Overview of BioCreative VI Kinome Track

Text-mining services for Kinome Curation

Julien Gobeill[1,2], Pascale Gaudet[1], Patrick Ruch[1,2]
[1] SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
[2] HES-SO / HEG Geneva, Information Sciences, Geneva, Switzerland

***Abstract*— The BioCreative VI Kinome Track proposed a competition to assess the effectiveness of text mining to perform literature triage, thanks to an unpublished curated dataset from the SIB Swiss Institute of Bioinformatics. This dataset contains comprehensive annotations for 300 human proteins kinases. For a given protein and a given curation axis (disease, or biological processes), participants' systems have to search relevant articles in a collection of 5.2M MEDLINE citations for the subtask 1, or 270,000 fulltext articles for the subfulltask 2. The list of explored strategies comprises Named Entity Recognition and machine learning frameworks. In particular, participants managed to derive a set of negative instances, as the databases typically do not store articles that were judged as irrelevant by curators. Final results shows significant improvements compared to the baseline established in a previous study, and compared to a basic PubMed search.**

Literature triage; protein curation; text mining

I. INTRODUCTION AND MOTIVATION

Curators play a key role to define the content and ensure the quality of the biomedical databases and to spotlight the major findings (1,2). Their mission consists of continuously collecting, verifying and annotating the literature, in order to fill reference databases. Most curation methods are based on manual approaches, which produce the most accurate knowledge, but are time-consuming (3). With the exponential growth of biomedical literature (4), biocurators need help from the text mining community in order to remain up-to-date. In particular, (5) estimates that about 7% of the curation time is assigned to the rejection of papers. (6) assumes that 15% of curators' time is spent on triage. Assisting curators in retrieving, filtering and/or prioritizing the literature can save productivity.

The **CALIPHO group** develops the neXtProt database (7-8), a flagship resource of the SIB Swiss Institute of Bioinformatics that integrates information on human proteins. The data in the neXtProt database comes from both integration of external resources and annotation within the group using an internal annotation tool, the BioEditor. In a project funded by Merck-Serono from 2011-2013, the CALIPHO group has annotated 300 human protein kinases from over 13,600 research articles, producing a data corpus of over 30,000 different annotations describing the function, substrates of the kinases, and diseases in which they have been implicated. This large data corpus is still unpublished (to be released in the upcoming months), providing a unique opportunity to use curated data to create a text mining task.

Literature triage is an Information Retrieval task; it aims at retrieving/filtering articles that are supposed to be relevant for curation. This is a basic task performed by virtually all curated molecular biology databases to initiate a curation workflow. The BioCreative VI Kinome Track proposed a competition in literature triage, thanks to the neXtProt data. Text mining groups were invited to develop and test approaches aiming at assisting database curators in the selection and ranking of relevant articles for the curation of human protein kinases. Two aspects were investigated in two subtask: abstracts triage, and fulltexts triage. All abstracts annotated in the neXtProt data are available via MEDLINE. The availability of fulltexts is more problematic, as only a tiny fraction (approximately 10%) is open access in services such as Europe PMC (Europe PMC consortium).

II. TASKS & DATA

A. *The Kinome Track dataset*

The BioCreative VI Kinome Track dataset contains comprehensive annotations about kinase substrates. It covers a significant fraction of the Human Kinome: 300 proteins out of approximately 500 human kinases. The dataset contains more than 30,000 annotations, all supported by an article (a PubMed identifier). The Kinome Track focuses on two different curation axes: diseases, and biological processes. This dataset is ready to be integrated in the neXtProt database by 2017, yet still hidden from public and participants during the competition period.

The dataset represents a total of 4,581 different articles annotated for diseases annotations, and 5,357 for biological processes ones. There is a slight overlap between both axes: only 6% of the curated articles contain an annotation for both a disease and a biological process. In total, 9,367 different articles, published in 862 different journals, are present in the dataset. Table 1 shows the 10 most represented journals in the dataset.

TABLE I. TOP 10 JOURNALS IN DATASET

Journal	# articles in dataset	Cumulative percentage
J Biol Chem	744	7.9%
Proc Natl Acad Sci U S A	314	11.3%
Cancer Res	301	14.5%
Blood	288	17.6%

Journal	# articles in dataset	Cumulative percentage
Mol Cell Biol	253	20.3%
Oncogene	228	22.7%
PLoS One	219	25.1%
J Immunol	208	27.3%
Nature	156	28.9%
Clin Cancer Res	156	30.6%

Fig. 1. The top 10 journals in the dataset, ranked by the number of articles annotated in the dataset. The cumulative percentage is computed for the whole collection (e.g. *the top 10 journals represent 30.6% of all the annotated articles in the dataset*).

B. The Kinome Track benchmark

In the Cranfield paradigm (9) for evaluation of Information Retrieval systems, benchmarks are composed of three parts: a collection of documents, a set of queries, and relevance judgements. In the Kinome Track, a query was a human kinase and an axis (biological processes, or diseases). Participants' systems had to search in the collection, and to propose a ranked list of articles that are relevant for the curation of this kinase and this axis. Systems were evaluated on their ability to propose the articles that were chosen by neXtProt curators.

a) Design of the collection: For a fair comparison, systems obviously had to search for relevant articles in a common collection. This collection had to satisfy two conditions: being small enough to be efficiently processed by all teams for the competition, but large enough to make the task realistic. For designing such a collection, we applied a journal-centric strategy. As mentioned previously, annotated articles in the dataset were published in 862 different journals (out of approximately 5,500 in MEDLINE). We chose to include all articles published in these 862 journals. We also decided to limit the collection to papers published before 2014. Thus, the final collection contained a total of 5.3M PMIDs. Unfortunately, for the fulltexts collection, only a small fraction of the PMIDs corresponded to an open-access article.

b) Queries: They were pairs made of: one of the 300 kinases (e.g. "Activin receptor type-1B" - P36896), and a curation axis (biological processes, or diseases). For each kinase, neXtProt collected synonyms were provided to the participants.

c) Relevance Judgements: For a given kinase and a given axis, all annotated articles in the dataset are considered as relevant. The rejections of articles after screening by a curator are not stored in the database. Yet, the neXtProt curation is assumed to be comprehensive. This assumption means that all potential articles were screened, and that articles that were not annotated are considered as non-relevant. Articles that could have never been screened by a curator are equally distributed among all participants runs, and thus comparisons between systems are still valid.

C. The Kinome Track subtasks

Figure 2 presents an overview of the triage process.

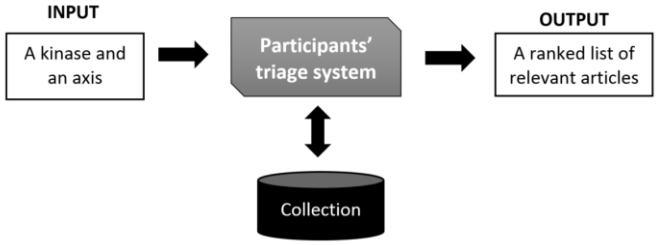

Fig. 2. Overview of literature triage for the Kinome Track. (*the collection differs depending on the subtask: abstracts, or fulltexts*).

The 300 proteins present in the dataset were randomly distributed in three different subsets: one third for the tuning set, one third for the subtask 1 test set, and one third for the subtask 2 test set. The tuning set contained 100 kinases, along with the PMIDs of the annotated articles for each axis (relevance judgements). The tuning set was made available in April 2017, thus participants were free to use it for analyses, and for tuning their system. Both test sets were delivered in May 2017. Obviously, test sets only contained the queries, while relevance judgements were kept for the official evaluation.

a) Subtask 1 – abstracts triage: This subtask focuses on abstracts triage. The collection (5.3M of articles) was given in the form of MEDLINE citations. Thus, systems had to perform triage only based on abstracts, and metadata (such as journal, publication year, publication type, etc). In the test set, for each kinase, this collection contained on average 16 relevant articles for the biological process axis, and 18 for the diseases axis. The collection was provided in BioC format.

b) Subtask 2 – fulltexts triage: This subtask focuses on fulltexts triage. The collection was given in the form of PubMed Central (PMC) fulltexts. Thus, systems had to perform triage based on fulltext contents. As only a fraction of PMC is open-access, the collection for the subtask 2 only contained 260,000 articles. Thus, in the test set, for each kinase, this collection only contained on average 2.6 relevant articles for the biological process axis, and 3.6 for the diseases axis. The collection was provided in XML format.

D. Metrics used for evaluation

TREC formats and metrics were used for evaluation (10) as follows:

- P10 or Precision at rank 10: among the top 10 articles returned by the system, how many are relevant. If the system returned 10 documents and only 4 are relevant, then P10 is 0.4. Idem for P30, P100 at ranks 30 and 100.
- R30 or Recall at rank 30: among all the relevant articles in the collection for a given query, how many are retrieved in the top 30 articles returned by the system. If for a given query there are 20 relevant documents in the

collection, and the system returns 10 of them in the top 30 documents, then R30 is 0.5. Idem for R100 at rank 100.

- P at R0: this is the maximum precision observed (for any rank value).
- Mean Average Precision (MAP): this is the average of all Precision at rank k, for ranks where a relevant article is retrieved (for no retrieved articles, 0 is counted).
- R-Prec: this is the Precision observed at rank r, where r is the number of relevant articles for a given query. If for a given query there are 20 relevant articles in the collection, R-Prec is Precision at rank 20.

III. RESULTS

More than twenty teams registered to the Kinome Track, but two finally submitted runs. In this section, we broadly describe strategies investigated by these participants. Then, we give results for both subtasks. All metrics were computed with the trec_eval reference program (http://trec.nist.gov/trec_eval/).

A. *Participants strategies*

Both participating teams exploited tuning data in order to train a machine learning system. They applied Named Entity Recognition system – such as the PubTator system (11) – in order to identify biological processes and diseases in the relevant articles, and thus obtained a set of positive pairs of kinase-concept. One team exploited these positive examples and the rest of the articles collection in order to bootstrap some pseudo negative samples. Then, both teams trained a machine learning system for triage based on several features, such as numbers and position of genes and axis.

B. *Subtask 1 – abstracts triage*

For the subtask 1, each team could submit up to ten runs. There were twenty runs for the disease axis, and nineteen for the biological process axis. Figures 3 and 4 present results for each run.

Team	Run name	MAP	R-Prec	P at R0	P10	P30	P100	R30	R100
383	DIS_team383run3	0.109	0.147	0.458	0.152	0.098	0.052	0.222	0.327
383	DIS_team383run8	0.109	0.145	0.453	0.148	0.097	0.052	0.223	0.327
383	DIS_team383run4	0.108	0.142	0.455	0.151	0.098	0.052	0.225	0.326
383	DIS_team383run5	0.088	0.125	0.351	0.119	0.081	0.044	0.203	0.304
383	DIS_team383run6	0.088	0.125	0.351	0.117	0.081	0.044	0.201	0.304
383	DIS_team383run1	0.081	0.098	0.37	0.103	0.075	0.042	0.184	0.286
383	DIS_team383run7	0.079	0.099	0.338	0.103	0.075	0.042	0.182	0.288
383	DIS_team383run2	0.073	0.084	0.338	0.094	0.064	0.038	0.166	0.269
383	DIS_team383run9	0.062	0.079	0.224	0.075	0.054	0.036	0.15	0.265
383	DIS_team383run10	0.06	0.079	0.227	0.065	0.057	0.034	0.154	0.259
392	run10_Abst_HPO_DIS_L-BOW	0.098	0.134	0.37	0.138	0.087	0.047	0.196	0.304
392	run6_Abst_NCIT_DIS_L-BOW	0.096	0.134	0.384	0.14	0.087	0.047	0.199	0.302
392	run7_Abst_NCIT_DIS_Rel	0.08	0.107	0.319	0.118	0.074	0.042	0.173	0.269
392	run26_Abst_NCIT_DIS_Ag	0.069	0.094	0.305	0.104	0.068	0.04	0.159	0.259
392	run28_Abst_HPO_DIS_Ag	0.053	0.073	0.235	0.079	0.054	0.033	0.128	0.23
392	run9_Abst_NCIT_DIS_S-ENG	0.051	0.077	0.176	0.069	0.059	0.034	0.146	0.237
392	run19_Abst_HPO_DIS_Rel	0.049	0.075	0.215	0.07	0.052	0.032	0.122	0.227
392	run11_Abst_HPO_L-ENGA	0.048	0.056	0.199	0.053	0.043	0.028	0.109	0.206
392	run8_Abst_NCIT_DIS_L-ENGA	0.044	0.059	0.153	0.06	0.044	0.028	0.117	0.209
392	run12_Abst_HPO_S-ENG	0.04	0.054	0.164	0.054	0.041	0.024	0.11	0.186

Fig. 3. Results for the abstracts triage subtask, disease axis.

Team	Run name	MAP	R-Prec	P at R0	P10	P30	P100	R30	R100
383	BP_team383run3	0.195	0.182	0.45	0.176	0.121	0.065	0.399	0.563
383	BP_team383run8	0.192	0.184	0.43	0.171	0.122	0.064	0.397	0.563
383	BP_team383run4	0.191	0.178	0.437	0.171	0.122	0.064	0.396	0.561
383	BP_team383run5	0.172	0.168	0.379	0.143	0.107	0.057	0.361	0.526
383	BP_team383run6	0.17	0.169	0.378	0.14	0.107	0.057	0.361	0.524
383	BP_team383run1	0.159	0.15	0.379	0.138	0.105	0.057	0.362	0.535
383	BP_team383run2	0.155	0.141	0.373	0.137	0.104	0.056	0.346	0.529
383	BP_team383run9	0.127	0.109	0.251	0.086	0.074	0.045	0.292	0.468
383	BP_team383run7	0.119	0.109	0.242	0.101	0.077	0.046	0.285	0.457
383	BP_team383run10	0.109	0.078	0.219	0.075	0.064	0.044	0.266	0.455
392	run1_Abst_GO_BP_L-BOWA	0.201	0.21	0.466	0.184	0.111	0.056	0.36	0.492
392	run31_Abst_GO_BP_Rel	0.197	0.189	0.433	0.17	0.117	0.06	0.379	0.529
392	run2_Abst_GO_BP_L-BOW	0.187	0.185	0.438	0.172	0.108	0.057	0.351	0.503
392	run30_Abst_GO_BP_Ag	0.181	0.172	0.383	0.158	0.112	0.058	0.372	0.519
392	run3_Abst_GO_BP_L-ENGA	0.18	0.174	0.399	0.152	0.098	0.052	0.347	0.494
392	run33_Abst_GO_BP_K	0.175	0.166	0.364	0.149	0.109	0.058	0.368	0.52
392	run32_Abst_GO_BP_Prx	0.154	0.156	0.355	0.145	0.095	0.051	0.335	0.482
392	run4_Abst_GO_BP_S-ENG	0.145	0.141	0.311	0.124	0.081	0.049	0.315	0.477
392	run34_Abst_GO_BP_A	0.13	0.115	0.293	0.092	0.062	0.035	0.272	0.39

Fig. 4. Results for the abstracts triage subtask, biological process axis.

In (12), the Text Mining group at the SIB describes the development of neXtA5, a curation service and interface, powered by different ontologies, and developed for the CALIPHO group. This system aims at assisting biocurators by prioritizing articles for the curation of a given protein and a given axis. The system was evaluated on the same dataset than in the Kinome Track. The study also compared the neXtA5 ranking with PubMed ranking (both Boolean and relevance-based). For the disease axis, the reported values P at R0 were between 0.12 and 0.13 for PubMed rankings, while they reached up to 0.41 for neXtA5. The best reported MAP was 0.04. For the biological process axis, the reported values P at R0 were 0.14 for both PubMed, and 0.45 for neXtA5. The best reported MAP was 0.11. These values can be considered as the baselines for the interpretation of results in the Kinome Track.

In terms of P at Ro values, for the disease axis, three runs (submitted by the same team) makes better than the baseline for P at R0 (+10%). For these runs, on the top 100 returned articles, 22% are relevant (P100), and they represent 33% of all available relevant articles (R100). For the biological process axis, best results observed are close to the baseline for P at R0, while the second team made slightly better. Even if we don't know yet the details of the implemented strategies, it is interesting to note that the best runs submitted by team 383 for both axes share the same names. In (Mottin 2016), separate strategies were implemented for each axis. The main advances are observed regarding the recall with a significant improvements at MAP values (+175% for disease, +80% for biological processes).

C. *Subtask 2 –triage of fulltext articles*

For the subtask 2, each team could submit up to ten runs. There were ten runs for the disease axis, and eight for the biological process axis. Figures 5 and 6 present results for each run.

Team	Run name	MAP	R-Prec	P at R0	P10	P30	P100	R30	R100
392	run27_FT_NCIT_DIS_Ag	0.118	0.111	0.278	0.065	0.028	0.011	0.233	0.293
392	run23_FT_NCIT_DIS_Rel	0.112	0.105	0.269	0.064	0.03	0.011	0.24	0.296
392	run25_FT_HPO_DIS_Rel	0.1	0.103	0.227	0.052	0.026	0.011	0.207	0.295
392	run29_FT_HPO_DIS_Ag	0.097	0.083	0.219	0.058	0.029	0.011	0.229	0.271
392	run24_FT_HPO_DIS_L-BOW	0.091	0.078	0.193	0.043	0.025	0.011	0.22	0.302
392	run22_FT_NCIT_DIS_L-BOW	0.088	0.076	0.193	0.038	0.025	0.011	0.219	0.306
392	run17_FT_HPO_DIS_L-ENGA	0.083	0.083	0.172	0.046	0.023	0.01	0.19	0.274
392	run16_FT_NCIT_DIS_S-ENG	0.041	0.028	0.076	0.018	0.013	0.007	0.124	0.212
392	run18_FT_HPO_DIS_S-ENG	0.041	0.024	0.09	0.025	0.015	0.007	0.141	0.217
392	run15_FT_NCIT_DIS_L-ENGA	0.029	0.017	0.056	0.012	0.013	0.007	0.127	0.194

Fig. 5. Results for the fulltexts triage subtask, disease axis.

Team	Run name	MAP	R-Prec	P at R0	P10	P30	P100	R30	R100
392	run35_FT_GO_BP_Ag	0.293	0.203	0.349	0.079	0.037	0.014	0.61	0.714
392	run38_FT_GO_BP_K	0.278	0.194	0.327	0.081	0.036	0.014	0.616	0.728
392	run37_FT_GO_BP_Prx	0.271	0.182	0.314	0.073	0.037	0.014	0.624	0.712
392	run40_FT_GO_BP_Prx50	0.267	0.165	0.313	0.078	0.036	0.014	0.608	0.712
392	run36_FT_GO_BP_Rel	0.264	0.169	0.312	0.075	0.037	0.014	0.631	0.724
392	run13_FT_GO_BP_L-ENGA	0.042	0.004	0.053	0.017	0.012	0.007	0.232	0.461
392	run39_FT_GO_BP_A	0.03	0.002	0.038	0.008	0.008	0.007	0.188	0.423
392	run14_FT_GO_BP_S-ENG	0.021	0	0.026	0.005	0.006	0.006	0.17	0.365

Fig. 6. Results for the fulltexts triage subtask, biological process axis.

Comparisons between abstracts and full text articles results must be made with great care. Indeed, the sizes of the collections made the direct comparison difficult (5.2M in subtask 1 versus 260,000 in subtask 2). The number of relevant articles was 16-18 per query in subtask 1 versus 2-3 per query in subtask 2. In this perspective, the best reported MAP values are higher with full text than abstracts, but the likeliness to miss some relevant articles in the subtask 1 is higher – and so to be penalized – than in subtask 2. A comparison remains possible using P at R0 since it focuses on the precision of the first relevant retrieved article. Best reported P at R0 values are higher with abstracts than fulltexts.

D. Subtask 3 – snippets extraction

A third subtask was initially considered: snippet extraction. In this subtask, the participants' system should extract from the fulltext a snippet of maximum 500 characters, which contains enough information to be "annotatable". Curators should judge snippets according to one of the three following values: 1 = Good (the snippet is sufficient for making an annotation without reading the paper); 0.5 = quite good (the curator thinks that there is a potential annotation, but needs to read the paper because the snippet is not sufficient for making the entire annotation); 0 = Irrelevant (nothing in the snippet indicates that an annotation is possible). Unfortunately, this task was cancelled due to no submitted runs.

IV. Discussion and Conclusion

Participants' papers and presentations will describe the respective methods. In particular, the impact of the bootstrapping strategy proposed by one of the competitor to acquire negative instances seems especially effective. Altogether, the precision at P0 improves slightly over (12) on the same tasks but with a different dataset, while the MAP is more significantly improved. Moreover, the triage of articles to support the curation of biological processes shows better results than for curation of diseases (+82% for best MAP values).

The low ratio of participation for registered teams is difficult to explain. The huge – yet realistic – size of the collection (5.2M of articles) could have been a serious obstacle. Moreover, data provided to participants did not contain any annotated concepts. Yet, both teams developed supervised strategies based on the potential concepts present in the articles. Thus, they had no gold standard for building their knowledge base, and had to use NER systems in order to generate a silver standard.

Further, data-driven approaches do need positive and negative examples for training. Unfortunately, curated databases usually do not store negative examples, such as articles, which are screened by the curator but not selected for curation. Hopefully, such a data stewardship "gap" should improve as FAIR principles become common practice.

Acknowledgment

This work was supported by the Swiss National Fund for Scientific Research, SNF Grant: [neXtPresso project, SNF #153437].

References

1. Mottin, L., Pasche, E., Gobeill, J., Rech de Laval, V., Gleizes, A., Michel, P. A., ... & Ruch, P. (2017). Triage by ranking to support the curation of protein interactions. Database, 2017(1).

2. Burge,S., Attwood,T.K., Bateman,A. et al. (2012) Biocurators and biocuration: surveying the 21st century challenges. Database (Oxford), 2012, bar059. doi: 10.1093/database/bar059.

3. Hirschman,L., Burns,G.A.P.C., Krallinger,M. et al. (2012) Text mining for the biocuration workflow. Database, 2012, bas020.

4. Lu,Z. (2011) PubMed and beyond: a survey of web tools for searching biomedical literature. Database, 2011, baq036.

5. Wiegers,T.C., Davis,A.P., Cohen,K.B. et al. (2009) Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD). BMC Bioinformatics, 10, 326.

6. Gaudet, P.; Michel, P. A.; Zahn-Zabal, M.; Cusin, I.; Duek, P. D.; Evalet, O.; Gateau, A.; Gleizes, A.; Pereira, M.; Teixeira, D.; Zhang, Y.; Lane, L.; Bairoch, A. The neXtProt knowledgebase on human proteins: current status. Nucleic Acids Res. 2015, 43, D764-70, doi:10.1093/nar/gku1178.

7.. Gaudet, P.; Michel, P.-A.; Zahn-Zabal, M.; Britan, A.; Cusin, I.; Domagalski, M.; Duek, P. D.; Gateau, A.; Gleizes, A.; Hinard, V.; Rech de Laval, V.; Lin, J.; Nikitin, F.; Schaeffer, M.; Teixeira, D.; Lane, L.; Bairoch, A. The neXtProt knowledgebase on human proteins: 2017 update. Nucleic Acids Res. 2017, 45, D177–D182, doi:10.1093/nar/gkw1062.

8. The Europe PMC Consortium (2014) Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic Acids Res., 43, D1042–D1048.

9.Cleverdon, C. W. (1960, December). The ASLIB Cranfield research project on the comparative efficiency of indexing systems. In Aslib Proceedings (Vol. 12, No. 12, pp. 421-431). MCB UP Ltd.

10. Manning, C., Raghavan, P., & Schütze, H. (2008). Evaluation in information retrieval. In Introduction to Information Retrieval (pp. 139-161). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511809071.009

11. Wei, C. H., Kao, H. Y., & Lu, Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. Nucleic acids research, 41(W1), W518-W522.

12. Mottin, L., Gobeill, J., Pasche, E., Michel, P. A., Cusin, I., Gaudet, P., & Ruch, P. (2016). neXtA 5: accelerating annotation of articles via

automated approaches in neXtProt. Database, 2016, baw098.Mottin et al.