Evaluation of chemical and gene/protein entity recognition systems at BioCreative V.5: the CEMP and GPRO patents tracks

Martin Pérez-Pérez[1,2], Obdulia Rabal[3], Gael Pérez-Rodríguez[1,2], Miguel Vazquez[4], Florentino Fdez-Riverola[1,2], Julen Oyarzabal[3], Alfonso Valencia[,5,6,7,8], Anália Lourenço*[1,2,9], Martin Krallinger*[4]

[1]ESEI - Department of Computer Science, University of Vigo, Ourense, Spain
[2]CINBIO - Centro de Investigaciones Biomédicas, University of Vigo, Campus Universitario Lagoas-Marcosende, 36310, Vigo, Spain
[3]Small Molecule Discovery Platform, Molecular Therapeutics Program, Center for Applied Medical Research (CIMA), University of Navarra, Pamplona, Spain
08010 Barcelona, Spain
[4]Biological Text Mining Unit, Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre, C/Melchor Fernández Almagro 3, E-28029 Madrid, Spain
[5]Life Science Department, Barcelona Supercomputing Centre (BSC-CNS). C/Jordi Girona, 29-31, E-08034 Barcelona, Spain
[6]Joint BSC-IRB-CRG Program in Computational Biology. Parc Científic de Barcelona. C/ Baldiri Reixac 10, E-08028 Barcelona, Spain
[7]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig de Lluís Companys 23, E-
[8]Spanish Bioinformatics Institute INB-ISCIII ES-ELIXIR, Madrid 28029, Spain
[9]CEB - Centre of Biological Engineering, University of Minho, Braga, Portugal

{martiperez,gaeperez,riverola,analia}@.uvigo.es; {orabal,julenoyarzabal}@unav.es;{mvazquezg,mkrallinger}@cnio.es;alfonso.valencia@bsc.es

Abstract. This paper presents the results of the BioCreative V.5 offline tasks related to the evaluation of the performance as well as assess progress made by strategies used for the automatic recognition of mentions of chemical names and gene in running text of medicinal chemistry patent abstracts.
A total of 21 teams submitted results for at least one of these tasks. The CEMP (chemical entity mention in patents) task entailed the detection of chemical named entity mentions. A total of 14 teams submitted 56 runs. The top performing team reached an F-score of 0.90 with a precision of 0.88 and a recall of 0.93. The GPRO (gene and protein related object) task focused on the detection of mentions of gene and protein related objects. The 7 participating teams (30 runs) had to detect gene/protein mentions that could be linked to at least one biological database, such as SwissProt or EntrezGene. The best F-score, recall and precision in this task were of 0.79, 0.83 and 0.77, respectively.

The CEMP and GPRO gold standard corpora included training sets of 21,000 records and test sets of 9,000 records. Similar to the previous BioCreative CHEMDNER tasks, evaluation was based on micro-averaged F-score. The *BeCalm* platform supported prediction submission and evaluation (http://www.becalm.eu).

Keywords. CEMP; GPRO; ChemNLP; BioCreative; Named Entity Recognition; Chemical compounds; Genes/proteins; Text Mining

1 Introduction

The BioCreative V.5 challenge encompassed two offline tasks, which followed the evaluation settings used for previous BioCreative competitions, in addition to a novel online task, which was geared towards the continuous evaluation of named entity annotation web servers. BioCreative is a community challenge with the aim of evaluating biomedical text mining efforts [1].

This paper describes the results obtained by participating teams for the offline tasks, which addressed the automatic extraction of chemical and biological data from medicinal chemistry patents. The CEMP (chemical entity mention in patents) and GPRO (gene and protein related object) tasks entailed the detection of chemical named entity mentions and mentions of gene and protein related objects in patent titles and abstracts, respectively.

Some of the general difficulties for such automatic name recognition in the scientific literature have been already highlighted in previous BioCreative CHEMDNER tasks [2, 3]. Indeed, the settings of the hereby described tasks were very similar to the counterparts in BioCreative V [2]. Briefly, given a set of patent documents, participating teams had to correctly detect the start and end indices corresponding to all the chemical entities (CEMP) and the gene and protein related objects (GPRO). All entities were manually annotated by domain experts using well-defined annotation guidelines [4]. In particular, the covered GPRO entities had to be annotated at a sufficient level of granularity to be able to determine whether the labelled mention could or could not be linked to a specific gene or gene product (represented by an entry of a biological annotation database such as SwissProt [5] or EntrezGene [6]).

The *BeCalm* Web metaserver platform supported prediction submission and evaluation. Participants could submit a total of five runs per task for final evaluation. The micro-averaged recall, precision and F-

score statistics were used for final prediction scoring, and F-score was selected as main evaluation metric.

2 Task description

The used patent abstract records were released in the form of plain-text, UTF8-encoded patent abstracts in a tab-separated format with the following three columns: (1) patent identifier, (2) title of the patent, (3) abstract of the patent. The annotated document sets used for training were produced with the intent of supporting the improvement of the automatic prediction tools enrolled in the challenge. Conversely, the test sets were used in the controlled comparison of the performance of the participating systems.

Table 1: CEMP and GPRO corpora overview.

	Training set	**Test set**	**Entire corpus**
Patent abstracts	21,000	9,000	30,000
CEMP mentions	99,632	44,486	144,118
GPRO mentions	17,751	8,998	26,749
GPRO type 1 mentions	12,422	5,330	17,752
GPRO type 2 mentions	5,329	3,668	8,997
Tokens	1,770,836	767,599	2,538,435

Furthermore, the annotation carried out for the GPRO task encompassed two types of GPRO entity mentions: GPRO *entity mention type 1*, i.e. covering those GPRO mentions that can be normalized to a bio-entity database record; GPRO *entity mention type 2*, i.e. covering those GPRO mentions that in principle cannot be normalized to a unique bio-entity database record (e.g. protein families or domains).

The *BeCalm* Web metaserver platform enabled both the examination of automatic predictions by participants and final submission benchmarking (Figure 1).

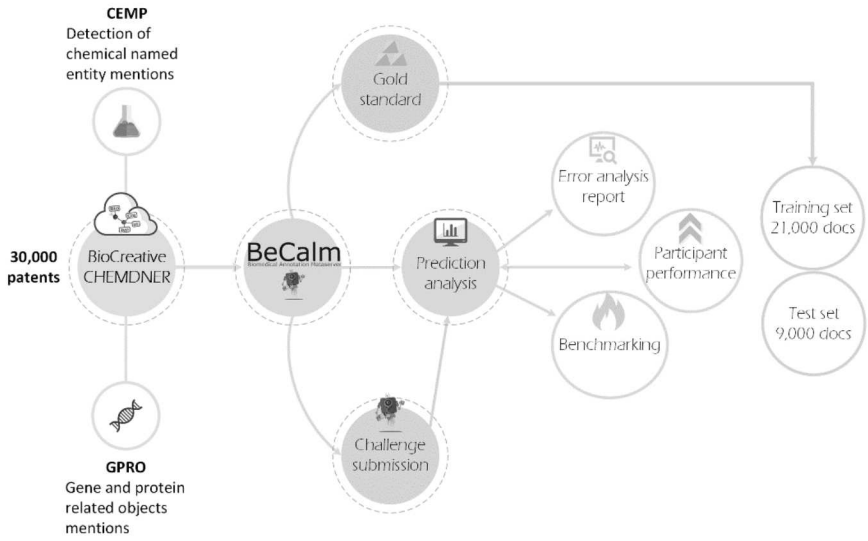

Figure 1: *BeCalm* support for chemical and gene entity recognition at BioCreative V.5 offline tasks.

BeCalm provided micro- and macro-average standard performance statistics, such as precision, recall and F-score [7, 8]. Furthermore, it enabled the examination of annotation mismatches, i.e. false positive annotations. In final evaluation, three main result types were examined: false negative (FN) results corresponding to incorrect negative predictions (i.e. cases that were part of the gold standard, but missed by the automatic system), false positive (FP) results being cases of incorrect positive predictions (i.e. wrong results predicted by the automatic system that had no corresponding annotation in the gold standard) and true positive (TP) results consisting of correct positive predictions (i.e. correct predictions matching exactly with the gold standard annotations).

Correspondingly, recall (Eq. 1) is the percentage of correctly labelled positive results over all positive cases, being a measure of the ability of a system to identify positive cases.

$$recall = \frac{TP}{(TP + FN)} \ (Eq. 1)$$

Precision (Eq. 2) represents the percentage of correctly labelled positive results over all positive labelled results, i.e. it is a measure of the reproducibility of a classifier of the positive results.

$$precision = \frac{TP}{(TP + FP)} \ (Eq. 2)$$

Lastly, F-score (or balanced F-measure) stands for the harmonic mean between precision and recall (Eq. 3).

$$F - score = 2 * \frac{(precision \times recall)}{(precision + recall)} \ (Eq. 3)$$

Partial hits, i.e. predictions that only in part overlapped with the manually defined gold standard annotations, were not taken into account in the analyses. Micro-average statistics were calculated globally by counting the total true positives, false negatives and false positives. Conversely, macro-average statistics were calculated by counting the true positives, false negatives and false positives on a per-document basis and then, averaged across documents.

During the test phase, teams were requested to generate automatic annotations (according to predefined evaluation format) for a blinded collection of documents, and submit them after a short period of time. Teams could submit for each of the tasks up to five predictions (runs). The micro-averaged recall, precision and F-score statistics were used for final prediction scoring, and F-score was selected as main evaluation metric. Furthermore, the statistical significance of each prediction with respect to the other final submissions was examined by means of a Bootstrap resampling simulation, in a similar way to what was done in previous CHEMDNER challenges [2, 3]. This statistical analysis was done for both the CEMP and GPRO tasks by taking 4,500 bootstrapped samples from all the annotated documents in the test sets (a total of 9,000 documents in each set). The micro-average F-scores for each team on each sample were calculated and these 2,500 resampled results were further used to calculate the standard deviation of the F-score of each team (SDs). Teams were grouped based on statistically significant difference (at two SD) between results.

The annotation guidelines (as well as the GPRO guidelines) were published together with the manually annotated corpora in order for teams to actually understand how the annotations were done and to make it possible to examine how their systems could consider the annotation rules.

3 Results

A total of 21 teams submitted results for at least one of the two offline tasks. For both tasks, the training set consisted of 21,000 patent records and the test set consisted of 9,000 patent records.

A total of 14 teams submitted 56 runs for the CEMP task. As illustrated in Table 2, the top performing team reached an F-score of 0.90 with a precision of 0.88 and a recall of 0.93. The top scoring run in terms of F-score was generated by team 121 (from a total of 5 runs). The three top performing teams, namely teams 121, 112 and 107, reached an F-score of over 0.90. The highest precision was obtained by team 107 (0.90) while the highest recall was obtained by team 116 (0.93).

Table 2. CEMP evaluation results (best runs per team only).

Row	Team	F-score	Precision	Recall	SD	Range	Group
A	121	90.42	88.32	92.62	0.25%	A-C	1
B	112	90.37	88.97	91.82	0.27%	A-C	1
C	107	90.32	90.02	90.62	0.27%	A-C	1
D	153	89.14	88.02	90.28	0.3%	D-E	2
E	116	88.47	84.39	92.97	0.23%	D-F	3
F	144	87.29	87.42	87.15	0.34%	E-G	4
G	102	86.59	89.01	84.29	0.33%	F-J	5
H	142	85.68	83.1	88.42	0.32%	G-J	6
I	117	85.44	88.42	82.64	0.32%	G-J	6
J	127	85.31	87.32	83.38	0.36%	G-J	6
K	135	83.95	85.68	82.28	0.37%	K	7
L	125	82.45	83.1	81.81	0.3%	L	8
M	110	59.24	52.93	67.26	0.35%	M	9
N	170	49.25	47.18	51.52	0.19%	N	10

The 7 teams that participated in the GPRO task submitted a total of 30 runs. Here, evaluation was two-fold: based only on annotations of type 1 (i.e. those that can be normalized to a bio-entity database record), and considering both annotation types (i.e. normalized or not to a bio-entity database record). For the GPRO type 1 evaluation, team 121 was the best performing team (achieved an F-score of 0.79), team 133 got the best recall (0.83) and team 144 obtained the best precision (0.77) (Table 3).

Table 3. GPRO type 1 evaluation results (best runs per team only).

Row	Team	F-score	Precision	Recall	SD	Range	Group
A	121	79.19	76.65	81.91	0.1%	A	1
B	112	76.34	75.23	77.49	0.08%	B-C	2
C	153	76.13	72.06	80.68	0.1%	B-C	2
D	133	73.73	66.53	82.68	0.1%	D	3
E	142	73.18	74.79	71.63	0.15%	E-F	4
F	144	73.07	76.86	69.62	0.17%	E-F	4
G	102	71.3	71.52	71.09	0.14%	G	5

For GPRO type 1 and type 2 evaluation, team 133 achieved top performing F-score (0.79) and recall (0.79) while team 153 obtained the best precision (0.84) (Table 4).

Table 4. GPRO type 1 and type 2 evaluation results (best runs per team only).

Row	Team	F-score	Precision	Recall	SD	Range	Group
A	133	78.66	78.63	78.7	0.05%	A	1
B	153	77.11	83.95	71.3	0.06%	B	2
C	112	75.91	80.41	71.89	0.04%	C	3
D	144	74.92	79.78	70.63	0.09%	D	4
E	121	72.28	81.56	64.89	0.12%	E	5
F	142	64.96	74.99	57.3	0.1%	F	6
G	102	62.24	77.75	51.89	0.1%	G	7

4 Discussion

A total of 14 teams have participated in BioCreative V.5. Compared to the systems that participated in the previous BioCreative V CHEMDNER task, the average results were better with 0.82 vs 0.76, 0.81 vs 0.77 and 0.83 vs 0.74, in terms of f-score, precision and recall, respectively. This time, the best f-score was 0.90 (three teams), slightly better than in BioCreative V CHEMDNER task (0.88). In view of the results, this task has reached the maximum performance one could expect taking into account the intrinsic difficulty of the task and the provided annotation quality.

Participation in the GPRO task has improved in BioCreative V.5. Considering GPRO *entity mention type 1*, i.e. GPRO mentions that can be normalized to a bio-entity database, the results for the best team were slightly worse than in BioCreative V CHEMDNER task, i.e. the f-score declined from 0.81 to 0.79, but the average team results were better with

0.74 vs 0.65, 0.73 vs 0.66 and 0.76 vs 0.64, in terms of f-score, precision and recall, respectively.
In the present GPRO task, *entity mention type 2*, i.e. non normalised mentions, were also evaluated. Comparing results for GPRO *entity mention type 1* to results for GPRO *entity mentions type 1 and 2*, it is observable that systems have a better average recall for *entity mention type 1* (0.76 vs 0.66) while precision is better when considering both types (0.73 vs 0.79).
Overall, the obtained results in CEMP and GPRO are considered competitive enough to derive in tools that not only could assist manual curation, but also could be used to automatic annotation extraction and patent abstract chemical indexing.

5 Acknowledgment

We acknowledge the OpenMinted (654021) and the ELIXIR-EXCELERATE (676559) H2020 projects, and the Encomienda MINETAD-CNIO as part of the Plan for the Advancement of Language Technology for funding. The Spanish National Bioinformatics Institute (INB) unit at the Spanish National Cancer Research Centre (CNIO) is a member of the INB, PRB2-ISCIII and is supported by grant PT13/0001/0030, of the PE I+D+i 2013-2016, funded by ISCIII and ERDF.

REFERENCES

1. Krallinger M, Leitner F, Vazquez M, Valencia A (2014) 6.04 – Text Mining. In: Compr. Biomed. Phys. pp 51–66
2. Krallinger M, Rabal O, Lourenço A, Perez Perez M, Perez Rodriguez G, Vazquez M, Leitner F, Oyarzabal J, Valencia A Overview of the CHEMDNER patents task.
3. Krallinger M, Leitner F, Rabal O, Vazquez M, Oyarzabal J, Valencia A (2015) CHEMDNER: The drugs and chemical names extraction challenge. J Cheminform 7:S1.
4. Krallinger M, Rabal O, Leitner F, Vazquez M, Salgado D, Lu Z, Leaman R, Lu Y, Ji D, Lowe DM, others (2015) The CHEMDNER corpus of chemicals and drugs and its annotation principles. J Cheminform 7:S2.
5. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158–D169. doi: 10.1093/nar/gkw1099
6. Maglott D, Ostell J, Pruitt KD, Tatusova T (2011) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 39:D52-7. doi: 10.1093/nar/gkq1237
7. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45:427–437. doi: 10.1016/j.ipm.2009.03.002
8. Van Rijsbergen CJ, Van CJ (1979) Information retrieval, 2nd ed. Butterworths