Mining Cell Cycle Literature Using Support Vector Machines

Theodoros G. Soldatos[1] and Georgios A. Pavlopoulos[2]

[1] Life Biosystems GmbH, Belfortstr. 2, 69115, Heidelberg, Germany
thsoldat@gmail.com
[2] ESAT-SCD / IBBT-K.U.Leuven Future Health Department, KatholiekeUniversiteit Leuven, KasteelparkArenberg 10, Box 2446, 3001, Leuven, Belgium
Georgios.Pavlopoulos@esat.kuleuven.be

Abstract. While biomedical literature is rapidly increasing, text classification remains a challenge for researchers, curators and librarians. In the context of this work, we use the Caipirini (http://caipirini.org) service to report on the exploration of a literature corpus related to the G1, S, G2 and M phases of the human cell cycle respectively. We use Support Vector Machines (SVMs) and a well-studied dataset to compare each of the cell cycle phases against all others in order to find abstracts that are related to one specific phase at a time. Finally we measure the performance of the results using the standard accuracy, precision and recall metrics. We find differences between the results of each of the four phases and we compare with previous findings of relevant work. We conclude that the results concur and help interpreting the observed classification performance.

Keywords: supervised machine learning, biomedical literature, cell cycle, support vector machines.

1 Introduction

Classifying literature and identifying a targeted set of articles of interest is frequently a bottleneck in biomedical research. As the number of papers produced per day increases rapidly, several tools that aim to help extracting information from biomedical literature have been developed [1,2]. For example, tools like ETBLAST [3], PubFinder [4], MScanner [5], BibGlimpse [6], Kleio [7], MedlineRanker [8], and Caipirini [9] help search and organise literature according to the interests of users. Mostly, such tools tackle the problem by trying to collect, classify and manipulate articles based on the biomedical terms or keywords that are mentioned in their texts. However, to our knowledge, only Caipirini [9] allows directly to compare and separate literature corpora according to relevance with gene sets. This task, i.e., distinguishing among a set of abstracts which are related more to one category and which are more relevant to another, can be useful in many ways. For example, many biomedical researchers often need to compare sets of genes which are expressed under different conditions or to compare gene lists produced in different ways, e.g., by using different

I. Maglogiannis, V. Plagianakos, and I. Vlahavas (Eds.): SETN 2012, LNAI 7297, pp. 278–284, 2012.

high-throughput experiments, or statistical analyses. Often one such researcher may want to identify literature focused on the specific conditions under consideration, e.g., to separate abstracts in groups that specifically discuss certain developmental stages, or abstracts that discuss the molecular (de)regulation of a plant's circadian rhythm specifically related to a certain season or time-period, or abstracts that discuss a specific disease for certain organisms only, and so on.

For the current study, we chose to work on a similar scenario: we wanted to explore literature corpora related to the human cell cycle and to identify abstracts related to each of the four phases, in specific. For this we relied on a previously studied dataset [10,11], for which a set of genes was assigned to each of the four phases (G1, S, G2 and M) – we used these gene lists as input to Caipirini [9] which allowed us to compare each phase against the other three, and to test the performance for each phase. We found that there were noticeable differences between the classifications of each case.

2 Materials and Methods

Caipirini: For this study, we took advantage of Caipirini [9], a service that allows researchers and curators to classify biomedical literature using support vector machines (SVMs). It mainly accepts as input two user-defined datasets (namely sets A and B). These can be imported directly as lists of PubMed [12] identifiers or as gene lists using Entrez[13] or Ensembl[14] gene identifiers. Sets A and B are used as examples for the training of the supervised learning method. The training relies on vectors extracted directly from the input abstracts or indirectly from the abstracts linked to the input genes. Next, abstracts from a third input set (called set C) are applied on the trained model which in turn assigns them either to set A or B. While Caipirini poses many advantages [9], one of its key features is that its automated pipeline enables users with no computational background to use SVMs, without having to take care of the underlying modelling complexities. This way, an experimental biologist who holds two sets of genes (A and B), can easily compare them directly and search among a relevant set of abstracts (set C) for the specific literature related with each of the sets of available genes. In its background, Caipirini [9] uses the SVM library LIBLINEAR [15], in accordance with the fact that linear SVM models have been found to perform well on text classification tasks. The service of Caipirini is described in detail at [9].

The dataset: We used a human gene set that was previously assigned to the four cell cycle phases by Martini [10]: 113 genes were assigned to the G1-phase, 154 to the S-phase, 82 to the G2-phase and 251 to the M-phase. The exact lists of gene identifiers can be found at the supplementary notes of [11], and at Martini's [10] webpage (under 'Example 1'), while the setup specific for S-Phase can be found in any of Caipirini's [9] examples '2', '3', or '4'.

Table 1. Number of abstracts that remained in Set C after removing from the results of the PubMed queries (a) first the overlap with the training set, and then also (b) the non indexed, by Caipirini's underlying dictionary, abstracts

Cell Cycle Phase	Number of abstracts		Query
	(a)	(b)	
G1	1337	1336	humans[MeSH Terms] AND ("G1 Phase"[MeSH Terms]) NOT ("S Phase"[MeSH Terms] OR "DNA Replication"[MeSH Terms] OR "G2 Phase"[MeSH Terms] OR "Prophase"[MeSH Terms] OR "Prometaphase"[MeSH Terms] OR "Metaphase"[MeSH Terms] OR "Anaphase"[MeSH Terms] OR "Telophase"[MeSH Terms] OR "Cytokinesis"[MeSH Terms]) AND ("2000/01/01"[PDAT] : "2008/06/31"[PDAT])
S	3904	3897	humans[MeSH Terms] AND ("S Phase"[MeSH Terms] OR "DNA Replication"[MeSH Terms]) NOT ("G1 Phase"[MeSH Terms] OR "G2 Phase"[MeSH Terms] OR "Prophase"[MeSH Terms] OR "Prometaphase"[MeSH Terms] OR "Metaphase"[MeSH Terms] OR "Anaphase"[MeSH Terms] OR "Telophase"[MeSH Terms] OR "Cytokinesis"[MeSH Terms]) AND ("2000/01/01"[PDAT] : "2008/06/31"[PDAT])
G2	1134	1134	humans[MeSH Terms] AND ("G2 Phase"[MeSH Terms]) NOT ("G1 Phase"[MeSH Terms] OR "S Phase"[MeSH Terms] OR "DNA Replication"[MeSH Terms] OR "Prophase"[MeSH Terms] OR "Prometaphase"[MeSH Terms] OR "Metaphase"[MeSH Terms] OR "Anaphase"[MeSH Terms] OR "Telophase"[MeSH Terms] OR "Cytokinesis"[MeSH Terms]) AND ("2000/01/01"[PDAT] : "2008/06/31"[PDAT])
M	1263	1260	humans[MeSH Terms] AND ("Prophase"[MeSH Terms] OR "Prometaphase"[MeSH Terms] OR "Metaphase"[MeSH Terms] OR "Anaphase"[MeSH Terms] OR "Telophase"[MeSH Terms] OR "Cytokinesis"[MeSH Terms]) NOT ("G1 Phase"[MeSH Terms] OR "S Phase"[MeSH Terms] OR "DNA Replication"[MeSH Terms] OR "G2 Phase"[MeSH Terms]) AND ("2000/01/01"[PDAT] : "2008/06/31"[PDAT])

The Comparisons: We performed four comparisons: we imported in Caipirini [9] the gene list assigned to each cell cycle phase as Set A and the three remaining gene lists (assigned to the other phases) as Set B. For Set C we used a literature corpus that allowed us to measure the performance for each classification and to evaluate the results, as described next.

The Evaluation: Following the example presented in [9], and in order to evaluate the classification results, we created a test set C by collecting abstracts known via Medical Subject Heading (MeSH) terms [16] to be related to each of the cell cycle phases. From [9] we reused only the S-phase query (that collects the S-phase related abstracts). In addition, we altered the S phase query so that we can assign abstracts specifically to each of the other three phases as well (see Table 1). As expected, the retrieved PubMed results for each query did not overlap - however, we processed these sets further and we removed any abstracts that belonged also to the training set (i.e., abstracts linked to the respective input genes). Last, we excluded abstracts that had not yet been indexed by Caipirini's underlying dictionary; only small differences were observed (see Table 1). Set C was used as a 'control', and was the same in all of the four classification tasks.

3 Results and Discussion

By assigning the genes that are related to each specific cell cycle phase to set A and the rest of the genes to set B we trained Caipirini [9] four times, as follows bellow:

- Set A = G1-Phase gene IDs; Set B = S,G2,M-Phase gene IDs
- Set A = S-Phase gene IDs; Set B = G1,G2,M-Phase gene IDs
- Set A = G2-Phase gene IDs; Set B = G1,S,M-Phase gene IDs
- Set A = M-Phase gene IDs; Set B = G1,S,G2-Phase gene IDs

In all cases, Set C was the same, i.e., the PubMed identifiers retrieved from the queries presented in Table 1. For each of the four cases, we calculated the accuracy, the precision and the recall of Caipirini's classification. The summarized results are presented in Table 2; the accuracy, the precision and the recall calculations were based on the standard formulas, defined next:

- Accuracy = (TP+TN) / (TP+TN+FP+FN)
- Precision = TP / (TP+FP)
- Recall = TP / (TP+FN)

For all measured metrics above, *TP* stands for 'true positives', *TN* stands for 'true negatives', *FP* stands for 'false positives', and *FN* stands for 'false negatives',

Table 2. Comparing each cell cycle-phase against the rest. Accuracy, precision and recall were calculated for each of the four experiments.

	Caiprini	Input Training Sets		Performance Measures		
	Task	Vectors of A	Vectors of B	Accuracy	Precision	Recall
1	G1 vs S,G2,M	5696	28621	0.806	0.404	0.220
2	S vs M,G1,G2	7920	26397	0.660	0.842	0.412
3	G2 vs G1,S,M	4805	29512	0.816	0.192	0.073
4	M vs G1,S,G2	15896	18421	0.634	0.237	0.548

For the S-phase, we used 154 genes associated with the S-phase as set A, 446 genes associated with the other three phases of the human cell cycle (G1, G2, and M) as set B, and as set C we used all abstracts known via MeSH terms to be related to the cell cycle (see Table 1). With this case we verified the performance of Caipirini presented in [9] (see Table 2), and we continued with the remaining three phases similarly:

- (a) We compared the G1 phase with the three other phases and in comparison to the S phase results we found lower precision and recall, but higher accuracy (see Table 2),
- (b) When comparing the G2 phase against the others, we observed that precision and recall were lower than all other three cases, although accuracy remained remarkably the highest (see Table 2).

- (c) In the last case, we compared the M phase against the other three and observed a mixed result: first, the accuracy was comparable to that of S-phase and lower than those for G1 and G2, whereas the precision was comparable to that of G2 and lower than both for G1 and S phases – also, in this case the best recall was achieved (see Table 2).

Comparing the results (see Figure 1), we believe that S-phase, represents the best-case scenario in this study. This can be a result of the distribution of abstracts in Set C (see Table 1). It can also be attributed to the distribution of vectors in each training set (see Table 2; the number of vectors represents the abstracts associated with each set – note that Caipirini does not remove multiple occurrences of abstracts in the training sets). For example, although for M phase there are more training vectors for set A than in the case of S phase, the latter seems to be more robust. Indeed, when comparing with the work from which we got the data set from it becomes clear that for S phase there are many more specific keywords [10]. This indicates that mining the literature for S phase has an advantage in comparison to the other three phases, because this way the SVM can learn better characteristic features and in turn to create a trained model that can separate better the classes.

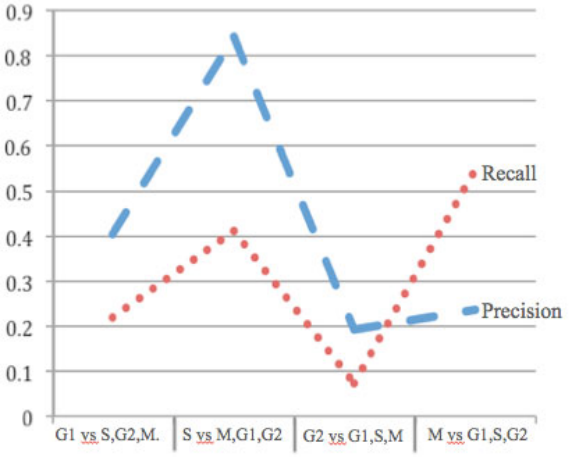

Fig. 1. Precision and Recall for the four experiments. Each case (*x* axis) is named after the phases to which the gene identifiers used in the input sets A and B belonged to.

However, the performance of this categorization can be tested further, along various dimensions: such examples include using different SVM configurations, using different combinations of term types (in the current study all term types were used) and dictionaries, by setting set C otherwise, or by using subsets and/or permutations of each cell cycle phase.

Last, in this work we do not try to interpret further the results since we believe that the comparison of Caipirini with the performance of another somewhat comparable tool [9] already indicated that for this dataset it can be difficult to achieve better results. Nevertheless, we expect that mining literature related to the different phases of

the human cell cycle can become a standard case-study used to evaluate and compare new methods and tools, which makes the human cell cycle dataset especially interesting for such tasks. Notably, in order to enable more such scenarios and with many classes of genes, Caipirini's future plans already include the development of an updated version in which multi-class SVM classifiers will also be feasible [9], e.g., in order to distinguish G1, S, G2 and M phases directly 'in one go'.

4 Conclusions

This work does not present novel methods, but rather reports on the performance of Caipirini in mining literature related to different phases of the human cell cycle. First, we verified previous results about S phase and then we expanded further to the remaining three phases. Last, we conclude that this gene set, as proposed in [9] and [10], indeed makes a good benchmark: the findings suggest that the chosen cell cycle data set possesses not only realistic biological scenarios, but also computational characteristics that are challenging for researchers interested in biomedical classification tasks.

Acknowledgements. ***Funding:*** Georgios A. Pavlopoulos would like to acknowledge support from: Research Council KUL:ProMeta, GOA Ambiorics, GOA MaNet, CoE EF/05/007 SymBioSys en KUL PFV/10/016 SymBioSys , START 1, several PhD/postdoc & fellow grants. Flemish Government: FWO: PhD/postdoc grants, projects G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); G.0733.09 (3UTR); G.082409 (EGFR) IWT: PhD Grants, Silicos; SBO-BioFrame, SBO-MoKa, TBM-IOTA3 FOD:Cancer plans, IBBT. Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007- 2011); EU-RTD: ERNSI: European Research Network on System Identification; FP7-HEALTH; CHeartED.

References

1. Krallinger, M., Valencia, A.: Text-mining and information-retrieval services for molecular biology. Genome Biol. 6(7), 224 (2005), doi:10.1186/gb-2005-6-7-224
2. Krallinger, M., Erhardt, R.A., Valencia, A.: Text-mining approaches in molecular biology and biomedicine. Drug Discov. Today 10(6), 439–445 (2005), doi:10.1016/S1359-6446(05)03376-3
3. Lewis, J., Ossowski, S., Hicks, J., Errami, M., Garner, H.R.: Text similarity: an alternative way to search MEDLINE. Bioinformatics 22(18), 2298–2304 (2006), doi:btl388
4. Goetz, T., von der Lieth, C.-W.: PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts. Nucleic Acids Res. 33, W774–W778 (2005)
5. Poulter, G.L., Rubin, D.L., Altman, R.B., Seoighe, C.: MScanner: a classifier for retrieving Medline citations. Bioinformatics 9, 108 (2008), doi:1471-2105-9-108
6. Tuchler, T., Velez, G., Graf, A., Kreil, D.P.: BibGlimpse: the case for a light-weight reprint manager in distributed literature research. BMC Bioinformatics 9, 406 (2008), doi:1471-2105-9-406

7. Nobata, C., Cotter, P., Okazaki, N., Rea, B., Sasak1, Y., Tsuruoka, Y., Tsujii, J.I., Ananiadou, S.: Kleio: A Knowledge-enriched Information Retrieval System for Biology. In: 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, pp. 787–788. Association for Computing Machinery (2008)
8. Fontaine, J.F., Barbosa-Silva, A., Schaefer, M., Huska, M.R., Muro, E.M., Andrade-Navarro, M.A.: MedlineRanker: flexible ranking of biomedical literature. Nucleic Acids Res. 37(Web Server issue), W141–W146 (2009), doi:gkp353
9. Soldatos, T.G., O'Donoghue, S.I., Satagopam, V.P., Barbosa-Silva, A., Pavlopoulos, G.A., Wanderley-Nogueira, A.C., Soares-Cavalcanti, N.M., Schneider, R.: Caipirini: using gene sets to rank literature. BioData Mining 5(1), 1 (2012), doi:10.1186/1756-0381-5-1
10. Soldatos, T., O'Donoghue, S.I., Satagopam, V.P., Brown, N.P., Jensen, L.J., Schneider, R.: Martini: using literature keywords to compare gene sets. Nucleic Acid Res. 38(1), 26–38 (2010), doi:10.1093/nar/gkp876
11. Jensen, L.J., Jensen, T.S., de Lichtenberg, U., Brunak, S., Bork, P.: Co-evolution of transcriptional and post-translational cell-cycle regulation. Nature 443(7111), 594–597 (2006), doi:10.1038/nature05186
12. PubMed, `http://pubmed.org`
13. Entrez gene database, `http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene`
14. Ensembl, `http://ensembl.org`
15. Fan, R.-E., Chang, K.W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
16. Medical Subject Headings (MeSH) Fact sheet. In: National Library of Medicine (2005)