Mining the biomedical literature using semantic analysis and natural language processing techniques

Ronen Feldman, Yizhar Regev, Eyal Hurvitz and Michal Finkelstein-Landau

The information age has made the electronic storage of large amounts of data effortless. The proliferation of documents available on the Internet, corporate intranets, news wires and elsewhere is overwhelming. Search engines only exacerbate this overload problem by making increasingly more documents available in only a few keystrokes. This information overload also exists in the biomedical field, where scientific publications, and other forms of text-based data are produced at an unprecedented rate. Text mining is the combined, automated process of analyzing unstructured, natural language text to discover information and knowledge that are typically difficult to retrieve. Here, we focus on text mining as applied to the biomedical literature. We focus in particular on finding relationships among genes, proteins, drugs and diseases, to facilitate an understanding and prediction of complex biological processes. The LitMiner™ system, developed specifically for this purpose; is described in relation to the Knowledge Discovery and Data Mining Cup 2002, which serves as a formal evaluation of the system.

Ronen Feldman
Dept of Computer Science
Bar-Ilan University
Ramat-Gan, Israel 52900
tel: +972 3 7350000
fax: +972 3 7350001
e-mail: feldman@cs.biu.ac.il
Yizhar Regev
Eyal Hurvitz
Michal Finkelstein-Landau
ClearForest Corporation
Or Yehuda
Israel 60376

▼ The most widely used biomedical literature database – NCBI's PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) – contains over 11,000,000 document abstracts. A query for documents mentioning the terms 'gene' or 'protein' returns ~2,800,000 documents of which over two-thirds were published just within the past decade. A more specific query for papers mentioning 'epidermal growth factor receptor' returns over 10,000 documents. Given that almost all current biomedical knowledge is published in scientific articles, researchers trying to make use of this information require automated tools that enable a search for the proverbial 'needle' of explicit knowledge in this 'haystack' of text.

Several disciplines involve the automated handling of text. These include: (1) information retrieval, which deals mostly with finding documents that satisfy a particular information need within a large database of documents [1–3]; (2) natural language processing (NLP)–a broad discipline concerned with all aspects of automatically processing both written and spoken language [4–7]; and (3) information extraction (IE), a sub-field of NLP, centered around finding explicit entities and facts in free text [8–15, T.R. Leek, MSc thesis, University of California, 1997], for example, identifying all the positions in free text in which a 'bank' is mentioned (entity extraction) or finding all acquisition relationships to populate a table of companies that acquired one another (relationship extraction). Text mining is the combined, automated process of analyzing unstructured, natural language text to discover information and knowledge that is typically difficult to retrieve.

This review focuses specifically on text mining of the biomedical literature ([16–33]), and describes the LitMiner™ system, developed for finding relationships among genes, proteins, drugs and diseases, to facilitate an understanding and prediction of complex biological processes. We focus in particular on the Knowledge Discovery and Data Mining (KDD) Cup 2002 ([34]), which serves as a formal evaluation of the LitMiner™ system.

Text mining

Text mining is a new and exciting research area that attempts to solve the information overload problem. It uses techniques from the general field of data mining [20,35–37] but, because it deals with unstructured data, a major part of the text mining process deals with the crucial stage of pre-processing the

1478-5282/03/$ – see front matter ©2003 Elsevier Science Ltd. All rights reserved. PII: S1478-5282(03)02314-6

document collections using techniques such as text categorization [38–40], term extraction [6] and IE [8–15, T.R. Leek, MSc thesis, University of California, 1997]. In addition to pre-processing of the document collection, the text mining process includes: (1) storage of the intermediate representations, (2) techniques to analyze these intermediate representations (such as distribution analysis, (3) clustering [41], (4) trend analysis [42], (5) association rules [35,36], and (6) visualization of the results [30,42].

A typical text mining system begins with collections of raw documents, that is, documents without any labels or tags. Documents are first automatically tagged by 'categories', 'entities' or 'relationships' that are extracted directly from the documents. Next, extracted categories, entities or relationships are used to support a range of data mining operations on the documents.

Text categorization [38–40,43,44], which is typically a sub-field of information retrieval [1–3], involves the partitioning of a large collection of documents into subsets that are interrelated by some pre-defined criteria. For example, PubMed currently offers subsets of documents for users to search through, such as AIDS literature and History of Medicine. A second example is the Yahoo! Homepage, which categorizes the whole web into areas such as 'News and Media', 'Science' and 'Arts'. Each document in the large collection is tagged by words characteristic of categories, which enables the association of the document (or website) with its relevant categories.

Limiting the set of documents for mining to certain relevant sub-categories simplifies the follow-up tasks for the mining tools, and increases the probability that these tools will extract the most on-target pieces of information from the text. The actual detection of facts within the text is typically performed through IE methods.

Information extraction

Information extraction [8–15, T.R. Leek, MSc thesis, University of California, 1997] is one of the most prominent techniques currently used in text mining. In particular, by combining NLP tools, lexical resources and semantic constraints, IE can provide effective modules for mining the biomedical literature. Complementary visualization tools enable the user to explore, check (and correct if required) the results of the text mining process effectively.

The first step in document tagging involves finding (extracting) entities and relationships from the documents that are likely to be meaningful and content-bearing. The term 'relationships' refer to facts or events involving certain entities. A possible 'event' might be that a company has entered into a joint venture. A 'fact' might be that a gene causes a certain disease. The extracted information provides more concise and precise data for the mining process than the more naive word-based approaches, such as those used for text categorization, and tends to represent concepts and relationships that are more meaningful and that relate directly to the domain of the examined document.

Box 1. Knowledge based learning: the CONSTRUE system

A typical rule in the CONSTRUE system.
If DNF (disjunction of conjunctive clauses) formula then category:

```
If    ((wheat & farm)                  or
      (wheat & commodity)              or
      (bushels & export)               or
      (wheat & tonnes)                 or
      (wheat & winter & ¬ soft))
      then Wheat
      else ¬ Wheat
```

Consequently, IE methods enable the mining of actual information present within the text, rather than the limited set of tags associated with the documents. Using the IE process, the number of different relevant entities and relationships on which the data mining is performed is unbounded (typically thousands or even millions), and far beyond the number of tags that any automated categorization system could handle.

Text categorization

Text categorization [38–40,43,44] is the activity of labeling natural language texts with thematic categories from a predefined set of categories. There are two main approaches to the categorization problem. The first approach is the knowledge engineering approach [38,43], where the user manually defines a set of rules that encode expert knowledge about how to classify documents for given categories. One example of the knowledge engineering approach is the CONSTRUE [38,43] system built by the Carnegie Group for Reuters (see Box 1). The main drawback of this approach is the knowledge acquisition bottleneck. Rules must be defined manually by a knowledge engineer interviewing a domain expert. If the set of categories is modified, then these two professionals must intervene again. Hayes *et al.* [38,43] reported a 90% break-even between precision and recall on a small subset of the Reuters test collection (723 documents). However, it took tremendous effort to develop the system (several human years), and the test set was not significant to validate the results. In addition, it is not clear whether these results would scale-up if a larger system were to be developed.

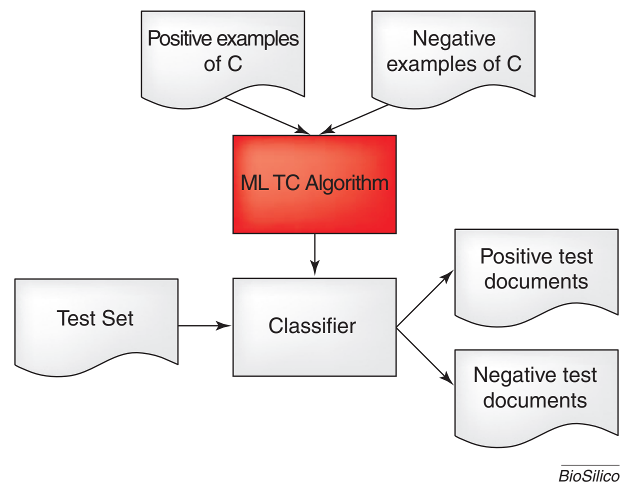


Figure 1. A typical machine learning (ML)-based categorization system.

The second approach is the machine learning (ML) approach [39,40,44], where a general inductive process automatically builds a text classifier by learning from a set of pre-classified documents. The ML-based approach is based on the existence of a training set of documents that are already pre-tagged using a pre-defined set of categories. A typical ML-based categorization system is shown in Figure 1.

There are two main methods for performing ML-based categorization. The first method is 'hard' (fully automated) classification where, for each pair of category and document, a truth value (either TRUE if the document belongs to the category or FALSE if otherwise) is assigned. The second method involves a ranking (semi-automated)-based classification. In this approach, rather than returning a truth value, the classifier returns a categorization status value (CSV), that is, a number between 0 and 1 that represents evidence supporting that the document belongs to the category. Documents are then ranked according to their CSV value. Specific text categorization algorithms are discussed later.

Comparison between text categorization and IE

In contrast to the IE approach, where the entities tagged in the document are based on actual terms extracted from the document, text categorization tags the document with concepts that are not necessarily mentioned in the document itself. The main advantage of using a categorization approach is that it is less time-consuming to prepare the training group, and there is no need to manually craft rules. However, the number of tags assigned to any given document would be less than six. These tags would capture only a few of the main topics of the document and certainly miss most of the important entities mentioned inside the document. By contrast, a document tagged by an IE system will average 20–50 tags (for a 2–3 page document). To summarize, IE was found to provide a much better infrastructure for text mining than text categorization.

Mining the biomedical literature

Text mining of the biomedical literature uses as an input to the system a set of biomedical articles, typically drawn from a broad sub-domain. These articles are analyzed by IE from a set of pre-defined entities and relationships. The following section describes the heuristics for extracting a variety of entities and relationships, followed by a description of the declarative information analysis language (DIAL) IE language. We also provide a detailed description of the LitMiner™ system (based on the DIAL language) that was used in the KDD Cup 2002 [45–47].

Main biomedical entities

The following section outlines the main entities that can be extracted from biomedical literature, giving examples and focusing on the appropriate heuristics for each entity.

Gene. Gene is the fundamental entity in the molecular biology and/or genetics literature. Gene names present many problems for IE systems [48]. The main problems include: (1) the huge number of genes recorded in databases and in the literature (e.g. thousands of genes are known for *Drosophila*); and (2) the substantial variability of gene names within the literature (e.g. different synonyms and different forms of capitalization). Although certain conventions have recently been set, names as common as 'clock' and 'Columbus', and varied as 'cp beta subunit' are still included in the *Drosophila* literature. Another major problem is the use of the same name for the gene and its corresponding protein.

To avoid such problems, we use evidence (terms) found within the text adjacent to the gene name. Examples of such evidence include:

- the use of action verbs such as 'activates' or 'inhibits';
- a multi-word phrase followed by an abbreviation in parenthesis on the first occurrence of the gene name in the text;
- a receptor with the same name as a gene in a signaling pathway; these are two separate genes, one of which includes 'receptor' in its name (e.g. epidermal growth factor and epidermal growth factor receptor).

In addition to such evidence we use several lexicons; however, these must be used carefully as they often include synonyms that are same as common English words (see earlier example).

Domain (motifs). Recognizable arrangements within proteins that are used elsewhere in other proteins (e.g. 'helix–loop–helix'). Here, again, there are substantial inconsistencies (occasionally they are named after one of the proteins that contain them) but the majority has independent names.

Enzymatic activity. The specific activity with which a protein domain is associated (e.g. 'ATPase').

Primers. Small sequences of DNA or RNA identified specifically and usually used for experimental purposes. In papers, primers occur simply as strings of letters, almost exclusively as 'a' 't' 'g' and 'c' (e.g. 'gatgaccggacttatgcgccgta') and are usually between 15 and 35 characters long. Such strings are thus relatively easy to extract.

Phenotype. The difference observed from the normal state; for example, 'inner ear defect' in the phrase 'Snell's waltzer mutants exhibit behavioral abnormalities suggestive of an inner ear defect'. From an IE viewpoint, phenotypes are usually nominal or verbal phrases described as a result of an abnormal state of a gene (described by a nominal phrase headed by gene name).

Signaling pathway. Typically a nominal phrase including a gene name and a following term such as 'pathway' or 'signal (e.g. 'map kinase pathway').

Mostly referred to as being activated, inhibited, used, or in a 'through' construct.

Lexicon-based entities

Organism. Lexicons including the more widely researched organisms, including 'mouse' and '*Drosophila*', although such a lexicon cannot, of course, cover all known organisms.

Tissue. A subsection of an organism, for example, 'wing'.

Organelle. A subunit of the cell, for example, 'mitochondrion'.

Chemicals and/or drugs. For example, 'tetracycline'.

Biomedical relationships

This section lists some of the relationships that can be induced from the biomedical literature, and provides actual sentences that exemplify each relationship.

Gene–gene pathways. Any chain of controlling relationships that appear together. For example, the phrases, 'GENE, a member of the...pathway' (and variations thereof), and 'GENE and its target(s)', give an indication of a pathway. The following verbs indicate that the second gene is controlled by the first, or that the second gene is downstream: regulate, upregulate, downregulate, activate, enhance, inhibit, induce and modulate.

Two genes bind. The following phrases give an indication that two genes bind: 'GENE forms a complex with GENE', 'GENE is found in a complex with GENE', 'GENE interacts with GENE'. (Note that the phrase 'has a complex interaction with' is not an indication of a bind.) The verbs 'bind' and (physically) 'associate' correlate two genes that bind. The verb 'associate' alone provides low confidence of a bind; the presence of 'physically' before 'associate' increases its confidence. 'Immunoprecipitate' and 'co-precipitate' are good clue words that two genes are related.

Genes are related. Indications that genes are from the same gene family. For example, 'GENE is a member of the... family'. An indication that genes are homologs is given by the example 'GENE1 is the ORGANISM homolog of GENE2'.

Gene–phenotype. Indicates a gene–phenotype relationship, for example, the phrase 'GENE is required for PHENOTYPE'. Verbs that correlate a gene and a phenotype include 'showed' (in association with an organism) and 'exhibited' (in association with an organism).

Gene–disease. Indicates a relationship between a gene and disease. For example, 'mutations in p53 were predominantly detected in Burkitt's lymphoma cells'.

Implementation in the DIAL Language

Our modules for extracting the entities and relationships described earlier are implemented in a language called DIAL. DIAL is a language designed specifically for writing IE rules [45–47]. Here, we describe the basic elements of the language (the complete syntax of DIAL is beyond the scope of this paper). Box 2 gives a full example of the DIAL rule.

Basic elements. The basic elements of the language are the syntactic and semantic elements of text, and sequences and patterns thereof. Among these elements the language uses: (1) pre-defined strings, for example, 'gene'; (2) word class elements, that is, a phrase from a pre-defined set of phrases that share a common semantic meaning, for example, 'wcOrganism', a list of organisms; and (3) skip patterns, that is, a pattern that matches (skips) up to a certain number of tokens (a series of characters that comprise a basic element of the document) followed by an instance of another element. For example, skip(')',10) matches up to 10 tokens until a ')' token.

Each of these elements can be optional (denoted by square brackets around the element).

Constraints. Constraints carry out on-the-fly Boolean checks for specific attributes. These can be applied to fragments of the original text, or to results obtained during processing of the extraction step. The syntax of constraints is verification of the keyword followed by the condition to be checked within brackets. The condition is typically implemented using a suitable Boolean function, implemented within DIAL infrastructure libraries. For example, InWC returns TRUE if the tested text segment is a member of the tested wordclass. For example, verify(InWC(Head,@wcHomolog)) means that the head pattern matching element must be a member of the word class wcHomolog.

Box 2. Declarative information analysis language

Declarative information analysis language (DIAL) is a language designed specifically for writing information extraction rules. The following example shows the syntax for the predicate ‘homologs’, which identifies a homology relationship between two genes. The code fragment starts by providing the formal definition of the predicate ‘homologs’ that has two arguments, followed by a definition of several word classes that will be used in the actual rule. The rule looks for a gene name. This is carried out by the predicate ‘GeneCandidate’, which returns the parameter ‘Gene1’ followed by an optional phrase delimited within parenthesis (usually an acronym of Gene1), and then an optional comma and optional occurrence of the wordclass ‘wcClauseConnector’. The system then looks for the word ‘is’ followed by the optional occurrence of an article. It then extracts a noun phrase (with its article, head and stem). Finally, the system looks for the word ‘of’ followed by an optional ‘wcOrganism’ and another gene (Gene2). In addition, the rule contains a constraint that ensures that the head of the noun phrase contains one of the phrases defined by the word class ‘wcHomolog’.

DIAL syntax example

```
predicate toplevel Homologs(STRING Gene1,STRING Gene2);

wordclass wcHomolog = homolog homologue paralog
paralogue ortholog orthologue;
wordclass wcOrganism = mammalian mammals drosophila
mouse rat xneopus yeast; //
wordclass wcClauseConnector = “which” “that”;

predicate OptParen();
OptParen():- “(“ skip(“)”,10) “)”;
OptParen():- TRUE;

Homologs(Gene1,Gene2):-
    GeneCandidate(Gene1)             //Decapentaplegic
    OptParen                         //(Dpp)
    [ “,” ]
    [ wcClauseConnector ]
    “is”
    [ wcArticle ]                    // a
    NounGroup(Article,Head,Stem)     // close homologue
    “of”
    [ wcOrganism ]                   // mammalian
    GeneCandidate(Gene2)             // BMPs
```

IE rule bases. The rule base can be viewed as a logic program. Thus, a rule base ‘Γ’ is a conjunction of definite clause, C_i: $H_i \leftarrow B_i$. Here, C_i is a clause tag; H_i (called the ‘head’) is a literal; and $B_i = [B_{i1}\ B_{i2}\ldots] = P_i \cup N_i$ (called the ‘body’) is a set of literals, where $P_i = [p_{ij}]$ is a set of pattern-matching elements, and $N_i = [n_{ij}]$ is a set of constraints operating on P_i. The clause C_i: $H_i \leftarrow B_i$ represents the assertion that H_i is implied by the conjunction of the literals in P_i while satisfying all the constraints in N_i.

KDD Cup 2002 text mining system

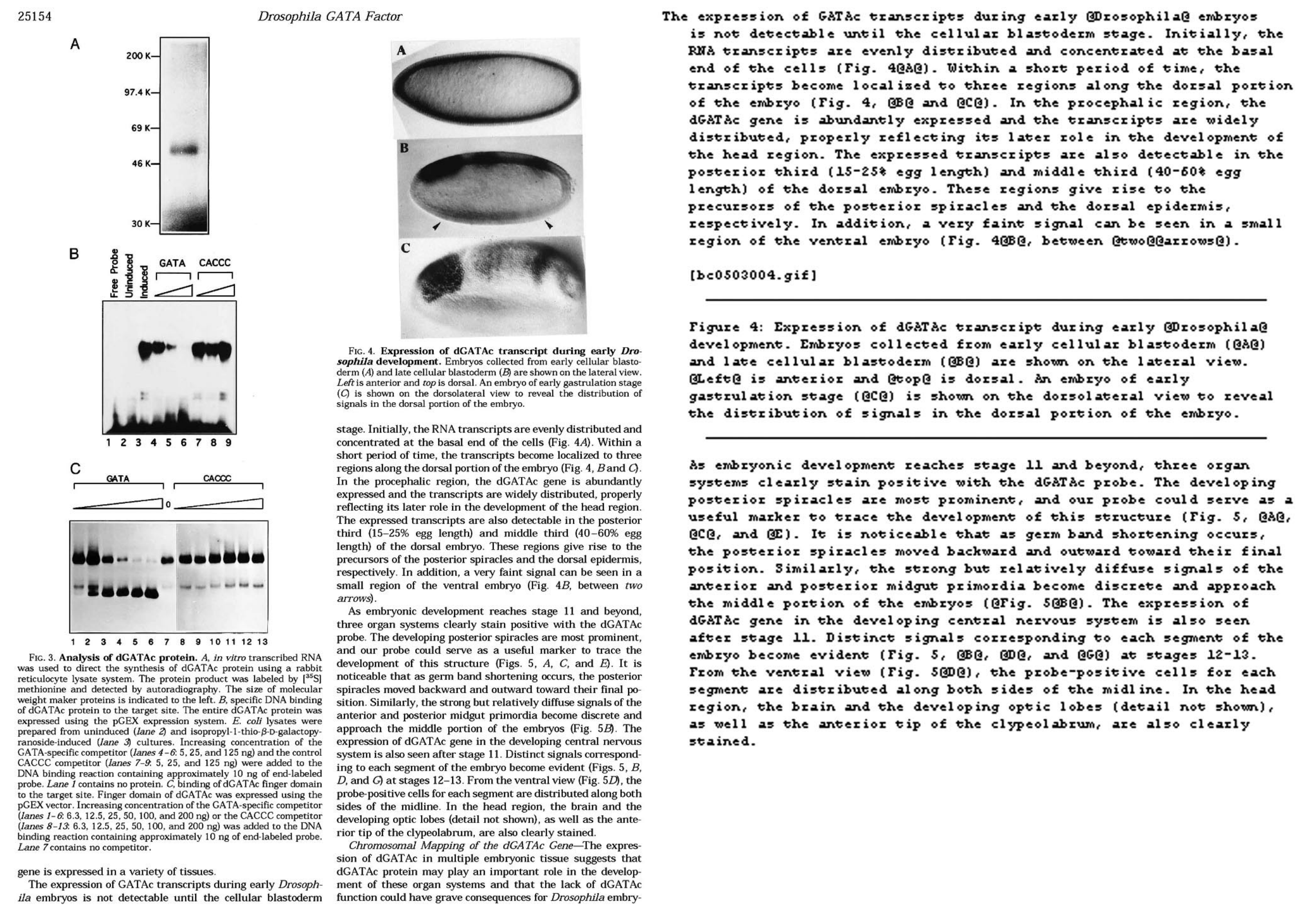
The KDD Cup competition was held in conjunction with the ACM SIGKDD International Conference on KDD held on July 23–26, 2002, in Edmonton, Alberta, Canada. The KDD Cup included two separated tasks, but we refer only to the first task. A full description of the two tasks is given in [34]. The following section gives a brief overview.

The *Drosophila* fruit fly is one of the most researched organisms in molecular biology and genetics. Although its genome has been already completely sequenced, research into the expression and function of various genes is ongoing. The number of papers discussing *Drosophila* is so high (tens of thousands) that a database dedicated to *Drosophila* genetics and molecular biology – FlyBase (http://www.flybase.org) – was constructed.

Given their number, a full review of all the published papers would take a significant amount of time. There was thus a demand for automated systems to perform at least some of the tasks. One of the basic tasks of the FlyBase curators is to identify which papers describe experimental results about the expression of *Drosophila* genes in natural conditions, and, for each such paper, for which genes the experimental evidence was provided. Experimental evidence can refer to the expression of transcripts (RNA) or to the expression of proteins.

Contrary to what might be expected, only about one-third of the papers include such evidence; many papers discuss results that were produced artificially (ectopically) or some indirect research related to *Drosophila* genes. Moreover, although a paper might mention many genes, the experimental results often refer only to a sub-set of the whole set of genes mentioned. The most well-known example is the *w* gene, which is responsible for the eye-color of the fly (red, dominant; white, recessive). This gene was identified as early as the 1920s, and because it is sex-linked and its prototype so easy to observe, it is often used as a tool for studying other genes. For the tasks discussed here, this would mean that the *w* gene alone is irrelevant. Unfortunately, the case for other genes is much more complex. Although one paper might investigate the expression of a gene at a certain stage of the development of the fly embryo, another paper might discuss unnatural expressions of the same gene, observed in conjunction to the action of other genes or processes.

Another problem inherent in this task is the format of the papers. Although the FlyBase curator reviews the full paper as originally published (typically in PDF format), the KDD Cup systems had to cope with a text version, without the figures and with little formatting. Figure 2 compares the two versions of a page of an article from the training set [49].


A
200 K–
97.4 K–
69 K–
46 K–
30 K–
B
Free Probe
Uninduced
Induced
GATA
CACCC
1 2 3 4 5 6 7 8 9
C
GATA
CACCC
0
1 2 3 4 5 6 7 8 9 10 11 12 13
A
B
C

25154 *Drosophila GATA Factor*

FIG. 3. **Analysis of dGATAc protein.** *A, in vitro* transcribed RNA was used to direct the synthesis of dGATAc protein using a rabbit reticulocyte lysate system. The protein product was labeled by [^{35}S] methionine and detected by autoradiography. The size of molecular weight maker proteins is indicated to the left. *B*, specific DNA binding of dGATAc protein to the target site. The entire dGATAc protein was expressed using the pGEX expression system. *E. coli* lysates were prepared from uninduced (*lane 2*) and isopropyl-1-thio-β-D-galactopyranoside-induced (*lane 3*) cultures. Increasing concentration of the GATA-specific competitor (*lanes 4–6*: 5, 25, and 125 ng) and the control CACCC competitor (*lanes 7–9*: 5, 25, and 125 ng) were added to the DNA binding reaction containing approximately 10 ng of end-labeled probe. *Lane 1* contains no protein. *C*, binding of dGATAc finger domain to the target site. Finger domain of dGATAc was expressed using the pGEX vector. Increasing concentration of the GATA-specific competitor (*lanes 1–6*: 6.3, 12.5, 25, 50, 100, and 200 ng) or the CACCC competitor (*lanes 8–13*: 6.3, 12.5, 25, 50, 100, and 200 ng) was added to the DNA binding reaction containing approximately 10 ng of end-labeled probe. *Lane 7* contains no competitor.

FIG. 4. **Expression of dGATAc transcript during early *Drosophila* development.** Embryos collected from early cellular blastoderm (*A*) and late cellular blastoderm (*B*) are shown on the lateral view. *Left* is anterior and *top* is dorsal. An embryo of early gastrulation stage (*C*) is shown on the dorsolateral view to reveal the distribution of signals in the dorsal portion of the embryo.

gene is expressed in a variety of tissues.

The expression of GATAc transcripts during early *Drosophila* embryos is not detectable until the cellular blastoderm stage. Initially, the RNA transcripts are evenly distributed and concentrated at the basal end of the cells (Fig. 4*A*). Within a short period of time, the transcripts become localized to three regions along the dorsal portion of the embryo (Fig. 4, *B* and *C*). In the procephalic region, the dGATAc gene is abundantly expressed and the transcripts are widely distributed, properly reflecting its later role in the development of the head region. The expressed transcripts are also detectable in the posterior third (15–25% egg length) and middle third (40–60% egg length) of the dorsal embryo. These regions give rise to the precursors of the posterior spiracles and the dorsal epidermis, respectively. In addition, a very faint signal can be seen in a small region of the ventral embryo (Fig. 4*B*, between *two arrows*).

As embryonic development reaches stage 11 and beyond, three organ systems clearly stain positive with the dGATAc probe. The developing posterior spiracles are most prominent, and our probe could serve as a useful marker to trace the development of this structure (Figs. 5, *A*, *C*, and *E*). It is noticeable that as germ band shortening occurs, the posterior spiracles moved backward and outward toward their final position. Similarly, the strong but relatively diffuse signals of the anterior and posterior midgut primordia become discrete and approach the middle portion of the embryos (Fig. 5*B*). The expression of dGATAc gene in the developing central nervous system is also seen after stage 11. Distinct signals corresponding to each segment of the embryo become evident (Figs. 5, *B*, *D*, and *G*) at stages 12–13. From the ventral view (Fig. 5*D*), the probe-positive cells for each segment are distributed along both sides of the midline. In the head region, the brain and the developing optic lobes (detail not shown), as well as the anterior tip of the clypeolabrum, are also clearly stained.

Chromosomal Mapping of the dGATAc Gene—The expression of dGATAc in multiple embryonic tissue suggests that dGATAc protein may play an important role in the development of these organ systems and that the lack of dGATAc function could have grave consequences for *Drosophila* embry-

```
The expression of GATAc transcripts during early @Drosophila@ embryos
   is not detectable until the cellular blastoderm stage. Initially, the
   RNA transcripts are evenly distributed and concentrated at the basal
   end of the cells (Fig. 4@A@). Within a short period of time, the
   transcripts become localized to three regions along the dorsal portion
   of the embryo (Fig. 4, @B@ and @C@). In the procephalic region, the
   dGATAc gene is abundantly expressed and the transcripts are widely
   distributed, properly reflecting its later role in the development of
   the head region. The expressed transcripts are also detectable in the
   posterior third (15-25% egg length) and middle third (40-60% egg
   length) of the dorsal embryo. These regions give rise to the
   precursors of the posterior spiracles and the dorsal epidermis,
   respectively. In addition, a very faint signal can be seen in a small
   region of the ventral embryo (Fig. 4@B@, between @two@@arrows@).

   [bc0503004.gif]

   ____________________________________________________________________

   Figure 4: Expression of dGATAc transcript during early @Drosophila@
   development. Embryos collected from early cellular blastoderm (@A@)
   and late cellular blastoderm (@B@) are shown on the lateral view.
   @Left@ is anterior and @top@ is dorsal. An embryo of early
   gastrulation stage (@C@) is shown on the dorsolateral view to reveal
   the distribution of signals in the dorsal portion of the embryo.
   ____________________________________________________________________

   As embryonic development reaches stage 11 and beyond, three organ
   systems clearly stain positive with the dGATAc probe. The developing
   posterior spiracles are most prominent, and our probe could serve as a
   useful marker to trace the development of this structure (Fig. 5, @A@,
   @C@, and @E). It is noticeable that as germ band shortening occurs,
   the posterior spiracles moved backward and outward toward their final
   position. Similarly, the strong but relatively diffuse signals of the
   anterior and posterior midgut primordia become discrete and approach
   the middle portion of the embryos (@Fig. 5@B@). The expression of
   dGATAc gene in the developing central nervous system is also seen
   after stage 11. Distinct signals corresponding to each segment of the
   embryo become evident (Fig. 5, @B@, @D@, and @G@) at stages 12-13.
   From the ventral view (Fig. 5@D@), the probe-positive cells for each
   segment are distributed along both sides of the midline. In the head
   region, the brain and the developing optic lobes (detail not shown),
   as well as the anterior tip of the clypeolabrum, are also clearly
   stained.
```

Figure 2. Original (left) and text (right) versions of an article from the Knowledge Discovery and Data Mining (KDD) Cup 2002.

Finding the right approach: IE vs categorization

At a first glance, the first and second tasks of the KDD Cup are clear categorization tasks: take a scientific paper and classify it either as a curatable (relevant, a paper that includes a experimental result evidence) or as non-curatable (irrelevant). Or, for the first task, simply order the documents according to their relevance. Even the third task, which required decisions regarding the product (transcript or protein) of each gene, can be presented as a categorization task (See [50]).

By contrast, as Yeh *et al.* presented [34], IE is a less intuitive approach to such tasks because it usually deals with the extraction of single instances of certain templates, without a global 'Yes/No' question regarding a whole document.

However, closer analysis of the tasks and actual experiments made with the task data led to the conclusion that IE is significantly more suitable than categorization for these tasks.

First, most papers are from the same domain (molecular biology and/or genetics) and thus use a relatively narrow vocabulary. As most categorization approaches use words from the document as the features inserted into the classification model, the representation of the various documents will be relatively similar; most papers in the KDD Cup collection include words such as '*Drosophila*', 'gene', and so on, with high frequency. The frequency of such words thus contributes little to information regarding curate or non-curate decisions.

Second, many curatable papers have both relevant results (e.g. wild-type expression) and irrelevant ones (e.g. mutations), and, as discussed previously, the same paper might include relevant results for some of the referenced genes, but not for others. Such distinction between the various genes cannot be achieved by classical categorization: relevant phrases (patterns) for the specific genes must be found within the papers.

Third, the third task required deciding, for each gene, between transcript and protein. Although the two products could theoretically be treated as independent, discussions with domain experts and checking of the actual training data proved that this was not the case. Evidence of a transcript leads, in most cases, to evidence of a protein, sometimes even when no direct expression of the protein is shown. Moreover, we found that certain forms of gene synonyms usually indicate reference to the protein (typically all-capital-case forms), and certain other forms indicate reference to the transcript of the same gene. Such distinctions and dependencies are much easier to deal with in an IE, rule-based system than in a categorization system.

Finally, many categorization approaches (i.e. SVM, support vector machine, used by the team from Imperial College London, UK [50]) are 'black box' approaches, supplying the final result but making it difficult for the user (certainly the end-user) to analyze the reasons for the result and to improve and/or fix the system if needed. By contrast, an IE system such as the one we constructed (see later) 'collects' various local evidences within the document and uses them to make the global decision for the whole document, and can therefore supply these evidences 'for free', both to the developer and to the end-user. We believe that this feature (although not required for the KDD Cup task itself) is crucial to curation and/or literature mining. An application that enables the end-user (in this case the FlyBase curators) to see the evidences extracted by our system and to check whether they are correct or incorrect (or add new evidences not found by the system) is described later in this review.

On a practical note, actual experiments that examined the use of classical categorization approaches for these tasks were not very successful (an F-measure score of only 62-64% for the second task on the training set, and an even lower score for the test set using the training set classifier). For the third task (decision for single genes), categorization results were extremely poor. These results are consistent with the reports of other teams [50].

Focus of the rule-based IE module

Following the analysis described earlier, we decided that, as it would be very difficult to search the whole paper, our focus should be on those elements of the paper that a human reader going through the paper quickly would focus. Following discussions with domain experts, we decided to focus on the following elements of the paper.

Figure legends. The figure legends best describe the experimental results found, using a relatively small set of patterns and vocabulary (see Fig. 2).

Title keywords and patterns. Relevant papers appear to have relevant keywords or patterns (e.g. 'Expression of a *Drosophila* GATA transcription factor'), whereas papers with title keywords identified as 'negative' (for this task) (e.g. 'Epidermal muscle attachment site-specific target gene expression and interference with myotube guidance in response to ectopic stripe expression in the developing *Drosophila* epidermis) are less likely to report experimental results about natural gene expressions.

Paper abstract. In relation to patterns found regarding gene expression, there were often notes indicating that the author submitted a novel sequence to GenBank (e.g. 'The nucleotide sequence(s) reported in this paper has been submitted to the GenBank(TM)/EMBL Data Bank with accession number(s) D50542[GenBank]).

Design of the rule-based IE module

Our IE Module was constructed using several layers as opposed to one unit.

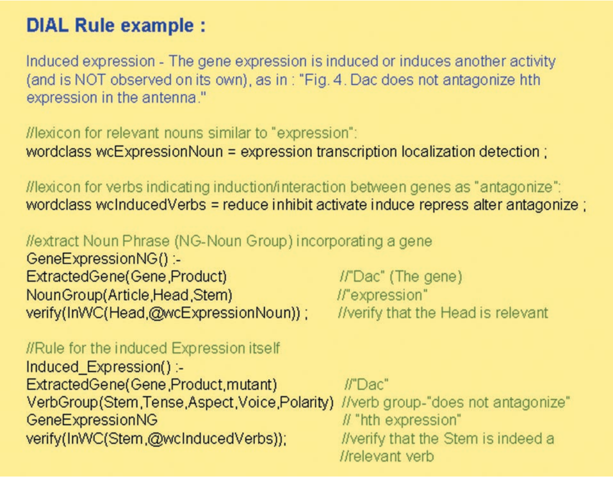
Infrastructure layer. This layer comprises general libraries developed not in conjunction with this specific project. These libraries provide basic utilities for the IE process itself and basic NLP (morphology) tools. The NLP tools themselves include several layers (see [45]), including the part-of-speech tagger, and noun phrase and verb phrase grouper.

Metadata management layer. This layer is responsible for: (1) identifying the gene candidates for a given paper and storing them in a suitable data structure (a map; an available data-type within our IE language); (2) updating the gene scores through the document as evidences regarding a specific gene are found (see later); and (3) checking the final score of each gene and of the whole paper, and writing the required results for the paper (at the post-processing stage). (These decisions are made using certain thresholds that can be adjusted as required.)

Gene identification layer. This is an auxiliary layer for extraction of gene occurrences with their synonyms. It performs normalization of gene names according to the gene thesaurus, and normalizes certain typographies (e.g. 'Dgc [alpha.gif] 1' to 'Dgc&agr ; 1'). This layer also attempts to identify whether the reference is to a protein or to a transcript (e.g. all-capital-case instances, such as APPL, are usually proteins), and whether the format of the gene itself and the following tokens indicate that it is a transgene (unnatural mutation), such as in 'Drab6[wt]'.

Structural analysis layer. Following our decision to focus on the sections listed earlier, and because we received text versions of the whole paper, it was essential to write rules to identify these sections based on clues such as carriage-returns and some keywords. Once such a section is found, it is sent for further processing by the fifth (main) layer.

Main layer. This layer extracts the required evidences using classical IE heuristics, that is, by extracting 'local' patterns found within the relevant sections, and updating the scores of the gene and the whole paper towards the final, 'global'



```
DIAL Rule example :

Induced expression - The gene expression is induced or induces another activity
(and is NOT observed on its own), as in : "Fig. 4. Dac does not antagonize hth
expression in the antenna."

//lexicon for relevant nouns similar to "expression":
wordclass wcExpressionNoun = expression transcription localization detection ;

//lexicon for verbs indicating induction/interaction between genes as "antagonize":
wordclass wcInducedVerbs = reduce inhibit activate induce repress alter antagonize ;

//extract Noun Phrase (NG-Noun Group) incorporating a gene
GeneExpressionNG() :-
ExtractedGene(Gene,Product)                  //"Dac" (The gene)
NounGroup(Article,Head,Stem)                 //"expression"
verify(InWC(Head,@wcExpressionNoun)) ;       //verify that the Head is relevant

//Rule for the induced Expression itself
Induced_Expression() :-
ExtractedGene(Gene,Product,mutant)            //"Dac"
VerbGroup(Stem,Tense,Aspect,Voice,Polarity)   //verb group-"does not antagonize"
GeneExpressionNG                              // "hth expression"
verify(InWC(Stem,@wcInducedVerbs));           //verify that the Stem is indeed a
                                              //relevant verb
```

Figure 3. Sample of the declarative information analysis language (DIAL) rule used in the Knowledge Discovery and Data Mining (KDD) Cup 2002. We used a lexicon of expression nouns ('wcExpressionNoun') and a lexicon of verbs ('wcInducedVerbs') to indicate a relationship between two genes. 'ExtractedGene' is the predicate implemented in the gene identification layer (matching a gene and finding its normalized form). Induced expression is the main rule. When a gene is found that is followed by a verb group whose main verb is in the 'wcInducedVerbs' wordclass, and then followed by an expression of another gene ('GeneExpressionNG'), then this an classed as induced expression. It should therefore be treated as negative evidence for either gene.

decision. We extracted both positive evidences (direct descriptions of relevant results or indicative keywords) and negative evidences (patterns suggesting that the document discusses results that are irrelevant). For example, positive evidence might be the phrase: 'Figure 2. Northern blot analysis of fruitless mRNA'. (Northern blot is a common technique for showing transcript expressions.) Negative evidence might be the phrase: 'Figure 3. Ectopic expression of dNSF2 in mesoderm is sufficient to rescue the lethality of dNSF2 mutations'.

The main layer combines lexical resources, NLP tools and semantic constraints. For example, if the *hsp70* gene is found within a phrase such as '@hsp70@-@white@ transgene', it is ignored. Similarly, if a gene expression phrase is found within a verb phrase that describes a functional dependency result, it is also ignored (e.g. 'Dac does not antagonize *hth* expression in the antenna'). The latter example relies directly on our infrastructure NLP layer, which extracts verb patterns such as 'Dac does not antagonize' (see [51] for more information).

Lexical resources. The system uses lexicons for key pattern elements such as analysis techniques (e.g. northern blot), positive headline keywords (e.g. homolog) and negative headline keywords (e.g. ectopic, unnatural).

Implementation in DIAL

As discussed earlier, DIAL provides built-in libraries for tokenizing, part of speech (POS)-tagging and noun- and verb-phrase extraction. DIAL also uses dynamic mapping from the metadata management layer, and the thesaurus for gene identification. We focus here on sample rules from the main (highest) layer that extracts the evidences themselves. Figure 3 shows a sample rule for 'induced expression' – expression of a gene reported under certain conditions influenced by another gene. As explained earlier, such expression should not be extracted as a 'natural expression'.

Evaluation of our system

We developed our system based on a training set of 862 full-text articles tagged by the FlyBase curators. The system was also tested on a separate set of 213 articles (see [34] for more information).

Our system achieved the best results (of 32 participating systems) in all three tasks. Our F-Measure result for the second task (Yes/No curate decision) was 78%, compared with a median of 58% for all the systems submitted. Our F-Measure result for the third task (Yes/No decision regarding each gene) was 67%, compared with a median of 35% for all the system submitted. As discussed earlier, these results clearly demonstrate the superiority of the rule-based IE approach for this task.

Visualization in text mining

One of the crucial requirements when developing a text mining system is the ability to browse through the document collection and be able to 'visualize' various elements within the collection. This type of interactive exploration enables the identification of new types of entities and relationships that can be extracted, and better exploration of the results from the IE phase [42,36].

Relationship maps provide a visual means for concise representation of the relationships among many terms in a given context. To define a relationship map, the user specifies: (1) a taxonomy category (e.g. 'genes') that determines the nodes of the graph; and (2) an optional context node (e.g. 'phosphorylation') that determines the type of connection the user wishes to find among the graph nodes.

If no context is provided, the system will revert to using co-occurrence information between entities. We say that two entities co-occur within a lexical unit (such as sentence, paragraph or document) if they are both contained within the same lexical unit. The most common lexical level for co-occurrence computation is the sentence. Entities that appear within the same sentence are said to be 'co-occurring in the sentence level'. Presenting co-occurrence maps is one of the main methods that enables the developer

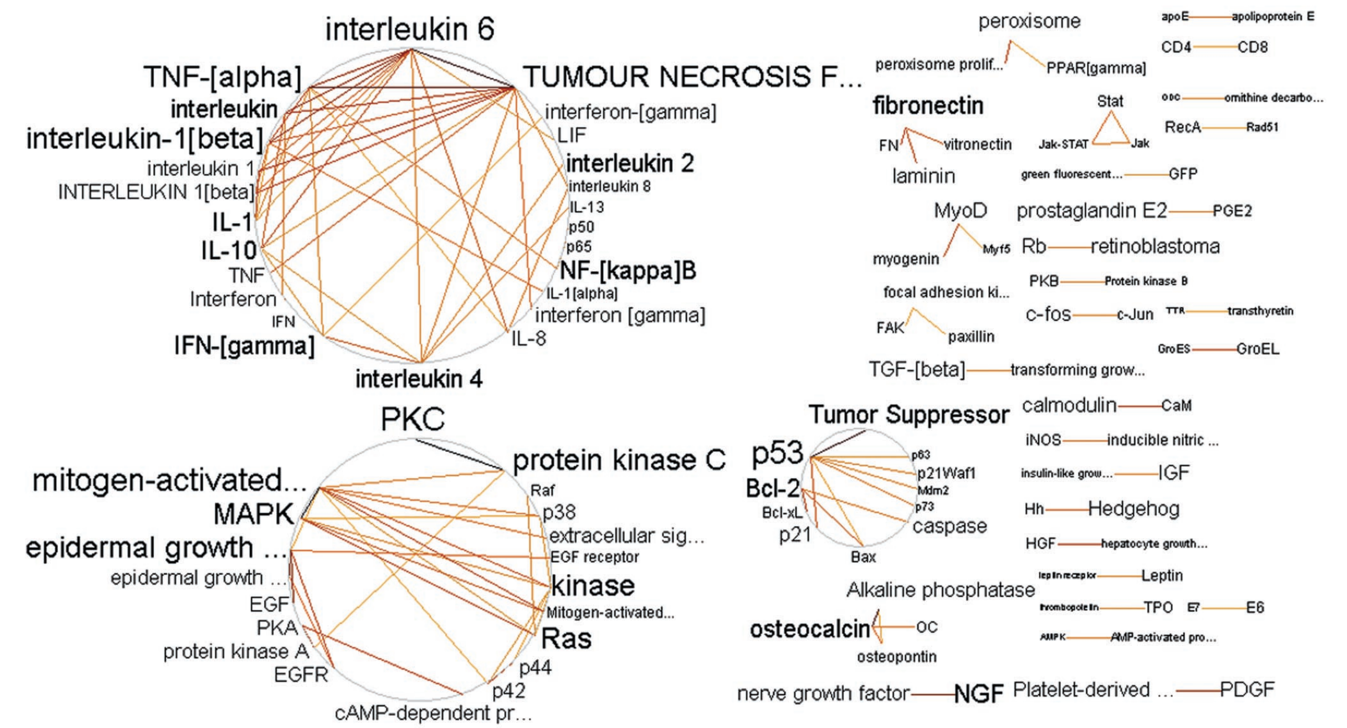


Figure 4. Co-occurrence between genes. Data is based on 30,000 MedLine abstracts.

to develop new IE rules. Such maps are used in the process that creates new semantic relationships. This semi-automatic process takes the output of such maps and creates IE-rule candidates that are refined by the user. Figure 4 shows a map created by finding co-occurrences of genes in PubMed articles. Figure 5 features a semantic relationship map showing links between genes that are either homologs or are associated by a gene-binding or gene-pathway relationship. The map in Figure 6 depicts relationships between genes and diseases.

Machine-assisted indexing

No IE system is 100% accurate. Regardless of the approach taken, there will always be instances of entities or relationships that the system will miss, as well as some incorrect (false-positive) instances that will nevertheless be extracted. The reason for this is the complex nature of the human language; a computerized system will never be able to trace all the possible phrasing and contexts used by humans or use all the domain expertise of humans. Therefore, for many applications, it is useful to give human experts the opportunity to review the results that are tagged or extracted by the IE system. This is particularly useful for areas where much domain expertise is required, such as the biomedical domain. For example, machine-assisted-indexing (MAI) (extraction) can be applied to the FlyBase curation task described earlier.

The results achieved by our system are not sufficient to achieve a totally automated process, but can certainly be used as the basis for an MAI system. Instead of having to carefully read all the papers, the system processes the papers

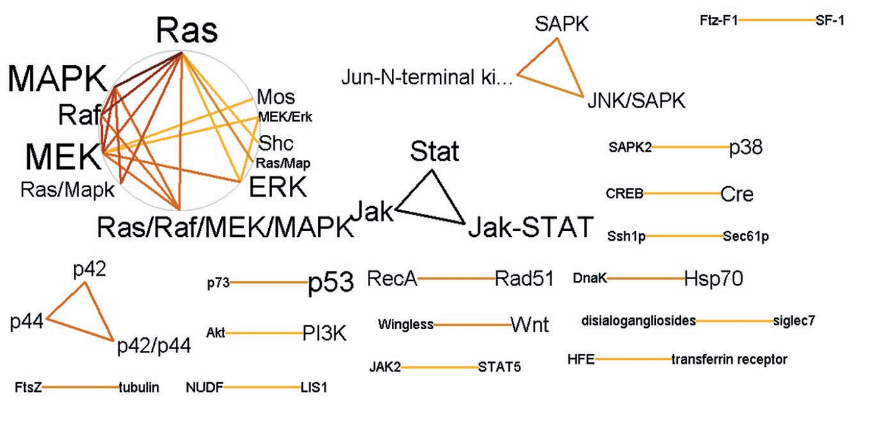


Figure 5. Relationship Map, Genes that are homologs, or form a GeneBinding or GenePathway Rerelationship, Data is based on 30,000 medline abstracts. The map is based on results of semantic extraction of the relationship.

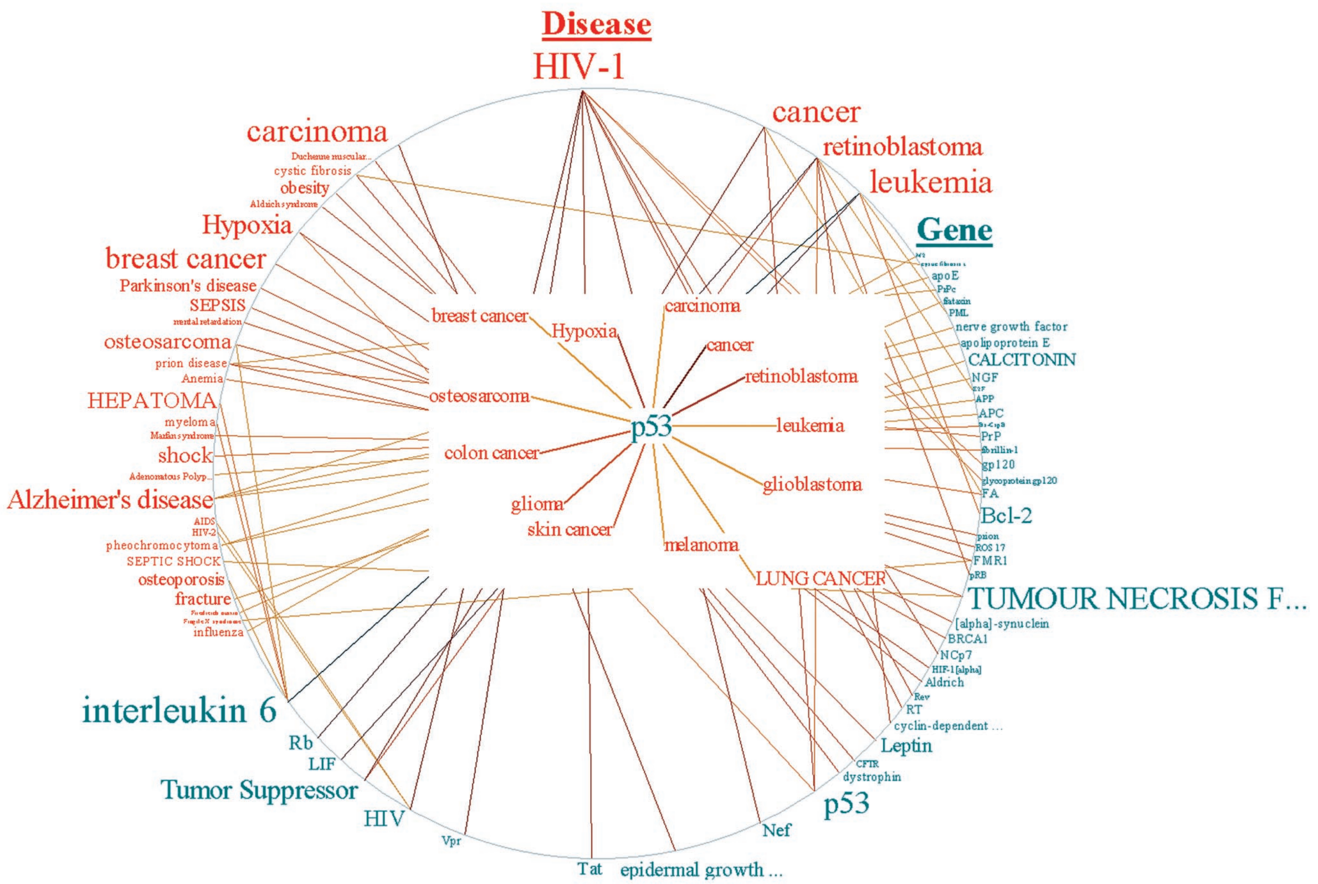


Figure 6. Relationships between genes and diseases. Data is based on 30,000 MedLine abstracts. Shown here is the gene encoding p53. Results are based on co-occurrence between genes and diseases within the same sentence.

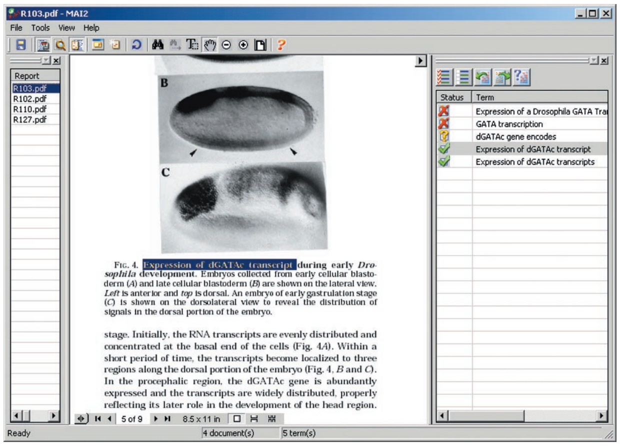
for the expert (i.e. the FlyBase curator), and suggests suitable results for each paper, that is, whether it should be curated, and for which genes the system found the required experimental results. The expert can then check the results of the system – a process significantly shorter than reading all the papers.

The MAI application is interactive. It performs the 'dirty work' and enables the user to go directly to the sections from which the relevant information was extracted.

The user can also view the location of the extracted information within the original paper in a PDF format (with the original figures) and then decide whether the instance was extracted correctly or incorrectly. All the tools associated with the PDF document (e.g. maximizing and minimizing the image) can also be used. In addition, the user can add or delete terms, or change the status of extracted terms. The MAI application can be used simultaneously by several users (each one checking different documents), recording any changes made by each user. Figure 7 shows the application of MAI to a sample of FlyBase articles (KDD Cup training documents; PDF versions are publicly available from NCBI's PubMed).

Concluding remarks

Owing to the abundance of available biomedical data in free text format, there is a growing need for efficient tools for text mining. Unlike structured data, where data mining



Report
R103.pdf
R102.pdf
R110.pdf
R127.pdf

FIG. 4. Expression of dGATAc transcript during early *Drosophila* development. Embryos collected from early cellular blastoderm (*A*) and late cellular blastoderm (*B*) are shown on the lateral view. *Left* is anterior and *top* is dorsal. An embryo of early gastrulation stage (*C*) is shown on the dorsolateral view to reveal the distribution of signals in the dorsal portion of the embryo.

stage. Initially, the RNA transcripts are evenly distributed and concentrated at the basal end of the cells (Fig. 4*A*). Within a short period of time, the transcripts become localized to three regions along the dorsal portion of the embryo (Fig. 4, *B* and *C*). In the procephalic region, the dGATAc gene is abundantly expressed and the transcripts are widely distributed, properly reflecting its later role in the development of the head region.

5 of 9 | 8.5 x 11 in

Status	Term
✗	Expression of a Drosophila GATA Tra
?	GATA transcription
?	dGATAc gene encodes
✓	Expression of dGATAc transcript
✓	Expression of dGATAc transcripts

Ready | 4 document(s) | 5 term(s)

Figure 7. Machine-assisted-indexing (MAI) for FlyBase articles.

algorithms can be applied directly to the underlying data, text mining requires some pre-processing before any mining algorithm can be successfully applied. IE has proven to be an efficient method for this pre-processing phase. Text mining using IE thus provides a useful middle-ground in the quest for tools to facilitate an understanding of the information captured in textual formats. The powerful combination of precise analysis of the biomedical documents with a set of visualization tools enables the user to navigate and use easily this abundance of biomedical document collections.

Acknowledgements

The KDD Cup work was performed in cooperation with a team from Celera Genomics (http://www.celera.com), headed by Hagit Shatkay, and with the help of Meir Shamay of the Weizmann Institute of Science, Rehovot, Israel (http://www.weizmann.ac.il).

References

1 Ponte, J.M. and Croft, W.B. (1998) A language modeling approach to information retrieval. *Proceedings of the 21st Annual International ACM SIGIR,* 275–281

2 Salton, G. (1989) *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*, Addison-Wesley

3 Shatkay, H. *et al.* (2002) Information retrieval meets gene analysis. *IEEE Intelligent Systems* 17, 45–53

4 Allen, J. (1995) *Natural Language Understanding*, Addison-Wesley

5 Charniak, E. (1993) *Statistical Language Learning*, MIT Press

6 Frantzi, T.K. (1997) Incorporating context information for the extraction of terms. *Presented at the conference of ACL-EACL, 7–11 July 1997, Madrid, Spain*

7 Manning, C.D. and Schutze, H. (1999) *Foundations of Statistical Natural Language Processing*, MIT Press

8 Cardie, C. (1997) Empirical methods in information extraction. *AI Magazine* 18, 65–80

9 Chang, J.T. *et al.* (2001) Including biological literature improves homology search. *Proceedings of the Pacific Symposium on Biocomputing 2001,* 374–383

10 Cowie, J. and Lehnert, W. (1996) Information extraction. *Comm. Assoc. Comput. Mach.* 39, 80–91

11 Fisher, D. *et al.* (1995) Description of the UMass systems as used for MUC-6. *Proceedings of the 6th Message Understanding Conference,* 127–140

12 Grishman, R. (1996) The role of syntax in information extraction. In *Advances in Text Processing: Tipster Program Phase II*, Morgan Kaufmann

13 Lehnert, W. *et al.* (1993) Description of the CIRCUS system as used for MUC-3. *Proceedings of the 3rd Message Understanding Conference,* 223-223

14 Ray, S. and Craven, M. (2001) Representing sentence structure in hidden Markov models for information extraction. *Presented at the International Joint Conference on Artificial Intelligence, August 4–10 2001, Seattle, WA, USA*

15 Riloff, E. and Lehnert, W. (1994) Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems* 12, 296–333

16 Akane, Y. *et al.* (2001) Event extraction from biomedical papers using a full parser. *Proceedings of the Pacific Symposium on Biocomputing 2001,* 408–419

17 Andrade, M.A. and Valencia, A. (1997) Automatic annotation for biological sequences by extraction of keywords from Medline abstracts. Development of a prototype system. *Presented at the AAAI Conference of Intelligent Systems in Molecular Biology, June 21–26, Halkidiki,* Greece

18 Blaschke, C. *et al.* (1999) Automatic extraction of biological information from scientific text: protein–protein interactions. *Proceedings of the AAAI Conference on Intelligent Systems in Molecular Biology,* 60–67

19 Craven, M. and Kumlien, J. (1999) Constructing biological knowledge bases by extracting information from text sources. *Proceedings of the Intelligent Systems for Molecular Biology,* 77–86

20 Ding, J. *et al.* (2002) Mining Medline: abstracts, sentences or phrases. *Proceedings of the Pacific Symposium on Biocomputing,* 326–337

21 Friedman, C. *et al.* (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology,* S74–S82

22 Fukuda, K. *et al.* (1998) Toward information extraction: identifying protein names from biological papers. *Proceedings of the Pacific Symposium on Biocomputing* 705–716

23 Humphreys, K. *et al.* (2000) Two applications of information extraction to biological science journal articles: interactions enzyme and protein structures. *Proceedings of the Pacific Symposium on Biocomputing,* 502–513

24 Krauthammer, M. *et al.* (2002) Truth of and pathways: chasing bits of information through myriads of articles. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology,* S249–S257

25 Park, J. *et al.* (2001) Bidirectional incremental parsing for automatic pathway identification with combinatory categorical grammar. *Proceedings of the Pacific Symposium on Biocomputing,* 396–407

26 Pearson, H. (2001) Biology's name game. *Nature* 411, 631–632

27 Pustejovsky, J. *et al.* (2002) Robust relational parsing over biomedical literature: extracting inhibit relations. *Proceedings of the Pacific Symposium on Biocomputing,* 362–373

28 Rindflesch, T.C. *et al.* (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Proceedings of the Pacific Symposium on Biocomputing,* 514–525

29 Spellman, P.T. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridisation. *Mol. Biol. Cell* 9, 3273–3297

30 Stapley, B.J. and Benoit, G. (2000) BioBibliometrics: retrieval information and visualization from co-occurrences of gene names in Medline abstracts. *Proceedings of the Pacific Symposium on Biocomputing,* 526–537

31 Tanabe, L. *et al.* (1999) MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. *BioTechniques* 27, 1210–1217

32 Thomas, J. *et al.* (2002) Automatic extraction of protein interactions from scientific abstracts. *Proceedings of the Pacific Symposium on Biocomputing,* 538–549

33 Tor-Jenssen, K. *et al.* (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet.* 28, 21–28

34 Yeh, A. *et al.* (2003) Background and overview for KDD Cup 2002 Task 1: information extraction from biomedical articles. *SIGKDD Explorations* 4, 87–89

35 Feldman, R. and Hirsh, H. (1997) Exploiting background information in knowledge discovery from text. *Journal of Intelligent Information Systems* 9, 83-97

36 Feldman, R. *et al.* (1997) Maximal association rules: a new tool for mining for keyword co-occurrences in document collections. *Presented at the 3rd International Conference on Knowledge Discovery, August 14–17 1997, Newport Beach, CA, USA*

37 Frawley, W.J. *et al.* (1991) Knowledge discovery in databases: an overview. In *Knowledge Discovery in Databases* (Piatetsky-Shapiro, G. and Frawley, W.J., eds), pp. 1–27, MIT Press

38 Hayes, P. (1992) Intelligent high-volume processing using shallow, domain-specific techniques. In *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*, pp. 227–242, Lawrence Erlbaum, Hillside, New Jersey

39 Joachims, T. (1998) Text categorization with support vector machines: learning with many relevant features. *Presented at the European Conference on Machine Learning, April 21–24 1998, Chemnitz, Germany*

40 Sebastiani, F. (2002) Machine learning in automated text categorization. *ACM Computing Surveys* 34, 1–47

41 Hofmann, T. (1999) The cluster-abstraction model: unsupervised learning of topic hierarchies from text data. *Proceedings of the Sixteenth Joint International Conference on Artificial Intelligence*, 682–687

42 Aumann, Y. *et al.* (1999) Circle graphs: new visualization tools for text-mining. *Proceedings of the 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases*, 277–282

43 Hayes, P.J and Weinstein, S.P. (1990) CONSTRUE: a system for content-based indexing of a database of news stories. *Presented at the Second Annual Conference on Innovative Applications of Artificial Intelligence, 1990, Washington DC, USA*

44 Yang, Y. and Liu, X. (1999) A re-examination of text categorization methods. *Proceedings of the ACM International Conference on Research and Development in Information Retrieval*, 42–49

45 Feldman, R. *et al.* (2002) A comparative study of information extraction strategies. *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, 349–359

46 Feldman, R. *et al.* (2001) A domain independent environment for creating information extraction modules. *Proceedings of the Tenth International Conference on Information and Knowledge Management*, 586–588

47 Feldman, R. *et al.* (2000) A framework for specifying explicit bias for revision of approximate information extraction rules. *Proceedings of the Sixth ACM International Conference on Knowledge Discovery and Data Mining*, 189–199

48 Cohen, K.B. *et al.* (2002) Contrast and variability in gene names. *Proceedings of Natural Language Processing in the BioMedical Domain*, 14–20

49 Lin, W.H. (1995) Expression of a *Drosophila* GATA transcription factor in multiple tissues in the developing embryos. Identification of homozygous lethal mutants with P-element insertion at the promoter region. *J. Biol. Chem.* 270, 25150–25158

50 Ghanem, M.M. *et al.* (2003) Automatic scientific text classification using local patterns: KDD CUP 2002 (Task 1). *SIGKDD Explorations* 4, 95–96

51 Regev, Y. *et al.* (2003) Rule-based extraction of experimental evidence in the biomedical domain – the KDD Cup (Task 1). *SIGKDD Explorations* 4, 90–92