Full-Text Mining: Linking Practice, Protocols and Articles in Biological Research

Full-Text Mining: Linking practice, protocols and articles in biological research

James M. Eales[1,*], Robert D. Stevens[2] and David L. Robertson[1]

[1]Faculty of Life Sciences, University of Manchester, Oxford Road, Manchester, UK, M13 9PL.

[2]School of Computer Science, University of Manchester, Oxford Road, Manchester, UK, M13 9PL.

ABSTRACT

Text mining has the potential of being of great utility to the life-sciences and particularly bioinformatics; it can support and inform studies with information that would otherwise be laborious to collect. The kind of information required is, however, often not found in article abstracts (the traditional source for text mining studies). This is fuelling the move to analysis of full-text articles, which have greater amounts of more detailed information. In this article we outline our approach to full-text mining of biological protocols and the subsequent linking of these with metrics of scientific quality. We also highlight the elements of full-text mining that we believe could benefit from the attention of the wider text mining community.

1 INTRODUCTION

Scientific articles are a rich source of information; they contain experimental methods, results, conclusions, arguments, proposals, algorithms, hypotheses, evidence, citations and many more elements. The information in most scientific articles is, however, only easily available to single human readers of the article. Text mining methods can allow us to capture information from text in a way that simulates elements of the action of human readers (albeit in a simplified and less flexible manner), thus making some of this information available to computational analysis [1-4]. Often the aim of text mining studies is to identify results in text; however the protocol used to derive those results can be just as biologically significant and can represent a valuable element of the study in its own right.

Protocols are important in biological research, not only for describing lab-based experiments but also computational analyses. However there are now so many different ways to complete the same task, that often the choice of methods or software can be more of a choice of habit than a choice of reason. We have been working on ways to inform decisions in the design of biological experiments by linking article-extracted protocols to metrics of scientific quality. We believe this can enable a new approach to experimental design centred on literature-based validation of protocol quality.

Here we present our work on a case study attempting to inform the practice of researchers in the discipline of molecular phylogenetics by automatically extracting and building representations of phylogenetic protocols from a large collection (21,866) of full-text articles. We then attempt to link these with quality metrics to allow suggestion of good quality protocols to the entire discipline. Our aim in this paper is not to present our work in the traditional style, but rather to highlight the aspects we believe could gain most

benefit from the text mining community and those that offer new opportunities for the future.

2 MINING FULL-TEXT

In the past most text mining studies have made use of text content that is more readily available (abstracts) rather than the more informative full-text [5-7]. Abstracts are indeed useful [3, 7], being widely available and providing a concise, information dense summary of the work, but they lack the richness of information required by some newly emerging text mining studies [1, 8-11]. Methods are very rarely mentioned in abstracts and only described in a very basic way if they are mentioned. Therefore we knew from the beginning of our work that access to and analysis of full-text articles would be required.

A major limiting factor in full-text mining is access to the electronic version of full-text articles [5, 6, 12]. There are still restrictions imposed by publisher license agreements that prevent large-scale content retrieval; this difficulty is often augmented by technical issues relating to the automation of document retrievals through interfaces designed for single human users. Furthermore, reaching individual agreements with publishers for access to content can be a prohibitively complex and time-consuming task, especially if your collection of articles spans many journals [1, 8, 10-12] and therefore many publishers. Recent trends in science publishing, especially in relation to free access to publicly-funded research have improved access to full-text articles considerably. This is mostly attributable to the open access publishing movement, the publishers who support it (BMC and PLoS) and the initiatives [13, 14] and public repositories (PMC, UKPMC and arXiv) who support open access to the output of publicly-funded scientific research. The amount of text mining research published that makes use of full-text article collections [1, 2, 8, 10, 15] is, however, still limited [1, 5], given the large range of useful information elements present in full-text but absent in abstracts (fig.1), this seems an unsustainable situation.

We collected a set of journal articles in PDF format identified by a PubMed search for "phylogen*[Title/Abstract] AND (full text[sb])". These were collected using automated download tools [16, 17], these typically used web spider methods to seek out full-text files from an original PubMed search. However due to download restrictions with Quosa and stability issues with BioRat we also had to develop our own download agent. Our original PubMed search resulted in 27,259 results, which yielded 24,494 different articles in PDF format. The difference is attributable to incorrect PubMed "link out" data and software difficulties with finding the PDF version of the article from the original link. Of
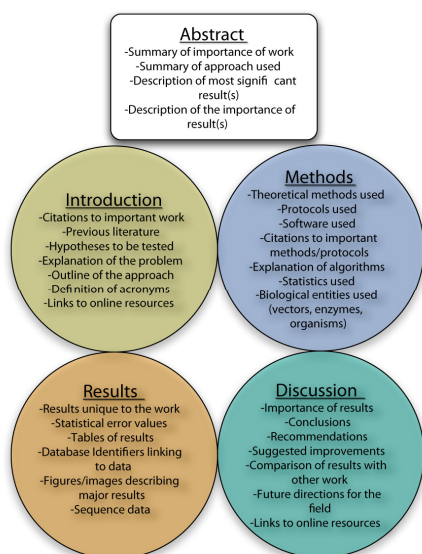
Figure 1. A model of a typical scientific article, describing the distribution of information among the abstract and article sections.

these 21,866 yielded usable plain text, after extraction with the executable 'pdftotext' [18].

3   SEARCHING IN THE RIGHT PLACE

Scientific articles follow a structure first developed to support the efficient communication of a piece of research based on its fidelity to the scientific method. In most cases it is therefore appropriate to include passages of text that can form the Introduction, Methods, Results and Discussion sections of the article. There are exceptions to this rule; for example a review article may not describe any Methods. It should also be stressed that some journals do not provide titles for their sections, this is especially common in compact format journals, such as Science and Nature. Finally it is also common for sections to be labelled according to their specific subject (such as this section) rather than the scientific aim of the section, this often depends on the broad aims of the paper (e.g. review, short article or research article) and where the article is to be published or presented.

Scientific authors do, in the majority of cases, follow the basic principles of the scientific article structure and assign information accurately to each section. We can profit from this accuracy of information assignment by exploiting what we know about the structure of the archetypal scientific article to direct text mining activities to sections of text that are most likely to contain the information that we are interested in [3, fig. 1, 7]. The Introduction or Background section will contain information that is not unique to the article, it will most often discuss previous research in the area often comparing and contrasting the results with subjective commentary on these results from the present authors. Furthermore the Introduction will commonly outline the aims and theoretical basis of the piece of research. The Methods or Materials and Methods section has a clear aim that involves the often highly technical description of the physical implementation of the piece of research it will also describe any assumptions or simplifications made. The Results section again has a clear aim; the presentation

of the new scientific information derived specifically from the piece of research in question, this section will on occasion discuss relationships between these results but will not refer to the wider meaning of the data. The discussion section attempts to derive meaning from the data presented in the Results section, it also discusses how these results fit in with current thinking in the field, that may have been discussed in the Introduction. Article sections are not merely structural elements of an article, but instead represent a valuable semantically rich link between the text of an article and the scientific method.

To help with our extraction of biological protocols, we created a text classifier [19] that would label bits of text according to their membership or textual similarity to one of the standard scientific article sections. In this case because we were only interested in methods we screened articles for sections that were classified as 'Methods' and then passed these on for further analysis. This has similarity with zone analysis [20-22], whereby fragments of text are classified into one of a detailed set of classes, described by an annotation scheme [20, 22]. We feel our approach, although less detailed, provides a useful starting point for many text mining studies, furthermore our classifier was trained in an automated manner, reducing difficulties with lengthy annotation tasks and inter-annotator agreement. Additionally, our classifier was trained on all currently available open access articles from PMC (48,105 articles, 330 journals) and therefore has a set of training data, applicable to most disciplines of biological science. Finally we have made our classifier available through a SOAP/web service interface, a browser-based interface for testing and as a downloadable Java application (with a GUI) for local use [23].

Using a training data set of half our PMC article collection, we assessed the accuracy of our classifier to label sections in the other half of the article collection. Verification of classification results was performed by surveying and counting all section titles from all articles, and manually grouping them together according to which of the archetypal section types that they correspond. For example the section titles "Methods", "Materials and methods" and "Implementation" can be reliably mapped to the Methods section type. All section titles that occurred more than 50 times and could be reliably mapped to a section type were included. This reduced our ability to test the classifier on all article sections (50,894 out of 92,647), but it did allow us to be confident in the calculated accuracy of the classifier. In total, 84.1% of all sections were correctly classified (15.9% incorrect). The best overall results come from the Methods class, with the highest f-measure (0.8807) and recall (0.9731) values. The best precision (0.9368) is achieved in the Discussion class which also exhibits the lowest recall (0.7472). The lowest precision (0.7649) and f-measure (0.8197) values came from the Introduction class.

Interestingly we found that the use of a stop word list when training our classifier was detrimental to its performance. With some words which commonly appear in standard stop word lists [24] being the most informative for discriminating between sections (table 1). This may also be a product of our use of counts of words rather than just a vector of unique words for training the classifier. This means you can derive the likelihood of each section label based on how many times a word occurs, which is far more suited to the common words included in a stop word list, rather than a binary present or not present system, which is less informative for common words.
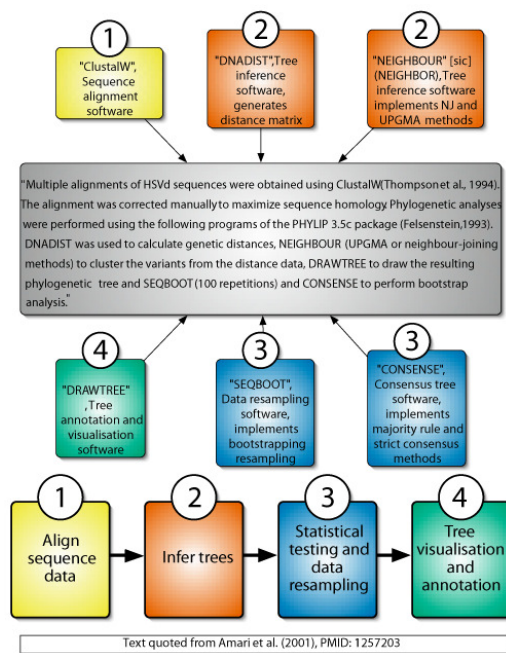
Figure 2: A model of the archetypal phylogenetic experiment, with an example representation of a protocol in text form. Protocol elements are coloured according to their stage (1 to 4) in the model.

One of the main reasons we had to limit our analysis to Methods section text was the large number of false positive matches, we detected in a preliminary analysis (without a section classifier) to terms important to the description of phylogenetic protocols. This kind of problem has been reported previously [3] and our solution was to limit our analysis to only Methods section text.

4 MODELLING BIOLOGICAL PROTOCOLS

Our approach to capturing protocols [8] from the phylogenetics literature was to first construct a model (fig.2) of a typical protocol and then use it to structure our extracted information and populate the model with matches to terms important in the description of phylogenetic protocols [25]. We take the set of important methodological terms found in any one article to be a direct description of the protocol employed in that piece of research.

We divide the methodological terms, found in the text, between four key stages: (i) sequence alignment, (ii) tree inference, (iii) statistical testing and data resampling, and (iv) tree visualisation and annotation (fig.2). These terms are represented in a structured controlled vocabulary that link the term with a manually tested and designed regular expression pattern to identify the term in text. The controlled vocabulary [25] contains 258 important names and terms representing theoretical methods, models, and the software that implements these methods and models. The controlled vocabulary is an XML document that arranges these terms according to whether they are methods, models or software. The software names were taken from Professor Joseph Felsenstein's page titled "Phylogeny programs" [26]. Other terms were manually created using the phylogenetics primary literature. Finally our controlled vocabulary is linked with our model of a phylogenetics protocol

(fig.2), so that we can order, analyse and present our term matches in a repeatable manner.

The individual protocols are thus a model of a scientific experiment that is inferred from the text of the methods described in an article. The phylogenetic terms found in the methods are inferred to describe a task or part of a task and the collection of these tasks is what we term the protocol. Our protocol model (fig.2) allows us to organise method terms according to the order in which they would have been used in the experiment rather than their order in the text. The model also allows us to collect partial protocols from articles where some of the method terms are either, not found by automated analysis, missing from the text entirely (due to poorly communicated methods) or are unnecessary for the analysis described (not all analyses require all 4 of the stages). We believe that our structured approach to capturing protocols from full-text articles could be applied to any discipline of science where the methods used can be broken down into individual sequential stages. For example, a simple task to sequence a genic region from a single fruit fly could be broken down into; DNA extraction, purification, amplification, sequencing and chromatogram analysis. As with a phylogenetic protocol several terms could map on to each one of these stages, for example, PCR or bacterial cloning could be used in the amplification step.

To test the accuracy of our term matching process we manually annotated the methods section or section of text most descriptive of methodological detail for 50 randomly chosen articles from our corpus. We annotated all pieces of text that referred to any of the phylogenetic entities that are present in the controlled vocabulary. By comparing the agreement between our annotations and those generated by our software we derived these values; precision 96.5%, recall 54.7% and f-measure (f-score) 69.8%. Clearly the information we are collecting is reliable (precision 96.6%), however the level of detail being captured seems very low (recall 54.7%), the most important cause of the low recall value is the text classifier missing section of text that are actually from a Methods section.

The result of our protocol gathering and extraction process was 527 unique phylogenetic protocols [8]. The usage of these was measured for the years 1996 to 2005. Before 1996 fewer than 300 articles per year yielded a protocol. There are several very popular protocols with most articles (60%) using one of the top 10 most used. This does, however, leave another 517 protocols that have on average only been used 5.8 times in the 10-year period. Additionally we identified 3 key communities within phylogenetics who use protocols that are very different from each other to produce very similar results [8]. This provided evidence to support our belief that practice in phylogenetics (the choice and use of certain protocols) is being heavily influenced by community structure in the discipline, and that new protocols, methods and software remain in a local community of researchers when they could be of benefit to the whole discipline.

We then used a measure of 'expertness' in the field of phylogenetics (individual contribution of literature to the field), to help us identify which protocols may be seen as 'good quality' or have certainly been subject to the highest level of peer review [8]. We then presented these and suggested that they could form the basis of a protocol for any researcher working in phylogenetics. Finally, we have also done some work on creating further quality metrics for ranking protocols; these include the use of citation data,

Table 1. Probabilities of occurrence (per word) of discriminatory words by section.

Word	Introduction	Methods	Results	Discussion
"figure"	0.00028	0.00053	0.00445†	0.00069
"table"	0.00016	0.00062	0.00290†	0.00038
"p"	0.00035	0.00113	0.00370†	0.00039
"="	0.00029	0.00214	0.00449†	0.00031
"-"	0.00265†	0.00050	0.00052	0.00081
"may"	0.00212	0.00036	0.00069	0.00373†
"were"	0.00196	0.01813†	0.00921	0.00355
"using"	0.00113	0.00494†	0.00140	0.00105
"each"	0.00075	0.00374†	0.00157	0.00058
"our"	0.00087	0.00066	0.00099	0.00327†

Probability of each of the ten most discriminatory words occurring in each section. †indicates the highest value in each row.

Journal impact factor and elements of experimental context (e.g. data set size) extracted from articles. In the end, what phylogeneticists and biologists in general want, is a protocol that gets the job done and has evidence to support its use over the many other alternatives available.

5   FUTURE DIRECTIONS

We see text mining of experimental methods and protocols as an area that offers many possible challenges to the text mining community. The first challenge that we encountered was how to capture elements of experimental context.

By context we mean elements of the work that can have a direct effect on the kind of protocol that would be appropriate for the work (e.g. data set size, level of detail required in results, the specific aims of the study). Some elements of context were relatively straightforward to capture; such as journal of publication and author seniority, however we would require more information to be able to truly capture the nature of an experiment. We see this kind of information being useful in automated protocol suggestion system, where a user would answer a series of question pertaining to their proposed experiment, and the system could then provide a protocol tailored to the users specific needs.

We also believe that phylogenetic term matching, which is essentially a named-entity recognition task, could be improved by more advanced text mining methods, especially with the use of part-of-speech tagging for assessing the linguistic context of term matches and possibly machine learning methods for identifying terms themselves, however this would require annotated training data.

Here we have illustrated our approach to address a real biological problem with text mining methods. We have also highlighted the areas of our work that we feel could greatly benefit from increased input from the text mining community, especially in reference to the analysis of experimental methods and protocols. We hope that full-text will increasingly be used by the text mining community for the simple reason that it conveys far more and varied information, however we are aware that some technical and theoretical difficulties do remain. Therefore we have attempted to highlight the specific technical issues associated with full-text and some ways that we have found to circumvent them.

REFERENCES

1. Natarajan J, Berrar D, Dubitzky W, Hack C, Zhang Y, DeSesa C, Van Brocklyn JR, Bremer EG: 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 2006, 7(1):373.
2. Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboue PA, Weng W, Wilbu WJ et al: GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. Journal of Biomedical Informatics 2004, 37(1):43-53.
3. Shah PK, Perez-Iratxeta C, Bork P, Andrade MA: Information extraction from full text scientific articles: Where are the keywords? BMC Bioinformatics 2004.
4. Yu H, Hatzivassiloglou V, Friedman C, Rzhetsky A, Wilbur WJ: Automatic Extraction of Gene and Protein Synonyms from MEDLINE and Journal Articles. In: Proc AMIA Symp: 2002; 2002: 23.
5. Cohen AM, Hersh WR: A survey of current work in biomedical text mining. Briefings in Bioinformatics 2005, 6(1):57-71.
6. Krallinger M, Valencia A: Text-mining and information-retrieval services for molecular biology. Genome Biology 2005.
7. Schuemie MJ, Weeber M, Schijvenaars BJA, van Mulligen EM, van der Eijk CC, Jelier R, Mons B, Kors JA: Distribution of information in biomedical abstracts and full-text publications. Bioinformatics 2004, 20(16):2597-2604.
8. Eales JM, Pinney JP, Stevens RD, Robertson DL: Methodology capture: discriminating between the 'best' and the rest of community practice. BMC Bioinformatics, (Submitted) 2008.
9. Crangle CE, Cherry JM, Hong EL, Zbyslaw A: Mining experimental evidence of molecular function claims from the literature. Bioinformatics 2007, 23(23):3232-3240.
10. Aerts S, Haeussler M, van Vooren S, Griffith O, Hulpiau P, Jones S, Montgomery S, Bergman C, The Open Regulatory Annotation C: Text-mining assisted regulatory annotation. Genome Biology 2008, 9(2):R31.
11. Muller H-M, Kenny EE, Sternberg PW: Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PLoS Biology 2004, 2(11):e309.
12. Dickman S: Tough Mining. PLoS Biology 2003, 1(2):e48.
13. Wellcome Trust position statement in support of open and unrestricted access to published research [http://www.wellcome.ac.uk/doc_WTD002766.html]
14. Research Councils UK' updated position statement on access to research outputs [http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/2006statement.pdf]
15. Tanabe L, Wilbur WJ: Tagging gene and protein names in full text articles. In: Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3: 2002; 2002: 9-13.
16. Quosa Information Manager [http://www.quosa.com/]
17. Corney DPA, Buxton BF, Langdon WB, Jones DT: BioRAT: Extracting Biological Information from Full-length Papers. Bioinformatics 2004, 20(17):3206-3213.
18. Xpdf Homepage [http://www.foolabs.com/xpdf/]
19. Eales JM, Stevens RD, Robertson DL: Searching in the right place: classifying manuscript sections. BMC Bioinformatics, (Submitted) 2008.
20. Mizuta Y, Collier N: An Annotation Scheme for a Rhetorical Analysis of Biology Articles. Proceedings of the Fourth Intl Conference on Language Resources and Evaluation (LREC2004) 2004.
21. Mizuta Y, Korhonen A, Mullen T, Collier N: Zone analysis in biology articles as a basis for information extraction. International Journal of Medical Informatics 2006, 75(6):468-487.
22. Wilbur WJ, Rzhetsky A, Shatkay H: New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinformatics 2006, 7(1):356.
23. Text Services, Bioinformatics, University of Manchester [http://jeales.smith.man.ac.uk:8080/TextServices/]
24. Stop word list at University of Glasgow, IR resources page [http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words]
25. The Phylontology [http://personalpages.manchester.ac.uk/postgrad/james.eales/phylontology.xml]
26. Phylogeny Programs [http://evolution.genetics.washington.edu/phylip/software.html]