BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 11 2009, pages 1404–1411
doi:10.1093/bioinformatics/btp175

Data and text mining

KiPar, a tool for systematic information retrieval regarding parameters for kinetic modelling of yeast metabolic pathways

Irena Spasić[1,2,*], Evangelos Simeonidis[1,3], Hanan L. Messiha[1,4], Norman W. Paton[1,2] and Douglas B. Kell[1,4,*]

[1]Manchester Centre for Integrative Systems Biology, [2]School of Computer Science, [3]School of Chemical Engineering and Analytical Science and [4]School of Chemistry, The University of Manchester, Manchester, UK

Received on August 12, 2008; revised on March 9, 2009; accepted on March 25, 2009

Advance Access publication March 31, 2009

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Most experimental evidence on kinetic parameters is buried in the literature, whose manual searching is complex, time consuming and partial. These shortcomings become particularly acute in systems biology, where these parameters need to be integrated into detailed, genome-scale, metabolic models. These problems are addressed by KiPar, a dedicated information retrieval system designed to facilitate access to the literature relevant for kinetic modelling of a given metabolic pathway in yeast. Searching for kinetic data in the context of an individual pathway offers modularity as a way of tackling the complexity of developing a full metabolic model. It is also suitable for large-scale mining, since multiple reactions and their kinetic parameters can be specified in a single search request, rather than one reaction at a time, which is unsuitable given the size of genome-scale models.

Results: We developed an integrative approach, combining public data and software resources for the rapid development of large-scale text mining tools targeting complex biological information. The user supplies input in the form of identifiers used in relevant data resources to refer to the concepts of interest, e.g. EC numbers, GO and SBO identifiers. By doing so, the user is freed from providing any other knowledge or terminology concerned with these concepts and their relations, since they are retrieved from these and cross-referenced resources automatically. The terminology acquired is used to index the literature by mapping concepts to their synonyms, and then to textual documents mentioning them. The indexing results and the previously acquired knowledge about relations between concepts are used to formulate complex search queries aiming at documents relevant to the user's information needs. The conceptual approach is demonstrated in the implementation of KiPar. Evaluation reveals that KiPar performs better than a Boolean search. The precision achieved for abstracts (60%) and full-text articles (48%) is considerably better than the baseline precision (44% and 24%, respectively). The baseline recall is improved by 36% for abstracts and by 100% for full text. It appears that full-text articles are a much richer source of information on kinetic data than are their abstracts. Finally, the combined results for abstracts and full text compared with the curated literature provide high values for relative recall (88%) and novelty ratio (92%), suggesting that the system is able to retrieve a high proportion of new documents.

Availability: Source code and documentation are available at: http://www.mcisb.org/resources/kipar/

Contact: i.spasic@manchester.ac.uk; dbk@manchester.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Systems biology (SB) has emerged as an approach to studying biological systems by understanding how the heterogeneous parts combine to form the whole (Henry, 2003) through systematic integration of technology, biology and computation (Hood, 2003; Kell, 2006). At the heart of SB is an iterative interplay between (i) mathematical/computational simulation of a biological system and (ii) experimental measurements of parameters and variables with which a model can be populated and/or validated. A principal goal of SB is to develop and exploit appropriate methods for detailed, genome-scale, metabolic modelling: the virtual cell, which can be used to simulate the dynamic aspects of biological systems (e.g. kinetic behaviour) *in silico*. While the structure of such systems is now relatively well known (Herrgård *et al*., 2008; Michal, 1999), access to the required kinetic parameters such as the Michaelis constant, K_m, and the catalytic constant, k_{cat}, remains a big challenge.

Most experimental evidence on kinetic parameters is buried in the scientific literature. However, its manual searching is time consuming and complex (Hull *et al*., 2008), and we anticipate that text mining (TM) technologies (Ananiadou *et al*., 2006; Jensen *et al*., 2006) may be appropriate for this. Currently, most TM applications for metabolic modelling focus on qualitative information (e.g. Ding, *et al*., 2002; Hoffmann *et al*., 2005; Humphreys *et al*., 2000; Rzhetsky *et al*., 2004; Yuryev *et al*., 2006). In addition, Humphreys *et al*. (2000) extract quantitative information (e.g. temperature and concentration) related to the participants of metabolic reactions. Only Hakenberg *et al*. (2004) focus on quantitative information related to the kinetics of metabolism. The main difficulty such applications face is the complexity of the problem: metabolic pathways involve various chemical alterations of relatively small molecules and are subject to stoichiometric constraints.

*To whom correspondence should be addressed.

Consequently, mining the literature for metabolic information cannot be tackled effectively with simple approaches based on co-occurrence information (Rzhetsky *et al.*, 2004). While such approaches produce good results for some types of relations between biological entities (e.g. protein–protein interactions), they are not applicable to relations between enzymes in metabolic pathways, because enzymes catalysing successive reactions in a pathway are rarely mentioned in the same text passage (Hoffmann *et al.*, 2005).

Even mining the literature for an individual metabolic reaction is difficult, since such a reaction is typically not named explicitly, hence it cannot be treated like a named entity. Reactions are complex events with different entities having specific roles (e.g. substrates, products or enzymes). Instead of names, reactions are often referred to by their descriptions, e.g. '*the breakdown of hydrogen peroxide into water and oxygen* ($2H_2O_2 \rightarrow H_2O + O_2$) *which is catalyzed by the enzyme catalase*'. Although the exact phrasing cannot be predicted reliably, search queries combining the names of the reaction participants may help retrieve documents relevant to the reaction, e.g.

'hydrogen peroxide' AND water AND oxygen AND catalase

However, the problem is further complicated by high levels of both *variation* (a single concept is expressed by a number of synonyms, e.g. *D-glucopyranose*, *dextrose*, *D-glucose* and *grape sugar* refer to the same compound) and *ambiguity* (a single name refers to multiple concepts, e.g. *reaction* may describe 'an interaction of chemical entities', but also 'a response to some treatment, situation, or stimulus') (Spasic *et al*., 2005). Therefore, it is unrealistic to expect a user to supply both knowledge and terminology concerned with every reaction s/he may be interested in, especially not in the case of genome-scale metabolic modelling. In addition, searching for information about 'one-concept-at-a-time' (e.g. a reaction) is not suitable for large-scale literature mining (Müller *et al*., 2004; Shatkay, 2005).

Hence, we developed KiPar as a dedicated information retrieval (IR) system to facilitate access to the literature relevant for the problem of kinetic modelling of metabolism. Given an input consisting of a metabolic pathway, a subset of its reactions and a set of required kinetic parameters, the system retrieves documents that are likely to contain a value of a given parameter applicable to a given reaction. As a result, the user is provided with a literature starter pack for studying the kinetic aspects of a particular metabolic pathway, where individual documents cover the kinetics of individual reactions of the pathway. To facilitate the navigation among and inside the retrieved documents, each document is annotated with relevant information, i.e. particular parameter(s) and reaction(s) that apply to it. By doing so, KiPar aims to reduce the time involved in the kinetic modelling of metabolic pathways.

The system is 'high-throughput' in two ways. First, the system allows multiple reactions in a single search request. Second, apart from specifying the reactions and parameters of interest, the user is freed from providing any other knowledge and terminology of relevance or formulating complex search queries. Instead, the system retrieves relevant knowledge and terminologies from public data resources on the fly, and combines the retrieved information automatically to perform complex literature searches that are completely transparent to the user.

2 METHODOLOGY

2.1 Problem specification

KiPar is a computer application for the retrieval of textual documents likely to contain parameters required for kinetic modelling of a given metabolic pathway in yeast. During the retrieval process different types of information are considered (Fig. 1). Based on the requirements of kinetic metabolic modelling, we identified the core information that needs to be supplied as user input (given in bold type set in Fig. 1). These include: (i) the *enzymes* catalysing the reactions of interest, (ii) a *pathway* to which these reactions belong, and (iii) the *parameters* whose values are required for kinetic modelling. Within the context of a metabolic reaction, the remaining types of information are enzyme dependent and can be retrieved from public data resources. Given an enzyme, the known information about (i) the *compounds* acting as substrates/products of the reaction catalysed, and (ii) the *genes* encoding the enzyme can be retrieved from the KEGG ENZYME database (DB) (Kanehisa *et al*., 2008). Such information is retrieved automatically, and therefore need not be specified explicitly as a part of the user input.

2.2 Input specification

Given the terminological variability (Spasic *et al*., 2005) of biomedical sublanguages (Friedman *et al*., 2002; Harris, 2002), KiPar does not accept free-text descriptions as input. Instead, the user is asked to specify identifiers used in relevant biomedical ontologies and DBs. Enzymes are specified by *EC numbers* as their identifiers in KEGG ENZYME (Kanehisa *et al*., 2008; KEGG, 2008), e.g. 2.7.1.1 is used to request information on hexokinase. Further, a pathway is specified by the *Gene Ontology (GO) terms* (Ashburner *et al*., 2000; GO, 2008), whose entries describe it as a biological process, e.g. GO:0006096 is used for glycolysis. Finally, the required kinetic parameters are specified by the corresponding *Systems Biology Ontology (SBO) terms* (Le Novere, 2006; SBO, 2008), e.g. SBO:0000025 is used to represent k_{cat}. By supplying widely recognized identifiers for the concepts, rather then their possibly ambiguous names, we facilitate subsequent integration of information acquired from disparate public data resources.

Input information together with the configuration parameters are specified in an XML format described using XML Schema language. This enabled us to deploy Pedro, a model-driven data capture tool (Garwood *et al*., 2004), for interactive manipulation of input data (see Supplementary Material 1 for a screenshot of the user interface).

2.3 Workflow

Figure 2 depicts the structure of the system, numbered in a logical sequence of the elementary acts that it performs. Given the high-level input specification (box 1), KiPar employs a range of integrated bioinformatics strategies (mostly based on web services) to harvest reaction-specific terms (e.g. an enzyme, compounds acting as substrates/products, and the genes encoding the enzyme) from publicly available biological DBs (box 2): KEGG, PubChem (PubChem, 2008), ChEBI (ChEBI, 2008; Degtyarenko *et al*., 2008), SGD (Cherry *et al*., 1998; SGD, 2008) and CYGD (CYGD, 2008; Güldener *et al*., 2005). The problem of terminological variability is further tackled by collecting additional synonyms from the UMLS (Bodenreider, 2004; UMLS, 2008) (box 3). The collected terms are used to support a transition from conceptual to textual space. In order to query the literature for

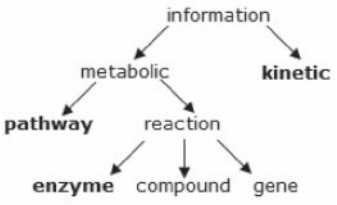

Fig. 1. Different types of information considered.

information required for a kinetic model of a given pathway (box 4), KiPar first indexes the literature with relevant concepts (i.e. pathway, enzymes and kinetic parameters specified by the user as well as related compounds and genes retrieved from KEGG). The indexing process involves mapping each concept to a query based on the synonyms acquired in the previous steps (see Supplementary Material 2 for an illustration). For example, the following query was generated automatically to search for information on enzyme with EC number 2.7.1.40:

"2.7.1.40"[RN] OR
"*pyruvate kinase*"[TEXT:noexp] OR
"*phosphoenolpyruvate kinase*"[TEXT:noexp] OR (Q1)
"*phosphoenol transphosphorylase*"[TEXT:noexp]

To keep indexing within a context of kinetic modelling for yeast, each query is further constrained with a user-specified combination of keywords. We have chosen the following constraint based on the analysis of a sample of relevant documents:

(((*brewer* OR *baker* OR *budding*) AND *yeast*) OR
((*S* OR *Saccharomyces*) AND *cerevisiae*)) AND
(*enzyme* OR *reaction* OR *substrate* OR *product*) AND (Q2)
(*kinetic* OR *parameter* OR *constant* OR *concentration* OR *rate*)

The indexing query (e.g. $Q1$ AND $Q2$) is then passed to Entrez (Entrez, 2008), an IR system that enables access to information from many NCBI DBs (Wheeler *et al*., 2008), including two literature DBs, PubMed (PubMed, 2008) and PubMed Central (PMC, 2008). By taking advantage of its search facilities (including indexing, document annotations, search term tagging, query expansion, etc.), we effectively avoid the need to locally store and manage a document collection (e.g. with an IR library such as Lucene). Instead, only information gathered about concepts, terms, documents (but not the documents themselves) and their relations (e.g. mapping of concepts to the documents that mention them) is stored in a local DB, which is then queried for relevant information within the indexed documents. Each document is scored using a weighted formula that combines different types of information considered (i.e. enzymes, compounds, genes, pathways and kinetic parameters) following the structure given in Figure 1:

$$\begin{aligned} S &= \omega_m \cdot S_m + \omega_k \cdot M(K) \\ S_m &= \omega_p \cdot M(P) + \omega_r \cdot \max_{e \in E} \{S_r(e)\} \\ S_r &= \omega_\varepsilon \cdot M(\{e\}) + \omega_c \cdot M(C_e) + \omega_g \cdot M(G_e) \end{aligned} \quad (1)$$

where ω_m and ω_k, are the *weights* given to metabolic and kinetic information, and ω_p, ω_r, ω_ε, ω_c and ω_g are the weights used for pathways, reactions, enzymes, compounds and genes. K, P and E represent the concepts specified as *user input*, i.e. kinetic parameters, pathway-related concepts and enzymes. Given e as an enzyme from E, C_e denotes a set of compounds involved in the reaction catalysed by the enzyme, whereas G_e is a set of *Saccharomyces cerevisiae* genes encoding the enzyme. $M(A)$ is the percentage of concepts from the set A matching the document.

The total score, S, is calculated as a weighted sum of the scores obtained for metabolic, S_m, and kinetic, $M(K)$, information. S_m is in turn calculated as a weighted sum of the scores obtained for pathway-related information, $M(P)$, and the maximal score obtained for an individual reaction, $S_r(e)$. The latter is maximized, since it is not likely that all relevant reactions will be discussed in a single document. In this way, we are looking for any of the

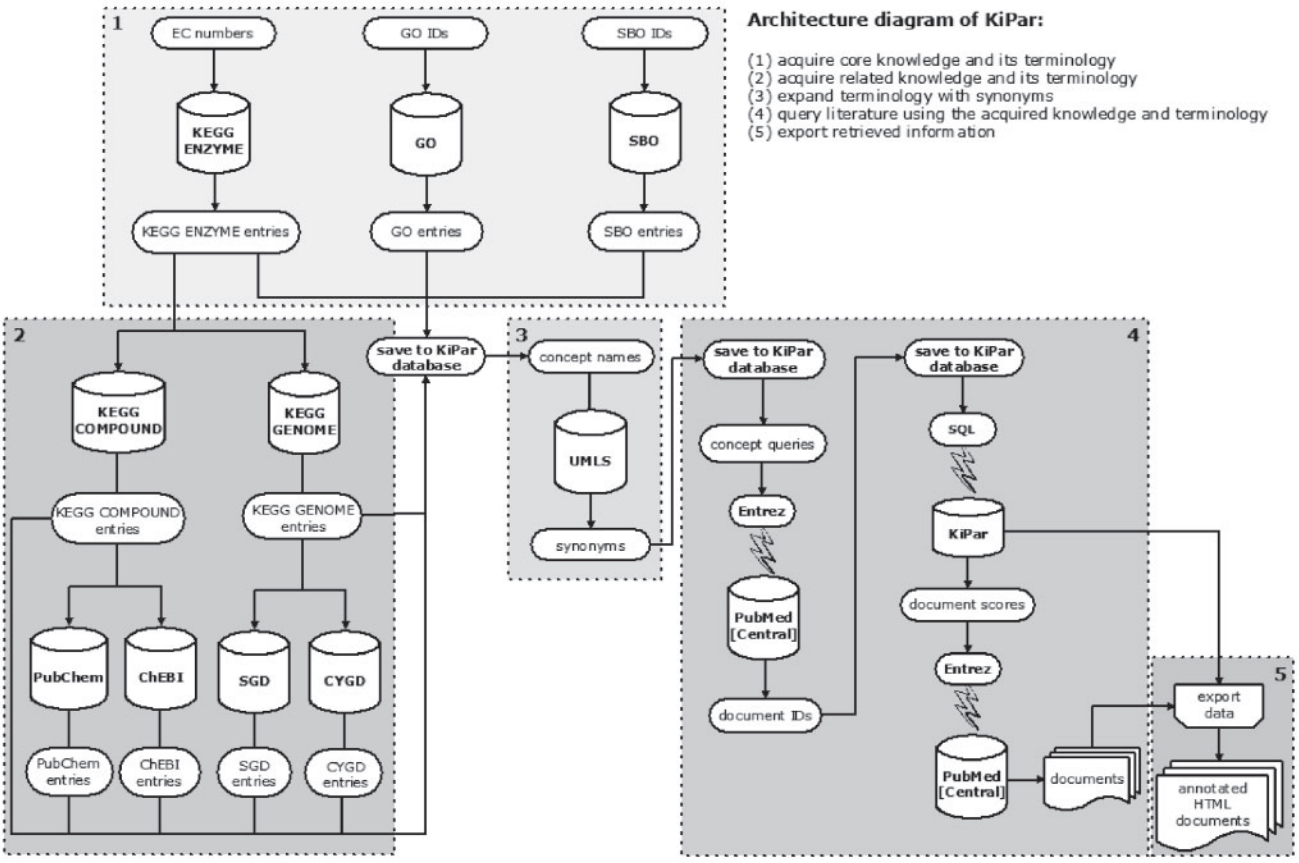

Fig. 2. Architecture diagram of KiPar.

relevant reactions individually. Finally, the score for a reaction catalysed by an enzyme e is calculated as a weighted sum of the scores obtained for the enzyme itself—$M(\{e\})$, the related compounds—$M(C_e)$, and the related genes—$M(G_e)$.

The actual scoring is implemented as an SQL query over the local DB. In this manner, the querying ability of the DB management system is effectively combined with that of Entrez in order to address typical drawbacks of *Boolean* querying (which combines search terms using logic operators AND, OR and NOT): (i) a suitable query may be difficult to formulate due to the complexity of information needs, (ii) the relative importance of the search terms cannot be specified and (iii) it is difficult to rank the retrieved documents by their relevance (Wiesman *et al*., 1997).

Finally, documents with the highest score are presented to the user in HTML format (box 5). The results produced represent links to the original documents annotated with the matching concepts (linked to their entries in the relevant DBs) and quantitative data. The annotation helps a user to determine which information each document contains.

2.4 Output

The retrieval results are summarized in an HTML file (see Supplementary Material 3 for the screenshots of output). The summary page provides: a legend explaining the annotations in the retrieved documents, the input information and the configuration used, the total number of documents retrieved and a link to the ranked list of retrieved documents. For each retrieved document, the following information is provided: citation details, an annotated abstract, a score with the matching concepts, and the PubMed ID (PMID). Additional information available for full-text documents includes: the PMC ID and an annotated local copy of the article. We use an HTML format of the full paper where available, because it preserves the logical structure of the paper as well as the formatting, which may provide clues for automatic text processing (e.g. italicized text can be used to indicate potential organism names). However, much of the work on enzyme kinetics dates back to the 1960s, and although the time coverage of PMC now reaches back to the 1950s, these publications are currently not available in HTML. For instance, out of the top 100 papers retrieved for glycolysis, 15 were not available in HTML format. As part of United Stated National Library of Medicine (NLM's) digitization project, archival content that is not yet available in electronic form is scanned and saved in the PDF 'image plus text' format, where the text obtained from a scanned page image using optical character recognition (OCR) is layered invisibly over the image. The OCR text can then be used for full-text searching. PMC users do not have direct access to the OCR text, but it can be extracted from the PDF files. In the absence of the preferred HTML format, we extract ASCII text from a PDF file. Both formats are further handled in the same manner.

The full text of a publication is annotated with the concepts of interest. Practically, this involves mapping the synonyms acquired for these concepts onto the text, annotating their occurrences with HTML <a> tags of specific classes, and linking them to their entries in the relevant external DBs, e.g.:

```
<a class='compound' href= 'http://www.genome.jp/dbget-bin/
    www_bget?compound+C00031'>glucose</a>
```

As a result, an annotated HTML version of the paper is presented to the user in which terms denoting relevant concepts are highlighted, colour-coded and clickable with links to their descriptions in public DBs.

2.5 Implementation details

KiPar is implemented in Java and is distributed as a standalone Java application. It is an open source product, which can be used by software developers interested in modifying the functionality of KiPar or simply reusing some of its components as part of different bioinformatics applications. The supporting web site provides the information necessary to install and use KiPar. The Java application uses a local DB for storing and processing relevant information harvested from external resources. Therefore, in order to run KiPar, it is required to set up a local DB for which the schema is provided. We tested KiPar with a DB hosted on a PostgreSQL (PostgreSQL, 2008) system, but other DB management systems supporting SQL should work with KiPar by changing the driver information in the configuration file. A local copy of the PubChem DB can be configured and used in the same manner. Other examples of configurable parameters include literature DB choice, UMLS connection details, the weights used for document scoring, export options, etc.

3 RESULTS

3.1 Evaluation

A primary goal of KiPar is to retrieve literature relevant for the kinetic modelling of a particular metabolic pathway. We evaluated the retrieval of abstracts from PubMed and of full-text articles from PMC separately, and contrasted the results. We evaluated the system using three pathways in yeast: glycolysis, the pentose phosphate pathway and the citrate cycle (see Supplementary Material 4–6). Here, we present a detailed evaluation using yeast glycolysis (Pritchard and Kell, 2002; Teusink *et al*., 2000) as an example of a well-studied pathway. Evaluation of the other two pathways is available in Supplementary Material.

Input used to specify information about glycolysis included 21 SBO concepts related to kinetic modelling, 2 GO concepts related to the pathway itself and 13 enzymes that catalyse individual reactions from the given pathway. In addition, information about 31 compounds and 31 genes related to the given enzymes was retrieved automatically from KEGG. Using the synonyms acquired automatically from the relevant DBs, all concepts were mapped to the matching documents. As a result, 4149 abstracts from PubMed were indexed as well as 28 587 articles from PMC. The indexed documents were scored automatically using the formula (1). The weights were chosen intuitively. Equal weights were given to metabolic ($\omega_m = 50$) and kinetic ($\omega_k = 50$) information; otherwise, it would be likely to retrieve documents that contain kinetic data for irrelevant reactions, or conversely documents describing relevant reactions with no kinetic data reported. The weights given to specific aspects of metabolic information have been distributed to emphasize individual reactions ($\omega_r = 80$) from a pathway ($\omega_p = 20$). The remaining weights given to enzymes ($\omega_\varepsilon = 60$), compounds ($\omega_c = 30$) and genes ($\omega_g = 10$) reflect the strength of their association to a reaction, an enzyme being most strongly associated, since specific compounds may take part in many different reactions (e.g. water) and genes encoding the enzyme may not even be mentioned when discussing a reaction. The choice of weights can significantly influence the results (Fagin and Wimmers, 2000; Shatkay, 2005), so further work is needed to determine the optimal weights. On this occasion, we performed a sensitivity analysis and concluded that small changes in weights (± 5 and ± 10) do not incur significant changes on the performance except in an extreme case when zero-weight is given to genes (see Supplementary Material 7).

The average score for abstracts and full papers was 15.26 and 10.73, respectively. Documents whose score was higher than the average were considered for retrieval: 45 out of 17 million abstracts from PubMed and 238 full papers out of 71 120 from PMC. A high disproportion between the relative numbers of abstracts and full papers retrieved can be observed immediately. An analysis of the distribution of indexed concepts across documents shows that the relative numbers of indexed documents differ by three

Table 1. Documents indexed with concepts of the five types (see Section 2)

Concept type	Abstracts			Full-text articles		
	Total number	Percentage of indexed abstracts	Percentage of PubMed abstracts	Total number	Percentage of indexed articles	Percentage of PMC articles
Pathway	212	5.11	0.01×10^{-3}	1470	5.14	0.02
Enzyme	1153	27.79	0.07×10^{-3}	2420	8.47	0.03
Compound	3681	88.72	0.22×10^{-3}	27971	97.85	0.39
Gene	182	4.39	0.01×10^{-3}	1858	6.50	0.03
Kinetics	233	5.62	0.01×10^{-3}	1146	4.01	0.02
All	4149	100	0.24×10^{-3}	28587	100	0.40

orders of magnitude between abstracts and full-text articles (Table 1, columns 4 and 7). These results are bound to be somewhat skewed, because PubMed and PMC differ significantly in their coverage of biomedical literature. For this reason, we compared full text against the abstract for the top 10 papers retrieved from PMC, 9 of which were judged to be relevant by domain experts. As expected, the average number of relevant concepts mentioned in the full text of a paper (16.8) significantly outweighs their number in the abstract (1.5). In other words, <9% of concepts relevant for kinetic metabolic modelling are mentioned in the abstract. Given the complexity of targeted information and consequently the formula (1) for scoring the documents that contain it, the fact that on average no more than two relevant concepts will be mentioned in the abstract points out that such documents probably will not be retrieved if only abstracts are searched. Indeed, the retrieval results for PubMed and PMC overlapped on just a single article (PMID 16584566), even though PubMed contains abstracts for a great majority of the articles from PMC.

A standard set of evaluation measures used to quantify the IR results includes *precision* (the percentage of retrieved documents that are relevant) and *recall* (the percentage of relevant documents retrieved) (Baeza-Yates and Ribeiro-Neto, 1999). Domain experts, presented with the retrieved documents ordered by the score calculated [see formula (1)], helped distinguish between *true positives* (i.e. relevant documents retrieved) and *false positives* (i.e. irrelevant documents retrieved) by reading the retrieved literature and judging the relevance of individual documents. Precision is obtained by dividing the number of true positives by the number of documents retrieved, while recall requires the number of true positives to be divided by the total number of relevant documents. Therefore, it is more difficult to estimate recall, as it requires a comprehensive set of relevant documents. For this purpose, we attempted to use information from SABIO-RK, a DB of biochemical reactions, their kinetic equations with their parameters and the experimental conditions under which these parameters were measured (Wittig *et al*., 2006). We searched the information on the required kinetic parameters for all reactions in the glycolysis pathway and collected the corresponding citations (given as PMIDs) provided by human curators at SABIO-RK. A total of 60 PubMed citations were collected in this manner, out of which only 6 existed in PMC. The small number of citations is bound to underestimate the recall. Therefore, we were not able to properly estimate the recall. However, we compared the results achieved by KiPar and a baseline method, which represents a search performed by Entrez using a *Boolean* query that combines the preferred names of enzymes, kinetic parameters and organism of interest, e.g.: *("alcohol dehydrogenase" OR … OR "phosphoglycerate mutase") AND ("Michaelis constant" OR … OR "equilibrium constant") AND "Saccharomyces cerevisiae"*. Given the top 50 retrieved documents, Figure 3 compares precision of the two methods and the numbers of true positives. Since the ratio between the numbers of true positives preserves the ratio between the recall values, they can be used to relate recall to that of the baseline method.

Overall, the evaluation of the approach used in KiPar helped to identify 30 relevant citations from PubMed and 24 relevant citations from PMC. From Figure 3, we can see that the numbers of relevant citations (i.e. true positives) retrieved by the baseline method from PubMed and PMC were 22 and 12, respectively. By comparing the numbers of true positives, we see improvement over the baseline performance: the recall achieved by KiPar is 36% better for abstracts and 100% better for full-text documents.

Combining abstracts and full-text documents gives a total of 53 citations (there was one citation in common), which is comparable with the number of citations provided at SABIO-RK. This means that the system's recall is comparable with that achieved by human curators. We can provide better insight into the retrieval results using user-oriented measures of IR performance such as relative recall and novelty ratio (Baeza-Yates and Ribeiro-Neto, 1999). In the absence of the total number of relevant documents, recall is approximated by *relative recall*, which is calculated as the ratio between the number of relevant documents retrieved and the number of relevant documents previously known to the user. *Novelty ratio* represents the ratio between the number of relevant documents retrieved that were previously unknown to the user and the number of all relevant documents retrieved. Overall, we achieved high values for both measures: 88.33% for the relative recall and 92.45% for the novelty ratio. High novelty ratio implies that by complementing information already available in specialized DBs, the system reveals many new documents previously unknown to the user, thus highlighting the need for such a TM tool. As for the types of information retrieved, the analysis of the results revealed that they represent a good coverage of the pathway: no relevant documents were retrieved for only 2 out of 13 reactions. By facilitating navigation through huge volumes of scientific literature and highlighting the key results within the relevant documents, we see that KiPar provides valuable support to those interested in kinetic modelling of metabolism. In this respect, precision and recall should be judged relative to the number of person-hours saved [e.g. other studies report a 70% reduction in curation time with the use of TM support (Donaldson *et al*., 2003)].

All processing, including indexing, is done on the fly. The execution time consumed depends on a pathway (i.e. the number of concepts needed to describe it). We averaged the times recorded for the three pathways used for evaluation to provide an estimation of the time required to retrieve and annotate the literature for an individual pathway. The time required for indexing of PubMed and PMC with pathway-related concepts is approximately the same: just over 14 min. However, a significant difference in the number of indexed documents in the two DBs (Table 1) is reflected in the amount of time needed to complete the subsequent operations (Table 2). Most of the run time is consumed to export an annotated version of a full-text article. This operation includes downloading an article, a possible conversion from PDF to ASCII, and annotation

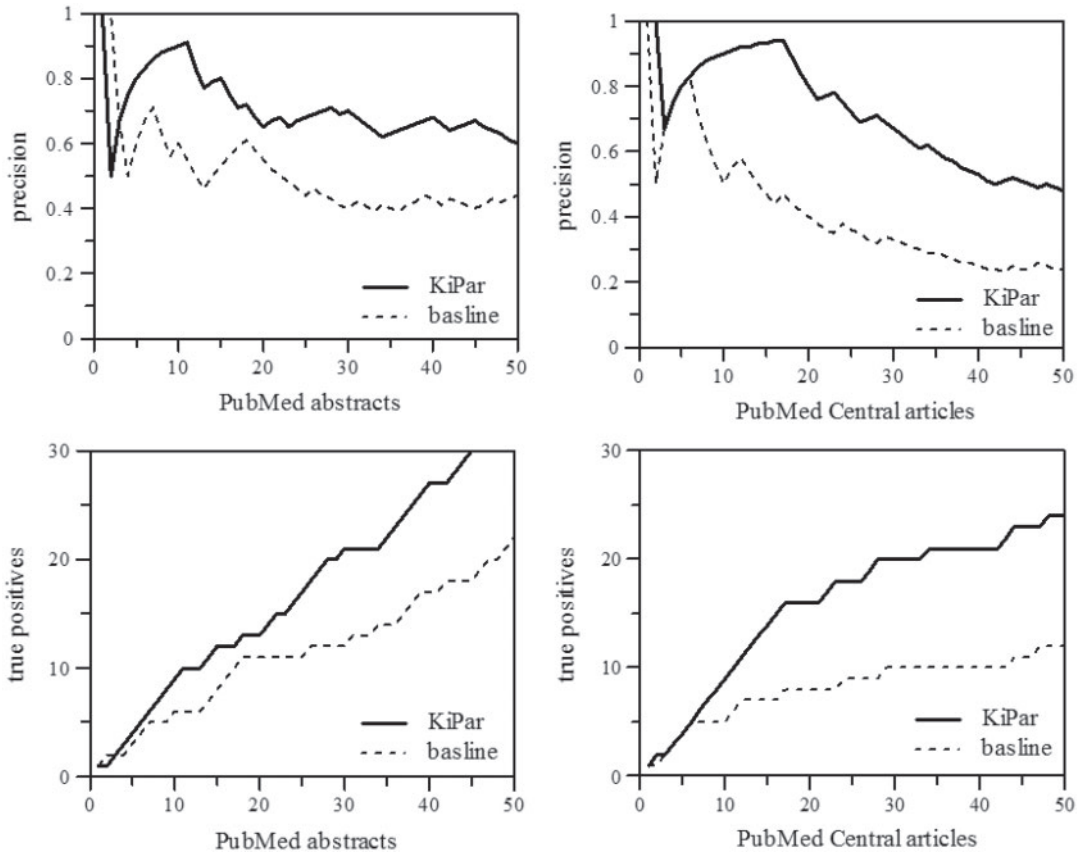

Fig. 3. Precision and number of true positives for the top 50 documents.

Table 2. Average run time for individual operations following indexing

Literature DB	PubMed		PMC	
Operation	Total	Per doc	Total	Per doc
Score documents	45 s	13 ms	7.7 min	16 ms
Retrieve document details	29 s	890 ms	3 min	1.8 s
Export results	1.3 min	2.4 s	80.8 min	48.5 s

of its content. The whole retrieval process of 100 full-text articles is completed in 1 h and 45 min. We emphasize again that the main goal of KiPar is to provide a literature starter pack for studying the kinetic aspects of a metabolic pathway, and as such KiPar is not an interactive system, i.e. it does not require sub-second response time.

3.2 Comparison to other systems

The idea of KiPar is to offer a way to retrieve information on complex entities (metabolic pathways and the corresponding reactions). In that respect, KiPar can be compared with MedBlast, an IR system searching for documents about a given biological sequence (Tu *et al*., 2004). MedBlast uses BLAST (Altschul *et al*., 1997) to find the corresponding nucleic acid and protein sequences and their homologues. This provides gene names, after which their synonyms are looked up in a local thesaurus. This information is then combined in a *Boolean* query and passed to Entrez, e.g. ATP5E AND ('Homo sapiens' OR human [mh]). KiPar does similar processing relevant for a pathway, but includes an additional layer to process information obtained from Entrez to perform non-*Boolean* search, which is more appropriate for the specific IR task at hand. An alternative way to overcome the limitations of *Boolean* search offered by Entrez is to launch multiple queries with different combinations of search terms. For example, PubMatrix initiates an Entrez query for each pair of search terms specified by a user and computes a matrix of counts of their co-occurrence in the literature as a way of measuring their associations (Becker *et al*., 2003). Such statistical information extracted from the literature is useful especially when explicit information about relations between domain-specific concepts is not readily available. However, the performance of biomedical TM applications depends on an active use of domain knowledge as a support for more sophisticated reasoning about the improvement of queries (Wiesman *et al*., 1997). Therefore, many IR systems utilize a comprehensive body of knowledge that is currently stored in biomedical ontologies. While the use of wide-encompassing semantic networks such as UMLS or MeSH was found to improve the results of IR (e.g. Aronson and Rindflesch, 1997; Swanson *et al*., 2006), many tasks require more fine-grained knowledge representation. Therefore, some IR systems rely upon customized ontologies and/or thesauri, which are most often maintained locally. For example, in order to support queries such as 'Given *X*, find all *Y*'s', where *X* or *Y* can be

diseases, tissues, cell compartments, etc., PolySearch maintains nine different thesauri, compendia or synonyms lists (Cheng *et al*., 2008). Other systems go a step further by maintaining a semantically annotated local corpus of documents in addition to local knowledge and terminology resources, e.g. Textpresso (Müller *et al*., 2004), EBIMed (Rebholz-Schuhmann *et al*., 2007). While improving the quality of the results of IR, the decisions to maintain knowledge and textual resources locally may also improve the efficiency of IR. However, the preprocessing of results incurs the need to keep local resources up-to-date, which may require frequent updates over large data collections. KiPar makes extensive use of domain knowledge (see Section 2.3), but it incorporates dynamic access to knowledge resources through their web services. This avoids the need for maintenance of complex local resources, but does require more time to pull information from external resources on demand. The nature of the specific IR task justifies such choice (see Section 3.1).

As for similar systems, we are aware of only one system that addresses the problem of retrieving documents relevant for kinetic modelling. (Hakenberg *et al*., 2004) developed a system that classifies documents in terms of whether or not they contain information regarding experimentally obtained kinetic parameters. Our approach differs in several respects.

While the other system looks for kinetic parameters in general, KiPar searches for kinetic data in the context of a defined 'pathway' and specific reactions therein. In that sense, KiPar tackles the complexity of developing a full kinetic model of (yeast) metabolism by a modular approach focusing on individual metabolic pathways.

The other system uses a machine learning (ML) approach, while KiPar is rule based. ML offers more flexible knowledge acquisition, since it does not require explicit knowledge elicitation through costly and time-consuming interaction with domain experts. Instead it can make use of knowledge implicitly embedded in the examples. However, this requires the provision of a training set of examples, which can again lead to substantial use of domain expertise. For example, Hakenberg *et al*. (2004) trained their system on a set of 791 manually annotated documents, which meant about two person-months of work. The knowledge that had to be explicitly specified in KiPar is rather generic and of a structural nature; e.g. that a pathway consists of reactions; reactions are catalysed by enzymes that convert substrates to products; enzymes are encoded by genes; etc. More specific knowledge (e.g. which compounds are involved in which reactions and how they are referred to) is reused from the existing data resources and as such is not an explicit part of the system. This fact reduces the explicit elicitation of knowledge from domain experts.

Hakenberg *et al*. (2004) retrieved papers from a locally stored corpus of 4582 publications selected randomly from 12 journals that focus on biological areas that make use of kinetic data. KiPar queries much larger literature DB (e.g. the presently 17M abstracts in PubMed) and stores locally only the indexing information and not the actual documents. Also, KiPar searches over a complete set of journals covered by PubMed and PMC. However, the choice of journals covered by PMC is rather limited at the moment due to not many journals taking part in this open access initiative.

The evaluation of the system of Hakenberg *et al*. (2004) reports 60% precision and 50% recall. In absolute numbers, the system retrieved 127 documents of which 77 were relevant. At the same time, it did not retrieve 78 out of 155 relevant documents. These results show a significant advancement over the results they reported for random selection (precision of 12%) and for simple keyword search (precision of 20%). The performance of this system and KiPar cannot be directly compared due to the differences in their problem specifications (kinetics data in general versus kinetics data for specific metabolic pathways and reactions). However, it can be said that both systems show consistent improvements in comparison to the keyword search and that they perform comparably well. There is also room for the integration of the two methodologies, since 'voting' systems are known to outperform their components (Hastie *et al*., 2001).

Finally, Hakenberg *et al*. (2004) report that they missed some relevant papers because kinetic data are often presented in figures and tables, which are inaccessible to their system. They used a local corpus of documents converted from PDF to ASCII. During this conversion all figures and tables included in PDF documents as images were lost. In contrast, KiPar accesses the full-text articles from PMC via Entrez, which does search through the captions/legends of figures and tables in an article. The use of Entrez for intermediary access to literature also eliminates the need for KiPar to explicitly support typical processing involved in IR such as tokenization, stemming, neutralization of spelling variants or differences in nomenclature usage, estimating a term's local weight (i.e. within a document), etc., since Entrez already implements such capability.

4 CONCLUSIONS

We have presented an integrative approach, combining a number of publicly available data and software resources, for the time- and cost-effective development of TM tools for IR, i.e. gathering and filtering of relevant documents (Baeza-Yates and Ribeiro-Neto, 1999). Our approach to TM consists of the following steps: (i) *input*: specify a set of concept identifiers used as entry points into public data resources, (ii) *acquire knowledge*: use input information to acquire other relevant information (other relevant concepts and relations between them) from these and other cross-referenced data resources, (iii) *acquire terminology*: use the data resources to map concepts to known synonyms, (iv) *indexing*: map concepts to matching textual documents using synonyms, (v) *query literature*: use indexing results and relations between concepts to search for information and (vi) *export information*: annotate a set of potentially relevant documents with concepts of interest and cross-link to their DB entries.

This approach has been demonstrated successfully in KiPar, a TM application developed for retrieving documents discussing enzyme kinetic parameters required for quantitative modelling of yeast metabolism. There are two groups of users of this specific application: (i) experimentalists who wish to compare experimentally estimated values of kinetic parameters to those reported in the literature, and (ii) mathematical modellers who wish to incorporate known values of kinetic parameters into metabolic models. These users are actively testing KiPar and their comments are used to further improve its performance. At the same time, their feedback will serve as a basis for adding information extraction (IE) functionality to KiPar. IE selects specific facts about prespecified types of entities and relationships of interest (Hobbs, 1993), and in KiPar it will be used to convert free-text descriptions of kinetic parameters into a structured format, which will enable importing their values directly into formal representations of metabolic models

[e.g. SBML (Hucka *et al.*, 2003; SBML, 2008)]. At the moment, a naïve approach is used to highlight potential descriptions in the text. These annotations, combined with users' feedback, will be used to identify patterns for extracting information regarding kinetic parameters.

ACKNOWLEDGEMENTS

We thank many colleagues in Manchester for the most useful discussions. Particular thanks to Goran Nenadic for his help in testing the software.

Funding: This work is funded by Biotechnology and Biological Sciences Research Council and Engineering and Physical Sciences Research Council. This is a contribution from The Manchester Centre for Integrative Systems Biology (http://www.mcisb.org/).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ananiadou,S. *et al.* (2006) Text mining and its potential applications in Systems Biology. *Trends Biotechnol.*, **24**, 571–579.

Aronson,A.R. and Rindflesch,T.C. (1997) Query expansion using the UMLS Metathesaurus. *proc of AMIA Annu. Fall Symp.*, 485–489.

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Baeza-Yates,R. and Ribeiro-Neto,B. (1999) *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA.

Becker,K. *et al.* (2003) PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics*, **4**, 61.

Bodenreider,O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, **32**, D267–D270.

ChEBI (2008) http://www.ebi.ac.uk/chebi/ (last accessed date January 1, 2009).

Cheng,D. *et al.* (2008) PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Res.*, **36**, W399–W405.

Cherry,J.M. *et al.* (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res.*, **26**, 73–79.

CYGD (2008) http://mips.gsf.de/genre/proj/yeast/ (last accessed date January 1, 2009).

Degtyarenko,K. *et al*. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.

Ding,J. *et al.* (2002) Mining MEDLINE: abstracts, sentences, or phrases. In *Proceedings of the 7th Pacific Symposium on Biocomputing (PSB 2002), Hawaii, USA*, World Scientific Press, pp. 326–337.

Donaldson,I. *et al.* (2003) PreBIND and Textomy: mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, **4**, 11.

Entrez (2008) http://www.ncbi.nlm.nih.gov/Entrez/ (last accessed date January 1, 2009).

Fagin,R. and Wimmers,E.L. (2000) A formula for incorporating weights into scoring rules. *Theor. Comput. Sci.*, **239**, 309–338.

Friedman,C. *et al.* (2002) Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J. Biomed. Inform.*, **35**, 222–235.

Garwood,K.L. *et al.* (2004) Pedro: a configurable data entry tool for XML. *Bioinformatics*, **20**, 2463–2465.

GO (2008) http://www.geneontology.org/ (last accessed date January 1, 2009).

Güldener,U. *et al*. (2005) CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res.*, **33**, D364–D368.

Hakenberg,J. *et al.* (2004) Finding kinetic parameters using text mining. *OMICS*, **8**, 131–152.

Harris,Z.S. (2002) The structure of science information. *J. Biomed. Inform.*, **35**, 215–221.

Hastie,T. *et al.* (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, Berlin.

Henry,C.M. (2003) Systems biology. *Chem. Eng. News*, **81**, 45–55.

Herrgård,M.J. *et al*. (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.*, **26**, 1155–1160.

Hobbs,J. (1993) The generic information extraction system. In Sundheim,B. (ed.) *Fifth Message Understanding Conference (MUC5)*. Morgan Kaufmann Publishers, Inc., San Francisco, CA.

Hoffmann,R. *et al.* (2005) Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci STKE*, **2005**, pe21.

Hood,L. (2003) Systems biology: integrating technology, biology, and computation. *Mech. Ageing Dev.*, **124**, 9–16.

Hucka,M. *et al*. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–531.

Hull,D. *et al.* (2008) Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Comput. Biol.*, **4**, e1000204.

Humphreys,K. *et al.* (2000) Two applications of information extraction to biological science journal articles: enzyme interactions and protein structures. In *Proceedings of the 5th Pacific Symposium on Biocomputing (PSB 2000), Hawaii, USA*, World Scientific Press, pp. 505–516.

Jensen,L.J. *et al.* (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, **7**, 119–129.

Kanehisa,M. *et al*. (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

KEGG (2008) http://www.genome.ad.jp/kegg/ (last accessed date January 1, 2009).

Kell,D.B. (2006) Metabolomics, modelling and machine learning in systems biology: towards an understanding of the languages of cells. The 2005 Theodor Bücher lecture. *FEBS J.*, **273**, 873–894.

Le Novere,N. (2006) Model storage, exchange and integration. *BMC Neurosci.*, **7**, S11.

Michal,G. (1999) *Biochemical Pathways: an Atlas of Biochemistry and Molecular Biology*. Wiley, Heidelberg.

Müller,H.-M. *et al.* (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, 2, e309.

PMC (2008) http://www.pubmedcentral.nih.gov/ (last accessed date January 1, 2009).

PostgreSQL (2008) http://www.postgresql.org/ (last accessed date January 1, 2009).

Pritchard,L. and Kell,D.B. (2002) Schemes of flux control in a model of *Saccharomyces cerevisiae* glycolysis. *Eur. J. Biochem.*, **269**, 3894–3904.

PubChem (2008) http://pubchem.ncbi.nlm.nih.gov/ (last accessed date January 1, 2009).

PubMed (2008) http://www.pubmed.gov/ (last accessed date January 1, 2009).

Rebholz-Schuhmann,D. *et al.* (2007) EBIMed—text crunching to gather facts for proteins from Medline. *Bioinformatics*, **23**, e237–e244.

Rzhetsky,A. *et al*. (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data, *J. Biomed. Inform.*, **37**, 43–53.

SBML (2008) http://www.sbml.org/ (last accessed date January 1, 2009).

SBO (2008) http://www.ebi.ac.uk/sbo/ (last accessed date January 1, 2009).

SGD (2008) http://www.yeastgenome.org/ (last accessed date January 1, 2009).

Shatkay,H. (2005) Hairpins in bookstacks: information retrieval from biomedical text. *Brief. Bioinform.*, **6**, 222–238.

Spasic,I. *et al.* (2005) Text mining and ontologies in biomedicine: making sense of raw text. *Brief. Bioinform.*, **6**, 239–251.

Swanson,D.R. *et al.* (2006) Ranking indirect connections in literature-based discovery: the role of Medical Subject Headings. *J. Am. Soc. Inform. Sci. Technol.*, **57**, 1427–1439.

Teusink,B. *et al*. (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur. J. Biochem.*, **267**, 5313–5329.

Tu,Q. *et al.* (2004) MedBlast: searching articles related to a biological sequence. *Bioinformatics*, **20**, 75–77.

UMLS (2008) http://umlsks.nlm.nih.gov/ (last accessed date January 1, 2009).

Wheeler,D.L. *et al*. (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.

Wiesman,F. *et al.* (1997) Information retrieval: an overview of system characteristics. *Int. J. Med. Inform.*, **47**, 5–26.

Wittig,U. *et al.* (2006) SABIO-RK: integration and curation of reaction kinetics data. *Lecture Notes in Bioinformatics*, **4075**, 94–103.

Yuryev,A. *et al.* (2006) Automatic pathway building in biological association networks. *BMC Bioinformatics*, **7**, 171.