Metabolomics
DOI 10.1007/s11306-010-0251-6

ORIGINAL ARTICLE

Mining metabolites: extracting the yeast metabolome from the literature

Chikashi Nobata · Paul D. Dobson · Syed A. Iqbal · Pedro Mendes · Jun'ichi Tsujii · Douglas B. Kell · Sophia Ananiadou

Received: 18 June 2010 / Accepted: 12 October 2010

Abstract Text mining methods have added considerably to our capacity to extract biological knowledge from the literature. Recently the field of systems biology has begun to model and simulate metabolic networks, requiring knowledge of the set of molecules involved. While genomics and proteomics technologies are able to supply the macromolecular parts list, the metabolites are less easily assembled. Most metabolites are known and reported through the scientific literature, rather than through large-scale experimental surveys. Thus it is important to recover them from the literature. Here we present a novel tool to automatically identify metabolite names in the literature, and associate structures where possible, to define the reported yeast metabolome. With ten-fold cross validation on a manually annotated corpus, our recognition tool generates an *f*-score of 78.49 (precision of 83.02) and demonstrates greater suitability in identifying metabolite names than other existing recognition tools for general chemical molecules. The metabolite recognition tool has been applied to the literature covering an important model organism, the yeast *Saccharomyces cerevisiae*, to define its reported metabolome. By coupling to ChemSpider, a major chemical database, we have identified structures for much of the reported metabolome and, where structure identification fails, been able to suggest extensions to ChemSpider. Our manually annotated gold-standard data on 296 abstracts are available as supplementary materials. Metabolite names and, where appropriate, structures are also available as supplementary materials.

Keywords Text mining · Named entity recognition · Yeast metabolome

1 Introduction

Modern molecular biology is a science dominated by very large quantities of data, yet most useful knowledge remains locked in the scientific literature. Though large this

Electronic supplementary material The online version of this article (doi:10.1007/s11306-010-0251-6) contains supplementary material, which is available to authorized users.

C. Nobata · P. Mendes · J. Tsujii · S. Ananiadou
School of Computer Science, The University of Manchester, Oxford Road, Manchester, UK

C. Nobata · S. A. Iqbal · J. Tsujii · S. Ananiadou
National Centre for Text Mining (NaCTeM), Manchester Interdisciplinary Biocentre (MIB), Manchester, UK

P. D. Dobson · D. B. Kell
School of Chemistry, The University of Manchester, Oxford Road, Manchester, UK

S. A. Iqbal
Plastic and Reconstructive Surgery Research (PRSR), Manchester Interdisciplinary Biocentre (MIB), Manchester, UK

P. Mendes
Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA

J. Tsujii
Department of Computer Science, University of Tokyo, Tokyo, Japan

C. Nobata (✉)
1.001 Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester M1 7DN, UK
e-mail: chikashi.nobata@manchester.ac.uk

Published online: 31 October 2010

literature can be efficiently explored through text mining, the application of computational methods to identify and extract entities and their relationships from text (Ananiadou and McNaught 2006). Already there are many text mining services for biology that enrich papers with semantic annotations for richer querying and also to extract relations between annotated entities. To illustrate some of these to the uninitiated, in current biological text mining it is possible to identify proteins within text (Rebholz-Schuhmann et al. 2007; Nobata et al. 2008), pull out their physical interactions (Miyao et al. 2009) and associations with disease states, phenotypes and other terms (Hoffmann and Valencia 2005; Tsuruoka et al. 2008). One can also resolve biological abbreviations (Okazaki et al. 2010), resolve species ambiguity (Wang et al. 2010), or make semantically rich queries over the literature ("what activates p53?" being a more meaningful search than simply "p53 activation") (Miyao et al. 2006). For most applications leading methods compare favourably to expert annotators, but of course can be applied on a much larger scale, which is simple using workflow systems (Kano et al. 2009, 2010). Indeed, owing to the increasing rate of scientific publication it is clear that increased automation through text mining is the only way to reach a useful understanding of the biological literature (Ananiadou et al. 2006, 2010).

In the post-genomic era we are beginning to be able to properly consider molecular biology as the integrated system it evidently is through the burgeoning discipline of systems biology (Kell 2009). Underpinning systems biology are ideas connecting data-rich experimental approaches and computational simulations (Mendes et al. 2009) of the underlying biochemistry to move toward ever more accurate depictions of how life operates at the molecular level. It is therefore necessary to understand the network of interactions and reactions that occur in the cell, most usefully in a standardized format such as SBML (Hucka et al. 2003, 2004). Such reconstructions have benefitted greatly from genome-driven identification of the metabolic enzymes and transporters that constitute the macromolecular 'parts list' of metabolism. The remaining molecular species required are the small endogenous molecules of the metabolome. While increasingly identified by high-throughput experiments, most knowledge of metabolites and their reactions is primarily reported in the scientific literature. Considerable manual efforts have extracted metabolite information from papers into metabolite databases such as HMDB (Wishart et al. 2009) and reaction databases such as KEGG (Kanehisa and Goto 2000; Kanehisa et al. 2006, 2010) and BioCyc (Karp et al. 2005) and major community efforts have led to robust and well-annotated reaction networks defined in SBML. It remains the case, however, that the very large literature around our organism of interest, the yeast *Saccharomyces cerevisiae*, still harbors many uncaptured metabolites. Through the application of methods of text mining and cheminformatics we have addressed this issue to detect and structurally identify novel metabolites to move toward defining the reported yeast metabolome for consideration in future metabolic network reconstructions.

Few small molecule resources are limited solely to metabolites (much less only to yeast metabolites) owing to a lack of consensus among biologists upon the definition of metabolism. Without a strict definition it is inevitable that drug, nutrient and other molecules at the boundary of metabolism are arbitrarily included or excluded from different metabolite resources according to idiosyncrasies of database requirements or annotator interpretation. Our challenge here is not to remedy this by the imposition of a particular definition of metabolism but to replicate the standard of manually-curated databases without too much focus upon the ill-defined region. In utilizing text mining approaches to emulate the somewhat subjective and rather ill-defined biologist's concept of metabolism the challenge is different to related work on chemical text mining (Banville 2006; Corbett and Copestake 2008; Jiao and Wild 2009; Pirkola 2008; Townsend et al. 2004; Wilbur et al. 1999; Wren 2006), the scope of which, by contrast, is rather easier to define and mark up in documents. The broader scope of these tools leads to the identification of many chemicals that are not metabolites. While they are of relevance it is the narrower focus of the metabolite mining task that precludes the direct application of existing chemical name recognition tools.

The identification of metabolite names within the scientific literature by text mining is only one part of the problem. Once a metabolite name is found it is most useful to identify its molecular structure. Various mechanisms for converting chemical names into structures exist. Some rely upon systematic naming conventions (for example, IUPAC nomenclature) that can be interpreted (Corbett and Murray-Rust 2006; Eller 2006; Klinger et al. 2008), yet metabolites do not tend to follow rigorously such approaches, while others are primarily underpinned by large chemical dictionaries (Hettne et al. 2009; Klinger et al. 2008; Brecher 1999; Goebels et al. 1991; Wisniewski 1990). Given the tendency for non-standard naming of metabolites, plus the size and machine accessibility through web services of ChemSpider (2007) it forms the basis of our name to structure conversions.

Detecting metabolites in text is here posed as a Named Entity Recognition (NER) task. NER is a technique that finds the boundary and the semantic category of specific terms in text. NER was originally defined for information extraction from news-wire articles (MUC6 1995), in which NE categories are proper nouns (such as person names and location names) and numeral expressions (such as date and the amount of money). The same technique has also been applied in the biomedical domain to annotate protein, gene or

organism names (Takeuchi and Collier 2002; Kim and Tsujii 2002; Kazama et al. 2002; Kim et al. 2003, 2004; Tsuruoka and Tsujii 2004; Cohen 2005; Finkel et al. 2005; Hirschman et al. 2005; Krallinger et al. 2005). In this paper, we report the manually annotated corpus we created for metabolite NER and the evaluation results of the NER system trained on the corpus. We also apply the NER system to a large set of unannotated abstracts from the yeast literature to extract novel metabolites. These entities are then mapped by name to ChemSpider to resolve their structures and move toward definition of the reported yeast metabolome.

2 Materials and methods

2.1 Construction of manually annotated data

Two domain experts (Annotator A and B) annotated metabolite expressions in the MEDLINE (2007) abstracts. The target documents are 296 MEDLINE abstracts included in the version 1 of the yeast metabolic network reconstruction (Herrgård et al. 2008). Each domain expert annotated metabolite and also enzyme names in the abstracts. The annotations were restricted to only those names that appear in the context of metabolic pathways. For example, in the sentence "glucose is an economically important chemical in the food industry" the role of glucose is not as a metabolite. When a metabolite name appears as a part of an enzyme name, the metabolite part is not annotated (e.g. "diadenosine hexaphosphate hydrolases" is annotated as enzyme, and the part of "diadenosine hexaphosphate" is not annotated as metabolite).

In this work we have focused on the annotation of metabolite names and therefore the enzyme annotations were used only to exclude a metabolite name when part of an enzyme name. The gold-standard (consensus) data are created by integrating these two manual annotations.[1] The evaluation results of two manual annotations compared to the gold-standard data are shown in Table 1. The *f*-score[2] of the data is 88.49 (Annotator A), 78.35 (Annotator B). The difference in annotations between Annotators was mainly due to Annotator B annotating more compared to Annotator A. This also demonstrates the ambiguity that can arise in annotating metabolites, as defining "a metabolite" is itself inherently difficult.

Table 1 Evaluation of manually annotated data

Data	# Metabolites	Recall	Precision	*f*-score
Annotation A	1743	85.86	91.28	88.49
Annotation B	1986	81.17	75.73	78.35
Gold-standard	1853	–	–	–

Table 2 Numbers of entries used in dictionary-based NER

DB	Types	Terms
Yeast consensus v.1	664	2,748
ENZYME	4,905	20,566

2.2 Methods of named entity recognition

Our method of recognizing NEs is similar to the system described in (Sasaki et al. 2008), which consists of two components. The first part, dictionary-based tagging, finds candidates for entities using a dictionary, and the second part is a supervised method trained with the results of dictionary-based NER and manually annotated data.

The first part uses dictionary information of metabolites from the yeast consensus metabolic reconstruction, and annotates metabolite names included in the abstracts. It also uses enzyme term lists (Bairoch 2000) so that the system can ignore metabolite names as a part of enzyme names (Table 2).[3]

The second part, statistical sequential labeling is a supervised method with manually annotated data. The module uses results of dictionary-based NER as well as word, orthographic and Part-of-Speech information as features to predict the NE labels. We also added the results of the dictionary lookup with ChEBI (Degtyarenko et al. 2008) and HMDB (Wishart et al. 2009) data as one of features. Table 3 shows the statistics of entries obtained from the databases used in annotating metabolites.

We use an open-source morphological analyzer Mecab (2008) as a POS tagger. Word features are the surface form of the word and the postfixes (the last two and four letters of the word). Orthographic features are the first letter and the last four letters of the word form, in which capital letters in a word are normalized to "A", lower case letters

[1] Both annotators A and B discussed and checked the gold-standard data. The annotator A is senior to annotator B in terms of experience of annotation and years in biochemistry and therefore made the final decision.

[2] The metrics are derived as follows from the so-called confusion matrix described in (Broadhurst and Kell 2006):

$\text{Recall (R)} = \text{TP}/(\text{TP} + \text{FN})$

$\text{Precision (P)} = \text{TP}/(\text{TP} + \text{FP})$

$\text{F-Score} = 2 * \text{P} * \text{R}/(\text{P} + \text{R})$ (Harmonic mean of P and R)

where TP (true positive) is the number of correct entities that are also annotated in the results, FP (false positive) the number of wrongly annotated entities, and FN (false negative) is the number of correct entities that are not annotated.

[3] To filter out some ambiguous terms that frequently appear we used a word list (all. 10–20 list) included in Spell Checking Oriented Word Lists (SCOWL) (Atkinson 2004), which is used in a spell checker program (GNU Aspell 2004).

Table 3 Numbers of entries extracted from databases

DB	Types	Terms
HMDB	7,983	76,191
ChEBI	454,455	529,189

Table 4 Evaluation of experimental results with the test corpus

System	Recall	Precision	*f*-score
Dict-NER	59.42	68.56	63.66
CRF-NER	74.42	83.02	78.49

are normalized to "a", and digits are replaced by "0", e.g., the word form of IL-2 is AA-0.

The NE labels adopts IOB2 format (Sang and Veenstra 1999), *i.e.* the first token of the target sequence is labeled with "B" of "Beginning" (e.g. B-metabolite), the intermediate and the last tokens in the target sequence are labeled with "I" of "Intermediate" (e.g. I-metabolite) and other tokens are labeled just as "O" of "Others". For instance, the sequence "7-keto-8-aminopelargonic acid" is annotated as "B-metabolite I-metabolite I-metabolite". Models of Conditional Random Fields (CRFs) (Lafferty et al. 2001) are used to predict the IOB2 labels with the above features.[4]

2.3 Cheminformatics workflows

All cheminformatics workflows were implemented in Pipeline Pilot (Accelrys, San Diego, CA). ChemSpider searches and retrieval utilized the ChemSpider search web service.[5] Clustering of non-unique hits was performed using the 'Cluster Molecules' component in a connectivity fingerprint space (FCFP4) at a maximum Tanimoto distance of 0.15 (Dobson et al. 2009). Lists of molecules were collapsed using the 'Merge molecules' component that acts upon canonical SMILES representations of molecular structure.

3 Results and discussion

3.1 Evaluation of the NER system on a test corpus

We conducted an experiment to evaluate our NER system with a manually annotated corpus. As stated before, the corpus contains 296 abstracts included in the yeast consensus metabolic reconstruction with annotations of metabolites and enzymes. Enzyme annotations are only used to select proper metabolite names; the system does not annotate enzyme names. Table 4 shows the evaluation results of the annotation. The dictionary-based NER system (Dict-NER) identifies metabolites purely by reference to dictionaries, whereas the supervised NER system (CRF-NER) extends this through learning from linguistic cues using the annotated corpus. The CRF-NER system is evaluated with[6] ten-fold cross validation. By applying CRF to the result of dictionary-based NER, the system was able to improve the *f*-score from 63.66 to 78.49.

We also compared evaluation results to those available through Whatizit pipelines (Rebholz-Schuhmann et al. 2007). Whatizit is a text processing system that identifies molecular biology terms in text. Whatizit pipelines relevant to metabolite annotations are three pipelines that annotate chemical entities (whatizitOscar3, whatizitChebiDict, and whatizitChemicals). The whatizitOscar3 pipeline provides annotations of chemical entities by Oscar3 (Corbett and Murray-Rust 2006; Batchelor and Corbett 2007), and whatizitChebiDict provides annotations of ChEBI entities based on a dictionary approach. The whatizitChemicals pipeline contains annotations of both whatizitOscar3 and whatizitChebiDict pipelines as well as Drugs and Protein names. The evaluation results are shown in Table 5.

For the results from Oscar3, we compared entities annotated as CM (Chemical Molecules) and as ONT (ontology terms) with the test corpus, and ignored other categories (i.e. CPR, RN, ASE) because they don't include metabolite annotations. For the results from chemicals, we showed two results. The best precision and *f*-score are achieved when only the CM and Drug names are used in evaluation (CM, DRUG). The best recall is achieved when all annotated categories except for Oscar 3's ignored categories are used in evaluation (CM, ONT, DRUG, PROTEIN).

We see that though their recalls are higher than our system, but their precisions and *f*-scores are lower than our system because of annotations of non-metabolites. Because these pipelines intend to annotate general chemical entities, the annotation results include more false positives than our system for metabolite annotations.

For example, underlined expressions in the following sentence are annotated with the whatizitChemicals pipeline, but these are not suitable as annotation for metabolite names: The YJR019C product is highly similar to *tesB*, a bacterial *acyl-CoA thioesterase*, and carries a *tripeptide*

[4] We use the CRF++ (2003) toolkit to acquire the CRF model for NER.

[5] http://www.chemspider.com/Search.asmx?WSDL.

[6] Ten-fold cross validation is a method to evaluate the system. First the data are split randomly into ten parts. nine parts are then used as training data, and the remaining part used as testing data. This procedure is repeated ten times so that each part is used as testing data. In this experiment, all results are gathered and compared with gold-standard data to evaluate as if all abstracts are one large document.

Table 5 Evaluation of Whatizit results with the test corpus

Whatizit pipeline	Recall	Precision	F-score
Oscar3 (CM)	85.97	37.65	52.37
Oscar3 (CM, ONT)	87.48	26.91	41.16
ChebiDict	76.96	31.53	44.73
Chemicals (CM, DRUG)	82.89	42.92	56.55
Chemicals (CM, ONT, DRUG, PROTEIN)	88.88	19.84	32.43

sequence, *alanine-lysine-phenylalanineCOOH*, that closely resembles the consensus *sequence* for type-1 peroxisomal targeting signals.

3.2 Name to structure via ChemSpider

The metabolite recognition tool was applied to the corpus covering *S. cerevisiae* (Gene Literature 2010) taken from Saccharomyces Genome Database (SGD) (2010), which contains about 53,000 MEDLINE abstracts. The probability generated from the CRF model is attached to each entity to indicate its most plausible annotation. If the same entity is annotated more than once in the set of documents its highest probability is assigned. Results are summarized in Table 6. The tool identifies 4,326 unique metabolite names. 2,441 (56%) of these are known from the dictionaries used in training. 1,885 (44%) are potentially novel metabolites not found in the dictionaries. All names were searched against the chemical database ChemSpider to identify appropriate structures. Of the 1,885 potentially novel metabolites ChemSpider searches return one or more hits for 1,245 names, the service fails on 7 names, and does not match any record for 633 names (33% of novel names). These 633 names therefore represent molecules that are not named in ChemSpider (as of 29 March 2010) and can be found in supplementary table S1. From the training dictionaries 2,441 names are detected. ChemSpider searches return one or more hits for 2,338 of these names; the service fails once, and does not match any record for 102 names. These 102 names can be found in supplementary table S2.

Table 6 Summary of name to structure efforts. 'All names' is the number of names detected by NER, including duplicates

All names	80650
Unique names	4326
Known names	2441
Potentially novel names	1885
No matches	735
Mismatches	1118
Unique matches	1887
Near matches	528
Service failures	7
Unique structures	1435

'Unique names' is this set with duplicates removed. 'No matches' is the set of names that do not match any ChemSpider record, 'mismatches' are those matches that appear to be incorrect (as judged by structurally clustering hits). Structural information is associated through unique (only one match) and near matches (matching a set of related molecules) with 2,415 names (55% of all unique names). Removing structural redundancy from the 1,887 unique matches yields 1,435 structures. (Rows 5–7 do not sum to 4,326 as some names match the same ChemSpider record)

As is to be expected, the majority of names from the dictionaries (96%) match some record in ChemSpider, unlike one-third of novel names that do not. Of the 1,245 novel and 2,338 dictionary matches against ChemSpider, 1,228 and 2,305 structural records were retrieved (with web service failures accounting for the remainder). 620 (50% of 1228) novel and 1,267 (55% of 2305) names match one and only one structure. Collapsing these down to remove structural redundancy (where identical structures are merged into one record on the basis of their canonical SMILES representation) the 620 novel names become 528 unique structures, and the 1,267 dictionary names become 1,003 unique structures. Removing redundancy from the union of both sets yields 1,435 unique chemical structures. These can be found in supplementary table S3.

Searches can hit multiple structures owing to duplicated synonyms on multiple ChemSpider records (for example, 'glucose' matches structures that only differ in the detail to which stereochemistry is specified), or because upon failing to match exactly the search service breaks the query string into parts for approximate matching and retrieves multiple inappropriate structures that only correspond to parts of the name. In the former case minor structural variants (charge, stereochemistry, and so on) are likely and not overly difficult to reconcile manually, whereas the latter case is not useful in the context of name to structure conversion. To discriminate between these scenarios the results are clustered by molecular structure (maximum cluster Tanimoto distance of 0.15 in FCFP4 space). 608 novel and 1,038 dictionary non-unique matches form one and only one tight cluster 139 and 389 times respectively. The remaining 469 and 649 names with hits that form more than one tight cluster are likely ChemSpider mismatches that may further extend the database in the same way as names that do not match at all. These incorrect structural matches are reported in tables S5 (dictionary names) and S6 (novel names). All 528 queries that generate only one cluster are reported with structures and ChemSpider identifiers in supplementary table S4.

3.3 Discussion

Metabolic network reconstructions for systems biology require knowledge of the list of 'molecular parts' involved.

Macromolecular components can largely be elucidated from the genome, yet the endogenous small molecules remain mostly understood through the scientific literature. Extraction of metabolites from the literature is an arduous task for a well-studied organism such as *S. cerevisiae* on which well over 100,000 papers have been published. Text mining approaches are absolutely required.

The task of metabolite recognition is inherently difficult as there is no real consensus among biologists on the definition of metabolites and metabolism. Despite this metabolic databases, even though potentially polluted with non-metabolites or liable to omit real metabolites, are of tremendous value in metabolomics and in guiding metabolic network reconstructions. Here we have constructed our metabolite recognition tool to emulate how these valuable databases are constructed by allowing the slack definition to remain.

The gold-standard corpus of 296 MEDLINE abstracts, created by two independent annotators working to agreed guidelines, contains 1,853 annotations. The corpus is provided as stand-off annotations in supplementary material S7. Agreement between annotators is slightly lower than for similar NER tasks, largely due to ambiguity in the definition of metabolites. Expert comparable performance on this corpus generates an *f*-score of 78.35 (the lower annotator's *f*-score). The metabolite NER reported here achieves slightly better with an *f*-score of 78.49.

It is not possible to draw very close comparisons to competing methods as the NER method is the first of its kind to focus specifically upon recognition of endogenous metabolites. There are, however, chemical recognition tools to detect any type of small molecule in text. Being more general means such tools tend to out-perform our method by recall but are considerably bettered in terms of precision. For example, as a metabolite recognition tool (which it is not) the Whatizit Chemicals pipeline 'erroneously' identifies 1,695 chemicals that are not metabolites in our corpus. Here the benefit of a metabolite-focused NER becomes apparent as a far more focused and manageable 220 false positives are generated, many of which, on closer analysis, appear to be attributable to misannotations (particularly where the definition of metabolism is problematic). In the sentence "<metabolite>Thioredoxin</metabolite> has been implicated in the reduction of PAPS in Saccharomyces cerevisiae.", for example, 'thioredoxin' has not been annotated as a metabojlite, presumably as it is a protein encoded in the genome, but in many ways it is functionally more akin to metabolites and, by some definitions, might be considered as such. Curation of metabolite NER false positives will be a more efficient way to extend metabolite dictionaries than general chemical recognition tools.

The metabolite recognition tool was applied to the abstracts of yeast papers collated by SGD to identify the reported yeast metabolome. Owing to its size and high level of curation the ChemSpider database formed the basis of efforts to annotate names with structural information. 735 identified names do not match any record and 1,118 names match improperly. These 1,853 names represent possible extensions to ChemSpider. 1,887 names match one and only ChemSpider record. 528 names match sets of very closely related structures, most likely due to similar structures improperly sharing synonyms, and require further curation to identify the correct form. In total some level of structural annotation has been attached to 2,415 names. The 1,887 uniquely-matched names collapse to 1,435 structures. Although 55% of unique names could be associated with some structural information, it is clear from NER performance on the manually annotated corpus and examination of names not matched by ChemSpider that many unmatched names are real yeast metabolites. To augment the outstanding 45% of names with structural information, improved name to structure methods are required. Given the non-standard nature of many metabolite names the most successful approach is likely to be based on extending ChemSpider (and similar databases) rather than algorithm-driven. Extensions will need to cover more metabolites and improve synonym listings. The ability to automatically mine metabolites from the literature in a robust and discriminating fashion is essential to this problem of efficiently extending metabolite databases and continuing to improve metabolic network reconstructions.

4 Concluding remarks

The author have created an NER system for metabolites using term lists from the yeast consensus metabolite reconstruction and trained with annotation data that we manually created from 296 MEDLINE abstracts. Our NER system generates an *f*-score of 78.49 with ten-fold cross validation, which is comparable to the lower annotator's *f*-score 78.35. We have also applied our NER to about 53,000 MEDLINE abstract corpus covering *S. cerevisiae*, and the recognized names are searched against ChemSpider, a major chemical database, to identify appropriate structures. We have identified structures for 55% of unique names (2,415/4,326), and also found many real yeast metabolites among unmatched names, which are good candidates to extend metabolite databases. Defining the reported yeast metabolome has created a useful resource for the yeast and metabolomics communities. It is anticipated that the metabolite NER will also be of value for other organisms, although should be used with caution. An important future direction is to automatically recognize reactions and metabolic pathways described in documents,

eventually moving toward a fully-automated network reconstruction platform for systems biology. NER for metabolites is a vital step toward this.

Acknowledgments The authors are grateful to Muhammad Imran for creating annotation data for our experiment.

Funding Biotechnology and Biological Sciences Research Council (grant code BBS/B/13640 and BB/F006039/1).

Conflict of interest None.

References

Ananiadou, S., Kell, D. B., & Tsujii, J. (2006). Text mining and its potential applications in systems biology. *Trends in Biotechnology, 24*(12), 571–579.

Ananiadou, S., & McNaught, J. (2006). *Text mining for biology and biomedicine*. City: Artech House.

Ananiadou, S., Pyysalo, S., Tsujii, J., & Kell, D. B. (2010). Event extraction for systems biology by text mining the literature. *Trends in Biotechnology, 28*, 381–390.

Atkinson, K. (2004). Spell checking oriented word lists (SCOWL). Available at http://wordlist.sourceforge.net/.

Bairoch, A. (2000). The ENZYME database in 2000. Nucleic Acids Res 28, 304–305. Available at http://www.expasy.org/enzyme/.

Banville, D. (2006). Mining chemical structural information from the drug literature. *Drug Discovery Today, 11*(1–2), 35–42.

Batchelor, C. R., & Corbett, P. T. (2007). *Semantic enrichment of journal articles using chemical named entity recognition* (pp. 45–48). Prague: Proceedings of the ACL 2007 Demo and Poster Sessions.

Brecher, J. (1999). Name = Struct: A practical approach to the sorry state of real-life chemical nomenclature. *Journal of Chemical Information and Computer Sciences, 39*(6), 943–950.

Broadhurst, D., & Kell, D. B. (2006). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics, 2*, 171–196.

ChemSpider (2007). Available at http://www.chemspider.com.

Cohen, A. (2005). Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. In *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics. Association for Computational Linguistics*, pp. 17–24.

Corbett, P., & Copestake, A. ((2008)). Cascaded classifiers for confidence-based chemical named entity recognition. *BMC Bioinformatics, 9*(Suppl 11), S4.

Corbett, P. & Murray-Rust, P. (2006). High-throughput identification of chemistry in life science texts. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in Bioinformatics). *LNBI, 4216*, 107–118.

CRF++ (2003). Available at http://crfpp.sourceforge.net/.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., et al. (2008). ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Research, 36*, D344–D350.

Dobson, P. D., Patel, Y., & Kell, D. B. (2009). 'Metabolite-likeness' as a criterion in the design and selection of pharmaceutical drug libraries. *Drug Discovery Today, 14*(1–2), 31–40.

Eller, G. A. (2006). Improving the quality of published chemical names with nomenclature software. *Molecules, 11*(11), 915–928.

Finkel, J., Dingare, S., Manning, C., Nissim, M., Alex, B., & Grover, C. ((2005)). Exploring the boundaries: gene and protein identification in biomedical text. *BMC Bioinformatics, 6*(Suppl 1), S5.

Gene Literature (2010). Retrieved from Feb 15, 2010. http://downloads.yeastgenome.org/literature curation/gene literature.tab GNU Aspell (2004). Available at http://aspell.net.

Goebels, L., Lawson, A., & Wisniewski, J. (1991). AUTONOM: System for computer translation of structural diagrams into IUPAC-compatible names. 2. Nomenclature of chains and rings. *Journal of Chemical Information and Computer Sciences, 31*(2), 216–225.

Herrgård, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M., Büthgen, N., Borger, S., Costenoble, R., Heinemann, M., Hucka, M., Novère, N. L., Li, P., Liebermeister, W., Mo, M. L., Oliveira, A. P., Petranovic, D., Pettifer, S., Simeonidis, E., Smallbone, K., Spasíc, I.,Weichart, D., Brent, R., Broomhead, D. S., Westerhoff, H. V., Kürdar, B., Penttilä, M., Klipp, E., Palsson, B. Ø., Sauer, U., Oliver, S. G., Mendes, P., Nielsen, J. & Kell, D. B. (2008). A consensus yeast metabolic reconstruction obtained from a community approach to systems biology. *Nature Biotechnology 26,* 1155–1160. Available at http://www.comp-sys-bio.org/yeastnet/.

Hettne, K. M., Stierum, R. H., Schuemie, M. J., Hendriksen, P. J. M., Schijvenaars, B. J. A., van Mulligen, E. M., et al. (2009). A dictionary to identify small molecules and drugs in free text. *Bioinformatics, 25*(22), 2983–2991.

Hirschman, L., Yeh, A., Blaschke, C. & Valencia, A. (2005). Overview of BioCreAtIvE: critical assement of information extraction for biology. *BMC Bioinformatics, 6* (Suppl 1), S1.

Hoffmann, R., & Valencia, A. (2005). Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics, 21*(suppl. 2), ii252–ii258.

Hucka, M., Finney, A., Bornstein, B. J., Keating, S. M., Shapiro, B. E., Matthews, J., et al. (2004). Evolving a lingua franca and associated software infrastructure for computational systems biology: The Systems Biology Markup Language (SBML) project. *System Biology (Stevenage), 1*(1), 41–53.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J.-H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Le Novère, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J. & Wang, J. (2003). The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics, 19*(4), 524–531. Available at http://sbml.org.

Jiao, D., & Wild, D. (2009). Extraction of cyp chemical interactions from biomedical literature using natural language processing methods. *Journal of Chemical Information and Modeling, 49*(2), 263–269.

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research, 28*, 27–30.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., & Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research, 38*, D355–D360.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K., Itoh, M., Kawashima, S., et al. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research, 34*, D354–D357.

Kano, Y., Baumgartner, W. A., Jr., McCrohon, L., Ananiadou, S., Cohen, K. B., Hunter, L., et al. (2009). U-Compare: Share and compare text mining tools with UIMA. *Bioinformatics, 25*(15), 1997–1998.

Kano, Y., Dobson, P., Nakanishi, M., Tsujii, J., & Ananiadou, S. (2010). Text mining meets workflow: linking U-compare with taverna. *Bioinformatics, 26*(19), 2486–2487.

Karp, P., Ouzounis, C., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., Tsoka, S., Darzentas, N., Kunin, V. & Lopez-Bigas, N. (2005). Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Research* 19, 6083–6089. Available at http://biocyc.org/.

Kazama, J., Makino, T., Ohta, Y. & Tsujii, J. (2002). Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 Workshop on Natular Language Processing in the Biomedical Domain*. pp. 1–8.

Kell, D. B. (2009). Journal club. A systems biologist ponders how disparate ideas can sometimes come together beautifully. *Nature, 460*(7256), 669.

Kim, J. D., Ohta, T., Tateisi, Y., & Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for biotextmining. *Bioinformatics, 19*(Suppl. 1), 180–182.

Kim, J., Ohta, T., Tsuruoka, Y., Tateisi, Y.& Collier, N. (2004). Introduction to the bio-entity recognition task at JNLPBA. In: Collier, N., Ruch, P. and Nazarenko, A. (Eds.), *Proceedings of the International JointWorkshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA)*, Geneva, Switzerland. held in conjunction with COLING'2004 (pp. 70– 75).

Kim, J. D., & Tsujii, J. (2002). *Corpus-based approach to biological entity recognition*. In: Text data mining SIG (ISMB2002).

Klinger, R., Kolarik, C., Fluck, J., Hofmann-Apitius, M., & Friedrich, C. M. (2008). Detection of IUPAC and IUPAC-like chemical names. *Bioinformatics, 24*(13), i268–i276.

Krallinger, M., Morgan, A., Smith, L., Leitner, F., Tanabe, L., Wilbur, J., et al. (2005). Overview of BioCreative: Critical assement of information extraction for biology. *BMC Bioinformatics, 6*(Suppl 1), S1.

Lafferty, J. D., McCallum, A. & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of International Conference on Machine Learning(ICML)*. pp. 282–289.

Mecab (2008). Available at http://mecab.sourceforge.net/.

MEDLINE (2007). Available at http://www.pubmed.gov/.

Mendes, P., Hoops, S., Sahle, S., Gauges, R., Dada, J., & Kummer, U. (2009). Computational modeling of biochemical networks using COPASI. *Methods in Molecular Biology, 500*, 17–59.

Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T. & Tsujii, J. (2006). Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. In *Proceedings of COLINGACL 2006*. Sydney, Australia, pp. 1017–1024.

Miyao, Y., Sagae, K., Sætre, R., Matsuzaki, T., & Tsujii, J. (2009). Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics, 25*(3), 394–400.

MUC6, 1995. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann, Columbia, MD, USA.

Nobata, C., Cotter, P., Okazaki, N., Rea, B., Sasaki, Y., Tsuruoka, Y et.al. (2008). Kleio: a knowledge-enriched information retrieval system for biology. In *Proceedings of the 31st Annual International ACM SIGIR Conference*. Singapore, pp. 787–788.

Okazaki, N., Ananiadou, S., & Tsujii, J. (2010). Building a high quality sense inventory for improved abbreviation disambiguation. *In: Bioinformatics, 26*(9), 1246–1253.

Pirkola, A. (2008). Extracting variant forms of chemical names for information retrieval. Information Research 13 (3).

Rebholz-Schuhmann, D., Arregui, M., Gaudan, M., Kirsch, H. and Jimeno, A. (2007). Text processing through web services: Calling whatizit. Bioinformatics 24 (2), 296–298. Available at http://www.ebi.ac.uk/webservices/whatizit/ Saccharomyces Genome Database (2010). Available at http://www.yeastgenome.org.

Sang, E. F. T. K. & Veenstra, J. (1999). Representing text chunks. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-99)*. Bergen, pp. 173–179.

Sasaki, Y., Tsuruoka, Y., McNaught, J., & Ananiadou, S. (2008). How to make the most of NE dictionaries in statistical NER. *BMC Bioinformatics, 9*(Suppl 11), S5.

Takeuchi, K. & Collier, N. (2002). Use of support vector machines in extended named entity recognition. In *Proceedings of the 6th conference on natural language learning (CoNLL-2002)*. pp. 119–125.

Townsend, J., Adams, S., Waudby, C., De Souza, V., Goodman, J., & Murray-Rust, P. (2004). Chemical documents: Machine understanding and automated information extraction. *Organic and Biomolecular Chemistry, 2*(22), 3294–3300.

Tsuruoka, Y., & Tsujii, J. (2004). Improving the performance of dictionary-based approaches in protein name recognition. *Journal of Biomedical Informatics, 37*, 461–470.

Tsuruoka, Y., Tsujii, J., & Ananiadou, S. (2008). FACTA: A text search engine for finding associated biomedical concepts. *Bioinformatics, 24*(21), 2559–2560.

Wang, X., Tsujii, J., & Ananiadou, S. (2010). Disambiguating the species of biomedical named entities using natural language parsers. *Bioinformatics, 26*(5), 661–667.

Wilbur, W., Hazard Jr., G., Divita, G., Mork, J., Aronson, A. & Browne, A. (1999). Analysis of biomedical text for chemical names: a comparison of three methods. *Proceedings/AMIA. Annual Symposium. AMIA Symposium*, 176–180.

Wishart, D. S., Knox, C., Guo, A., Eisner, R., Young, N., Gautam, B., Hau, D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J., Lim, E., Sobsey, C., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H. & Forsythe, I. (2009). HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research 37(Database issue)*, D603–D610.

Wisniewski, J. (1990). AUTONOM: System for computer translation of structural diagrams into IUPAC-compatible names. 1. General design. *Journal of Chemical Information and Computer Sciences, 30*(3), 324–332.

Wren, J. (2006). A scalable machine-learning approach to recognize chemical names within large text databases. *BMC Bioinformatics, 7*(Suppl 2), S3.