ICBO: International Conference on Biomedical Ontologies · Buffalo, NY, USA
Working with Multiple Biomedical Ontologies Workshop, July 26, 2011

Use of Multiple Ontologies to Characterize the Bioactivity of Small Molecules

Ying Yan[1], Janna Hastings[1], Jee-Hyub Kim[1], Stefan Schulz[2], Christoph Steinbeck[1], Dietrich Rebholz-Schuhmann[1]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[2]Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria

Abstract. ChEBI is an ontology of biologically interesting chemicals. Biological activities of chemical entities comprise interactions with biological entities such as proteins and anatomical structures such as the cell membrane. Currently, ChEBI represents these biological activities of small molecules within a 'role' ontology which includes terms such as 'cyclooxygenase inhibitor'. However, this 'role' ontology is not complete, and is not directly interlinked with the biological ontologies which serve as the main source of concepts describing biological entities. This makes it difficult to reason over the relationships between chemical entities and their biological targets. To address this issue, we propose a model for interrelating multiple ontologies and controlled vocabularies in the biomedical domain in order to formally characterise the bioactivity of small molecules. In support of this work, we have developed a method for analysing the scientific literature for textual descriptions of bioactivity events linked to chemical entities. We examine the distribution of terms from various controlled vocabularies (biological processes, proteins, organisms and organ/body parts) in combination with the chemical entities in the literature, to better understand reports of bioactivity. We find that proteins are the most commonly reported type of target of small molecule bioactivity, and that organisms and organs are most commonly reported in the literature as locational constraints rather than as targets.

1 Introduction

ChEBI is an ontology of chemical entities of biological interest [3]. It describes chemical entities such as molecules and ions together with their structural and biologically relevant properties. As of June 2011, it consists of around 25,000 entities, divided into a structure-based classification and a role-based classification. The role-based classification includes terms describing the biological activities of chemical entities, such as 'cyclooxygenase inhibitor' and 'immunomodulator'. These terms describe small molecule *bioactivity*: the combined influence of a small molecular entity on the components of a living organism and on the organism as a whole.

On a molecular level, small molecule bioactivity corresponds to the binding of the molecule to a macromolecular receptor, resulting in some observable physiological effect on the biological systems involving that macromolecule [4]. Bioactive molecules can have positive effects, such as repressing the development of disease, or they can have negative (toxic) effects, leading to illness or even death. The differentiation of bioactive molecules from non-bioactive molecules is one of the core requirements for *in silico* drug discovery approaches [12], as are delineating molecules which share similar activity profiles [9].

To properly formalise the description of activities of chemical entities in biological contexts requires reference to multiple terminological sources, some of which fulfill the requirements for formal ontologies (such as, e.g. the OBO Foundry ontologies [15]), whereas other ones are better characterised as thesauri, databases, or controlled vocabularies. For example, to formalise a description of enzymatic inhibitor activity requires reference to the enzyme which is being inhibited; to formalise participation in a in a particular biological process requires reference to the process; and bioactivity descriptions may require reference to the exact location of the

activity and the organism within which, or against which, the activity took place.

The ChEBI role ontology does allow the categorisation of chemical entities by their bioactivities. However, in its present form it suffers from two key limitations:

1. Role assertions are relatively *sparse* as compared to the full ontology of chemical entities (just less than 3000 chemical entities are mapped to just less than 500 roles, ca. 10% of the full chemical entity ontology). The result is that many of the chemical entities included in the ontology are not adequately described in terms of their biological context.
2. Bioactivity descriptions in the role hierarchy of the ontology are not explicitly linked to a primary reference source for the biological entities themselves. For example, the term 'cyclooxygenase inhibitor' describes the inhibition of a cyclooxygenase enzyme, yet this term is not explicitly linked to a reference for enzymes such as UniProt.

The aim of the present work is to use the automated analysis of literature as a means to address these limitations. The remainder of this paper is organised as follows. Firstly, we present our methods, which include the definition of a language model for bioactivity description and its application to extract mentions of bioactivity events from publicly available literature resources, in Section 2. Section 3 describes and discusses our results, including the implications of our literature analysis on the ontology model for interrelating chemical entities and biological objects and processes. In the final section we conclude with the relevance of this work both for biomedical research generally and for improved curation tools in the context of the ChEBI project.

2 Methods and Models

We first defined a language model for bioactivity terminology based on the examination of relevant portions of the Metathesaurus of the Unified Medical Language System (UMLS) [1] and the ChEBI biological roles. This is described further in Section 2.1.

We then used this language model to extract bioactivity descriptions for ChEBI entities from MEDLINE abstracts. The text mining methods used are described in Section 2.2. After examining the sentences returned, we defined an ontology model for characterising the formal relationships between ChEBI entities and other biological entities.

2.1 Bioactivity Language Model

Basic Phraseal Patterns. Bioactivity of a chemical entity (CE) is described using given a set of language features: "inhibitor" and "activator", "modulator", "agonist" and "antagonist", "toxin", "regulator", "suppressor", "adaptor", "stimulator", "factor", "messenger" and "blocker"; these will be called *trigger words*.

Any of these features can occur as a head noun in a phrase structure leading to the following type of phrasal patterns for their identification: a head noun preceded by a noun phrase, as follows: <modifier> <head>.

Ideally, the phrase composing (<modifier>) is constituted by one or more tokens which denote the *target* of the bioactivity, whereas the head word specifies the *nature* of the interaction between the small molecule and the target. For example, 'beta-adrenergic receptor inhibitor' has as modifier 'beta-adrenergic receptor' (the target) and as head word 'inhibitor' (the nature of the interaction is inhibition).

The basic language model was further extended to include alternative, compatible language patterns such as 'inhibition of X', where 'X' corresponds to the modifier and 'inhibition' the head word [8]. We identified four different syntactical structures for bioactivity descriptions, namely:

1. Noun phrase or adjective/adverb compositions as modifier. This is the most commonly seen structure of the basic noun phrase, and we find a considerable number of bioactivity terms presented in this way. For example:

   HIV transcriptase inhibitor

2. Prepositional phrase as modifier. Prepositional phrases are generally formed by a preposition followed by a prepositional complement. We also find this structure is often represented in bioactivity terms. For example:

Suppressor of fused protein Oct-1 CoActivator in S phase protein, human

3. Verb phrase as noun phrase modifier. When the verb phrase functions as a modifier in a bioactivity noun phrase, it presents the way in which the activation of the described subject results in a kind of influence to its object. For example:

   TIR domain containing adaptor inducinginterferon-beta protein

4. Relative clauses as modifier. Relative clauses are defined as subordinate clauses that consist of a clause beginning with a relative pronoun. This type of modifier is also used in the bioactivity presentation. For example:

   Factor that binds to inducer of short transcripts protein 1

2.2 Bioactivity Term Extraction from MEDLINE Abstracts

The method used to extract bioactivity descriptions from MEDLINE abstracts is a simple procedure which identifies noun phrase structures by matching *syntactical language patterns*. These hand-crafted language patterns form an alternative to syntactic parsing, which requires significant compute resources and is still error prone in several extraction tasks [7].

After bioactivity noun phrases were identified using the above patterns, we pruned outliers which had the trigger word as other parts of the phrase. For example, *Mononuclear cell growth inhibitor assay* is not considered to represent a valid bioactivity phrase because the activity term *inhibitor* is not the head noun (which in this case is *assay*). The solution for the identification of the noun phrases is based on hierarchically organised language patterns developed for the extraction of protein noun phrases in the protein-protein interaction pipeline [14]. The syntactical structures of the matching patterns have been tailored to fit the language model used in the approach of this manuscript.

The purpose of this analysis was to investigate the target types for bioactivity descriptions. To this end, four taggers for named entity and concept label identification (UniProtKB [10], Organ [6], Organisms [16] and GO [13]) were applied on the modifier of candidates extracted from MEDLINE. The unique count of tagging was cross analysed by features provided in Section 2.1. We collected tabulated statistics which are presented in the results section.

Classifying Bioactivity Terms. After extracting bioactivity descriptions from MEDLINE, we aimed to find an efficient method of classifying all the candidates with a high rate of recall.

When the process results in the entire modifier being annotated by a tagger, this consequently indicates its semantic type. For example, *CaM kinase I activator* is easily classified as a *protein* activator since the modifier has been annotated as a protein from making reference to UniProtKB.

However, in the majority of cases, we found that the result is a nested case, in which the semantic tagger annotates just part of the modifier, i.e. the tagged result resides within the boundaries of the whole phrase for the modifier. For instance, *Agkistrodon blomhoffi ussuriensis protein C activator*. In this case, *ussuriensis protein C* is the authentic target of the activation, though *Agkistrodon blomhoffi* is identified. As previously mentioned, a simple method to rule out un-associated tagging is used. We retain the tag which is in the last position within the modifier, ignoring other tags. In this example, the target type is not species but protein.

2.3 Text Mining Methods for Bioactivity Triple Extraction

We used a dictionary-based approach to extract names of small molecules and their targets together with their relation types from the whole MEDLINE resource. The approach processed text on a sentence level, extracting triples which contain (1) a small molecule term, (2) the 'feature' trigger word, which presents the relation type, and (3) a term representing the target of the small molecule.

To identify the small molecules, we compared results from using two different chemical taggers, namely the newest versions of Oscar3 [2] and Jochem [5]. Jochem, being dictionary-based, has the advantage that all chemical entities it recognises are known entities, whereas Oscar3 can recognise non-known strings that resemble syntactical

structures denoting chemical entities (higher recall).

All the possible combinations of small molecule terms, features and target terms in each sentence are generated. We found that false positive cases were significant, and therefore applied three stages of rule-based filtering:

1. Remove triples from the candidate list when the putative small molecule term is actually a role term according to the ChEBI ontology (e.g. 'antibiotic')
2. Filter out those triples where the small molecule term has the suffix "-ase", since these terms are normally enzyme names.
3. Remove triples when the string that supposedly denotes a small molecule has less than three characters.

3 Results and Discussion

3.1 Evaluation of Language Model

The evaluation of our approach is ongoing work and requires a gold standard corpus (GSC). The GSC would enable us to test supervised learning methods against our existing feature-based extraction method. However, the named entity recognition methods have all been evaluated. The identification of proteins and genes performs at *52.37%/61.63%/56.62%* (Rec,Prec,F-Meas) on PennBioIE and *50.26%/61.63%/56.62%* (Rec,Prec,F-Meas) on BC-II [11]. The method applied for the identification of genes and proteins was based on the UniProtKB dictionary with basic disambiguation and was not trained on one of the different gold standard corpora, since methods trained on gold standards show high differences in their performance when being tested against other gold standard corpora. Trained methods for gene mention identification are available and show higher performance, but do not allow linking results to data from biomedical data resources, e.g. UniProtKB and EntrezGene.

3.2 Results of Running Language Model on MEDLINE Abstracts

Table 1 shows an overview of target type associated with feature trigger words. Each cell shows the unique count of semantic tagging for a certain feature. Both nested and exact matching on the modifier of bioactivity terms are considered.

Feature	Protein	Organ	Organism	Biological Process
stimulator	2,526	3,303	500	1,808
adaptor	3,729	100	133	1,016
modulator	7,847	1,468	536	4,204
messenger	10,056	1,186	1,151	3,876
agent	10,522	10,292	19,374	8,744
blocker	13,588	1,371	9,235	4,203
toxin	16,890	1,583	10,265	3,276
suppressor	18,534	1,301	2,382	2,988
regulator	27,724	5,469	2,802	27,270
factor	40,427	21,959	11,152	77,670
agonist	48,973	3,633	13,154	12,353
activator	71,165	1,745	3,895	19,376
antagonist	80,932	9,483	11,740	19,486
inhibitor	336,420	12,102	30,839	142,289

Table 1. Identifying target type of small molecule on MEDLINE abstracts.

From this table, the main target types of bioactivity are identified based on a two-feature driven method.

In general, protein names are mostly nested in the modifier of bioactivity terms. UniProtKB tagging and 'inhibitor' gives a high number of hits: 336,420 unique combinations. This suggests that bioactivity descriptions in text usually refer to activities against a protein or enzyme. Two such examples are:

- *Other lysosomal hydrolases are not inhibited by N-bromoacetyl-beta-D-galactosylamine, with the exception of 'neutral' beta-glucosidase glucohydrolase.*
- *At the biochemical level cardiac guanyl cyclase activity is enhanced 2–3 times with acetylcholine and this enhancement is completely blocked by atropine.*

There are not as many hits in the Organ and Organisms groups. We can find a few true positive examples such as *bothrops jararaca inhibitor* and *thyroid stimulator*. However, there are many examples in which the organ or organism appears in the sentence only to denote the location of the bioactivity being described. For example:

1. Caesium ion antagonism to chlorpromazine - and L-dopa-produced

behavioural depression *in mice*.

2. The changes in the contents of glycolytic intermediates in the livers indicate that the phosphoenolpyruvate carboxykinase [EC 4.1.1.32] reaction is inhibited by tryptophan administration *in all groups of rats*.

3. The oral administration of meta-proteranol increased the leukocyte adenyl cyclase activity which was stimulated by NaF and decreased the count of peripheral eosinophils *in some of the monkeys*.

We conclude that in the literature, organ and organism most commonly provide the contextual information about where a bioactivity takes place, rather than being themselves the target of the bioactivity. This will influence our ontology model, described in Section 3.4.

We also analysed the case where GO terms were tagged in bioactivity terms. For example, *inhibitor of DNA transcription*. Here, a biological process is the target of the bioactivity term.

Limitations. As is the norm in this type of text mining approach, there are also typical 'noisy' false positives in the result, such as 'hand' being tagged as a body part in the sentence 'On the other hand, ...', and 'dialysis' being tagged as a species in the sentence 'Influence of peritoneal dialysis on factors affecting oxygen transport.' (Dialysis is, indeed, a species: a kind of bug.). Care also needs to be taken in that some of the results reflect sentences in which the bioactivity being described in the extracted triple is explicitly *not* reported as taking place, such as:

1. *Without* influence on WDS were: physotigmine, atropine, ganglionic-or adrenergic-blocking drugs, Dopa, MAO-inhibitors, serotonin- and histamin-antagonists and nonnarcotic analgesics.

2. The cellulase component was *not* markedy inhibited by most metal ions tested.

3.3 Comparison of Chemical Taggers

To identify chemical entities, we compared a dictionary-based approach using Jochem with the results generated using Oscar3 which is able to identify novel chemical names in text using a machine learning approach.

Table 2 shows the frequency of each triple mentioned in text together with the unique count of triples before and after the rule-based filtering described in Section 2.3.

Oscar3 yields many more triples than Jochem does. This is expected, since Oscar3 recognises any chemical-like string. However, Oscar3's approach also results in a considerable number of false positives due to its recognition of chemical-like nomenclature appearing as a component in larger strings (such as protein names). Furthermore, we can observe a smaller number of triples identified by UniProtKB and Oscar3 compared to the set identified by UniProtKB and Jochem. This is because Oscar3 produces annotations that nest within a protein mention in the sentence and thus lowers the subsequent annotation protein mentions. Jochem performs more long-form matching than Oscar3 does, therefore the following protein identification has a higher likelihood of identifying a protein term within the sentence, hence yielding a greater number of triples.

The comparison of before and after filtering show whether the triple mention is by chance and the association between the chemical and the other semantic group is more than contextually related. Between chemicals and proteins the ratio is smaller than the other groups. The non-unique number of triples is less than twice the number of unique ones, while it is more than this ratio in other groups (specifically in the chemical organ group). The number of non-unique triples identified by Jochem after filtering is almost three times the unique count.

		UniProtKB		Organ		Organism		GO	
Chemical tagger	Filtering	uniq	non uniq	uniq	non uniq	uniq	non uniq	uniq	non uniq
Jochem	before	4,114,286	7,853,314	2,666,468	7,148,677	1,785,771	4,076,253	1,244,099	2,947,289
	after	2,912,756	5,457,529	1,632,855	5,302,115	1,394,310	3,085,056	935,864	2,089,163
Oscar3	before	11,599,131	23,988,686	4,344,247	11,855,944	2,672,206	5,836,725	1,864,403	4,607,315
	after	7,827,737	12,776,542	2,222,450	4,598,353	1,347,442	2,338,487	945,320	1,804,411

Table 2. Triples analysis from MEDLINE

3.4 Ontology Model for Interrelating Small Molecules and Biological Entities

The relationship between the chemical entity and its bioactivity which is already used in ChEBI is *has_role*.

Based on our analysis of bioactivity phrases in the literature, we have identified macromolecules and biological processes as the most common types of targets for the bioactivity of small molecules. We could therefore introduce a *has_target* relationship to relate a bioactivity description to either a macromolecule or a biological process. However, strictly speaking, the range of the *has_target* relationship should be restricted to those entities with which the chemical entity can physically interact – macromolecules. We can assume that biological processes are mentioned where the exact macromolecular target is unknown. In the same way, anatomical or subcellular locations may be mentioned when the exact target is unknown. Therefore, we can further formalise the *has_target* relation link to processes: in this case the target is a *macromolecule* **and** *participant_of* **some** *Process* (Manchester syntax).

Examples:

m1 is a *betaadrenergic receptor inhibitor*:
m1 **subclassOf** *bearer_of* **some**
(*realized_by* **only**
(*Inhibition* **and**
(*has_target* **some** *BetaAdrenergicReceptor*)))

m2 is a *mitosis stimulator*:
m2 **subclassOf** *bearer_of* **some**
(*realized_by* **only**
(*Stimulation* **and**
(*has_target* **some**
(*participant_of some Mitosis*))))

m3 is a *thyroid stimulator*:
m3 **subclassOf** *bearer_of* **some**
(*realized_by* **only**
(*Stimulation* **and**
(*has_target* **some**
(*has_locus* **some** *ThyroidGland*))))

m4 is a *mouse thyroid stimulator*:
m4 **subclassOf** *bearer_of* **some**
(*realized_by* **only**
(*Stimulation* **and**
(*has_target* **some**
(*has_locus* **some** (*ThyroidGland* **and** *part_of* **some** *Mouse*)))))

We have noted that organisms, organs and bodily parts appear frequently as contextual, locational modifiers for the bioactivity descriptions in the literature. In these cases, the above formalism is too strict, since the location is assumed to contribute to the definition of the bioactivity. We therefore introduce an additional relationship, *has_context*, which may hold between a bioactivity description and an organism, bodily organ or component to express *non-definitional* information: the bioactivity *can* take place in many organisms, but was *discovered* through investigations in one specific organism.

An important limitation of Description Logic-based ontology representation formalisms is that they are unable to elegantly express the fact that the context applies not to a bioactivity description *per se*, but rather to a small molecule-bioactivity association. This would require a ternary relationship. However, for our purposes it will be sufficient to assume that we can get around this problem through the standard method of reification.

Finally, we note that the different head nouns used in our analysis (inhibitor, antagonist and so on) correspond to different types of bioactivity, such as are delineated by upper-level distinctions in the ChEBI role ontology.

4 Conclusions

We have presented a language model for bioactivity descriptions which we have used to examine the distribution of bioactivity descriptions in the scientific literature. From this analysis we derive insights into the model needed to accurately formalise an ontology for bioactivity, appropriately distinguishing between bioactivity targets and contextual (locational) information. Such an ontology will serve as a bridge between small molecules, their biological targets, and the locations and contexts in which they act, allowing automated reasoning about the activities of chemical entities in a biological context.

This work should be understood as a first step in the direction of such a formalisation, a pressing goal in the context of ChEBI's participation in the OBO Foundry effort to interrelate ontologies in the biomedical domain. Future work will develop our text analysis platform further as a support utility

for ChEBI curation, and aim to incorporate the increased formalisation described here directly into the ChEBI ontology. Since ChEBI is a manually curated resource, we cannot prepopulate ChEBI with extracted relationships based on the text mining methods described here. However, such automatically identified bioactivity descriptions in the literature can be used to provide semantically enriched information in our ontology curation workbench, which allows a much improved and more rapid curation experience.

References

1. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucl. Acids Res. 32 (suppl 1), D267–270 (Jan 2004).
2. Corbett, P., Murray-Rust, P.: High-Throughput Identification of Chemistry in Life Science Texts, Lecture Notes in Computer Science, vol. 4216, chap. 11, pp. 107–118. Springer, Berlin (2006).
3. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., Mcnaught, A., Alcántara, R., Darsow, M., Guedj, M., Ashburner, M.: ChEBI: a database and ontology for chemical entities of biological interest. Nucl Acids Res 36(suppl 1), D344–D350 (2008).
4. Gohlke, H., Klebe, G.: Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew. Chem. Int. Ed 41, 2644–2676 (2002)
5. Hettne, K.M., Stierum, R.H., Schuemie, M.J., Hendriksen, P.J.M., Schijvenaars, B.J.A., Mulligen, E.M., Kleinjans, J., Kors, J.A.: A dictionary to identify small molecules and drugs in free text. Bioinformatics 25(22), 2983–2991 (Nov 2009).
6. Hishiki, T., Ogasawara, O., Tsuruoka, Y., Okubo, K.: Indexing anatomical concepts to omim clinical synopsis using umls metathesaurus. In Silico Biology 4 (2003)
7. Hobbs, J.R., Appelt, D.E., Bear, J., Israel, D.J., Kameyama, M., Stickel, M.E., Tyson, M.: Fastus: A cascaded finite-state transducer for extracting information from natural-language text. CoRR cmp-lg/9705013 (1997)
8. Kirsch, H., Gaudan, S., Rebholz-Schuhmann, D.: Distributed modules for text annotation and IE applied to the biomedical domain. International Journal of Medical Informatics In Press.
9. Lipinski, C., Hopkins, A.: Navigating chemical space for biology and medicine. Nature 432 (2004)
10. Magrane, M., Consortium, U.: UniProt knowledgebase: a hub of integrated protein data. Database: the journal of biological databases and curation 2011 (Mar 2011).
11. Morgan, A., Lu, Z., Wang, X., Cohen, A., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., et al.: Overview of biocreative ii gene normalization. Genome biology 9 (Suppl 2), S3 (2008)
12. Oprea, T.I., Tropsha, A.: Target, chemical and bioactivity databases – integration is key. Drug Discovery Today: Technologies 3, 357–365 (2006)
13. Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., Yepes, A.J.: Text processing through Web services: calling Whatizit. Bioinformatics (Oxford, England) 24(2), 296–298 (Jan 2008).
14. Rebholz-Schuhmann, D., Jimeno-Yepes, A., Arregui, M., Kirsch, H.: Measuring prediction capacity of individual verbs for the identification of protein interactions. Journal of biomedical informatics (Oct 2009).
15. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Gold-berg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 25(11), 1251–1255 (Nov 2007).
16. Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T.A., Rapp, B.A.: Database resources of the NCBI. Nucl acids res 28(1), 10–14 (Jan 2000).