *tmVar: a text mining approach for extracting sequence variants in biomedical literature.* Bioinformatics, 2013. **29**(11): p. 1433-9.

11. Wei, C.H., et al., *tmVar 2.0: Integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine.* Bioinformatics, 2017.

12. Cer, D., et al., *Parsing to Stanford Dependencies: Trade-offs between Speed and Accuracy*. 2010.