An Exploration of Mining Gene Expression Mentions and their Anatomical Locations from Biomedical Text

Martin Gerner
Faculty of Life Sciences
University of Manchester
Manchester, UK
martin.gerner@postgrad.manchester.ac.uk

Goran Nenadic
School of Computer Science
University of Manchester
Manchester, UK
g.nenadic@manchester.ac.uk

Casey M. Bergman
Faculty of Life Sciences
University of Manchester
Manchester, UK
casey.bergman@manchester.ac.uk

Abstract

Here we explore mining data on gene expression from the biomedical literature and present Gene Expression Text Miner (GETM), a tool for extraction of information about the expression of genes and their anatomical locations from text. Provided with recognized gene mentions, GETM identifies mentions of anatomical locations and cell lines, and extracts text passages where authors discuss the expression of a particular gene in specific anatomical locations or cell lines. This enables the automatic construction of expression profiles for both genes and anatomical locations. Evaluated against a manually extended version of the BioNLP '09 corpus, GETM achieved precision and recall levels of 58.8% and 23.8%, respectively. Application of GETM to MEDLINE and PubMed Central yielded over 700,000 gene expression mentions. This data set may be queried through a web interface, and should prove useful not only for researchers who are interested in the developmental regulation of specific genes of interest, but also for database curators aiming to create structured repositories of gene expression information. The compiled tool, its source code, the manually annotated evaluation corpus and a search query interface to the data set extracted from MEDLINE and PubMed Central is available at http://getm-project.sourceforge.net/.

1 Introduction

With almost 2000 articles being published daily in 2009, the amount of available research literature in the biomedical domain is increasing rapidly. Currently, MEDLINE contains reference records for almost 20 million articles (with about 10 million abstracts), and PubMed Central (PMC) contains almost two million full-text articles. These resources store an enormous wealth of information, but are proving increasingly difficult to navigate and interpret. This is true both for researchers seeking information on a particular subject and for database curators aiming to collect and annotate information in a structured manner.

Text-mining tools aim to alleviate this problem by extracting structured information from unstructured text. Considerable attention has been given to some areas in text-mining, such as recognizing named entities (e.g. species, genes and drugs) (Rebholz-Schuhmann *et al.*, 2007; Hakenberg *et al.*, 2008; Gerner *et al.*, 2010) and extracting molecular relationships, e.g. protein-protein interactions (Donaldson *et al.*, 2003; Plake *et al.*, 2006; Chowdhary *et al.*, 2009). Many other areas of text mining in the biomedical domain are less mature, including the extraction of information about the expression of genes (Kim *et al.*, 2009). The literature contains a large amount of information about where and when genes are expressed, as knowledge about the expression of a gene is critical for understanding its function and has therefore often been reported as part of gene studies. Gene expression profiles from genome-wide studies are available in specialized databases such as the NCBI Gene Expression Omnibus (Barrett *et al.*, 2009) and FlyAtlas (Chintapalli *et al.*, 2007), but results on gene expression from smaller studies remain locked in the primary literature.

Previously, a number of data-mining projects have combined text-mining methods with structured genome-wide gene expression data in order

Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 72–80,
Uppsala, Sweden, 15 July 2010.

to allow further interpretation of the gene expression data (Natarajan *et al.*, 2006; Fundel, 2007). However, only recently has interest in text-mining tools aimed at extracting gene expression profiles from primary literature started to grow. The 2009 BioNLP shared task (Kim *et al.*, 2009) aimed at extracting biological "events", where one of the event types was gene expression. For this event type, participants were asked to determine locations in text documents where authors discussed the expression of a gene or protein and extract a *trigger* keyword (e.g. "expression") and its associated gene *participant* (the gene whose expression is discussed). The group that achieved the highest accuracy on the "simple event" task (where gene expression extraction was included) achieved recall and precision levels of 64.2% and 77.5%, respectively (Björne *et al.*, 2009). A key limitation of the 2009 shared task was that all genes had been annotated prior to the beginning of the task, making it difficult to anticipate the accuracy of tools that do not rely on pre-annotated entities.

Biologists are interested not only in finding statements of gene expression events, but also in knowing where and when a gene is expressed. However, to the best of our knowledge, no effort has previously been made to extract and map the expression of genes to specific tissues and cell types (and vice versa) from the literature. Thus, we have taken preliminary steps to construct a software tool, named Gene Expression Text Miner (GETM), capable of extracting information about what genes are expressed and where they are expressed. An additional goal of this work is to apply this tool to the whole of MEDLINE and PMC, and make both the tool and the extracted data available to researchers.

We anticipate that the data extracted by GETM will provide researchers an overview about where a specific gene is expressed, or what genes are expressed in a specific anatomical location. Moreover, GETM will aid in the curation of gene expression databases by providing text passages and identifiers to database curators for verification.

2 Methods

An overview of the workflow of GETM is given in Figure 1. Articles are initially scanned for mentions of gene entities, anatomical entities and keywords indicating the discussion of gene expression (called *triggers* following BioNLP terminology, e.g. "expression" and "expressed in"). After the detection of the entities and triggers, abbreviations are detected and entities are grouped in the cases of enumerations. Finally, sentences are split and each sentence is processed in order to associate triggers with gene and anatomical entities. Each step is described below in more detail.

2.1 Named entity recognition and abbreviation detection

In order to extract information on the expression of genes and their anatomical locations, a key requirement is the accurate recognition and normalization (mapping the recognized terms to database identifiers) of both the genes and anatomical locations in question. In order to locate and identify gene names, we utilized GNAT (Hakenberg *et al.*, 2008), an inter-species gene name recognition software package. Among the gene name recognition tools capable of gene normalization, GNAT is currently showing the best accuracy (compared to the BioCreative corpora (Hirschman *et al.*, 2005; Morgan *et al.*, 2008)). The species identification component of GNAT, used to help disambiguate gene mentions across species, was performed by LINNAEUS (Gerner *et al.*, 2010).

In order to perform named entity recognition (NER) of anatomical locations, we investigated the use of various anatomical ontologies. A key challenge with these ontologies is that the terms

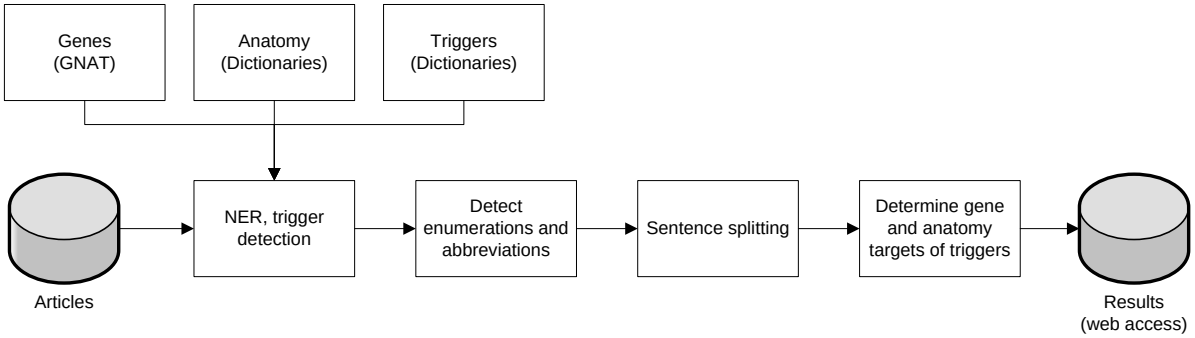

Figure 1. Schematic overview of the processing workflow of GETM.

vary significantly from one species to another. For example, fruit flies have wings while humans do not, and humans have fingers, while fruit flies do not. Efforts have been made in creating unified species-independent anatomical ontologies, such as Uberon (Haendel *et al.*, 2009; Mungall *et al.*, 2010). However, in preliminary experiments we found that the coverage of Uberon was not extensive enough for this particular application (data not shown), motivating us to instead use a combination of various species-specific anatomical ontologies hosted at the OBO Foundry (Smith *et al.*, 2007). These ontologies (n = 13) were chosen in order to cover terms from the main model organisms that are used in research (e.g. human, fruit fly, mouse, *Caenorhabditis elegans*) and a few larger groups of organisms such as e.g. amphibians and fungi. It is worth noting that the more general terms, such as e.g. "brain", are likely to match anatomical locations in other species as well. In total, the selected ontologies contain terms for 38,459 different anatomical locations.

We also utilized an ontology of cell lines (Romano *et al.*, 2009), containing terms for a total of 8,408 cell lines (ranging across 60 species), as cell lines can be viewed as biological proxies for the anatomical locations that gave rise to them. For example, the HeLa cell line was derived from human cervical cells, and the THP1 cell line was derived from human monocytes (Romano *et al.*, 2009).

The anatomical and cell line NER, utilizing the OBO Foundry and cell line ontologies, was performed using dictionary-matching methods similar to those employed by LINNAEUS (Gerner *et al.*, 2010).

After performing gene and anatomical NER on the document, abbreviations were detected (using the algorithm by Schwartz and Hearst (2003)) in order to allow the detection and markup of abbreviated entity names in the cases where the abbreviations do not exist in any of the ontologies that are used.

2.2 Trigger detection

The trigger keywords indicating that an author is discussing the expression of one or several genes, such as e.g. *"expression"* and "*expressed in"* were detected using a manually created list of regular expressions. The regular expressions were designed to match variations of a set of terms, listed below, that were identified when inspecting documents not used when building the gold-standard corpus (see Section 3.1).

The terms used to construct the trigger regular expressions were orthographical, morphological and derivational variations of "expression", "production" and "transcription". Descriptions of the level of expression were also considered for the different terms, such as "over-expression," "under-expression," "positively expressed," "negatively expressed," *etc*.

Each gene expression mention that has been extracted by GETM contains information about the trigger term used by the author, allowing researchers to look only at e.g. the "negative" mentions (where genes are e.g. "under-expressed" or "negatively expressed") or the "positive" mentions (where genes are e.g. "over-expressed").

2.3 Association of entities to the trigger

To help associate triggers with the correct gene and anatomical entities, articles were first split into sentences, allowing each sentence to be processed in turn. In order to reduce the number of false positives and preserve a high level of precision, any sentences that did not contain a trigger, at least one gene mention and at least one anatomical mention were ignored. For the sentences that did contain a combination of all three requirements (trigger, gene and anatomical mention), the following pattern- and distance-based rules were employed in order to associate each trigger with the correct gene and anatomical mention:

1. If there is only one gene mention and only one anatomical mention in the sentence, the trigger is associated with those mentions.
2. If there is one gene mention (G) and one anatomical mention (A) in the sentence such that they match one of the patterns "<G> is expressed in <A>", "expression of <G> in <A>", "<A> transcribes <G>" or "<A> produces <G>", the gene mention <G> and anatomical mention <A> are associated with the trigger (variations of the triggers, such as "over-expressed" and "negative expression" are considered as well). Additional gene or anatomical mentions that fall outside the pattern are ignored.
3. If neither of the above rules applies, the trigger is associated with the gene and anatomical mentions that are closest to the trigger.

For the purposes of these rules, an enumeration of several genes or anatomical locations was

handled as if it was only a single mention. For example, Rule 1 might trigger even if there are several genes mentioned in the same sentence, as long as they are mentioned together as part of an enumeration.

In order to detect these enumerations, a rule-based algorithm for connecting enumerated gene and anatomical entity mentions (as in e.g. "... RelB and DC-CK1 gene expression ...") was also implemented. Being able to detect enumerations allowed the rules described above to recognize that a particular gene expression mention do not refer to only e.g. "RelB" or "DC-CK1", but both of them at the same time.

Each trigger was processed independently, allowing the potential extraction of multiple gene expression statements from a single sentence.

Initially, experiments were performed using stricter rules where only variations of Rule 2, requiring gene and anatomical mentions to conform to certain patterns, were used. However, recall was in these cases found to be extremely low (below 5%, data not shown). The current rules are more permissive, allowing higher recall.

The fact that the method requires a combination of a trigger, a gene and an anatomical location makes it susceptible to false negatives: if any one of them cannot be found by the NER or trigger detection methods, the whole combination is missed.

3 Evaluation

3.1 Extending the BioNLP shared task gold-standard corpus

In order to make a meaningful evaluation of the accuracy of text-mining applications, a gold-standard corpus, consisting of manually annotated mentions for a set of documents, is required. Previously, no such corpus existed that was suitable for this problem (providing annotations linked to mentions of both gene and anatomical locations). However, the BioNLP corpus (Ohta *et al.*, 2009) which is based on the GENIA corpus (Kim *et al.*, 2008), does contain annotations about gene expression. Annotations in the corpus contain trigger terms that are linked to genes (or gene products) where the authors discuss gene expression. However, anatomical locations have not been annotated in this corpus.

In order to allow evaluation of the accuracy of our software, we extended the annotations of gene expression events in part of the BioNLP corpus. Each gene expression entry in the corpus was linked to the anatomical location or cell line that the author mentioned. In cases where gene expression was only discussed generally without referring to expression in a particular location, no association to an anatomical location could be made (these entries were ignored during evaluation). Note that named entities were only linked to their locations in the text, not to unique database identifiers (such as Entrez Gene or OBO Foundry identifiers). Because of this, subsequent evaluation in this extended corpus is limited to the accuracy of recognition (locating the entities in the text), but not normalization (linking the entities to database identifiers).

In total, annotations for 150 abstracts (constituting the development set of the BioNLP corpus) were extended to also include anatomical locations. These abstracts contained 377 annotated gene expression events, of which 267 (71%) could be linked to anatomical locations. These results demonstrate that the majority of gene expression mentions include reference to an anatomical location. For a few cases where the author described the expression of a gene in several cell types, a single gene expression event gave rise to several distinct "entries" in the extended corpus, creating a total of 279 final gene expression entries that are linked to anatomical locations.

4 Results

In order to evaluate the accuracy of GETM, it was first run on the 150 abstracts in the gold-standard corpus, after which the extracted results were compared against the annotations of the corpus. GETM was also applied to the whole of MEDLINE and PMC, in order to extract a searchable and structured data set of gene expression mentions in published biomedical articles.

4.1 Accuracy

The gene expression mentions extracted by GETM from the corpus were compared against the manually created annotations in order to estimate the accuracy of the software. After inspecting the false positives and false negatives, we noted that a number of the false positives actually were correctly identified by our system and had been marked as false positives only because of incomplete annotations in the corpus. Because of this, all false positives were manually examined in order to determine the "correct" number of false positives. For one of the corrected expression mentions, two anatomical locations were enumerated, with GETM only locat-

ing one of them. This introduced both a new true positive (for the one that was recognized) and a new false negative (for the one that was not). The number of true positives, false positives, false negatives, precision and recall (before and after correction) are shown in Table 1.

	Original	Corrected
TP	53	67
FP	61 (p = 46.5%)	47 (p = 58.8%)
FN	214 (r = 19.8%)	215 (r = 23.8%)

Table 1. The number of true positives (TP), false positives (FP), false negatives (FN) and levels of precision (p) and recall (r) for GETM when compared against the gold-standard corpus.

4.2 Analysis of false negatives

In order to determine the causes of the relatively high number of false negatives, the gene entities, anatomical entities and triggers identified by GNAT and GETM were compared to the extended corpus, allowing us to determine the number of corpus entities that could not be found by the GNAT and GETM NER tools. An analysis was also performed in order to determine the number of corpus entries that were spread across several sentences, as any expression mentions spread over several sentences are missed by GETM.

The analysis results can be seen in Table 2, showing that virtually all false negatives are caused either by incomplete NER or multi-sentence entries. Only considering the NER, 68% of the gold-standard corpus annotated entries contain either a trigger (example FN: "detected"), gene (example FN: CD4) or anatomical location (example FN: "lymphoblastoid cells") that could not be located automatically. GETM was further limited by entities being spread across several sentences (n=66, 23.6%). In total, 74.3% of all entries could not be extracted correctly due to either incomplete NER, incomplete trigger detection or the entities being spread across multiple sentences. This limited recall to 25.7%, even if the rule-based method was working perfectly.

4.3 Analysis of false positives

Manual inspection of the false positives (after adjusting the false positives caused by incomplete annotations) allowed the identification of one clear cause: if the NER methods fail to recognize the entity associated with a manually annotated expression entry, but there are other entities (that have been recognized) in the sentence, those entities might be incorrectly associated with the trigger instead. For example, in the sentence "In conclusion, these data show that IL-10 induces *c-fos expression* in human *B-cells* by activation of tyrosine and serine/threonine kinases." (Bonig et al., 1996) (the correct entities and trigger are italicized), a correctly extracted entry would link c-fos to B-cells through the trigger expression. However, the gene NER component failed to recognize c-fos but did recognize IL-10, causing GETM to incorrectly associate IL-10 with B-cells. Either increasing the accuracy of the NER methods or performing deeper grammatical parsing could potentially reduce the number of false positives of this type. We note that the number of cases for this category (n = 15; 34%) only make up a minority of the total number of false positives, and the remainder have no easily identifiable common cause.

4.4 Application to MEDLINE and PMC documents

GETM was applied to the whole set of 10,240,192 MEDLINE entries from the 2010 baseline files that contain an abstract (many MEDLINE entries do not contain an abstract). From these abstracts, 578,319 statements could be extracted containing information about the expression of a gene and the location of this expression. In addition, GETM was also applied to the set of 186,616 full-text articles that make up the open-access portion of PMC (downloaded February 5th, 2010). The full-text articles allowed the extraction of 145,796 statements (an 18-fold increase in entries per article compared

Problem type	Number of occurrences
Trigger not found	58 (20.7%)
Gene not found	139 (49.6%)
Anatomical location not found	74 (26.4%)
Any of the entities or trigger not found	190 (67.9%)
Total number of entities not contained in a single sentence	66 (23.6%)
Total number of entities either not found or not in the same sentence	208 (74.3%)

Table 2. Breakdown of the causes for false negatives in GETM, relative to the total number of entries in the gold-standard corpus.

Gene	Anatomical location	Number of mentions
Interleukin 2	T cells	3511
Interferon, gamma	T cells	2088
CD4	T cells	1623
TNF	Macrophages	1596
TNF	Monocytes	1539
Interleukin 4	T cells	1323
Integrin, alpha M	Neutrophils	1063
Inteleukin 10	T cells	971
ICAM 1	Endothelial cells	964
Interleukin 2	Lymphocytes	876

Table 3. The ten most commonly mentioned combinations of genes and anatomical locations

to the MEDLINE abstracts). In total, 716,541 statements were extracted, not counting the abstracts in MEDLINE that also appear in PMC. Overall, the combined extracted information ranges across 25,525 different genes (the most common being *tumor necrosis factor (TNF superfamily, member 2)* in human) and 3,655 different anatomical locations (the most common being *T cells*). The most common combination concerns the expression of human *interleukin 2* in *T cells*. The 10 most commonly mentioned combinations of genes and anatomical locations are shown in Table 3. Overall, these results suggest that studies on gene expression in the field of mammalian immunology are the dominant signal in MEDLINE and PMC. The genes that were recognized and normalized range across 15 species, out of the 23 supported by GNAT (Hakenberg *et al.*, 2008). The most common species is human, as expected (Gerner *et al.*, 2010), followed by mouse, rat, chicken and cow.

The majority of statements were associated to anatomical locations from the OBO Foundry ontologies (n=649,819; 89.7%), while the remainder were associated to cell lines (n=74,294; 10.3%). This result demonstrates the importance of taking cell lines into account when attempting to identify anatomical entities.

Finally, a total of 73,721 (11.7%) of the statements extracted from MEDLINE contained either genes or anatomical locations that had been enumerated by the author, underscoring the importance of considering enumerations when designing text-mining algorithms.

4.5 Availability

GETM is available under an open source license, and researchers may freely download GETM, its source code and the extended gold-standard corpus from http://getm-project.sourceforge.net/. Also available on the web site is a search query interface where researchers may search for extracted gene expression entries relating to a particular gene, anatomical location or a combination of the two and view these in the context of the surrounding text.

5 Discussion

5.1 Overview of design philosophy

When constructing text-mining applications, a balance between precision (reflecting the relative number of false positives) and recall (reflecting the relative number of false negatives) is often used to optimize system performance. Accordingly, a measure which often is used to evaluate the accuracy of software is the F-score (the harmonic mean of the precision and recall). In this work, we have decided that rather than trying to maximize the F-score, we have put more focus on precision in order to ensure that the data extracted by GETM are of as high quality as possible. This typically leads to lower recall, causing the software to detect a relatively smaller number of relevant passages. Nonetheless, we believe that for this particular application, a smaller amount of data with higher quality would be more useful to curators and biologists than a larger amount of data that is less reliable.

5.2 Comparison with previous work

It is difficult to compare the precision and recall levels of GETM (at 58.8% and 23.8%, respectively) against other tools, as GETM is the first tool aiming to perform this particular task. The closest comparison that can be made is against the software evaluated in the BioNLP shared task (Kim *et al.*, 2009). However, software developed for the BioNLP shared task did not attempt to extract the anatomical location of gene expression mentions, nor did they need to identify the component entities involved. The tool with the highest accuracy for the simple event task (where gene expression extraction was included) showed

precision and recall levels of 77.5% and 64.2%, respectively (Björne *et al.*, 2009). It is not clear how tools evaluated in the 2009 BioNLP shared task would perform if they identified entities themselves rather than using pre-annotated entities.

5.3 Limits on accuracy

When investigating the cause of the low level of recall, the main reason that emerged for the high number of false negatives was the high number of annotated entries that could not be automatically extracted due to at least one of the gene, anatomical or trigger mentions not being recognized. This fact underscores the importance of accurate NER for applications that rely on the extracted entity mentions, especially those that attempt to extract information from multiple entity types, like GETM. The results also demonstrate that NER, particularly in the case of gene name normalization, continues to pose a challenging problem. It is possible that using a combination of GNAT and other gene NER tools would improve the overall gene NER accuracy.

We further explored the effects of "perfect" gene NER on the accuracy of GETM by using the manual gene mention annotations supplied in the BioNLP corpus. Using the pre-annotated gene names increased the number of gene expression mentions recognized and the number of true positives, significantly improving recall (from 23.8% to 37.8%; data not shown). However, a number of additional false positives were also introduced, causing precision to decrease very slightly from 58.8% to 58.5% (data not shown). This demonstrates the complexity of gene expression mentions in text, indicating that a combination of accurate trigger detection, accurate NER (for both genes and anatomical locations) and deeper NLP methods are needed in order to accurately capture gene expression profiles in text.

A secondary cause of false negatives was a relatively high number of annotated corpus entries that spanned several sentences. The high proportion (23%) of multi-sentence entries in our extended corpus differs from previously reported results. For the event annotations in the BioNLP corpus, previous analyses showed that only 5% of all entries spanned several sentences (Björne *et al.*, 2009). This suggests that the mentions of anatomical locations are located outside of the "trigger sentence" more often than gene mentions or other entities in the BioNLP corpus.

6 Conclusions

In this paper, we have explored integrated mining of gene expression mentions and their anatomical locations from the literature and presented a new tool, GETM, which can be used to extract information about the expression of genes and where they are expressed from biomedical text. We have also extended part of a previously existing gold-standard corpus in order to allow evaluation of GETM. When evaluated against the gold-standard corpus, GETM performed with precision and recall levels of 58.8% and 23.8%, respectively.

The relatively low level of recall was primarily caused by incomplete recognition of individual entities, indicating that – in order to increase the recall of GETM – future work would primarily need to focus on increasing the accuracy of the NER methods. With more accurate NER, while increasing recall, the higher number of recognized entities is also expected to increase the number of false positives, causing a need for deeper NLP methods in order to preserve and increase the level of precision.

While having a low level of recall, GETM was nonetheless able to extract 716,541 statements from MEDLINE and PMC, constituting a large and potentially useful data set for researchers wishing to get an overview of gene expression for a particular gene or anatomical location. The high number of mentions extracted from MEDLINE can give an indication of the amount of data available in MEDLINE: if the recall on the BioNLP corpus is representative for MEDLINE as a whole, a tool with perfect accuracy might be able to extract almost 2.5 million entries.

The level of precision (p = 58.8%) will most likely not be high enough for researchers to rely on the extracted data for high-throughput bioinformatical experiments without some kind of verification. However, we believe that it nonetheless will be of high enough quality that researchers and curators will not feel inconvenienced by false positives, as currently the only alternatives are multi-word free text searches through PubMed or Google. Additionally, we provide an interface with the text context surrounding gene expression statements, making it easier for researchers to quickly locate relevant results.

In the future, we will aim to evaluate the normalization of entities detected by GETM in order to quantify the level to which the identifiers assigned to the entities are correct. In addition,

both the gene and anatomical NER components could be improved in order to both reduce the number of false negatives and cover gene and anatomical terms for a wider range of species, beyond the common model organisms. We also believe that extending this work by utilizing deeper NLP methods (e.g. dependency parsers) could further improve the accuracy of GETM and related approaches to mining the abundance of data on gene expression in the biomedical literature.

Acknowledgements

We thank Jörg Hakenberg (Arizona State University) for providing access to GNAT. We also thank members of the Bergman and Nenadic groups for helpful comments and suggestions throughout the project, and three anonymous reviewers of this article for valuable comments that helped improve the manuscript. This work was funded by the University of Manchester and a BBSRC CASE studentship (to M.G.).

References

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Muertter, R. N. and Edgar, R. (2009). "NCBI GEO: archive for high-throughput functional genomic data." Nucleic Acids Res 37(Database issue): D885-90.

Björne, J., Heimonen, J., Ginter, F., Airola, A., Pahikkala, T. and Salakoski, T. (2009). "Extracting complex biological events with rich graph-based feature sets." In Proceedings of the Workshop on BioNLP: Shared Task Boulder, Colorado: 10-18.

Bonig, H., Korholz, D., Pafferath, B., Mauz-Korholz, C. and Burdach, S. (1996). "Interleukin 10 induced c-fos expression in human B cells by activation of divergent protein kinases." Immunol Invest 25(1-2): 115-28.

Chintapalli, V. R., Wang, J. and Dow, J. A. T. (2007). "Using FlyAtlas to identify better Drosophila models of human disease." Nature Genetics 39: 715-720.

Chowdhary, R., Zhang, J. and Liu, J. S. (2009). "Bayesian inference of protein-protein interactions from biological literature." Bioinformatics 25(12): 1536-42.

Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G. D., Michalickova, K., Pawson, T. and Hogue, C. W. (2003). "PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine." BMC Bioinformatics 4: 11.

Fundel, K. (2007). Text Mining and Gene Expression Analysis Towards Combined Interpretation of High Throughput Data. Dissertation. Faculty of Mathematics, Computer Science and Statistics. München, Ludwig-Maximilians Universität.

Gerner, M., Nenadic, G. and Bergman, C. M. (2010). "LINNAEUS: a species name identification system for biomedical literature." BMC Bioinformatics 11: 85.

Haendel, M. A., Gkoutos, G. V., Lewis, S. E. and Mungall, C. J. (2009). "Uberon: towards a comprehensive multi-species anatomy ontology." In International Conference on Biomedical Ontology Buffalo, NY.

Hakenberg, J., Plake, C., Leaman, R., Schroeder, M. and Gonzales, G. (2008). "Inter-species normalization of gene mentions with GNAT." Bioinformatics 24(16): i126-i132.

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

Kim, J. D., Ohta, T., Pyysalo, S., Kano, Y. and Tsujii, J. i. (2009). "Overview of BioNLP'09 Shared Task on Event Extraction." In Proceedings of the Workshop on BioNLP: Shared Task, Boulder, Colorado, Association for Computational Linguistics: 1-9.

Kim, J. D., Ohta, T. and Tsujii, J. (2008). "Corpus annotation for mining biomedical events from literature." BMC Bioinformatics 9: 10.

Morgan, A., Lu, Z., Wang, X., Cohen, A., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H., Torres, R., Krauthammer, M., Lau, W., Liu, H., Hsu, C., Schuemie, M., Cohen, K. and Hirschman, L. (2008). "Overview of BioCreative II gene normalization." Genome Biology 9(Suppl 2): S3.

Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E. and Ashburner, M. (2010). "Integrating phenotype ontologies across multiple species." Genome Biol 11(1): R2.

Natarajan, J., Berrar, D., Dubitzky, W., Hack, C., Zhang, Y., DeSesa, C., Van Brocklyn, J. R. and Bremer, E. G. (2006). "Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line." BMC Bioinformatics 7: 373.

Ohta, T., Kim, J.-D., Pyysalo, S., Wang, Y. and Tsujii, J. i. (2009). "Incorporating GENETAG-style annotation to GENIA corpus." In Workshop on BioNLP, Boulder, Colorado: 106-107.

Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J. and Leser, U. (2006). "AliBaba: PubMed as a graph." Bioinformatics 22(19): 2444-5.

Rebholz-Schuhmann, D., Arregui, M., Gaudan, M., Kirsch, H. and Jimeno, A. (2007). "Text processing through Web services: Calling Whatizit." Bioinformatics 23(2): e237-e244.

Romano, P., Manniello, A., Aresu, O., Armento, M., Cesaro, M. and Parodi, B. (2009). "Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines." Nucl. Acids Res. 37(suppl_1): D925-932.

Schwartz, A. S. and Hearst, M. A. (2003). "A simple algorithm for identifying abbreviation definitions in biomedical text." Pac Symp Biocomput: 451-62.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, 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. and Lewis, S. (2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." Nat Biotechnol 25(11): 1251-5.