BIOINFORMATICS

Vol. 20 no. 17 2004, pages 3206–3213
doi:10.1093/bioinformatics/bth386

BioRAT: extracting biological information from full-length papers

*David P. A. Corney, Bernard F. Buxton, William B. Langdon and David T. Jones**

Bioinformatics Unit, Department of Computer Science, University College London, Gower Street, London, WC1E 6BT, UK

Received December 19, 2003; revised on June 4, 2004; accepted on June 25, 2004
Advance Access publication July 1, 2004

ABSTRACT

Motivation: Converting the vast quantity of free-format text found in journals into a concise, structured format makes the researcher's quest for information easier. Recently, several information extraction systems have been developed that attempt to simplify the retrieval and analysis of biological and medical data. Most of this work has used the abstract alone, owing to the convenience of access and the quality of data. Abstracts are generally available through central collections with easy direct access (e.g. PubMed). The full-text papers contain more information, but are distributed across many locations (e.g. publishers' web sites, journal web sites and local repositories), making access more difficult.

In this paper, we present BioRAT, a new information extraction (IE) tool, specifically designed to perform biomedical IE, and which is able to locate and analyse both abstracts and full-length papers. BioRAT is a Biological Research Assistant for Text mining, and incorporates a document search ability with domain-specific IE.

Results: We show first, that BioRAT performs as well as existing systems, when applied to abstracts; and second, that significantly more information is available to BioRAT through the full-length papers than via the abstracts alone. Typically, less than half of the available information is extracted from the abstract, with the majority coming from the body of each paper. Overall, BioRAT recalled 20.31% of the target facts from the abstracts with 55.07% precision, and achieved 43.6% recall with 51.25% precision on full-length papers.

Availability: The software and documentation can be found at http://bioinf.cs.ucl.ac.uk/biorat

Contact: d.corney@cs.ucl.ac.uk; dtj@cs.ucl.ac.uk

1 INTRODUCTION

The rapid and ongoing growth in the number of biological and medical publications means that researchers can no longer read more than a small proportion of the literature in their field. Yet interesting and useful information, relevant to the researcher, could appear in papers they have not read and therefore be missed entirely. Accompanying this growth in literature is the increasing proportion of electronically available papers, as most publishers now produce on-line versions of their journals. But while this may ease access, there is still a vast quantity that a researcher may feel they should read, with no concomitant increase in their ability to do so.

Information retrieval helps researchers to find papers, but it still leaves a large amount of reading to be done. Information extraction (IE) goes one stage further, and analyses the papers on behalf of the researcher. IE systems achieve this by identifying semantic structures in the text, and in so doing, distill an entire document down to the key facts.

BioRAT can be regarded as a research assistant that is given a query and, autonomously, finds a set of papers, reads them and highlights the most relevant facts in each. BioRAT uses natural language processing techniques and domain-specific knowledge to search for patterns in documents, with the aim of identifying interesting facts. These facts can then be extracted to produce a database of information, which has a higher 'information density' than a pile of papers. This is similar to an information extraction system that has recently been developed by Blaschke and Valencia (2001, 2002), and which will be discussed in more detail below.

There have been several attempts to apply IE techniques to scientific papers, but these have used only the abstract of each paper. Example applications include protein–protein interactions (Thomas *et al*., 2000); using machine learning to classify biological relationships (Craven and Kumlien, 1999); and protein structure and residues (Gaizauskas *et al*., 2003).

Abstracts are readily available in large numbers (e.g. through PubMed, http://www.ncbi.nlm.nih.gov/entrez/), are available in plain text, and typically have no superscript or subscript characters, no footnotes and so forth. This avoids potential difficulties in interpreting unusual symbols, Greek letters, etc. However, the abstract is only a summary of the paper in question; the full text will typically include more detail that may be of direct interest to the reader. BioRAT is designed to extract information both from abstracts and

*To whom correspondence should be addressed.

from full text, and in this work, we use BioRAT to compare information extraction from abstracts and from full-length papers.

A 'challenge evaluation' has recently been proposed, to encourage researchers to focus on a particular task, allowing a direct comparison of their systems. As described by Yeh *et al*. (2003), full-length articles were used in the challenge, after they had been manually 'cleaned' to convert Greek letters, superscripts and subscripts and italics, into marked-up plain text. Furthermore, a list of genes mentioned in each paper was also available to entrants. While necessary for a formal evaluation, such resources are not generally available to text mining systems, so we have not used them here. Also, Yeh *et al*. (2003) state that PDF papers 'were not suitable for processing by most text mining systems', and so the contest was limited to those papers that were available in the HTML format. BioRAT uses the full-length paper whenever it is available, instead of just the abstract and uses PDF files directly from the Internet. PDF is one of the most widely used formats for research papers on the Internet.

The rest of this paper is organized as follows. In the next section, we describe the BioRAT system and its key components. We then discuss two experiments which evaluate BioRAT using the DIP database and discuss the results. We follow the advice of Blaschke and Valencia (2001), who specifically recommend the use of DIP as a 'realistic scenario for the comparison of IE systems'.

2 SYSTEM OUTLINE

We designed BioRAT to give people with no IE experience a powerful tool to help them locate and analyse research papers. The system therefore combines tools to locate papers, to download full-length papers, to extract information from papers and to design templates to allow this extraction.

Typically, the user enters a query into BioRAT, which is then passed on to PubMed. The user is then presented with a list of papers, from which they can choose to download abstracts or, where available, full-length papers. Having obtained some text, the user can then apply some pre-existing templates or create their own. In either case, the templates match patterns in the text that contains 'useful' information, which is extracted for display to the user and for possible incorporation into a database. Figure 1 shows screenshots of this process.

2.1 Web spidering

One distinctive feature of BioRAT is that it automatically locates and acquires full-length papers wherever possible, instead of just using abstracts. It does this via the Internet, by following a series of hyperlinks to find each target paper. To find a particular paper, BioRAT starts with a URL (web address) provided by the PubMed database. It then goes to that web page and identifies the hyperlinks there, and recursively follows links until it finds the target paper, in PDF format. This is downloaded and converted into a text-only version, ready for the IE engine.

Finding the target paper is non-trivial for such a tool. The URL provided by PubMed (and ultimately, by the journal publishers) does not point to the paper itself, but rather to a web page from which the paper can be accessed. The spider's task is to find the target paper by following a series of hyperlinks.

The system works by downloading the web page, identifying and evaluating all the links in it, and iteratively following the highest-scoring link, with scores based on simple keyword matching. Having located and downloaded a PDF file, it is converted into plain text for later analysis. To ensure that the correct paper has been identified, and that the text conversion process has succeeded, the first part of the plain text file is compared with the corresponding abstract obtained directly from PubMed, using a fuzzy string matching routine.

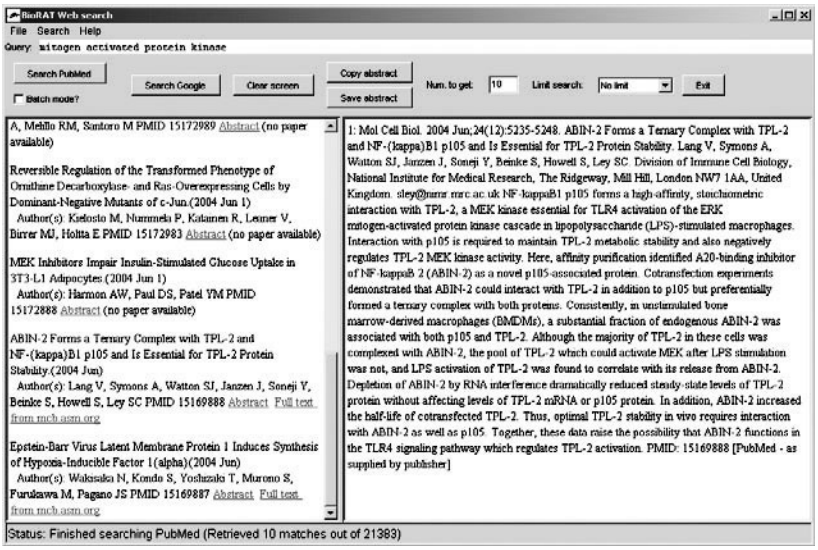
BioRAT only attempts to locate and download PDF papers, as this is by far the most widely used format. Although some have suggested a move towards using XML for distributing research papers (Murray-Rust and Rzepa, 2002), papers in this format are not generally available to biological researchers. It is also unlikely that existing archives would be marked-up manually.

Having obtained some relevant documents, the system then attempts to extract interesting facts from them.

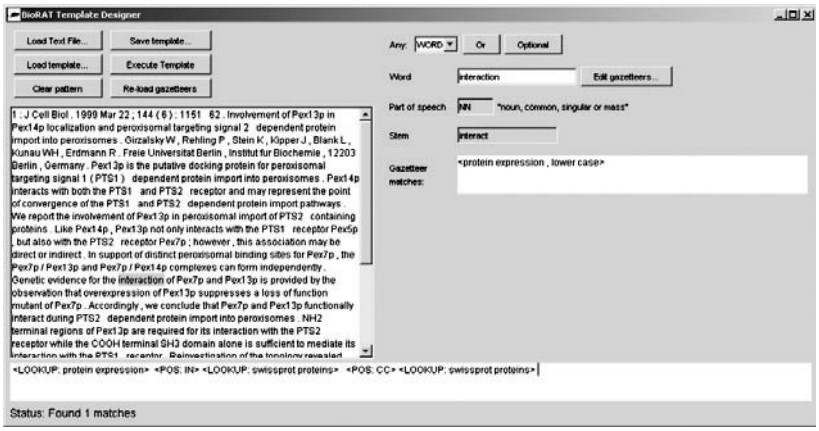
2.2 Information extraction engine

Information extraction (IE) is a key part of BioRAT's functionality. The aim of IE is to extract from a set of documents the key facts about prespecified types of events, objects and relationships. These facts are then used automatically to populate a database. This can then be used to ease on-line access.

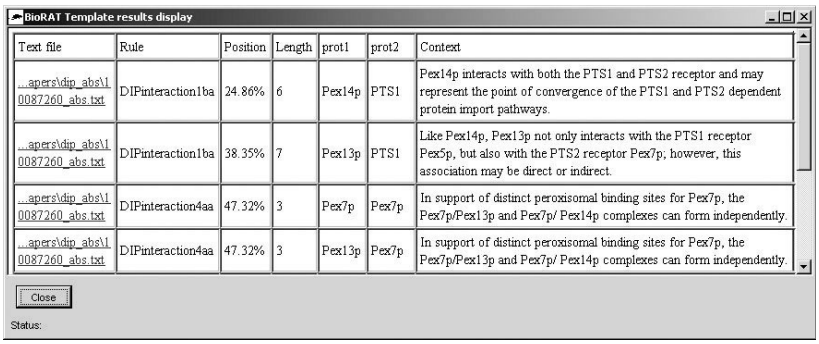
The heart of BioRAT is an IE engine, based on the GATE toolbox (General Architecture for Text Engineering, http://gate.ac.uk/), produced at Sheffield University (Cunningham *et al*., 2002). GATE is a general purpose text engineering system, whose modular and flexible design allows us to use it to create a more specialized biological IE system. One issue common to any biological information extraction system is that many protein and gene names are easily mistaken for common words. For example, the Swiss-Prot database includes entries with names 'mice', 'was' and 'alpha', as well as 26 single-letter gene names. The problem is to distinguish whether the word 'was' refers to a gene or is simply the past tense of the verb 'to be', for example. Sometimes, this can be resolved by considering the case of the letters, but this is not reliable. Instead, BioRAT uses GATE to label words according to their parts of speech, and then applies a filter that rejects determinants verbs, etc. as not being proteins. This provides one possible advantage of BioRAT. Two components of GATE that must be modified for our domain-specific application are gazetteers and templates, which we shall now discuss in turn.



(a) Finding papers on the web



(b) Designing a template



Text file	Rule	Position	Length	prot1	prot2	Context
...apers\dip_abs\10087260_abs.txt	DIPinteraction1ba	24.86%	6	Pex14p	PTS1	Pex14p interacts with both the PTS1 and PTS2 receptor and may represent the point of convergence of the PTS1 and PTS2 dependent protein import pathways.
...apers\dip_abs\10087260_abs.txt	DIPinteraction1ba	38.35%	7	Pex13p	PTS1	Like Pex14p, Pex13p not only interacts with the PTS1 receptor Pex5p, but also with the PTS2 receptor Pex7p; however, this association may be direct or indirect.
...apers\dip_abs\10087260_abs.txt	DIPinteraction4aa	47.32%	3	Pex7p	Pex7p	In support of distinct peroxisomal binding sites for Pex7p, the Pex7p/Pex13p and Pex7p/ Pex14p complexes can form independently.
...apers\dip_abs\10087260_abs.txt	DIPinteraction4aa	47.32%	3	Pex13p	Pex7p	In support of distinct peroxisomal binding sites for Pex7p, the Pex7p/Pex13p and Pex7p/ Pex14p complexes can form independently.

(c) Extracting information

Fig. 1. Screen shots showing BioRAT in use. (**a**) The document search interface. The user enters a query at the top and BioRAT accesses PubMed via the Internet. A list of matching titles (with date of publication, author, etc.) is shown on the left and the user can select any item to view the abstract, on the right or to download the full-length paper. (**b**) The BioRAT template design component. The user can view a document, select target words (or a phrase) from it, and then define a template in terms of parts of speech, gazetteer headings, word stems or the words themselves. Gazetteers can also be viewed and edited through the same interface. (**c**) The results from templates designed to recognize protein–protein interactions. Four interactions are shown, with the context quoted from the source text. A command line interface is also available.

2.2.1 Gazetteers One task in IE is 'named entity recognition', which aims to identify key items within text. For example, we may want to identify words that are people's names, company names, proteins, genes and so on. Once identified, these words or phrases can then be matched by the templates. One simple approach that we adopt is to use a gazetteer.

A gazetteer is a list of words identifying members of a particular category. For example, one gazetteer may list names of proteins, while another lists names of people. BioRAT incorporates gazetteers from three sources, namely MeSH (Medical Subject Hierarchy, http://www.nlm.nih.gov/mesh/), Swiss-Prot (http://www.expasy.org/) and hand-made lists.

The top two levels of the MeSH hierarchy contain a total of approximately 120 entries, each of which was used to define a separate gazetteer. Each of the almost 22 000 entries in MeSH was extracted and added to the appropriate gazetteer(s). Further gazetteers were derived from Swiss-Prot. Each entry from Swiss-Prot describes a single protein, but proteins often have many synonyms, all of which are included in the relevant gazetteer. Also, some authors refer to proteins in terms of the genes that encode them, so the gene names were also extracted, and used to create another gazetteer.

To supplement these two sources, two further gazetteers were created by hand. These comprised words that covered concepts of interest that were not already in other gazetteers. One consisted of 30 words describing the interaction of proteins (e.g. 'bind', 'down-regulate', 'interact' and so on). The other consisted of a few further synonyms of proteins not already covered by the other gazetteers. These hand-made gazetteers were initially created following domain expert advice, and subsequently modified as required.

2.2.2 Templates A template is a representation of a text pattern that allows us to extract information automatically. It consists of a number of predefined slots to be filled by the system from information contained in the text. One of the simplest templates from BioRAT is:

'interaction of' (PROTEIN_1) 'and' (PROTEIN_2)

Here, 'PROTEIN_1' and 'PROTEIN_2' are slots to be filled with names of proteins, as defined by a gazetteer. The contextual phrase ('interaction of') is a fixed string: only phrases containing those exact words will be matched by this particular template. For example, the template shown would identify the sentence 'Genetic evidence for the interaction of Pex7p and Pex13p is provided . . .' and extract from it the interaction (Pex7p ↔ Pex13p)[1].

[1]We use the format 'X ↔ Y' to represent any form of interaction between two proteins, X and Y.

A slightly more complicated template is:

(EXPRESSION) 'of' (PROTEIN_1)
(WORD)? (WORD)? (WORD)?
('by' | 'to' | 'with')
(PROTEIN_2) 'and' (PROTEIN_3)

Here, 'EXPRESSION' refers to a gazetteer containing words relating to protein expression and interaction, such as 'bind' and 'inhibit'. The slot (WORD)? is a wildcard that matches any word, but is optional, so the sequence (WORD)? (WORD)? (WORD)? matches between zero and three consecutive words of any type. As before, the three (PROTEIN_x) slots match protein names, and the quoted strings must be matched exactly. The | character is a logical 'OR'. For example, this template matches part of the sentence 'Specific binding of Rna15 in complex with Hrp1 and Rna14 creates a polymerase pause site . . .', and identifies two interactions: (Rna15 ↔ Hrp1) and (Rna15 ↔ Rna14), with the expression type 'binding'.

As with comparable IE systems, such as those mentioned in the introduction, the templates in BioRAT are written by hand. There have been attempts at automatic template creation (Collier, 1998), but these have not been broadly applicable. Although template design takes time and requires some practice, it does allow the user to maintain full control over what information is extracted, and allows experts to incorporate their knowledge within the system. Because of this, BioRAT incorporates a template design tool, designed to allow ordinary users to create their own templates with little effort, as discussed in the next section.

BioRAT produces data in XML format, which can be readily imported into existing database query systems. The same data are produced simultaneously as HTML and as a comma-separated list, for viewing in applications such as a browser or a spreadsheet, if that is more convenient for the user. Each record in the resulting database represents a single completed template.

2.3 Template design tool

One feature that BioRAT shares with several other text mining systems is the need for a set of templates to be developed for each task. This is often a time-consuming process that requires expertise in both text mining and the problem domain. BioRAT includes a template design tool with a graphical user interface, which allows non-expert users to develop templates without having to learn a complex new language. To use it, the user first selects a document that is then displayed. The user can then click on individual words in that document, whose properties are then shown on the screen. The properties used are: part-of-speech tag; gazetteer headings; the word stem; and the word itself. The user can click on these properties to append them to the current template pattern, along with various wildcard and boolean options, and build up a sequence of terms. This can then be applied as a template to the current

document, and the results displayed. The user can then cycle between editing the template and viewing the results, until satisfied. Once saved, the template can then be applied to a large set of papers using the main BioRAT template matching interface. Alternatively, the user can select an entire phrase, and the system will create a default template based on that phrase, which the user can subsequently edit as required. The tool can also be used to view and edit gazetteers.

3 USING DIP TO EVALUATE BioRAT

Having described the BioRAT system, and considered the documents that it can be used to analyse, we now turn to a particular study to test the usefulness of the system. For this, we used the Database of Interacting Proteins (DIP, http://dip.doe-mbi.ucla.edu) (Xenarios *et al*., 2002). Blaschke and Valencia (2001) recommend using DIP as a way of evaluating biological IE systems, because it represents a realistic problem of practical interest to biological researchers. IE researchers can use their systems to extract protein-protein interactions, and then compare these with the records in DIP. This does not rely on the interpretation of the authors, and so gives greater confidence in the results. By re-creating (a manageable subset of) DIP, we can calculate the recall and precision of different systems, and compare the results. The recall (or 'sensitivity') is the fraction of target records that the IE system correctly re-creates[2]. Precision is a measure of how much of the output of an IE system is correct, and is defined as the ratio of the number of correct positive predictions to the total number of positive predictions made[3].

Each record in DIP defines a pair of proteins that interact with each other, and provides citations of papers that describe the interaction. Proteins are defined by entry keys to Swiss-Prot, GenBank or PIR. For simplicity, we only consider DIP records containing two Swiss-Prot identifiers.

For each experiment, we started by selecting a subset of DIP. BioRAT can analyse papers rapidly, typically taking just a few seconds to complete its analysis of each abstract. However, for our experiments, the results need to be manually checked in order to calculate the recall and precision rates, and this time-consuming task forced us to limit the targets to a manageable subset of DIP.

Having selected some DIP records, as detailed below, we then used BioRAT to process the corresponding papers, using both the abstract and full-text versions. We manually compared the predictions made by BioRAT to the source DIP records to measure the recall. For each record in DIP, we search through the output of BioRAT corresponding to the same paper, and checked to see if the interaction mentioned in DIP had been identified. Similarly, we measured the precision by manually counting how many of the records produced by BioRAT were correctly extracted from the text. Throughout this work, we used the January 2003 version ('dip20030105') of DIP, the March 2003 version of Swiss-Prot and the 2003 edition of MeSH.

[2] $\text{Recall} = \frac{\text{No. of true positives}}{\text{No. of true positives} + \text{No. of false negatives}}$

[3] $\text{Precision} = \frac{\text{No. of true positives}}{\text{No. of true positives} + \text{No. of false positives}}$

4 EXPERIMENTS

4.1 Comparison with SUISEKI

In this section, we compare BioRAT with the existing SUISEKI information extraction engine described by Blaschke and Valencia (2001, 2002). We compare the performance of BioRAT to that of their system by measuring the recall of BioRAT on a sample of papers from DIP that were also used by Blaschke and Valencia (2001). This provides a suitable benchmark for BioRAT.

The SUISEKI system, like BioRAT, uses gazetteers derived from Swiss-Prot and DIP to identify protein names. To extract information, it uses 'frames', which are similar to BioRAT's templates in that they define patterns of language that form the basis for IE. However, the frames in SUISEKI make less use of linguistic knowledge, but more use of statistics. For example, the frames in SUISEKI distinguish between nouns and verbs, but do not recognize conjunctions, adjectives or any other parts of speech. Also, they count the number of words occurring in a phrase, and favour short phrases over long ones.

There were 389 records from DIP, which were used by Blaschke and Valencia (2001) and have a DIP record that refers to two Swiss-Prot records. These 389 DIP records relate to 229 PubMed citations. We applied BioRAT to all 229 abstracts, and then analysed the results by hand.

We used a total of 19 templates, initially derived from the SUISEKI frames and subsequently modified by hand; and 127 gazetteers, derived from MeSH and other sources, as described earlier. The templates and gazetteers used here can be accessed from the same website as the BioRAT software, http://bioinf.cs.ucl.ac.uk/biorat. Initial trials revealed weaknesses in both the templates and the gazetteers, which were subsequently improved.

Table 1 shows the recall from these abstracts by BioRAT, namely 20.31%. This is a similar recall to that achieved by SUISEKI. The results can be compared with the larger study reported by Blaschke and Valencia (2002), where 190 DIP interactions were correctly detected, from a possible set of 851 interactions, giving a recall score of 22.33%.

We can compare the 'abstract' results (Table 1) to the results in Blaschke and Valencia (2002), if we assume the results follow a binomial distribution. Our recall rate of 20.31% from 389 trials gives a variance of $\sigma^2 = 389 \times 0.2031 \times (1 - 0.2031) = 62.96$ and hence a SD of $\sigma = 7.934$. Blaschke and Valencia quote a recall of 190 cases from 851 trials, giving a recall rate of $190/851 = 0.2233$. If they had achieved the same rate on our smaller sample, we would expect them to achieve $389 \times 0.2233 = 86.86$ successes. This is within 1 SD

Table 1. Comparison of BioRAT and SUISEKI on recall from abstracts

Result	BioRAT		SUISEKI	
	Cases	Percent	Cases	Percent
Match	79	20.31	190	22.33
No match	310	79.69	661	77.67
Totals	389	100.00	851	100.00

BioRAT results from 389 DIP records, derived from 229 abstracts. SUISEKI results from 851 DIP records, derived from 514 abstracts. The former set of records is a subset of the latter.

of our success score, so we can say that both systems are performing with approximately the same recall.

4.2 Abstracts versus full-length papers

In the second experiment, we want to assess the benefits of using the full-length version of a paper, rather than just the abstract. Clearly, one would expect to extract more information from the full paper, than just the abstract. However, obtaining full-length papers requires extra time and resources, in terms of locating and downloading them, processing the extra text, storing extra files and so on. If the gain in recall is small, this may not be worth the extra effort. Also, we need to discover whether the conversion of PDF papers to text loses too much information, such as Greek letters and superscript or subscript information, and to discover the effect on precision of having a lot of extra text.

We took a random sample of 211 DIP records, based on 130 different documents, where full text and abstract are both available. We used BioRAT to extract protein–protein interactions from both, and then compared the results. We were, of course, limited to articles that are available electronically. For example, this excluded most papers that were published before the mid-1990s, when most journals were paper-only. Also, the experiments described here were carried out using computers at UCL, and so we could only access full-length papers from journals to which UCL subscribes or that are freely available.

Table 2 shows the results. The information extraction rate obtained from full-length papers was 43.6%, with more than half of the information coming from the body of the paper, and the rest from the abstract. This clearly shows the benefit of locating and analysing the full text of a paper, rather than restricting information extraction to just the abstract.

Using a similar binomial analysis to that described earlier, we can also test whether this improvement is significantly better than the information extracted from just the abstracts. The SD of the recall from the abstracts is $\sigma = \sqrt{211 \times 0.1800 \times (1 - 0.1800)} = 5.582$. Thus, the recall score using abstracts is more than 7 SD below the recall score using full-length papers, clearly a significant result.

Table 2. Recall results from 211 DIP records, derived from 130 full-length papers; the total recall from full-length papers is $18.0 + 25.6 = 43.6\%$

Result	Cases	Percent
Match in abstract	38	18.00
Match in full text (but not in abstract)	54	25.60
No match	119	56.40
Totals	211	100.00

Table 3. Precision analysis

Result	Abstracts		Full-length	
	Cases	Percent	Cases	Percent
Correct	239	55.07	205	51.25
Protein id	125	28.80	119	29.75
Template	70	16.13	76	19.00
Totals	434	100.00	400	100.00

Here, 'correct' refers to records where the interaction information was extracted correctly from the text, regardless of whether that interaction is in DIP. 'Template' refers to failures caused by imperfect templates and 'protein id' refers to failures to recognize proteins.

4.3 Precision

Having analysed the recall of BioRAT in the previous sections, we now turn to precision. In our experiments, precision is somewhat harder to measure than recall, because we need an estimate of the number of false positives. If a record produced by BioRAT is not found in DIP, it could be that (a) it is a false-positive example, reducing the precision of BioRAT; or (b) the record is missing from DIP. The latter case consists of interactions that are mentioned in papers, but have not (yet) been added to DIP.

For the first experiment described earlier, BioRAT produced 434 interaction records, derived from 229 abstracts. We manually re-analysed these records with no reference to DIP but instead we counted how many of BioRAT's predictions were correctly extracted from the text, and what sort of mistakes it made. We repeated this for a sample of 400 from the total of over 10 000 records produced in the analysis of the corresponding full-length papers. Table 3 shows the results.

Around half of all records produced by BioRAT are correct, in the sense that the information contained in the papers was correctly extracted, whether or not the information is in DIP. In order to understand where BioRAT fails, we analysed the output when BioRAT failed to extract the correct information from the documents, also shown in Table 3. Around two-thirds of the mistakes are caused by failure to identify the correct proteins. Each protein is typically known by several different names, and may also be referred to by its associated gene, which itself may have several distinct names. Furthermore, long names may be abbreviated by the

authors, producing further non-standard ways of referring to the protein. The gazetteer used in these experiments included more than 230 000 gene names and more than 99 000 protein names, but still failed to recognize a large number of proteins.

One example of this protein identification failure comes from DIP 'edge' record DIP:43E. The corresponding Swiss-Prot entry (P15172) refers to the protein 'Myoblast determination protein 1', and lists synonyms 'Myogenic factor 3' and 'Myf-3', with gene names 'MYOD1' and 'MYF3'. However, the paper in question (PMID 9184158) refers repeatedly to 'MYOD'. While this is clearly the same protein, a slightly different abbreviation has been used by the author compared to those included in Swiss-Prot. The gazetteer used by BioRAT is derived principally from Swiss-Prot, and so BioRAT failed to recognize this protein, and hence failed to extract this interaction.

Most of the remaining failures are due to imperfections in the set of templates used by BioRAT. Although these errors could no doubt be reduced by improving the templates, there is no clear way to achieve this without a significant manual effort, even with BioRAT's template design tool. Thus, template design remains a major issue in information extraction research (Cowie and Wilks, 2000).

Even when BioRAT fails to extract the relevant information, it may still highlight the correct piece of text. For example, DIP record DIP:800E defines an interaction between proteins p53 and UBE2I. BioRAT failed to identify this interaction, but did extract this sentence (PMID 8921390):

> Since the tumor suppresser protein p53 and a newly identified ubiquitin-like protein (UBL1) are implicated in the RAD51/RAD52 complex . . ., we further tested their associations with UBE2I.

Note that BioRAT correctly identified the above sentence as defining the interaction between RAD51 and RAD52, even though it missed the target interaction.

4.4 Example output

From PubMed ID 9012827, BioRAT found the interaction (Swi6 ↔ Hrr25), which corresponds to the DIP 'edge' record DIP:250E. BioRAT quoted the following sentence:

> These observations show that Swi6 is phosphorylated by a kinase with the expected properties of Hrr25.

A similar, but slightly more complex template can recognize two interactions at once. The following sentence (PMID 11689698) correctly lead BioRAT to produce two records for the interactions (Pcf11 ↔ Rna14) and (Pcf11 ↔ Rna15).

> Since Pcf11 interacts simultaneously with Rna14 and Rna15, its role *in vivo* may also be to stabilize their interaction.

A less successful example comes from this sentence:

> Many interactions between nucleoporins and nuclear transport receptors have already been identified; however, we were unable to detect a biochemical interaction between Cse1p and Nup2p.

BioRAT incorrectly predicted that Cse1p interacted with Nup2p, whereas the text is less conclusive.

4.5 Speed and memory

The time it takes BioRAT to analyse a piece of text depends on the size of the text, the size of gazetteers, and the complexity of the templates. In the work described here, BioRAT typically took 3–5 s to analyse each abstract, and 6–10 min to analyse each full-length paper, running on a standard desktop PC (a single 1.7 GHz CPU), and used ∼500 Mb of RAM. Given that each paper can be analysed independently, large-scale applications of text mining lend themselves well to distributed processing, although we have yet to use BioRAT in that way. BioRAT can also be used from the command line, allowing non-interactive batch processing, and potentially reducing the impact of a slow execution time for full-length papers. Since it is written in Java, BioRAT can be run on almost any platform, and has been tested successfully under Linux, Solaris, MacOS and MS Windows.

5 DISCUSSION

As expected, the density of 'interesting' facts found in the abstract is much higher than the corresponding density in the full text. This is at least in part because full-length papers include background discussion, a description of the method, references and so on. While these are necessary to set the work in context, and to provide supporting evidence, they may not contain the kind of information that BioRAT is attempting to extract.

Figure 2 is one view of information density. It shows the location of each fact extracted from the set of full-length papers used earlier. As different journals divide papers into sections in different ways, we only consider the location of the information relative to the entire paper. The peak on the left shows that a lot of information is found at the start of the paper, corresponding approximately to the title and abstract. The dip in the graph ∼10–30% shows that relatively little information is extracted from the next section, typically the Introduction and Methods sections. There is another larger peak ∼40–80%, corresponding to the results and discussion sections, which contain a large amount of relevant information, before tailing off towards the end of the paper, which is typically a citation list. Note that many interactions were found more than once, through repetition within or between papers, and the graph shows the location of all the extracted information, including duplicates.

These peaks show from where most of the information has been extracted, but the troughs are also of interest. Even the

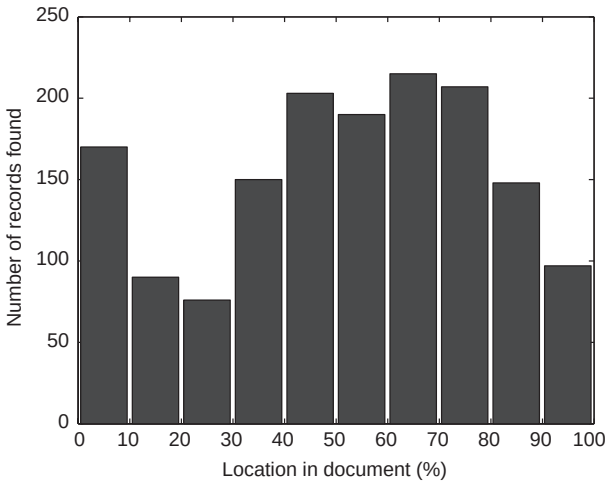


Fig. 2. Location of information extracted from full-length papers. Location 0% is the start of the paper; location 100% is the end. The peak on the left corresponds to the abstract; the larger peak in the middle corresponds to the results and discussion sections of the source papers.

least informative parts of papers still contain considerable amounts of information. This strongly suggests that the entire paper should be analysed, wherever possible, and not just a few selected sections. Although the task is different, this contrasts with the behaviour of some of the teams described by Yeh *et al*. (2003), who restricted analysis to certain sections of the papers.

Even when BioRAT (or any other IE system) fails to find a particular relationship, or incorrectly predicts a relationship not mentioned in the text, it is quite possible that it has found an interesting part of an interesting document. In this way, using IE to guide a literature search is perfectly feasible, even if the recall and precision are a long way from the ideal 100%.

The template design tool allows biological researchers, with no text mining experience, to design, test and use a sophisticated template-based information extraction system. This flexibility allows BioRAT to be applied to a wide range of problems without a large overhead, in contrast to many comparable systems, which require both biological and text mining expertise for them to be used fully.

The results that BioRAT produces can be stored and retrieved using a variety of interfaces, easing the user's access to the information. Furthermore, BioRAT also provides quotes from the source texts, and links directly to the source papers and related databases. In this way, BioRAT behaves like a virtual research assistant, guiding the user towards interesting papers.

6 CONCLUSIONS

In this paper, we have presented BioRAT, an information extraction system specially designed to process biological research papers. A distinguishing feature of BioRAT is that it uses full-length papers, rather than being limited to abstracts as previous studies have been. The recall and precision performance of BioRAT was assessed by use of the DIP database of protein–protein interactions, and the recall was compared with that of a previous system, SUISEKI, which processed only the abstracts. The recall performance of BioRAT on the abstracts alone (20%) was similar to that of SUISEKI. Overall, BioRAT achieved 43% recall and over 50% precision on full-length papers. Extra time is required to obtain the full-length papers, and there are difficulties in converting them into a usable plain text format. However, these costs are outweighed by the fact that more than twice as much relevant information can then be extracted automatically.

ACKNOWLEDGEMENTS

This work was sponsored by GlaxoSmithKline.

REFERENCES

Blaschke,C. and Valencia,A. (2001) Can bibliographic pointers for known biological data be found automatically? Protein interactions as a case study. *Comp. Funct. Genomics*, **2**, 196–206.

Blaschke,C. and Valencia,A. (2002) The frame-based module of the SUISEKI information extraction system. *IEEE Intell. Syst.*, **17**, 14–20.

Collier,R. (1998) Automatic template creation for information extraction. PhD thesis, Department of Computer Science, University of Sheffield, UK.

Cowie,J. and Wilks,Y. (2000) Information extraction. In Dale,R., Moisl,H. and Somers,H. (eds) *Handbook of Natural Language Processing*. Marcel Dekker, New York.

Craven,M. and Kumlien,J. (1999) Constructing biological knowledge-bases by extracting information from text sources. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*. Heidelberg, Germany, pp. 77–86.

Cunningham,H., Maynard,D., Bontcheva,K. and Tablan,V. (2002) GATE: a framework and graphical development environment for robust NLP tools and applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, Philadelphia, USA.

Gaizauskas,R., Demetriou,G., Artymiuk,P. and Willett,P. (2003) Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics*, **19**, 135–143.

Murray-Rust,P. and Rzepa,H.S. (2002) STMML. A markup language for scientific, technical and medical publishing. *Data Science*, **1**, 1–65.

Thomas,J., Milward,D., Ouzounis,C., Pulman,S. and Carroll,M. (2000) Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.*, **5**, 538–549.

Xenarios,I., Salwinski,L., Duan,X., Higney,P., Kim,S. and Eisenberg,D. (2002) DIP: the database of interacting proteins. A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.

Yeh,A., Hirschman,L. and Morgan,A. (2003) Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics*, **19** (Suppl. 1), i331–i339.