

Names of chemical compounds within drug discovery context

Elzbieta Dura, Ola Engkvist, Sorel Muresan

Abstract— Drug discovery is costly and time consuming, mainly due to very high attrition rates. To remedy this, the INFUSIS project tries to improve predictive modeling for ADMET by fusing information from various sources. Unstructured texts are among the most important sources and we use text corpus technology to uncover information on toxicity of specific compounds in articles on diabetes research. A selection of 200,000 abstracts on diabetes from PubMed was turned into a corpus. This allows extracting patterns of the actual use of chemical nomenclature in texts. The task of proper identification of chemical compounds in texts is not trivial despite availability of large compound libraries. There is a significant difference in how terms are registered in lexicons and how they are actually used.

Index Terms—text mining, chemical compound, drug discovery, named entity recognition

I. INTRODUCTION

A large part of the relevant information in the drug discovery process is found in unstructured text (journal, patents, reports, etc.). Extracting and processing this information to a format suitable for further analysis is not trivial and manual or automated processes can be used. Manual curation¹ has the advantage of providing high quality data, where the relations between different entities (document, compound, target, biological activity, and assay) can be unambiguously mapped by expert curators [1]. Text mining becomes useful when large volume of unstructured text needs to be processed and when the relation between compounds and their toxicity is not obvious. This work package from the INFUSIS project aims at using a text mining approach in combination with a chemical dictionary to extract and model ADMET properties for bioactive compounds.

II. INFUSIS

A. ADMET problems

One of the most intensive areas of research within the pharmaceutical industry today is to collect and analyse data on

¹“Curation of biological databases means basically the manual extraction of biological information from the literature by a domain expert. The aim is to transform information contained in free text (scientific literature) to information stored in form of a structured database record (biological databases).” Quoted after the BioCreative Glossary at: http://biocreative.sourceforge.net/biocreative_glossary.html

absorption, distribution, metabolism, excretion and toxicity (ADMET) [1], [2]. The overall purpose is to learn how various compounds interact with the human body in order to guide drug development projects in the search for promising compounds. Specifically, compounds unsuitable as drug candidates because of toxicity should be detected as early as possible. Hence, a lot of effort is spent on “front loading” the drug development projects (i.e. a substantial amount of analysis is invested in the very early phases of the projects).

B. INFUSIS overall plan

The project aims at developing new techniques for predictive ADMET modelling that result in substantial improvements in predictive performance by incorporating multiple sources of information, including textual annotations, and by robustly combining multiple classifiers. The project furthermore aims at developing techniques that allow computational chemists to gain understanding of the reasons behind predictions. The project will contribute to the fields of information fusion, machine learning and data mining, by addressing limitations of current techniques, providing new technologies and analyses of their performance. The project will also contribute to the field of cheminformatics by exploring alternative ways of representing chemical information to allow for effective analysis.

III. WORK PACKAGE 1.1

This work package from the INFUSIS project aims at using a text mining approach in combination with a chemical dictionary to extract and model ADMET properties for bioactive compounds. Patterns from relevant PubMed data will be extracted with Culler, a text mining software. A chemical dictionary will be used to identify and map chemical terms in the text. Datasets containing ADMET assertions will be extracted and the chemistry information will be expanded with calculated physicochemical properties. These will then be used to model ADMET properties.

Our work is in the initial phase in which we have separated the subtasks as consecutive or simultaneous steps, and identify problems in each of the steps. Here we present the problems of identification of names of chemical compounds within drug discovery context.

Considering the availability of truly large chemical databases, one could be lead to thinking that the recognition of chemical names in texts should be a rather trivial task: just a matter of looking up an item in a database. However, this is

not the case, mainly because of ambiguity and variation, both in chemical nomenclature and in actual usage; both of these problems are exemplified below. The huge sizes of chemical databases (20,000,000 unique structures in PubChem) and text collections (250,000,000,000 tokens² in PubMed) do not make the task easier.

IV. TURNING SELECTED TEXTS INTO A CORPUS

About 200,000 abstracts mentioning diabetes have been selected from PubMed, processed with natural language tools, indexed and turned into a text corpus in the Culler system [3].

A. Corpus technology

Text excerpts of similar content can be aligned using in-depth text analysis techniques. Corpus technology involves a combination of statistics and natural language analysis, which can provide insights into text patterns unavailable otherwise, such as a bird's eye view on some purposefully chosen content in the whole text collection.

The specific trait of the Culler system is the possibility to add knowledge sources, such as special lexica. The system makes use of them in analysis, indexing and at the same time makes them available for querying. We have provided Culler with a list of constituents of chemical compounds.

B. The Diabetes Corpus in Culler

Different biocorpora are available in Culler as an open public service at the University of Skövde.³ The biggest one is the Gene Corpus - a selection of abstracts mentioning *gene*, counting more than 250,000,000 text tokens. The selection of abstracts mentioning *diabetes* in the Diabetes Corpus has about 57,000,000 text tokens. The typically scientific language of the collection has almost twice the number of nouns and half the number of verbs compared to general English (these statistics are available for the corpus on clicking *Select corpus*).

V. COMPOUNDS IN LEXICONS AND IN TEXTS

A. Repositories, public and proprietary

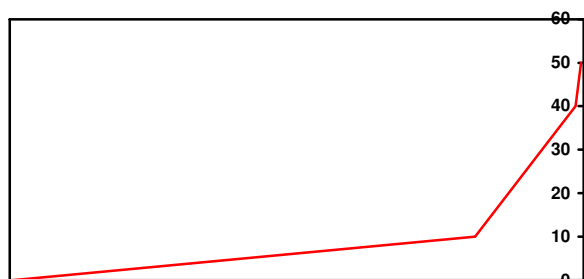


Figure 1. Registration of new chemical substances by CAS from 1957 to 2009.

The 50th millionth unique chemical substance was registered by CAS in August 2009. Its 40th millionth substance was registered only nine months before, while it took 33 years to register 10 million compounds up to 1990.⁴

TABLE 1. LIBRARIES OF CHEMICAL SUBSTANCES

PubChem	over 20 million compounds
ChemSpider	over 20 million compounds
Zinc	4.6 million compounds
ChemDB	4.1 million compounds
BioCyc	3500 compounds involved as enzyme substrates, products, inhibitors and activators
DrugBank	4300 drug data with target information
ChemBank	36000 biological assays of small molecules
Web Reactions	391000 organic reactions

The steeply growing number of substances registered in general repositories, as shown in Figure 1, create the need for more specialized collections, some of which are listed in Table 1. In addition, extensive compound collection enhancement programs have been described in the literature to address the problem of sampling this vast chemical space [4], [5].

B. Compound names and their contexts

We have selected a list of drug names and synonyms from several sources. These were turned into a variable *&chem* in Culler, counting about 4000 items. Using this variable it is possible to extract compound clusters defined by common contexts. The example in Figure 2 shows how such a variable can be used to elicit subsets of compounds sharing some features, here occurring in the context of *toxic effects of...* If needed, terms may be directly extracted as an input for further processing, because a query can easily distinguish terms in focus from their context.

Figure 2. A screen dump showing the results of the query using the list of compounds in the context of *toxic effects of...*

The Diabetes Corpus discloses varieties of patterns used by scientists in reporting their results. The insights will assist us both in better recognition of compounds and identifying

²A text token is a sequence of characters considered a single unit.

³<http://bergelmir.iki.his.se/culler>.

⁴<http://www.cas.org/newsevents/releases/casregistry50m081609.html>

features that are supposed to lead to improvement of the decision process when considered together with molecular descriptors. The first step in this process is the recognition of compounds from a dictionary in their actual realizations in texts.

C. The nomenclature in actual use

The main problem is to determine whether a new name and compound is meant or whether an alternative notation is used, or whether a word is simply misspelled.

Are the two chemical entities below merely variants or are they different substances:

3-hydrox 3-methylglutaryl (HMG)-Co-enzyme A (CoA)

3-hydroxy-3-methyl glutaryl coenzyme A (HMGCoA)

When a name identified in the text is absent in our lexicon it is sometimes fairly obvious that it is a misspelling of some name, such as *atorvstatin* and *atrovastatin*, *simvastatin* and *simvstatin*. Similarity measures or phonological rules can prompt here. In other cases the decision is uncertain, for instance *antistomatostatin* and *anatistomatostatin* or *somastostatin* and *somatostatin*.

Expediency always prevails in every day language use over directions promulgated by some academy, be it the French Academy, UPAC, etc. The phenomenon is valid in scientific language as well. Words can get very lengthy and shorthand variants are preferred whenever they are non-equivocal. For instance, when *atorvastatin* is mentioned in previous context, not only *this statin* but also *therapy with statins* may refer to the former. In this case compound identification requires anaphora resolution - identification of a referent based on prior context and knowledge of the domain. Anaphora resolution is considered to be one of the most difficult tasks in the semantic analysis of a text.

Traditional names of chemical substances, or simply short names, are often preferred to their systematic names, because this enhances readability. For instance, *α -D-glucopyranosyl-(1,2)- β -D-fructofuranoside* is not once used in the corpus in which *sucrose* occurs 1,779 times. The number of occurrences in the Diabetes Corpus in Table 2 shows a comparison: total number occurrences of a popular/shorter and of a longer name.

TABLE 2. NUMBER OF TOTAL OCCURRENCES HINTS ON THE MORE POPULAR NAME VARIANTS.

Short name		Long name	
<i>Sucrose</i>	1779	<i>β-D-fructofuranosyl-(2\rightarrow1)-α-D-glucopyranoside</i>	0
<i>Vitamin C</i>	1384	<i>Ascorbic acid</i>	859
<i>Aspirin</i>	3282	<i>Acetylsalicylic acid</i>	215
<i>Naproxen</i>	63	<i>(S)-6-methoxy-α-methyl-2-naphthaleneacetic acid</i>	0
<i>dUMP</i>	196	<i>2'deoxyuridine-5'monophosphate</i>	3
<i>ASS</i>	297	<i>argininosuccinate synthetase</i>	306

VI. TASKS AHEAD

Compound identification is the nearest goal for which we shall

test how different compound lexicons impact precision and recall. Then we need to test which is important for the goal of capturing toxicity information. Dependent on the use, named entity recognition with high precision may be more valuable than one with high overall F-score [5].

So far we have only touched upon the problems of extracting data which can be turned into features relevant in decision support models. The Diabetes Corpus can assist us in sharpening our intuition on how relevant facts are actually stated in research articles.

Named entity recognition is the most basic text extraction problem in life sciences. It has attracted the attention of several researchers and has seen significant progress, yet the results are still not satisfactory. Despite the existence of rich terminologies and ontologies, the problems of language variation and ambiguity remain a challenge. Simple matching against terminologies is actually far from simple. The promising path which will be followed in our future work in order to cope with the variation problem is fusing soft string matching with normalization rules [8].

ACKNOWLEDGMENT

This work was supported by the INFUSIS project (<http://www.his.se/infusis>) at the University of Skövde, Sweden, in partnership with AstraZeneca, Lexware Labs and the Swedish Knowledge Foundation under grant 2008/0502.

REFERENCES

- [1] C. Southan, P. Varkonyi, S. Muresan, "Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds", in *J. Cheminfo*. 2009, 1 (1), 10.
- [2] H. van de Waterbeemd and E. Gifford, "ADMET in silico modelling: towards prediction paradise?", in *Nat Rev Drug Discov.*, Vol. 2, No. 3 (2003) 192-204
- [3] E. Dura, B. Olsson. "Information retrieval combined with extraction to assist decision making", in *Proceedings of the second Skövde Workshop on Information Fusion Topics SWIFT 2008*, H. Boström, R. Johansson and J. van Laere, Eds. Available: <http://www.his.se/english/research/infusion/news-events/workshops/past-workshops/swift-2008/proceedings/>
- [4] E. Jacoby, A. Schuffenhauer, M. Popov, K. Azzaoui, B. Havill, U. Schopfer, C. Engeloch, J. Stanek, P. Acklin, P. Rigollier, F. Stoll, G. Koch, P. Meier, D. Orain, R. Giger, J. Hinrichs, K. Malagu, J. Zimmermann, H. J. Roth, "Key aspects of the Novartis compound collection enhancement project for the compilation of a comprehensive chemogenomics drug discovery screening collection", in *Curr. Top. Med. B. Chem.*, 2005, 5, 397-411.
- [5] T. Cooper and K. Andrews-Cramer, "Designed chemical libraries for hit/lead optimisation", in *Innovations in Pharmaceutical Technology 2000*, The Pharmaceutical Technology Journal, Samedan Pharmaceutical Publishers LTD. Available: <http://www.iptonline.com/synopsis.asp?cat=2&article=127>
- [6] R. B Altman, C. M. Bergman, J. Blake, C. Blaschke, A. Cohen, F. Gannon, L. Grivell, U. Hahn, W. Hersh, L. Hirschman, L. J. Jensen, M. Klrallinger, B. Mons, S. I. O'Donoghue, M. C Peitsch, D. Rebholz-Schuhmann, H. Shatkay and A. Valencia, "Text mining for biology - the way forward: opinions from leading scientists," in *Genome Biology - The BioCreative II - Critical Assessment for Information Extraction in Biology Challenge*, vol. 9, L. Hirschman, M. Krallinger, A. Valencia, Eds. Available: <http://genomebiology.com/2008/9/S2/S7>.
- [7] P. Corbett and A. Copestake, "Cascaded classifiers for confidence-based chemical named entity recognition", in *BMC Bioinformatics 2008*, 9. Available: <http://www.biomedcentral.com/1471-2105/9/S11/S4>

- [8] Y.Tsuruoka, J. McNaught, S. Ananiadou, Normalizing biomedical terms by minimizing ambiguity and variability. In *BMC Bioinformatics* (2008, 9).