

# **Literature mining for the biologist**

**From information retrieval to biological discovery**

**Lars Juhl Jensen**

European Molecular Biology Laboratory, D-69117 Heidelberg, Germany

**Jasmin Šarić**

EML Research gGmbH, D-69118 Heidelberg, Germany

**Peer Bork**

European Molecular Biology Laboratory, D-69117 Heidelberg, Germany

Max-Delbrück-Centre for Molecular Medicine, D-13092 Berlin, Germany

## Contents

<b>Introduction</b>	<b>3</b>
<b>Information retrieval</b>	<b>5</b>
<b>Entity recognition and identification</b>	<b>7</b>
<b>Information extraction</b>	<b>9</b>
Statistical co-occurrence methods . . . . .	9
Natural language processing (NLP) . . . . .	10
An example NLP system . . . . .	11
Applications of IE . . . . .	14
<b>Text mining</b>	<b>17</b>
Mining text for overlooked “golden nuggets” . . . . .	17
Discovery of global correlations from literature . . . . .	19
<b>Text/data integration</b>	<b>20</b>
<b>Outlook</b>	<b>24</b>
<b>Acknowledgments</b>	<b>24</b>
<b>Online tools and resources</b>	<b>25</b>

For most biologists hands-on literature mining is currently limited to keyword searches in PubMed. However, methods for extracting biomedical facts from literature have improved considerably, and the associated tools will likely soon be used by many researchers in bioinformatics as well as wet-lab biology. Advanced literature mining tools will be crucial to successfully analyze the deluge of high-throughput experimental data sets in the context of the rapidly increasing body of scientific text. New tools will need to that can automatically propose new hypotheses and thereby catalyze the discovery process. This will require that high-throughput data and literature become tightly integrated, which encourages close collaborations between biologists, bioinformaticians, and computational linguists.

## Introduction

The focus in biology is shifting from individual genes and proteins to entire biological systems, and biologists must therefore be able to systematically compare large-scale data sets with currently known, that is the scientific literature. As the numbers of articles published each year is increasing exponentially, it is no longer possible for a researcher to read all the relevant articles manually, not even on a single, specialized topic such as the cell cycle (Figure 1).

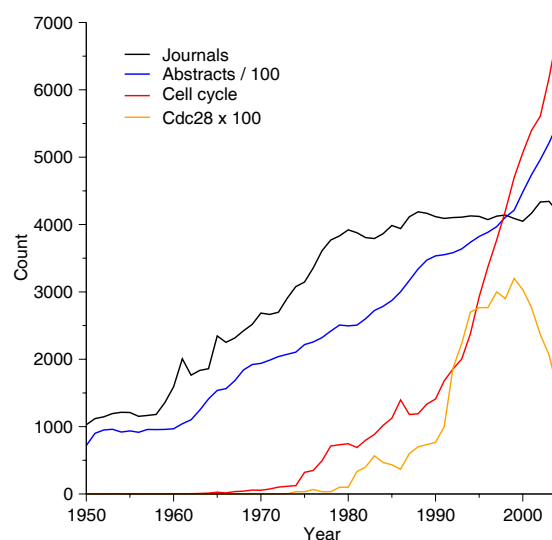


Figure 1: **Growth of MEDLINE.** The counts reflect the number of papers and journals published per year; a running average of three years was calculated for the Cdc28 curve due to the much lower counts. The number of new papers published each year continues to increase, especially on certain topics such as the cell cycle, for which it is no longer possible to read all new papers that are published. In contrast, specific proteins that are “hot” at one point in time tend to later lose their popularity, as exemplified by Cdc28.

Because of these changes, literature mining tools are becoming essential to researchers in the fields of biology and bioinformatics. First, they enable researchers to identify relevant papers (information retrieval, IR). They also allow the biological entities (for example, genes and proteins) mentioned in these papers to be recognized (entity recognition, ER) and enable specific facts from papers to be pulled out (information extraction, IE). IR tools like PubMed have long been used on a regular basis by most biologists to find papers of interest. In contrast, automatic methods for extracting facts from text (IE) have only very recently become sufficiently accurate to be useful in practice (Rebholz-Schuhmann, 2005). It is obvious how both IR and IE can be used for curation efforts; however, they are often dismissed as being useless for discovery purposes as they can only extract what has already been published.

More advanced tools based on these methods facilitate systematic searches of the scientific literature for overlooked connections (text mining) and integration of the literature with other data types to make new discoveries. Although some text mining methods are indeed capable of making novel hypotheses by combining information from multiple papers, we believe that the full discovery potential will only be realized with data mining approaches that integrate literature and data from high-throughput experiments such as genome sequencing, microarray expression studies, or protein–protein interaction screens.

Here, we will briefly describe the aim of each field described above, give an overview of the methods employed, and discuss what can currently be achieved. We first give an overview the most important IR, ER and IE methodologies subsequently give examples of how the results can be mined and integrated with other data types to make new biological discoveries. For more details, the reader is referred to the numerous reviews on these topics (Manning and Schütze, 1999; Andrade and Bork, 2000; Hirschman et al., 2002; Yandell and Majoros, 2002; Krallinger and Valencia, 2005; Scherf et al., 2005; Shatkay, 2005; Skusa et al., 2005; Cimiano et al., 2006; Jensen et al., 2006).

To exemplify the goals of each sub-field within biomedical literature mining, we will use the following example sentence: “Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homolog) directly phosphorylated Swe1 and this modification served as a priming step to promote subsequent Cdc5-dependent Swe1 hyperphosphorylation and degradation” (Asano et al., 2005). Its context is the cell cycle of the yeast *Saccharomyces cerevisiae*, and it allows us illustrate the powers and pitfalls of current literature mining approaches.

## Information retrieval

Information retrieval (IR) systems aim to identify the text segments (be it full articles, abstracts, paragraphs, or sentences) pertaining to a certain topic—in our example, the yeast cell cycle. The topic can be defined in one of two ways: either by a user-provided query (*ad hoc* IR) or by a set of papers that has been manually selected as being relevant to the topic (text categorization). Both types of IR system should ideally recognize our example sentence as being related to the yeast cell cycle, although neither “yeast” nor “cell cycle” is explicitly mentioned.

The best known biomedical IR system, PubMed, is an *ad hoc* system that uses two well established IR methodologies, the Boolean model and the vector model. The former enables the user to retrieve all documents that contain certain combinations of terms, for example “yeast AND cell cycle”. In contrast, the vector model represents each document by a term vector, in which each term is assigned a value according to a frequency-based weighting scheme. These document vectors can subsequently be compared to a query vector that specifies the relative importance of each query term (Wilbur and Yang, 1996). Alternatively, they can be compared to each other to calculate document similarity, which is used by PubMed’s *related articles* function (Wilbur and Coffee, 1994) and other document clustering methods (Renner and Aszodi, 2000; Iliopoulos and Ouzounis, 2001; Glenisson et al., 2003). The vector representation is also used as input for machine learning methods, which are trained to discriminate between known relevant (positive) and irrelevant (negative) papers based on their word content (Usuzaka et al., 1998; Marcotte et al., 2001; Bhalotia et al., 2003; Donaldson et al., 2003; Kayaalp et al., 2003; Aronson et al., 2004; Goetz and von der Lieth, 2005; Shah et al., 2005; Suomela and Andrade, 2005). Such methods are able to learn fairly complex rules; for example, a method trained to identify sentences related to the yeast cell cycle would have learned that the word “Cdc28” in our example sentence is a strong hint, whereas the words “Cdk1” and “Clb2” could be related to the cell cycle of other organisms as well.

*Ad hoc* IR systems like PubMed generally have more difficulty than text categorization systems in dealing with the many abbreviations, synonyms, and ambiguities in the biomedical terminology, although blind assessments have shown that most of the lessons learned from IR in other research fields carry over to biomedicine (Glenisson et al., 2003; Hersh and Bhuptiraju, 2003; Hersh et al., 2004a,b). These include removing so-called *stop words* like “the” and “it”, which occur in almost every document, and truncating common word endings like “-ing” and “-s” to allow different forms of the same word to be matched, for example “yeast” and “yeasts” (Bhalotia et al., 2003). PubMed and many other good biomedical IR systems also make use of thesauri to automatically expand the query with additional related terms (Bhalotia et al.,

2003; Hersh and Bhuptiraju, 2003; Büttcher et al., 2004; Hersh et al., 2004b). For example the Boolean query “yeast AND cell cycle” might be expanded to “(yeast OR *Saccharomyces cerevisiae*) AND cell cycle”). Many advanced methods, such as Med-Miner (Tanabe et al., 1999) and Textpresso (Muller et al., 2004), also use ER methods (see below) to better identify documents that mention a certain gene or protein and/or part-of-speech tagging to disambiguate whether a word such as “wingless” occurs as a noun or an adjective. As many documents may be retrieved by a single query, simply presenting them as a long list gives poor overview. Alternative ways to present and summarize IR results are thus being actively explored (Perez-Iratxeta et al., 2001, 2003; Hoffmann and Valencia, 2004; Doms and Schroeder, 2005; Hoffmann et al., 2005).

Even with these improvements, current *ad hoc* IR systems are not able to retrieve our example sentence given the query “yeast cell cycle”. This could be achieved by realizing that “yeast” is a synonym for *S. cerevisiae*, that “cell cycle” is a Gene Ontology term, that the word “Cdc28” refers to a *S. cerevisiae* protein, and finally looking up the Gene Ontology terms of Cdc28 to connect the two. Although this will by no means be easy to make work, we see this type of ontology-based reasoning as the next logical step for *ad hoc* IR.

## Entity recognition and identification

The seemingly modest goal of ER is to find the biological entities mentioned within a text, in particular the names of genes and proteins. This task is often divided into two sub-tasks: i) the *recognition* of words that refer to entities and ii) the unique *identification* of the entities in question. In our example sentence, the terms “Cib2”, “Cdc28”, “Cdk1”, “Swe1”, and “Cdc5” should thus all be recognized as gene/protein names and uniquely identified by, for example, their respective *Saccharomyces Genome Database* (SGD) accession numbers.

While ER may at first glance appear neither challenging nor particularly useful, it is possibly the most difficult task in biomedical text mining and is prerequisite for constructing both IE and advanced IR systems. When used alone, ER is useful for molecular biologists to search and structure the biomedical literature based on the genes of interest (Hoffmann and Valencia, 2004).

The early ER methods relied on hand-crafted rules that look for typical features of names, such as letters followed by numbers or the ending “-ase”, as well as contextual information from nearby words like “gene” or “receptor” (Fukuda et al., 1998; Proux et al., 1998; Franzen et al., 2002; Tanabe and Wilbur, 2002; Bhalotia et al., 2003; Narayanaswamy et al., 2003; Tamames, 2005). As several corpora in which gene and protein names have been tagged are now available for download (Kim et al., 2003; Tanabe et al., 2005), this approach is no longer so attractive and most newer systems instead rely on machine learning algorithms such as hidden Markov models (HMM), support vector machines (SVM), or a mixture of both to recognize names based on their characteristic features (Coller et al., 2000; Hatzivassiloglou et al., 2001; Collier et al., 2002; Tanabe and Wilbur, 2002; Chang et al., 2004; Zhou et al., 2004; Hakenberg et al., 2005; McDonald and Pereira, 2005; Settles, 2005; Zhou et al., 2005).

In contrast to these systems, a number of dictionary-based methods instead rely on a comprehensive list of synonymous gene names that are matched against the documents using algorithms that allow variation in how the names are written, for example “CDC28”, “Cdc28”, “Cdc28p”, or “cdc-28” (Krauthammer et al., 2000; Leonard et al., 2002; Bhalotia et al., 2003; Hanisch et al., 2003; Chang et al., 2004; Mika and Rost, 2004; Finkel et al., 2005; Crim et al., 2005; Fundel et al., 2005; Hanisch et al., 2005). These methods are aided by the availability of databases of synonymous gene/protein names (Bussey et al., 2003; Pillet et al., 2005; Shi and Campagne, 2005) as well as tools that can help in constructing these (Pustejovsky et al., 2001; Yoshida et al., 2000; Yu and Agichtein, 2003). Many systems combine dictionary matching with either rule-based or statistical methods to reduce the number of false positive hits caused by homonyms (Leonard et al., 2002; Chang et al., 2004; Seki and Mostafa, 2005; Tsu-ruoka and Tsujii, 2003; Mika and Rost, 2004; Finkel et al., 2005; Kou et al., 2005;

Mitsumori et al., 2005).

The many different methods for doing named entity recognition were evaluated in the blind assessment BioCreAtIvE task 1 (Colosimo et al., 2005; Hirschman et al., 2005; Yeh et al., 2005). It revealed that best performing ER methods all rely on careful curation of the gene name lists to remove aliases that cause many false positives (Fündel et al., 2005; Hanisch et al., 2005). In addition, dictionary-based approaches have the crucial advantage over feature-based ones that they not only recognize names as such, but also identify the accession number of the genes or proteins to which they refer.

The major difficulty in ER arises from the lack of standardization of names. Each gene or protein typically has several different names and abbreviations thereof (“Cdc28” is also known as “Cyclin-dependent kinase 1” or just “Cdk1”), some of are also common English words (“hairy”), biological terms (“SDS”), or names of other genes (“Cdc2” refers to two completely unrelated genes in budding and fission yeast) (Chen et al., 2005). The recent development of methods for disambiguating gene/protein names is thus an important advance for ER as well as IR (Gaudan et al., 2005; Hanisch et al., 2005; Schijvenaars et al., 2005).

Instead of focusing on this important problem, many methods have instead attempted to recognize whether a particular mention of a name refers to a gene or its protein product (Fukuda et al., 1998; Coller et al., 2000; Mika and Rost, 2004). However, this distinction is not always clear as, for example, “Cdc5-dependent Swe1 hyperphosphorylation” depends on both the Cdc5 protein but also the gene that encodes it. Indeed, human annotators only agree with each other in 77% of cases when asked to distinguish between genes, RNAs, and proteins (Tanabe et al., 2005). Fortunately, the ability to discriminate between genes and proteins is of little consequence for down-stream IE applications.



## Information extraction

In contrast to IR systems that identify texts concerning certain topics, IE systems aim to extract pre-defined types of facts, in particular relations between biological entities. From our example sequence, an IE system should deduce that i) Cdc28 binds Clb2, ii) Swe1 is phosphorylated by the Cdc28–Clb2 complex, and iii) Cdc5 is involved in Swe1 phosphorylation. These facts can subsequently be stored in a database, with the option of being verified by a curator reading the paper in question. Two fundamentally different approaches to extracting relations from biological texts are currently being used extensively, namely co-occurrence and natural language processing (NLP).

### Statistical co-occurrence methods

The simplest approach is to identify entities that co-occur within abstracts or sentences. As two entities might be mentioned together without being in any way related, most systems use a frequency-based scoring scheme to rank the extracted relations (Donaldson et al., 2003; Hoffmann and Valencia, 2004; Craven, 1999; Cooper and Kershenbaum, 2005; Ramani et al., 2005; Stephens et al., 2001; Blaschke and Valencia, 2002; Stapley and Benoit, 2000; Jenssen et al., 2001; Becker et al., 2003; Bowers et al., 2003; Chen and Sharp, 2004; von Mering et al., 2005; Schlitt et al., 2003; Wren and Garner, 2004; Alako et al., 2005; Maier et al., 2005; Tiffin et al., 2005). If two entities are repeatedly mentioned together, it is highly likely that they are somehow related, although the type of relation is not known (Jenssen et al., 2001; Stephens et al., 2001). Co-occurrence methods tend to give better recall but worse precision than NLP methods (Ding et al., 2002; Wren and Garner, 2004). Due to their ability to identify relations of almost any type, co-occurrence methods are well suited as parts of exploratory tools (Bowers et al., 2003; von Mering et al., 2005).

Co-occurrence methods can also be used to extract only relations of a certain type, such as physical protein–protein interactions, by combining them with a customized text categorization system to identify the relevant abstracts or sentences (Craven, 1999; Stephens et al., 2001; Blaschke and Valencia, 2002; Donaldson et al., 2003; Cooper and Kershenbaum, 2005; Ramani et al., 2005; Ray and Craven, 2005). This setup is particularly attractive for database curation as the custom-made text categorization system can also be used on its own, and because high coverage can be attained (Donaldson et al., 2003; Ramani et al., 2005). However, complex sentences that contain multiple relations give rise to additional, erroneous relations (our example sentence might link Cdc5 to Clb2). This approach is also unable to extract directional relations (is Cdc5 involved in Swe1 phosphorylation or vice versa) and has difficulty distinguishing between direct and indirect relations, for example whether or not Swe1

is directly phosphorylated by Cdc5).

## Natural language processing (NLP)

These issues can all be addressed by NLP methods that combine analysis of syntax and semantics. The text is first tokenized to identify sentence and word boundaries, and a part-of-speech tag (noun, verb, etc.) is assigned to each word. A syntax tree is then derived for each sentence to delineate noun phrases (for example, “Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homolog)”) and represent their interrelations. ER methods and simple dictionaries are subsequently used to semantically tag the relevant biological entities (genes, proteins, etc.) and other keywords (activation, repression, phosphorylation, etc.). Finally, a rule set is used to extract relations based on the syntax tree and the semantic labels. Very few NLP systems attempt to resolve anaphoric relations and most systems are thus unable to extract relations that span multiple sentences (Narayanaswamy et al., 2005). This is not as big a limitation as it might seem since most relations are in fact mentioned within a single sentence (Ding et al., 2002; Cooper and Kershenbaum, 2005).

Several programs exist for tokenization and part-of-speech tagging of English texts, most of which are easily adapted to biomedical texts by retraining them on a manually tagged corpus such as GENIA or PennBioIE (Saric et al., 2004a,b, 2006; Finkel et al., 2005). Semantic tagging is more complicated, but it can be greatly simplified by using existing ER methods. In contrast, development of grammars and extraction rules that can correctly parse the sentences and extract the facts remains challenging.

The idealized workflow described above suggests that syntactic parsing of the sentences and their semantic interpretation is performed as two separate steps (Rindfleisch et al., 2000; Proux et al., 2000; Yakushiji et al., 2001; Novichkova et al., 2003; Daraselia et al., 2004). However, most generic English parsers perform poorly if applied directly to biomedical texts due to the technical terminology, and particularly the use of long complex noun phrases. Better results can be obtained by first tagging the noun phrases (Yakushiji et al., 2001). However, many biomedical NLP systems have merged the syntactic parser and the semantic extraction rules in a customized partial parser that specifically targets only the relevant parts of sentences and directly extracts the facts (Blaschke et al., 1999; Friedman et al., 2001; Ono et al., 2001; Wong, 2001; Leroy and Chen, 2002; Pustejovsky et al., 2002; Gaizauskas et al., 2003; Koike and Takagi, 2004; Rzhetsky et al., 2004; Saric et al., 2004a,b, 2006; Temkin and Gilder, 2003). The major drawback of this approach is that a large number of extraction rules is needed to cover the many slightly different ways of expressing a certain on. These rules may be either developed manually (Blaschke et al., 1999; Friedman et al., 2001; Ono et al., 2001; Wong, 2001; Leroy and Chen, 2002; Gaizauskas et al., 2003; Huang

et al., 2004; Koike and Takagi, 2004; Rzhetsky et al., 2004; Temkin and Gilder, 2003; Saric et al., 2004a,b, 2006) or learned automatically from a corpus (Craven, 1999; Hao et al., 2005). Both approaches are very labor intensive as the latter requires the prior manual tagging of a large training corpus.

### An example NLP system

To give a more detailed idea of how an NLP-based IE system works, we here show by examples how relations are extracted by our own system (Saric et al., 2004a,b, 2006). The system is organized in cascaded modules where the output of one module is the input of the next module. The input text is first segmented into sentences and tokens using a tokenizer developed by Helmut Schmid. Each token is subsequently assigned a part-of-speech tag using TreeTagger, which correctly tagged 96.4% of tokens after being retrained on a corrected/revised version of the GENIA corpus (Saric et al., 2004a). Terms of particular interest (for example, *kinase* or *phosphorylates*) were subsequently assigned semantic tags based on a lookup table.

To be able to recognize gene/protein names as such, and to associate them with the appropriate database identifiers, a synonyms list was compiled from UniProt (Bairoch et al., 2005) and SGD (Christie et al., 2004). The name lists is expanded to include orthographic variants of each name and is then matched against the text. Noun-phrases containing one or more named entities are subsequently identified using finite state automata in the form of a CASS grammar (Abney, 1996). The following simplified example shows how we recognize and semantically categorize a complex, nested noun chunk:

```
[nx_expr
  [expr expression] [of of]
    [nx_geneprod
      [nx_gene
        [dt the] [nnpg argF] [gene gene]]
        [prod product]]]
```

Various type of relations between genes and proteins are subsequently extracted using separate grammar modules, which work on top of the entity recognition module just described. The following series of examples illustrates how the rules operate to extract the relations shown in Figure 2. All examples show a simplified bracketed structure illustrating the major principles of our rules; the internal structure is highly complex and derives from a pass through a number of cascading finite state transducers. Within the following examples the first line always indicates the type of relation

that we extract, which is either phosphorylation, dephosphorylation, or regulation of expression.

The first example shows a phosphorylation relation phrased in active voice. The participating proteins are shown in bold-faced letters, the relational word is underlined, and the selective negation is also marked by the *negation*-bracket. The NLP system correctly extracts that **Lyn** phosphorylates **CrkL** from the following example:

[*phosphorylation\_active*  
**Lyn**, [*negation* but not **Jak2** ]  
phosphorylates  
**CrkL** ]

This active-voice phosphorylation construct below is detected through the relational noun *phosphorylation* as argument of *participates*. The *phosphorylation* bracket is triggered through the key word *phosphorylation*, enabling the system to extract that

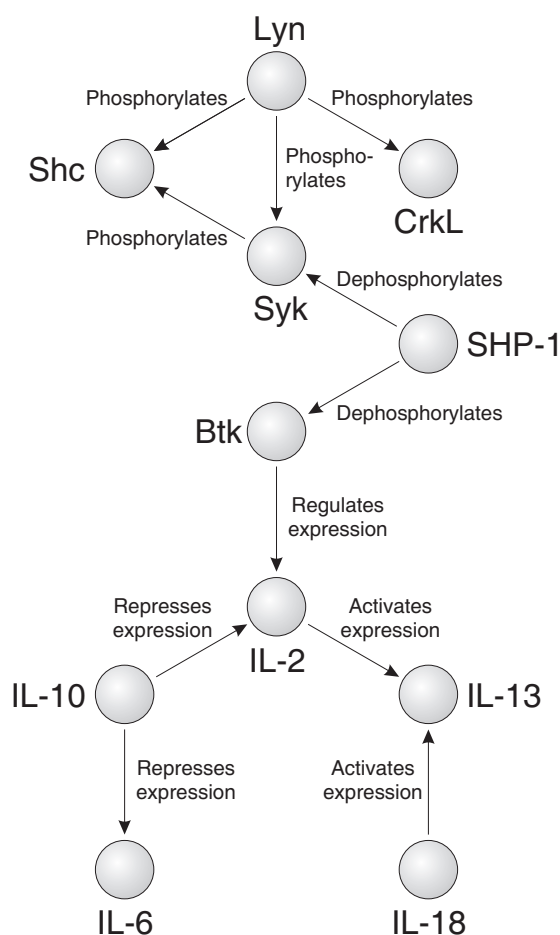


Figure 2: **An example network for mouse proteins.** The network exemplifies the multiple types of relations extracted by our rule based approach; the phrases from which these relations were extracted are discussed in the main text.

**Lyn** phosphorylates **syk**:

[*phosphorylation\_active*

**Lyn**

also participates in

[*phosphorylation* the tyrosine *phosphorylation*  
and *activation* of **syk** ]]

The following two examples illustrate nominalization for phosphorylation. The arguments are attached through the *of* and *by* prepositional phrases, where the latter identifies the agent role:

[*phosphorylation\_nominal*

the phosphorylation of

the adapter *protein* **SHC**

by the Src-related *kinase* **Lyn** ]

[*phosphorylation\_nominal*

phosphorylation of **Shc** by

the hematopoietic cell-specific

tyrosine *kinase* **Syk** ]

The system is also able to identify dephosphorylation relations, as exemplified by the following nominalisation example, from which we extract that both **Syk** and **Btk** are dephosphorylated by **SHP-1**:

[*dephosphorylation\_nominal*

Dephosphorylation of

**Syk** and **Btk**

mediated by

**SHP-1** ]

The following examples shows gene expression relations. The first of these illustrates the ability of our system to deal with passive voice. Based on the verb (“induce”) and the relational noun (“expression”) we conclude that **IL-2** and **IL-18** activate expression of **IL-13**:

```
[expression_activation_passive
  [expression IL-13 expression ]
  induced by
  IL-2 + IL-18 ]
```

Repression of gene expression relation be the next example, where one protein (**IL-10**) represses the expression of two other genes (**IL-2** and **IL-6**):

```
[expression_repression_active
  IL-10
  also decreased
  [expression mRNA expression of
    IL-2 and IL-6 cytokine receptors ]]
```

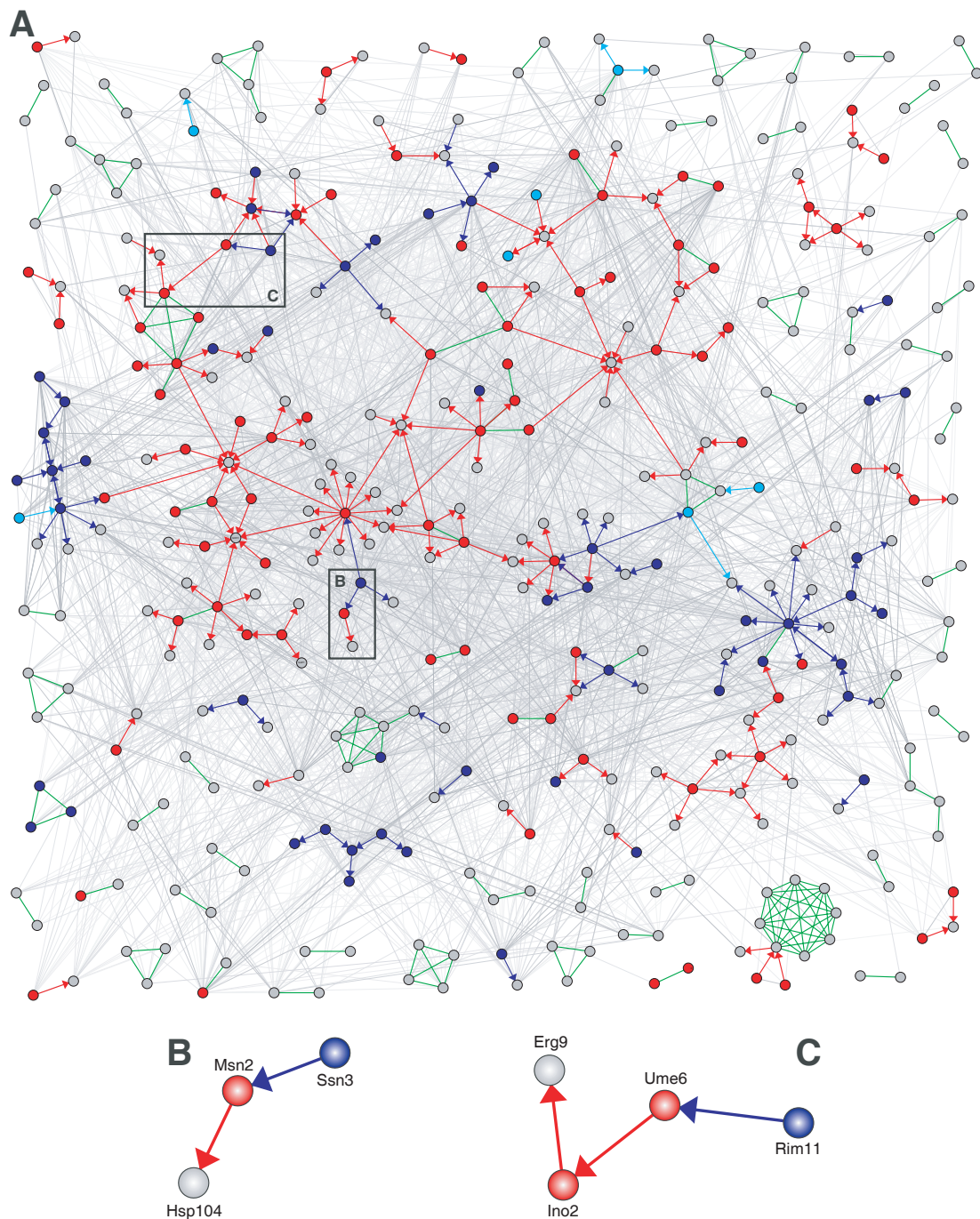
In the final example, it is only possible to extract that **Btk** regulates the expression of the **IL-2** gene, not whether it activates or represses it:

```
[expression_regulation_active
  Btk
  regulates
  [expression the transcription of
    the IL-2 gene ]]
```

## Applications of IE

Most studies so far have focused on extracting very few types of relations. These include physical protein–protein interactions (Blaschke et al., 1999; Thomas et al., 2000; Friedman et al., 2001; Ono et al., 2001; Stephens et al., 2001; Yakushiji et al., 2001; Donaldson et al., 2003; Temkin and Gilder, 2003; Novichkova et al., 2003; Daraselia et al., 2004; Huang et al., 2004; Rzhetsky et al., 2004; Cooper and Kershenbaum, 2005; Hao et al., 2005; Ramani et al., 2005) and interactions that involve unspecified molecular mechanisms among proteins (Sekimizu et al., 1998; Blaschke et al., 1999; Proux et al., 2000; Stapley and Benoit, 2000; Friedman et al., 2001; Jenssen et al., 2001; Stephens et al., 2001; Yakushiji et al., 2001; Blaschke and Valencia, 2002; Pustejovsky et al., 2002; Novichkova et al., 2003; Schlitt et al., 2003; Bowers et al., 2003; Chen and Sharp, 2004; Chiang et al., 2004; Daraselia et al., 2004; Hoffmann and Valencia, 2004; Koike and Takagi, 2004; Rzhetsky et al., 2004; Domedel-Puig and Wernisch, 2005; von Mering et al., 2005). Relations have also been extracted for concepts such as disease names, Gene Ontology terms or nouns in general (Craven,





**Figure 3: A literature derived network for yeast.** A) The complete yeast network. The protein network was derived from MEDLINE using both a statistical co-occurrence method (von Mering et al., 2005) and an NLP-based one (Saric et al., 2004a,b, 2006). Functional associations derived from co-occurrence are shown in shades of gray according to the level of confidence. The NLP method extracts four types of relations: stable physical interactions (green), regulation of expression (red), phosphorylation (dark blue), and dephosphorylation (light blue). The proteins (circles) are colored according to their functional annotation: (co-)regulators of expression (red), kinases and cyclins (dark blue), phosphatases (light blue), and other proteins (gray). A version of this figure that includes all protein names is available as supplementary information. B+C) Examples of unpublished relations that can be inferred from the network. From the network we can infer that Ssn3 likely influences Hsp104 expression through phosphorylation of Msn2, that Ume6 likely regulates Erg9 expression, and that Rim11 regulates the expression of both Ino2 and Erg9. None of these hypotheses have been tested experimentally.

1999; Humphreys et al., 2000; Rindflesch et al., 2000; Hahn et al., 2002; Leroy and Chen, 2002; Raychaudhuri et al., 2002a; Becker et al., 2003; Chen and Sharp, 2004; Wren and Garner, 2004; Alako et al., 2005; Bajdik et al., 2005; Couto et al., 2005; Ehrler et al., 2005; Krallinger et al., 2005; Maier et al., 2005; Ray and Craven, 2005; Rice et al., 2005; Tiffin et al., 2005; Verspoor et al., 2005). Recently, NLP methods have been developed for extracting information on gene regulation (Saric et al., 2004a,b, 2006), protein phosphorylation (Friedman et al., 2001; Rzhetsky et al., 2004; Hu et al., 2005; Narayanaswamy et al., 2005; Saric et al., 2006), and tissue specificity of alternative transcripts (Shah et al., 2005). Probably because of the inherent complexity of the task, only a few systems have been designed that are able to extract multiple types of relations (Friedman et al., 2001; Novichkova et al., 2003; Daraselia et al., 2004; Rzhetsky et al., 2004; Saric et al., 2006).

Using the NLP-based system described in the previous section, all the relations mentioned in our example sentence can be correctly extracted (Saric et al., 2006). To illustrate how IE can be used at a larger scale, we have applied this method to all MEDLINE abstracts, extracting more than 5000 binary relations (which may each be mentioned multiple times) of which 370 are among yeast proteins. These are shown as a network in Figure 3A along with the interactions identified by co-occurrence (von Mering et al., 2005). The latter method identifies almost 3000 interactions among these proteins, however, only 150 are of comparable reliability to those obtained by NLP. With the growing interest in systems biology, IE will likely become a mainstream tool for biologists in the near future, as it is one of the only ways to identify diverse types of relations on a large scale.



## Text mining

Often used as a catch-all term for computational text analysis, *text mining* is more strictly defined as “the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources” (M. Hearst, personal communication; see also Ref. (Hearst, 1999)). IE methods do thus not qualify as text mining tools themselves since they can only extract what has already been published; rather, they form the basis for text mining much like ER does for IE (Hearst, 1999).

### Mining text for overlooked “golden nuggets”

It may at first seem impossible to have a computer make discoveries based on literature alone; afterall, IE is only able to extract the facts that have already been published. The trick is to use facts extracted from several different publications (A leads to B, B leads to C) to infer new, indirect relations (A leads to C). Since the literature is so vast that each researcher can only read a small subset, it may well be that no person is aware of all the facts required to make this logical inference. This is plausible especially if the facts were published within two disconnected research areas (Swanson, 1986b,a; Hearst, 1999) or if an overwhelming number of papers is published on a single topic (Blagosklonny and Pardee, 2002).

For almost two decades, Don Swanson has argued along these lines and used a simple semi-automated method (ARROWSMITH (Smalheiser and Swanson, 1998; Swanson and Smalheiser, 1999)) to infer the following novel relations: fish oil can help patients suffering from Reynaud’s disease (Swanson, 1986a), magnesium deficiency plays a role in migraine headache (Swanson, 1988b), arginine intake has an effect on somatomedin C blood levels (Swanson, 1990), and that estrogen protects against Alzheimer’s disease (Smalheiser and Swanson, 1996). These predictions have since then been re-examined by others (Gordon and Lindsay, 1996; Lindsay and Gordon, 1999) and the two first have both been experimentally confirmed (Swanson, 1988a; Smalheiser and Swanson, 1994). However, these early predictions were all made using a “closed” framework where the user provides the hypothesis (A is related to C), which is then tested by a computational search for shared, related words (B) that could support the hypothesis; it can thus be argued that the computer did not actually make the discovery.

The corresponding “open” discovery problem is more challenging, but also potentially more rewarding, as one starts from only a single entity (A, for example a disease) and attempts to find indirect, undiscovered relations to other entities (C, for example chemicals or genes). Several different methods exist that all rely on the same strat-

egy: i) identify the terms B that co-occur with A, ii) identify the terms C that co-occur with B but not with A (Weeber et al., 2000; Hristovski et al., 2001; Srinivasan and Libbus, 2004; Wren, 2004; Wren et al., 2004; Hristovski et al., 2005). More recently, alternative methods based on latent semantic indexing (Homayouni et al., 2005) or cross-subspace analysis (Matsunaga and Muramatsu, 2005) have been proposed. The major problem with all of these approaches is that inferences are made from undirected relations of unknown type, for which reason causality cannot be taken for granted. For example, many Cdc28 cooccurs with many of its substrates in MEDLINE abstracts, which would cause most existing methods to propose novel but incorrect relations between unrelated Cdc28 substrates.

To our knowledge, no published studies have made use of NLP-based IE as the basis for text mining, although this could ensure that the novel relations are inferred from causal chains of relations. A likely reason is that very few NLP systems are able to accurately extract a sufficiently large number of directed relations to enable this approach. By using the yeast network of phosphorylation and gene expression that we derived using IE (Figure 3A) to indirectly link 64 pairs of proteins that do not co-occur in MEDLINE abstracts, we here show the feasibility of using NLP-based text mining to discover novel relations. Manual inspection of the literature suggests that over 90% of the inferred relations are correct. For example, the network suggests that the cyclin-dependent kinase Ssn3 (also known as Srb10) influences expression of the stress response protein Hsp104 through phosphorylation of Msn2 (Figure 3B). It is known that Hsp104 expression is activated by the zinc finger protein Msn2 (Grably et al., 2002) and that Msn2 is phosphorylated by Ssn3 (Chi et al., 2001). Ssn3 was recently shown to be a repressor of general stress response, however, it remains controversial if and how this is mediated by Msn2 phosphorylation (Figure 3B) (Bose et al., 2005; Lenssen et al., 2005). It is thus likely that Ssn3 regulates Hsp104 expression, although it has not been experimentally verified. Similarly, it is known that Rim11 phosphorylates Ume6 (Xiao and Mitchell, 2000) that regulates the expression of another transcription factor, Ino2 (Eiznhamer et al., 2001), which in turn regulates Erg9 expression (Kennedy et al., 1999). It can thus be inferred that Ume6 likely regulates Erg9 expression and that Rim11 regulates the expression of both Ino2 and Erg9. Remarkably, however, neither of these relations appear to have been described in the published literature (Figure 3C).

While correct, the vast majority of the inferred relations in our study of yeast interactions turn out to be well known, despite the proteins never having been mentioned together in any abstract. Without full text access to all published papers, it is unfortunately impossible to rule out that an inferred relation has already been published. Also, some relations are likely considered to be so trivial that no one ever published them. To avoid flooding the user with trivial hypotheses, text mining methods need

to integrate other data sources than the text itself, in particular databases of curated knowledge.

### **Discovery of global correlations from literature**

An established data mining methodology, which has not previously been utilized in text mining, is to search for correlated events as exemplified by Amazon's "Customers who bought this item also bought ..." function. In the field of biology, this can be used to discover fundamental properties of, for example, regulatory networks.

To test the feasibility of this approach, we compared the lists of yeast proteins shown in Figure 3 that are regulated through expression and those that are regulated through phosphorylation. The overlap between the two sets is over four-fold larger than expected by chance ( $P < 5 \cdot 10^{-4}$ ), suggesting that phosphorylation and regulation of expression tend to target the same proteins, as was recently proposed by de Lichtenberg et al. through integration of several large-scale experimental data sets (de Lichtenberg et al., 2005). Similarly, data mining of the relations in Figure 3 reveals that protein kinases preferentially phosphorylate each other ( $P < 9 \cdot 10^{-9}$ ) and that transcription factors regulate the expression of each other ( $P < 2 \cdot 10^{-7}$ ), reflecting the existence of signaling cascades and transcriptional networks, respectively.

The individual pieces of information required for making other such discoveries are likely to be present in the literature, and could be combined using a similar systematic, computational method. A drawback of this methodology is that statistically significant correlations can arise easily due to study biases; however, this can be overcome by correlating IE results with genome-wide data sets. We believe that the latter type of data mining will likely play an important role in unveiling systems-level properties.

## Text/data integration

Although text mining can be used to uncover hitherto overlooked relations, data mining approaches that integrate literature with other data types have much greater potential for making biological discoveries. An illustrative example of how this could be achieved is to use sequence similarity searches to transfer the relations extracted from text to orthologous proteins (Yandell and Majoros, 2002). Text mining methods can then be used to make inferences based on relations from multiple model organisms, and hence bridge communities of researchers who work on different model organisms. To test this approach, we combined the fruit fly and mouse equivalents of Figure 3 (Saric et al., 2004a,b, 2006) using orthology assignments from the STRING database (von Mering et al., 2005), whereby we discovered the following indirect relation. In fruit fly, Suppressor of Hairless (Su(H)) has been shown to be a direct transcriptional repressor of *single-minded* (Morel and Schweisguth, 2000). Since the mouse Single-Minded 1 protein is a transcriptional activator of EPO (Woods and Witelaw, 2002), we make the hypothesis that one or more of the murine Su(H) orthologs down-regulate EPO expression, although none of them co-occur with EPO in MEDLINE abstracts. The power of such approaches will only improve with the growth of both the literature and the availability of large-scale dataset.

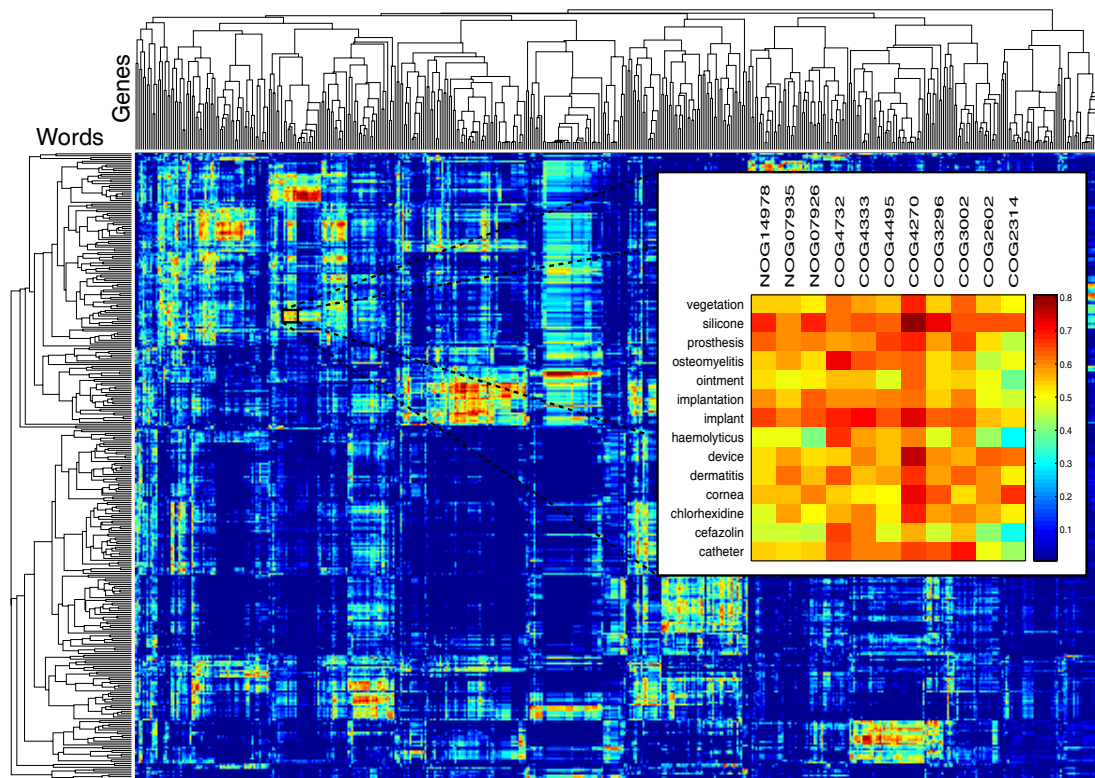
Very early on, researchers attempted to augment sequence similarity searches with literature mining in order to improve the detection of homologous proteins (Liu and Rost, 2000; Chang et al., 2001); despite fairly promising results, however, this methodology never really took off. The reason for this is likely that these methods fail to deliver novel results since homologies that are also supported by literature are precisely the ones that are already known. Recently, literature mining has instead been proposed as a means to help researchers get an overview of the results of a sequence similarity search (Dieterich et al., 2005).

Today, most attempts to integrate literature and biological data are instead directed towards the annotation of data obtained from functional genomics studies as manual in-depth analysis is not feasible due to the amount of data (Andrade and Valencia, 1998; Shatkay et al., 2000; Blaschke et al., 2001; Jenssen et al., 2001; Masys et al., 2001; Masys, 2001; Chaussabel and Sher, 2002; Raychaudhuri et al., 2002b; Raychaudhuri and Altman, 2003; Raychaudhuri et al., 2003; Glenisson et al., 2004; Djebbari et al., 2005). Most approaches use ER methods or database cross-references to first retrieve the MEDLINE abstracts that are associated with one or more genes, for example a protein family or a cluster of genes that are co-expressed in a microarray experiment. These abstracts are subsequently used either i) to identify significant overrepresentation of keywords within the text (Andrade and Valencia, 1998; Shatkay et al., 2000; Blaschke et al., 2001) or of annotated GO/MeSH terms that characterize the genes

in question (Masys, 2001; Masys et al., 2001; Chaussabel and Sher, 2002; Djebbari et al., 2005; ?), ii) to evaluate the cluster coherence (a measure of functional similarity for a group of genes) (Raychaudhuri et al., 2002b; Raychaudhuri and Altman, 2003; Raychaudhuri et al., 2003; Glenisson et al., 2004), or iii) to construct a functional association network of the genes based on either co-occurrence (Jenssen et al., 2001; Schlitt et al., 2003) or document similarity (Shatkay et al., 2000). Again, although these methods may be useful tools, they have little to offer in terms of making biological discoveries.

Through their ability to bring together many different types of data, networks have the potential to form the basis for text and data integration. Several web-based tools exist that provide access to protein networks based on both IE and high-throughput experiments (Bowers et al., 2003; Hoffmann and Valencia, 2004; von Mering et al., 2005), which have proven valuable both as exploratory tools and as a basis onto which, for example, expression data can be mapped to visualize how the synthesis of protein complexes is regulated at the transcript level (de Lichtenberg et al., 2005). Such networks can also be combined with other types of data to provide insight into the molecular basis of a disease. For example, literature-based protein networks have been integrated with genetic linkage to identify candidate genes for Alzheimer's disease from within a region, based on their interactions with genes that are already known to have a causal role in the disease (Iossifov et al., 2004; Krauthammer et al., 2004).

The types of networks described above only consider relations at the molecular level; however, the possibility of making discoveries is greatly improved by integrating relations at multiple levels. This is exemplified by several literature mining tools used to prioritize candidate genes with potential roles in inherited diseases for further study. The first such system, G2D, was published in 2002 (Perez-Iratxeta et al., 2002). It combines the MeSH annotation in MEDLINE with the Gene Ontology annotation in RefSeq entries to infer logical chains of connections from disease names, via chemicals and drugs, to molecular functions. Combined with functional annotation inferred from sequence similarity, this allows the genes within a mapped region to be ranked based on their association score with the disease in question. The BITOLA method instead relies purely on text mining to find candidate genes that are indirectly linked to a given disease, and subsequently filters these based on chromosomal mapping data on the disease (Hristovski et al., 2005). A third approach identifies co-occurring disease and tissue names in MEDLINE and combines this with tissue expression annotation from Ensembl to link the tissues to candidate disease genes (Tiffin et al., 2005). While the original G2D method was limited to Mendelian diseases, these approaches have recently been shown to also work for complex genetic diseases (Perez-Iratxeta et al., 2005; Tiffin et al., 2005).



**Figure 4: Correlating phenotypes with genotypes.** Integration of gene occurrence in genomes and keywords that are overrepresented in the literature in association with certain species (Korbel et al., 2005). The two trees show the individual clustering of species profiles for genes and keywords, respectively. The association scores between genes and keywords is visualized as a heat map. The insert shows a cluster that contains 11 groups of orthologous genes with unknown function that are only present in Staphylococci and certain other hospital bacteria. All these genes are strongly associated with words that preferably occur in abstracts on those species such as osteomyelitis (a disease related to Staphylococci), cornea (a part of the eye that can be infected by Staphylococci), Cefazolin (an antibiotic often used against Staphylococci), and Chlorhexidine (a disinfectant against which Staphylococci are resistant). As both genes and words seem associated with this species subset, the genes are likely to be directly or indirectly associated to the corresponding phenotypes. The genes might be directly involved in disease phenotypes or might only indirectly be involved by contributing to the lifestyle. In any case, the specificity of these genes to a limited set of infectious bacteria makes them candidates as drug targets.

Even broader in scope is a recent study that correlates text mining for phenotypic information with gene occurrences across species (genotype information) to infer phenotypic roles for genes of unknown function (Korbel et al., 2005). MEDLINE was first systematically searched for keywords associated with each prokaryote for which the genome has been sequenced. The resulting species distributions of keywords were then matched against the species distributions of genes in order to associate keywords to genes (Figure 4). The set of keywords associated with a group of genes can reveal the phenotypic characteristics caused by these genes. For example, genes unique to Staphylococci and other hospital bacteria clusters together with descriptive keywords such as “osteomyelitis” (a disease related to Staphylococci) and less



obvious ones like “Chlorhexidine” (a disinfectant against which *Staphylococci* are resistant) (Figure 4). This suggests putative roles for these genes of unknown function and highlights them possible drug targets. When applied globally, the approach recaptured many known genotype–phenotype relations and also predicted several novel ones, such as enzymes involved in plant degradation and genomic determinants for food poisoning (Korbel et al., 2005).

## Outlook

The peer reviewed scientific literature will continue to be a prime resource for accessing the worldwide scientific knowledge and its ongoing growth and diversification will require tremendous systematic and automated efforts to utilize the information therein. In the near future, tools for mining this knowledge base will likely play a pivotal role in systems biology. So far, more than 90% of all biomedical literature mining has been based on MEDLINE, mainly because it is freely available in a convenient format. To realize the full potential, future methods will need to work on full text papers, including context such as the citation network. This will require some methodological improvements as not all sections of a paper are equally relevant (Shah et al., 2003; Schuemie et al., 2004) and because some information must be extracted from figures and tables. However, it is the restricted access to full text papers and citation information, not the technology, that is currently the biggest limitation despite encouraging open access initiatives like PubMed Central and Highwire Press (Yandell and Majoros, 2002; Dickman, 2005).

Bridging the gap between biologists and computational linguists will be crucial to the success of biomedical literature mining in general its integration with high-throughput experimental data in particular. The field is currently dominated by researchers with a computational background, however, only biologists possess the knowledge required to properly evaluate methods, to identify specific tasks for which tools are needed, and to point out other data sources that would be valuable to integrate with literature. To bring more biologists into the field, tool developers need to focus more on designing user interfaces that make the tools accessible to non-specialists. Finally, both sides need to contribute to the diversity and novelty within this field, where too many researchers currently use the same few methods to solve the same few tasks. We hope that this review will make more biologists aware of the importance of literature mining, and that it will inspire the development of new tools for making the most of the growing bodies of both scientific literature and experimental data.

## Acknowledgments

The author would like to thank Jasmin Šarić, Rossitza Ouzounova, Jan Korbel, Tobias Doerks, Sean Hooper, Isabel Rojas, and Peer Bork, who were all involved in work presented in this tutorial as well as other group members at EMBL and EML Research for valuable discussions. This work was supported by grants from the German Ministry for Education and Science and by the BioSapiens Network of Excellence, contract number LSHG-CT-2003-503265, funded by the European Commission FP6 Programme.



## Online tools and resources

### Web-based applications

#### Information retrieval (IR)

E-BioSci (<http://www.e-biosci.org>)  
EBIMed (<http://www.ebi.ac.uk/Rebholz-srv/ebimed/>)  
GeneInfoMiner (<http://brainarray.mbni.med.umich.edu/GIM.asp>)  
Google Scholar (<http://scholar.google.com>)  
GoPubMed (<http://www.gopubmed.org>)  
MedMiner (<http://discover.nci.nih.gov/textmining/>)  
PubMed (<http://www.pubmed.org>)  
PubFinder (<http://www.glycosciences.de/tools/PubFinder/>)  
Textpresso (<http://www.textpresso.org>)  
XplorMed (<http://www.ogic.ca/projects/xplormed/>)

#### Entity recognition (ER)

iHOP (<http://www.pdg.cnb.uam.es/UniPub/iHOP/>)  
Whatizit (<http://www.ebi.ac.uk/Rebholz-srv/whatizit/>)

#### Information extraction (IE)

BioIE (<http://umber.sbs.man.ac.uk/dbbrowser/bioie/>)  
iProLINK (<http://pir.georgetown.edu/iprolink/>)  
JournalMine (<http://textmine.cu-genome.org>)  
MedLEE (<http://lucid.cpmc.columbia.edu/medlee/>)  
PreBIND (<http://prebind.bind.ca>)  
Protein Corral (<http://www.ebi.ac.uk/Rebholz-srv/pcorral/>)  
PubGene (<http://www.pubgene.org>)

#### Text mining

ARROWSMITH (<http://arrowsmith.psych.uic.edu>)

#### Integration BITOLA (<http://www.mf.uni-lj.si/bitola/>)

G2D (<http://www.ogic.ca/projects/g2d2/>)  
ProLinks (<http://dip.doe-mbi.ucla.edu/pronav/>)  
STRING (<http://string.embl.de>)

## Text collections

### Full text corpora

HighWire Press (<http://highwire.stanford.edu>)

PubMed Central (<http://www.pubmedcentral.org>)

### Tagged corpora

FetchProt (<http://fetchprot.sics.se>)

GENETAG (<ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe/>)

GENIA (<http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/>)

PennBioIE (<http://bioie ldc.upenn.edu>)

Yapex (<http://www.sics.se/humle/projects/prothalt/>)

## IE modules

### Entity taggers

ABNER (<http://www.cs.wisc.edu/~bsettles/abner/>)

GAPSCORE (<http://bionlp.stanford.edu/gapcore/>)

### Part-of-speech taggers

Brill Tagger (<http://www.cs.jhu.edu/~brill/>)

TNT Tagger (<http://www.coli.uni-saarland.de/~thorsten/tnt/>)

TreeTagger (<http://www.ims.uni-stuttgart.de/~schmid/>)

### Parsers

CASS (<http://www.vinartus.net/spa/>)

Collins Parser (<http://people.csail.mit.edu/mcollins/>)

LTG Software (<http://www.ltg.ed.ac.uk/software>)

SNoW (<http://l2r.cs.uiuc.edu/~cogcomp/software.php>)

Stanford Parser (<http://nlp.stanford.edu/software/>)

## References

- Abney, S. (1996). Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop*, pages 8–15, Prague, Czech Republic.
- Alako, B. T., Veldhoven, A., van Baal, S., Jelier, R., Verhoeven, S., Rullmann, T., Polman, J., and Jenster, G. (2005). CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics*, 6:51.
- Andrade, M. A. and Bork, P. (2000). Automated extraction of information in molecular biology. *FEBS Letters*, 476:12–17.
- Andrade, M. A. and Valencia, A. (1998). Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, 14:600–607.
- Aronson, A. R., Demner, D., Humphrey, S. M., Ide, N. C., Kim, W., Liu, H., Loane, R. R., Mork, J. G., Smith, L. H., Tanabe, L. K., Wilbur, W. J., and Xie, N. (2004). Knowledge-intensive and statistical approaches to the retrieval and annotation of genomics MEDLINE citations. In *Proceedings of TREC 2004*, volume 13.
- Asano, S., Park, J. E., Sakchaisri, K., Yu, L. R., Song, S., Supavilai, P., Veenstra, T. D., and Lee, K. S. (2005). Concerted mechanism of swe1/wee1 regulation by multiple kinases in budding yeast. *EMBO J.*, 24:2194–2204.
- Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L. S. (2005). The universal protein resource (UniProt). *Nucleic Acids Res.*, 33:D154–D159.
- Bajdik, C. D., Kuo, B., Rusaw, S., Jones, S., and Brooks-Wilson, A. (2005). CGMIM: Automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and candidate genes. *BMC Bioinformatics*, 6:78.
- Becker, K. G., Hosack, D. A., Dennis, G., J., Lempicki, R. A., Bright, T. J., Cheadle, C., and Engel, J. (2003). PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics*, 4:61.
- Bhalotia, G., Nakov, P. I., Schwartz, A. S., and Hearst, M. A. (2003). BioText team report for the TREC 2003 genomics track. In *Proceedings of TREC 2003*, volume 12.
- Blagosklonny, M. V. and Pardee, A. B. (2002). Unearthing the gems. *Nature*, 416:373.
- Blaschke, C., Andrade, M. A., Ouzounis, C., and Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein–protein interactions. In

- Proc. Int. Conf. Intell. Syst. Mol. Biol.*, volume 7, pages 60–67, Menlo Park, CA. AAAI Press.
- Blaschke, C., Oliveros, J. C., and Valencia, A. (2001). Mining functional information associated with expression arrays. *Funct. Integr. Genomics*, 1:256–268.
- Blaschke, C. and Valencia, A. (2002). The frame-based module of the SUISEKI information extraction system. *IEEE Intelligent Systems*, 17:14–20.
- Bose, S., Dutko, J. A., and Zitomer, R. S. (2005). Genetic factors that regulate the attenuation of the general stress response of yeast. *Genetics*, 169:1215–1226.
- Bowers, P. M., Pellegrini, M., Thompson, M. J., Fierro, J., Yeates, T. O., and Eisenberg, D. (2003). Prolinks: a database of protein functional linkages derived from coevolution. *Nucleic Acids Res.*, 5:R35.
- Bussey, K. J., Kane, D., Sunshine, M., Narasimhan, S., Nishizuka, S., Reinhold, W. C., Zeeberg, B., Ajay, W., and Weinstein, J. N. (2003). MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biology*, 4:R27.
- Büttcher, S., Clarke, C. L. A., and Cormack, G. V. (2004). Domain-specific synonym expansion and validation for biomedical information retrieval. In *Proceedings of TREC 2004*, volume 13.
- Chang, J. T., Raychaudhuri, S., and Altman, R. B. (2001). Including biological literature improves homology search. In *Pac. Symp. Biocomput.*, volume 6, pages 374–383, Hawaii. World Scientific.
- Chang, J. T., Schutze, H., and Altman, R. B. (2004). GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics*, 20:216–225.
- Chaussabel, D. and Sher, A. (2002). Mining microarray expression data by literature profiling. *Genome Biol.*, 3:RESEARCH0055.
- Chen, H. and Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, 5:147.
- Chen, L., Liu, H., and Friedman, C. (2005). Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21:248–256.
- Chi, Y., Huddleston, M. J., Zhang, X., Young, R. A., Annan, R. S., Carr, S. A., and Deshaies, R. J. (2001). Negative regulation of Gcn4 and Msn2 transcription factors by Srb10 cyclin-dependent kinase. *Genes Dev.*, 15:1078–1092.
- Chiang, J.-H., Yu, H.-C., and Hsu, H.-J. (2004). GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics*, 20:120–121.

- Christie, K. R., Weng, S., Balakrishnan, R., Costanzo, M. C., Dolinski, K., Dwight, S. S., Engel, S. R., Feierbach, B., Fisk, D. G., Hirschman, J. E., Hong, E. L., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C. L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Botstein, D., and Cherry, J. M. (2004). Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.*, 32:D311–D314.
- Cimiano, P., Reyle, U., and Saric, J. (2006). Status of text-mining techniques applied to biomedical text. *Drug Discovery Today*, 11:315–325.
- Coller, N., Nobata, C., and Tsujii, J. (2000). Extracting the names of genes and gene products with a hidden Markov model. In *Int. Conf. Comput. Linguistics*, volume 18, pages 201–207.
- Collier, N., Nobata, C., and Tsujii, J. (2002). Automatic acquisition and classification of terminology using a tagged corpus in the molecular biology domain. *Terminology*, 7:239–257.
- Colosimo, M. E., Morgan, A. A., Yeh, A. S., Colombe, J. B., and Hirschman, L. (2005). Data preparation and interannotator agreement: BioCreAtIvE Task 1B. *BMC Bioinformatics*, 6:S12.
- Cooper, J. W. and Kershenbaum, A. (2005). Discovery of protein–protein interactions using a combination of linguistic, statistical and graphical information. *BMC Bioinformatics*, 6:143.
- Couto, F. M., Silva, M. J., and Coutino, P. M. (2005). Finding genomic ontology terms in text using evidence content. *BMC Bioinformatics*, 6:S21.
- Craven, M. and Kumlien, J. (1999). Constructing biological knowledge bases by extracting information from text sources. In *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, volume 7, pages 77–86, Menlo Park, CA. AAAI Press.
- Crim, J., McDonald, R., and Pereira, F. (2005). Automatically annotating documents with normalized gene lists. *BMC Bioinformatics*, 6:S13.
- Daraselia, N., Yuruev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics*, 20:604–611.
- de Lichtenberg, U., Jensen, L. J., Brunak, S., and Bork, P. (2005). Dynamic complex formation during the yeast cell cycle. *Science*, 307:724–727.

- Dickman, S. (2005). Tough mining. *PLoS Biology*, 1:144–147.
- Dieterich, G., Kärst, U., Wehland, J., and Jänsch, L. (2005). MineBlast: a literature presentation service supporting protein annotation by data mining of BLAST results. *Bioinformatics*, 21:3450–3451.
- Ding, J., Berleant, d., Nettleton, D., and Wurtelle, E. (2002). Mining Medline: Abstracts, sentences, or phrases? In *Pac. Symp. Biocomput.*, volume 7, pages 326–337, Hawaii. World Scientific.
- Djebbari, A., Karamycheva, S., Howe, E., and Quackenbush, J. (2005). MeSHer: identifying biological concepts in microarray assays based on PubMed references and MeSH terms. *Bioinformatics*, 21:3324–3326.
- Domedel-Puig, N. and Wernisch, L. (2005). Applying GIFT, a Gene Interactions Finder in Text, to fly literature. *Bioinformatics*, 21.
- Doms, A. and Schroeder, M. (2005). GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res.*, 33:W783–W786.
- Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G. D., Michalickova, K., Pawson, T., and Hogue, C. W. V. (2003). PreBIND and Textomy—mining the biomedical literature for protein–protein interactions using a support vector machine. *BMC Bioinformatics*, 4:art. 11.
- Ehrler, F., Geissbühler, A., Jimeno, A., and Ruch, P. (2005). Data-poor categorization and passage retrieval for gene ontology annotation in swiss-prot. *BMC Bioinformatics*, 6:S23.
- Eiznhamer, D. A., Ashburner, B. P., Jackson, J. C., Gardenour, K. R., and Lopes, J. M. (2001). Expression of the INO2 regulatory gene of *Saccharomyces cerevisiae* is controlled by positive and negative promoter elements and an upstream open reading frame. *Mol. Microbiol.*, 39:1395–1405.
- Finkel, J., Dingare, S., Manning, C. D., Nissim, M., Alex, B., and Grover, C. (2005). Exploring the boundaries: gene and protein identification in biomedical text. *BMC Bioinformatics*, 6:S5.
- Franzen, K., Eriksson, G., Olsson, F., Asker, L., Liden, P., and Coster, J. (2002). Protein names and how to find them. *Int. J. Med. Inform.*, 67:49–61.
- Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17 Suppl. 1:S74–S82.

- Fukuda, K., Tamura, A., Tsunoda, T., and Takagi, T. (1998). Toward information extraction: identifying protein names from biological papers. In *Pac. Symp. Biocomput.*, volume 3, pages 707–718, Hawaii. World Scientific.
- Fundel, K., Güttler, D., Zimmer, R., and Apostolakis, J. (2005). A simple approach for protein name identification: prospects and limits. *BMC Bioinformatics*, 6:S15.
- Gaizauskas, R. J., Demetriou, G., Artymiuk, P. J., and Willett, P. (2003). Protein structures and information extraction from biological texts: The PASTA system. *Bioinformatics*, 19:135–143.
- Gaudan, S., Kirsch, H., and Rebholz-Schuhmann, D. (2005). Resolving abbreviations to their senses in Medline. *Bioinformatics*, 21.
- Glenisson, P., Antal, P., Mathys, J., Moreau, Y., and De Moor, B. (2003). Evaluation of the vector space representation in text-based gene clustering. In *Pac. Symp. Biocomput.*, volume 8, pages 391–402, Hawaii. World Scientific.
- Glenisson, P., Coessens, B., Van Vooren, S., Mathys, J., Moreau, Y., and De Moor, B. (2004). TXTGate: profiling gene groups with text-based information. *Genome Biol.*, 5:R43.
- Goetz, T. and von der Lieth, C.-W. (2005). PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts. *Nucleic Acids Res.*, 33:W774–W778.
- Gordon, M. D. and Lindsay, R. K. (1996). Toward discovery support systems: A replication, re-examination, and extension of Swanson’s work on literature-based discovery of a connection between Raynaud’s and fish oil. *J. Am. Soc. Inf. Sci.*, 47:116–128.
- Grably, M. R., Stanhill, A., Tell, O., and Engelberg, D. (2002). HSF and Msn2/4p can exclusively or cooperatively activate the yeast HSP104 gene. *Mol. Microbiol.*, 44:21–35.
- Hahn, U., Romacker, M., and Schulz, S. (2002). Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. In *Pac. Symp. Biocomput.*, volume 7, pages 338–349, Hawaii. World Scientific.
- Hakenberg, J., Bickel, S., Plake, C., Brefeld, U., Zahn, H., Faulstich, L., Leser, U., and Scheffer, T. (2005). Systematic feature evaluation for gene name recognition. *BMC Bioinformatics*, 6:S9.
- Hanisch, D., Fluck, J., Mevissen, H. T., and Zimmer, R. (2003). Playing biology’s name game: identifying protein names in scientific text. In *Pac. Symp. Biocomput.*, volume 8, pages 403–414, Hawaii. World Scientific.



- Hanisch, D., Fundel, K., Mevissen, H. T., Zimmer, R., and Fluck, J. (2005). ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6:S14.
- Hao, Y., Zhu, X., Huang, M., and M., L. (2005). Discovering patterns to extract protein-protein interactions from the literature: part II. *Bioinformatics*, 21.
- Hatzivassiloglou, V., Duboue, P. A., and Rzhetsky, A. (2001). Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics*, 17 Suppl. 1:S97–S106.
- Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.
- Hersh, W., Bhupatiraju, R. T., and Corley, S. (2004a). Enhancing access to the bibliome: the TREC genomics track. *Medinfo.*, 11:773–777.
- Hersh, W. and Bhuptiraju, R. T. (2003). TREC genomics track overview. In *Proceedings of TREC 2003*, volume 12.
- Hersh, W. R., Bhuptiraju, R. T., Ross, L., Johnson, P., Cohen, A. M., and Kraemer, D. F. (2004b). TREC 2004 genomics track overview. In *Proceedings of TREC 2004*, volume 13.
- Hirschman, L., Colosimo, M., Morgan, A., and Yeh, A. (2005). Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics*, 6:S11.
- Hirschman, L., Park, J. C., Tsujii, J., Wong, L., and Wu, C. H. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, 18:1553–1561.
- Hoffmann, R., Krallinger, M., Andres, E., Tamames, J., Blaschke, C., and Valencia, A. (2005). Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci. STKE*, 283:pe21.
- Hoffmann, R. and Valencia, A. (2004). A gene network for navigating the literature. *Nature Genetics*, 36:664.
- Homayouni, R., Heinrich, K., Wei, L., and Berry, M. W. (2005). Gene clustering by Latent Semantic Indexing of MEDLINE abstracts. *Bioinformatics*, 21:104–115.
- Hristovski, D., Peterlin, B., Mitchell, J. A., and Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inform.*, 74:289–298.
- Hristovski, D., Stare, J., Peterlin, B., and Dzeroski, S. (2001). Supporting discovery in medicine by association rule mining in MEDLINE and UMLS. *Medinfo.*, 10:1344–1348.



- Hu, Z. Z., Narayanaswamy, M., Ravikumar, K. E., Vijay-Shanker, K., and Wu, C. H. (2005). Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*, 21:2759–2765.
- Huang, M., Zhu, X., Hao, Y., Payan, D. G., Qu, K., and Li, M. (2004). Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, 20:3604–3612.
- Humphreys, K., Demetriou, G., and Gaizauskas, R. (2000). Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. In *Pac. Symp. Biocomput.*, volume 5, pages 505–516, Hawaii. World Scientific.
- Iliopoulos, I., Enright, A. J. and Ouzounis, C. A. (2001). Textquest: document clustering of medline abstracts for concept discovery in molecular biology. In *Pac. Symp. Biocomput.*, volume 6, pages 384–395, Hawaii. World Scientific.
- Iossifov, I., Krauthammer, M., Friedman, C., Hatzivassiloglou, V., Bader, J. S., White, K. P., and Rzhetsky, A. (2004). Probabilistic inference of molecular networks from noisy data sources. *Bioinformatics*, 20:1205–1213.
- Jensen, L. J., Saric, J., and Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.*, 7:119–129.
- Jenssen, T. K., Lægreid, A., Komorowski, J., and Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28:21–28.
- Kayaalp, M., Aronson, A. R., Humphrey, S. M., Ide, N. C., Tanabe, L. K., Smith, L. H., Demner, D., Loane, R. R., Mork, J. G., and Bodenreider, O. (2003). Methods for accurate retrieval of MEDLINE citations in functional genomics. In *Proceedings of TREC 2003*, volume 12.
- Kennedy, M. A., Barbuch, R., and Bard, M. (1999). Transcriptional regulation of the squalene synthase gene (ERG9) in the yeast *Saccharomyces cerevisiae*. *Biochim. Biophys. Acta*, 1445:110–122.
- Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 suppl. 1:i180–i182.
- Koike, A. and Takagi, T. (2004). PRIME: automatically extracted PRotein Interactions and Moolecular Information databasE. *In silico Biology*, 5:0004.

- Korbel, J. O., Doerks, T., Jensen, L. J., Perez-Iratxeta, C., Kaczanowski, S., Hooper, S. D., Andrade, M. A., and Bork, P. (2005). Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol.*, 3:e134.
- Kou, Z., Cohen, W. W., and Murphy, R. F. (2005). High-recall protein entity recognition using a dictionary. *Bioinformatics*, 21:i266–i273.
- Krallinger, M., Padron, M., and Valencia, A. (2005). A sliding window approach to extract protein annotations from biomedical articles. *BMC Bioinformatics*, 6:S19.
- Krallinger, M. and Valencia, A. (2005). Text-mining and information-retrieval services for molecular biology. *Genome Biology*, 6:224.
- Krauthammer, M., Kaufmann, C. A., Gilliam, T. C., and Rzhetsky, A. (2004). Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A.*, 101:15148–15153.
- Krauthammer, M., Rzhetsky, A., Morozov, P., and C., F. (2000). Using blast for identifying gene and protein names in journal articles. *Gene*, 259:245–252.
- Lenssen, E., James, N., Pedruzzi, I., Dubouloz, F., Cameroni, E., Bisig, R., Maillet, L., Werner, M., Roosen, J., Petrovic, K., Winderickx, J., Collart, M. A., and De Virgilio, C. (2005). The Ccr4-Not complex independently controls both Msn2-dependent transcriptional activation—via a newly identified Glc7/Bud14 type I protein phosphatase module—and TFIID promoter distribution. *Mol. Cell. Biol.*, 25:488–498.
- Leonard, J. E., Colombe, J. B., and Levy, J. L. (2002). Finding relevant references to genes and proteins in Medline using a Bayesian approach. *Bioinformatics*, 18:1515–1522.
- Leroy, G. and Chen, H. (2002). Filling preposition-based templates to capture information from medical abstracts. In *Pac. Symp. Biocomput.*, volume 7, pages 350–361, Hawaii. World Scientific.
- Lindsay, R. K. and Gordon, M. D. (1999). Literature-based discovery by lexical statistics. *J. Am. Soc. Inf. Sci.*, 50:574–587.
- Liu, J. and Rost, B. (2000). SAWTED: Structure Assignment With Text Description—enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons. *Bioinformatics*, 16:125–129.
- Maier, H., Döhr, S., Grote, K., O'Keeffe, S., Werner, T., Hrabé de Angelis, M., and Schneider, R. (2005). LitMiner and WikiGene: identifying problem-related key play-

- ers of gene regulation using publication abstracts. *Nucleic Acids Res.*, 33:W779–W782.
- Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA.
- Marcotte, E. M., Xenarios, I., and Eisenberg, D. (2001). Mining literature for protein–protein interactions. *Bioinformatics*, 17:359–363.
- Masys, D. R. (2001). Linking microarray data to the literature. *Nature Genetics*, 28:9–10.
- Masys, D. R., Welsh, J. B., Lynn Fink, J., Gribskov, M., Kłacansky, I., and Corbeil, J. (2001). Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17:319–326.
- Matsunaga, T. and Muramatsu, M. (2005). Knowledge-based computational search for genes associated with the metabolic syndrome. *Bioinformatics*, 21:3146–3154.
- McDonald, R. and Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*, 6:S6.
- Mika, S. and Rost, B. (2004). Protein names precisely peeled off free text. *Bioinformatics*, 20:i241–i247.
- Mitsumori, T., Fation, S., Murata, M., Doi, K., and Doi, H. (2005). Gene/protein name recognition based on support vector machine using dictionary as features. *BMC Bioinformatics*, 6:S8.
- Morel, V. and Schweisguth, F. (2000). Repression by Suppressor of Hairless and activation by Notch are required to define a single row of *single-minded* expressing cells in the *Drosophila* embryo. *Genes Dev.*, 14:377–388.
- Muller, H. M., Kenny, E. E., and Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, 2:e309.
- Narayanaswamy, M., Ravikumar, K. E., and Vijay-Shanker, K. (2003). A biological named entity recognizer. In *Pac. Symp. Biocomput.*, volume 8, pages 427–438, Hawaii. World Scientific.
- Narayanaswamy, M., Ravikumar, K. E., and Vijay-Shanker, K. (2005). Beyond the clause: extraction of phosphorylation information from medline abstracts. *Bioinformatics*, 21:i319–i327.

- Novichkova, S., Egorov, S., and Daraselia, N. (2003). MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics*, 19:1699–1706.
- Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. (2001). Automated extraction of information on protein–protein interactions from the biological literature. *Bioinformatics*, 17:155–161.
- Perez-Iratxeta, C., Bork, P., and A., A. M. (2001). XplorMed: a tool for exploring MEDLINE abstracts. *Trends Biochem. Sci.*, 26:573–575.
- Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2002). Association of genes to genetically inherited diseases using text mining. *Nature Genetics*, 31:316–319.
- Perez-Iratxeta, C., Perez, A. J., Bork, P., and A., A. M. (2003). Update on XplorMed: a web server for exploring scientific literature. *Nucleic Acids Res.*, 31:3866–3868.
- Perez-Iratxeta, C., Wjst, M., Bork, P., and Andrade, M. A. (2005). G2D: A tool for mining genes associated to disease. *BMC Genetics*, 6:45.
- Pillet, V., Zehnder, M., Seewald, A. K., Veuthey, A. L., and Petrak, J. (2005). GPSDB: a new database for synonyms expansion of gene and protein names. *Bioinformatics*, 21:1743–1744.
- Proux, D., Rechenmann, F., and Julliard, L. (2000). A pragmatic information extraction strategy for gathering data on genetic interactions. In *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, volume 8, pages 179–285, Menlo Park, CA. AAAI Press.
- Proux, D., Rechenmann, F., Julliard, L., Pillet, V. V., and Jacq, B. (1998). Detecting gene symbols and names in biological texts: A first step towards pertinent information extraction. *Genome Inform. Ser. Workshop Genome Inform.*, 9:72–80.
- Pustejovsky, J., Castano, J., Cochran, B., , and Morrell, M. (2001). Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Medinfo.*, 10:371–375.
- Pustejovsky, J., Castaño, J., Zhang, J., Kotecki, M., and Cochran, B. (2002). Robust relational parsing over biomedical literature: Extracting inhibit relations. In *Pac. Symp. Biocomput.*, volume 7, pages 362–373, Hawaii. World Scientific.
- Ramani, A. K., Bunescu, R. C., Mooney, R. J., and Marcotte, E. M. (2005). Consolidating the set of known human protein–protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 6:R40.
- Ray, S. and Craven, M. (2005). Learning statistical models for annotating proteins with function information using biomedical text. *BMC Bioinformatics*, 6:S18.

- Raychaudhuri, S. and Altman, R. B. (2003). A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics*, 19:396–401.
- Raychaudhuri, S., Chang, J. T., Imam, F., and Altman, R. B. (2003). The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res.*, 31:4553–4560.
- Raychaudhuri, S., Chang, J. T., Sutphin, P. D., and Altman, R. B. (2002a). Associating genes with Gene Ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res.*, 12:203–214.
- Raychaudhuri, S., Schutze, H., and Altman, R. B. (2002b). Using text analysis to identify functionally coherent gene groups. *Genome Res.*, 12:1582–1590.
- Rebholz-Schuhmann, D. (2005). Facts from text—is text mining ready to deliver. *PLoS Biology*, 3:e65.
- Renner, A. and Aszodi, A. (2000). High-throughput functional annotation of novel gene products using document clustering. In *Pac. Symp. Biocomput.*, volume 5, pages 50–68, Hawaii. World Scientific.
- Rice, S. B., Nenadic, G., and Stapley, B. J. (2005). Mining protein function from text using term-based support vector machines. *BMC Bioinformatics*, 6:S22.
- Rindflesch, T. C., Tanabe, L., Weinstein, J. N., and Hunter, L. (2000). EDGAR: extraction of drugs, genes and relations from the biomedical literature. In *Pac. Symp. Biocomput.*, volume 1, pages 517–528, Hawaii. World Scientific.
- Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., Yu, H., Dubou'e, P. A., Weng, W., Wilbur, W. J., Hatzivassiloglou, V., and Friedman, C. (2004). GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *J. Biomed. Inform.*, 37:43–53.
- Saric, J., Jensen, L. J., Ouzounova, R., Rojas, I., and Bork, P. (2004a). Extracting regulatory gene expression networks from pubmed. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*.
- Saric, J., Jensen, L. J., Ouzounova, R., Rojas, I., and Bork, P. (2006). Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*, 22:645–650.
- Saric, J., Jensen, L. J., and Rojas, I. (2004b). Large-scale extraction of gene regulation for model organisms in an ontological context. *In silico Biology*, 5:0003.
- Scherf, M., Eppler, A., and Werner, T. (2005). The next generation of literature analysis: Integration of genomic analysis into text mining. *Brief Bioinform.*, 6:287–297.

- Schijvenaars, B. J. A., Mons, B., Weeber, M., Schuemie, M. J., van Mulligen, E. M., Wain, H. W., and Kors, J. A. (2005). Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics*, 6:149.
- Schlitt, T., Palin, K., Rung, J., Dietmann, S., Lappe, M., Ukkonen, E., and Brazma, A. (2003). From gene networks to gene function. *Genome Res.*, 13:2568–2576.
- Schuemie, M. J., Weeber, M., Schijvenaars, B. J. A., van Mulligen, E. M., van der Eijk, C. C., Jelier, R., Mons, B., and Kors, J. A. (2004). Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20:2597–2604.
- Seki, K. and Mostafa, J. (2005). An approach to protein name extraction using heuristics and a dictionary. *Proceedings of the American Society for Information Science and Technology*, 40:71–77.
- Sekimizu, T., Park, H. S., and Tsujii, J. (1998). Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. *Genome Inform. Ser. Workshop Genome Inform.*, 9:62–71.
- Settles, B. (2005). ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 21:3191–3192.
- Shah, P. K., Jensen, L. J., Boue, S., and Bork, P. (2005). Extraction of transcript diversity from scientific literature. *PLoS Comput. Biol.*, 1:e10.
- Shah, P. K., Perez-Iratxeta, C., Bork, P., and Andrade, M. A. (2003). Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics*, 4:20.
- Shatkay, H. (2005). Hairpins in bookstacks: Information retrieval from biomedical text. *Brief Bioinform.*, 6:222–238.
- Shatkay, H., Edwards, S., Wilbur, W. J., and Boguski, M. (2000). Genes, themes and microarrays: using information retrieval for large-scale gene analysis. In *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, volume 8, pages 317–328, Menlo Park, CA. AAAI Press.
- Shi, L. and Campagne, F. (2005). Building a protein name dictionary from full text: a machine learning term extraction approach. *BMC Bioinformatics*, 6:88.
- Skusa, A., Ruegg, A., and Kohler, J. (2005). Extraction of biological interaction networks from scientific literature. *Brief Bioinform.*, 6:263–276.
- Smalheiser, N. R. and Swanson, D. R. (1994). Assessing a gap in the biomedical literature: Magnesium deficiency and neurological disease. *Neuroscience Research Communications*, 15:1–9.



- Smalheiser, N. R. and Swanson, D. R. (1996). Linking estrogen to Alzheimer's disease: An informatics approach. *Neurology*, 47:809–810.
- Smalheiser, N. R. and Swanson, D. R. (1998). Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput. Methods Programs Biomed.*, 57:149–153.
- Srinivasan, P. and Libbus, B. (2004). Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, 20:i290–i296.
- Stapley, B. J. and Benoit, G. (2000). Biobibliometrics: Information retrieval and visualization from co-occurrence of gene names in Medline abstracts. In *Pac. Symp. Biocomput.*, volume 5, pages 529–540, Hawaii. World Scientific.
- Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R., and Mostafa, J. (2001). Detecting gene relations from Medline abstracts. In *Pac. Symp. Biocomput.*, volume 6, pages 483–495, Hawaii. World Scientific.
- Suomela, B. P. and Andrade, M. A. (2005). Ranking the whole MEDLINE database according to a large training set using text indexing. *BMC Bioinformatics*, 6:75.
- Swanson, D. R. (1986a). Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.*, 30:7–18.
- Swanson, D. R. (1986b). Undiscovered public knowledge. *Library Quarterly*, 56:103–118.
- Swanson, D. R. (1988a). Intervening in the life cycle of scientific knowledge. *Library Trends*, 41:606–631.
- Swanson, D. R. (1988b). Migraine and magnesium: Eleven neglected connections. *Perspect. Biol. Med.*, 31:526–557.
- Swanson, D. R. (1990). Somatomedin C and arginine: implicit connections between mutually isolated literatures. *Perspect. Biol. Med.*, 33:157–186.
- Swanson, D. R. and Smalheiser, N. R. (1999). Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery. *Library Trends*, 48:48–59.
- Tamames, J. (2005). Text Detective: a rule-based system for gene annotation in biomedical texts. *BMC Bioinformatics*, 6:S10.
- Tanabe, L., Scherf, U., Smith, L. H., Lee, J. K., Hunter, L., and Weinstein, J. N. (1999). MedMiner: An internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, 27:1210–1217.

- Tanabe, L. and Wilbur, W. J. (2002). Tagging gene and protein names in biomedical text. *Bioinformatics*, 18:1124–1132.
- Tanabe, L., Xie, N., Thom, L. H., Matten, W., and Wilbur, W. J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6:S3.
- Temkin, J. M. and Gilder, M. R. (2003). Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics*, 19:2046–2053.
- Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll, M. (2000). Automatic extraction of protein interactions from scientific abstracts. In *Pac. Symp. Biocomput.*, volume 5, pages 707–709, Hawaii. World Scientific.
- Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B., and Hide, W. A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.*, 33:1544–1552.
- Tsuruoka, Y. and Tsujii, J. (2003). Boosting precision and recall of dictionary-based protein name recognition. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine*, pages 41–48.
- Usuzaka, S., Sim, K. L., Tanaka, M., Matsuno, H., and Miyano, S. (1998). A machine learning approach to reducing the work of experts in article selection from database: A case study for regulatory relations of *S. cerevisiae* genes in MEDLINE. *Genome Inform. Ser. Workshop Genome Inform.*, 9:91–101.
- Verspoor, K., Cohn, J., Josly, C., Mniszewski, S., Rechtsteiner, A., Rocha, L. M., and Simas, T. (2005). Protein annotation as term categorization in the gene ontology using word proximity networks. *BMC Bioinformatics*, 6:S20.
- von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. (2005). STRING: Known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, 33:D433–D437.
- Weeber, M., Klein, H., Aronson, A. R., Mork, J. G., de Jong-van den Berg, L. T. W., and Vos, R. (2000). Text-based discovery in biomedicine: The architecture of the DAD-system. *Proc. AMIA Symp.*, 20 Suppl.:903–907.
- Wilbur, W. J. and Coffee, L. (1994). The effectiveness of document neighboring in search enhancement. *Inf. Process. Manage.*, 30:253–266.
- Wilbur, W. J. and Yang, Y. (1996). An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput. Biol. Med.*, 26:209–222.



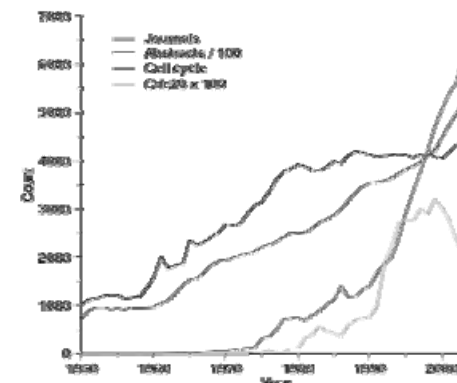
- Wong, L. (2001). PIES, a protein interaction extraction system. In *Pac. Symp. Biocomput.*, volume 6, pages 520–531, Hawaii. World Scientific.
- Woods, S. L. and Witelaw, M. L. (2002). Differential activities of Murine Single Minded 1 (SIM1) and SIM2 on a hypoxic response element. *J. Biol. Chem.*, 277:10236–10243.
- Wren, J. D. (2004). Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics*, 5:145.
- Wren, J. D., Bekeredijan, R., Stewart, J. A., Shohet, R. V., and Garner, H. R. (2004). Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics*, 20:389–398.
- Wren, J. D. and Garner, H. R. (2004). Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics*, 20:191–198.
- Xiao, Y. and Mitchell, A. P. (2000). Shared roles of yeast glycogen synthase kinase 3 family members in nitrogen-responsive phosphorylation of meiotic regulator Ume6p. *Mol. Cell. Biol.*, 20:5447–5453.
- Yakushiji, A., Tateisi, Y., Miyao, Y., and Tsujii, J. (2001). Event extraction from biomedical papers using a full parser. In *Pac. Symp. Biocomput.*, volume 6, pages 408–419, Hawaii. World Scientific.
- Yandell, M. D. and Majoros, W. H. (2002). Genomics and natural language processing. *Nat. Rev. Genet.*, 3:601–610.
- Yeh, A., Morgan, A., Colosimo, M., and Hirschman, L. (2005). BioCreAtIvE task 1A: gene mention finding evaluation. *BMC Bioinformatics*, 6:S2.
- Yoshida, M., Fukada, K., and Takagi, T. (2000). PNAD-CSS: a workbench for constructing a protein name abbreviation dictionary. *Bioinformatics*, 16:169–175.
- Yu, H. and Agichtein, E. (2003). Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, 19:i340–i349.
- Zhou, G., Shen, D., Zhang, J., Su, J., and Tan, S. (2005). Recognition protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics*, 6:S7.
- Zhou, G., Zhang, J., Su, J., Shen, D., and Tan, C. (2004). Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20:1178–1190.



## Biological literature mining from information retrieval to biological discovery

*Lars Juhl Jensen*  
 EMBL, Germany  
 jensen@embl.de

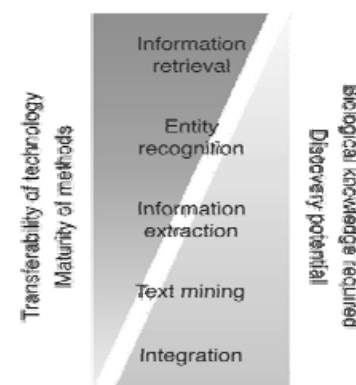
## Why do we need it?



## Overview

- Information retrieval and entity recognition
  - Methodologies for finding and classifying texts
  - Identification of gene/protein names in text
- Information extraction and text/data mining
  - Statistical co-occurrence and NLP methods for relation extraction
  - Making discoveries from text alone
  - Integration of text and other data types

## Status



- IR, ER, and simple IE methods are fairly well established
- NLP-based IE systems are rapidly improving
- Methods for text mining and text/data integration are still in their infancy

## Example

Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homolog) directly phosphorylated Swe1 and this modification served as a priming step to promote subsequent Cdc5-dependent Swe1 hyperphosphorylation and degradation

## Information Retrieval and Entity Recognition

*Lars Juhl Jensen*

EMBL, Germany

*jensen@embl.de*

## Overview

- Ad hoc information retrieval
  - The user enters a query
  - The system attempts to retrieve the relevant texts from a large text corpus
- Text categorization
  - A set of manually classified texts is created
  - A machine learning methods is trained and subsequently used to classify other texts

## Ad hoc IR

- Very flexible – any query can be entered
  - Boolean queries (yeast AND cell cycle)
  - A few systems instead allow the relative weight of each search term to be specified
- The goal is to find all the relevant papers
  - Ideally our example sentence should be identified by the query “yeast cell cycle” although none of these words are mentioned

NCBI PubMed National Library of Medicine NLM My NCBI (Sign In) (Register)

Search PubMed for [ ] Go Clear

Limits Preview/Index History Clipboard Details

• To get started, enter one or more search terms.  
• Search terms may be topics, authors or journals.

**News flash: deliver PubMed search results directly to your desktop with an RSS feed.**

To set up an RSS feed:  
(1) Run your search in PubMed.  
(2) Select **RSS Feed** from the **Send to** menu.  
(3) Click **Create Feed** and copy the XML icon into your RSS Reader.  
Read the [PubMed Help](#) to explore other options for automated e-mail updates using My NCBI.

PubMed is a service of the National Library of Medicine that includes over 15 million citations from MEDLINE and other life science journals for biomedical articles back to the 1950s. PubMed includes links to full text articles and other related resources.

Related Resources  
Order Documents  
NLM Mobile  
NLM Catalog  
NLM Gateway  
TOXNET  
Consumer Health  
Clinical Alerts  
ClinicalTrials.gov  
PubMed Central

IMB 2006 Fortaleza, Brazil August 6-10, 2006 177

NCBI PubMed National Library of Medicine NLM My NCBI (Sign In) (Register)

Search PubMed for yeast cell cycle Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Send to

All: 13713 Review: 1480

Items 1 - 20 of 13713 Page 1 of 686 Next

1: Homma MK, Wada J, Suzuki T, Yamaki J, Krebs EG, Homma Y.  
CK2 phosphorylation of eukaryotic translation initiation factor 5 potentiates cell cycle progression.  
Proc Natl Acad Sci U S A. 2005 Oct 14; [Epub ahead of print]  
PMID: 16227438 [PubMed - as supplied by publisher]

2: Leevers SJ, McNeill H.  
Controlling the size of organs and organisms.  
Curr Opin Cell Biol. 2005 Oct 11; [Epub ahead of print]  
PMID: 16226450 [PubMed - as supplied by publisher]

3: Wu JO, Holland TD.  
Counting cytokinesis proteins globally and locally in fission yeast.  
Science. 2005 Oct 14; 310(5746):110-4.  
PMID: 16224022 [PubMed - in process]

4: David-Plaut T.  
The flexible evolutionary anchorage-dependent Pardee's restriction point of mammalian cells. How its deregulation may lead to cancer.  
Biochim Biophys Acta. 2005 Sep 20; [Epub ahead of print]  
PMID: 16219425 [PubMed - as supplied by publisher]

5: Anskonda TS, Reddy PH.  
Neuronal protection by sirtuins in Alzheimer's disease.  
J Neurochem. 2005 Oct 7; [Epub ahead of print]  
PMID: 16219030 [PubMed - as supplied by publisher]

IMB 2006 Fortaleza, Brazil August 6-10, 2006 178

Google Scholar BETA

Search Advanced Scholar Search Scholar Preferences Scholar Help

Stand on the shoulders of giants

Inspired by the abstract? See if your library gives you access to the whole paper.

Google Home - About Google - About Google Scholar

©2005 Google

IMB 2006 Fortaleza, Brazil August 6-10, 2006 179

Google Scholar BETA

Search Advanced Scholar Search Scholar Preferences Scholar Help

Scholar Results 1 - 10 of about 256,000 for yeast cell cycle. (0.08 seconds)

Genomic binding sites of the yeast cell cycle transcription factors SBF and MBF  
VR Iyer, CE Horak, CS Scafe, D Botstein, M Snyder, ... - Nature, 2001 - nature.com  
... 15. Koch, C., Schieffler, A., Ammerer, G. & Nasmyth, K. Switching transcription on and off during the yeast cell cycle: Cln/Cdc28 kinases activate bound ...  
Cited by 312 - Web Search - nature.com - biomed.com - ncbi.nlm.nih.gov - all 5 versions »

Control of the yeast cell cycle by the Cdc28 protein kinase  
K Nasmyth - Curr. Opin. Cell Biol. 1993 - ncbi.nlm.nih.gov  
Control of the yeast cell cycle by the Cdc28 protein kinase. Nasmyth K.  
Research Institute of Molecular Pathology, Vienna, Austria. ...  
Cited by 207 - Web Search - ncbi.nlm.nih.gov

Morphogenesis in the yeast cell cycle: regulation by Cdc28 and cyclins  
DJ Lew, SI Reed - J. Cell Biol. 1993 - dx.doi.org  
... The Rockefeller University Press. ARTICLES. Morphogenesis in the yeast cell cycle: regulation by Cdc28 and cyclins. DJ Lew and SI Reed ...  
Cited by 187 - Web Search - info.ncbi.org - pubmed.ncbi.nlm.nih.gov - all 5 versions »

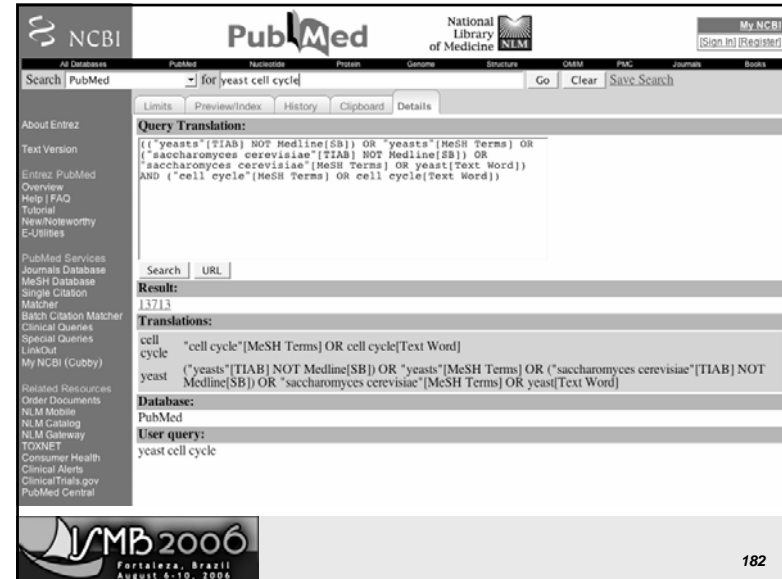
Mutation of fission yeast cell cycle control genes abolishes dependence of mitosis on DNA ...  
T Enoch, P Nurse - Cell. 1990 - ncbi.nlm.nih.gov  
Mutation of fission yeast cell cycle control genes abolishes dependence of mitosis on DNA replication. Enoch T, Nurse P. Department ...  
Cited by 157 - Web Search

At the heart of the budding yeast cell cycle  
K Nasmyth - Trends Genet. 1996 - ingentaconnect.com  
... At the heart of the budding yeast cell cycle. Author: Nasmyth K. 1. ... This article presents one view as to what lies at the heart of the budding yeast cell cycle. ...  
Cited by 153 - Web Search - ingentaconnect.com - ncbi.nlm.nih.gov

IMB 2006 Fortaleza, Brazil August 6-10, 2006 180

## Automatic query expansion

- The user will typically not provide all relevant words and variants thereof
- Query expansion can improve recall
  - Stemming of the words (yeast / yeasts)
  - Use of thesauri deal with synonyms and/or abbreviations (yeast / *S. cerevisiae*)
  - The next step is to use ontologies to make complex inferences (yeast cell cycle / Cdc28)



The screenshot shows the PubMed interface. The search bar contains 'yeast cell cycle'. Below the search bar, the 'Query Translation' section displays the expanded query: `((("yeasts"[TIAB] NOT Medline[SB]) OR "yeasts"[MeSH Terms] OR ("saccharomyces cerevisiae"[TIAB] NOT Medline[SB]) OR "saccharomyces cerevisiae"[MeSH Terms] OR yeast[Text Word]) AND ("cell cycle"[MeSH Terms] OR cell cycle[Text Word]))`. The 'Result' section shows 13713 results. The 'Translations' section lists: 'cell cycle' as '"cell cycle"[MeSH Terms] OR cell cycle[Text Word]', 'cycle' as '"yeasts"[TIAB] NOT Medline[SB]) OR "yeasts"[MeSH Terms] OR ("saccharomyces cerevisiae"[TIAB] NOT Medline[SB]) OR "saccharomyces cerevisiae"[MeSH Terms] OR yeast[Text Word]', and 'yeast' as '"yeasts"[TIAB] NOT Medline[SB]) OR "yeasts"[MeSH Terms] OR ("saccharomyces cerevisiae"[TIAB] NOT Medline[SB]) OR "saccharomyces cerevisiae"[MeSH Terms] OR yeast[Text Word]'. The 'Database' is PubMed and the 'User query' is 'yeast cell cycle'.

## Document similarity

- The similarity of two documents can be defined based on their word content
  - Represent each document by a word vector
  - Words should be weighted based on their frequency and background frequency
- Document similarity can be used in IR
  - Include the k nearest neighbors when matching queries against documents

## Document clustering

- Unsupervised clustering algorithms can be applied to a document similarity matrix
  - Calculate all pairwise document similarities
  - Apply a standard clustering algorithm
- Practical uses of document clustering
  - The “related documents” function in PubMed
  - Organizing the documents found by IR

## Text categorization

- These systems are less flexible than ad hoc systems but give better accuracy
  - The document classes are pre-defined
  - Needs manual classification of training data
- Methods
  - Rules can be manually crafted
  - Machine learning methods can be trained



185

## Example

Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homolog) directly phosphorylated Swe1 and this modification served as a priming step to promote subsequent Cdc5-dependent Swe1 hyperphosphorylation and degradation

### Hints in the text

- Yeast cell cycle: Cdc28 and Swe1
- Cell cycle: mitotic cyclin, Clb2, and Cdk1



186

## Machine learning

- Input features
  - Word content or bi-/tri-grams
  - Part-of-speech tags
  - Filtering (stop words, part-of-speech)
- Training
  - Support vector machines are best suited
  - Separate training and evaluation sets



187

## Entity recognition

- An important but boring problem
  - Find the entities (genes/proteins) mentioned within a given text
- Recognition vs. identification
  - Recognition: find the words that are names
  - Identification: identify the entities they refer to
  - Recognition alone is of limited use



188



## Example

Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homolog) directly phosphorylated Swe1 and this modification served as a priming step to promote subsequent Cdc5-dependent Swe1 hyperphosphorylation and degradation

### Entities identified

Clb2 (YPR119W), Cdc28 (YBR160W), Swe1 (YJL187C), and Cdc5 (YMR001C)



189

## Recognition

- Features
  - Morphological: mixes letters and digits
  - Context: followed by “protein” or “gene”
  - Grammar: should occur as a noun
- Methodologies
  - Manually crafted rule-based systems
  - Machine learning (SVMs)



190

## Identification

- A good synonyms list is the key
  - Combine many sources
  - Curate to eliminate stop words
- Orthographic variation
  - Case variation: CDC28, Cdc28, and cdc28
  - Prefixes and postfixes: c-myc and Cdc28p
  - Spaces and hyphens: cdc28 and cdc-28
  - Latin vs. Greek letters: TNF-alpha and TNFA



191

## Disambiguation

- The same word may mean different things
  - Entity names may also be common English words (hairy), technical terms (SDS) or refer to unrelated proteins in other species (cdc2)
- The meaning can be found from the context
  - ER can distinguish names from other words
  - Disambiguation of non-unique names is a hard problem



192

**iHOP**  
Information Hyperlinked  
Over Proteins

Search Gene

Gene Model  
Developer's Zone **new**  
Contact  
Help

COCCON

JAVA

site powered by pdg

Huffmann, R., Valencia, A. A Gene Network for Navigating the Literature. Nature Genetics 36, 664 (2004)

Search for a gene synonym or accession number... (Click here for an example: SNF1)

all fields in all organism

[SEARCH]

**UMB 2006**  
Fortaleza, Brazil  
August 6-10, 2006

193

Symbol	Name	Synonyms	Organism
<b>CDC28</b>	Cell division control protein 28	CDK1, HSL5, SRM5, YBR1211, YBR160W	Saccharomyces cerevisiae

UniProt P00546  
IntAct P00546  
NCBI Gene 852457  
NCBI RefSeq NP\_009718  
NCBI Accession CAA25065, CAA56509, CAA85119

Homologues of CDC28 ... **new**

Interaction information for this gene ...

Enhanced PubMed/Google query ... **new**

WARNING: Please keep in mind that gene detection is done automatically and can exhibit a certain error. Read more.

Find in this Page

These events require activation of Cdc28 kinase by G1 [cyclins](#).

Saccharomyces cerevisiae cell cycle: cdc28 and the G1 [cyclins](#).

Analysis of the Cdc28 [protein kinase](#) complex by dosage suppression.

Mitotic role for the Cdc28 [protein kinase](#) of Saccharomyces cerevisiae.

Morphogenesis in the yeast cell cycle: regulation by Cdc28 and [cyclins](#).

The CDC28 [mRNA](#) had been previously estimated at 7.0 +/- 2 copies per cell.

Invariant phosphorylation of the Saccharomyces cerevisiae Cdc28 [protein kinase](#).

The role of CDC28 and [cyclins](#) during mitosis in the budding yeast S. cerevisiae.

We show that DNA replication also requires activation of Cdc28 by B-type (Clb) [cyclins](#).

**UMB 2006**  
Fortaleza, Brazil  
August 6-10, 2006

194

Symbol	Name	Synonyms	Organism
<b>CDC28</b>	Cell division control protein 28	CDK1, HSL5, SRM5, YBR1211, YBR160W	Saccharomyces cerevisiae

UniProt P00546  
IntAct P00546  
NCBI Gene 852457  
NCBI RefSeq NP\_009718  
NCBI Accession CAA25065, CAA56509, CAA85119

Homologues of CDC28 ... **new**

Definitions for CDC28 ...

Enhanced PubMed/Google query ... **new**

WARNING: Please keep in mind that gene detection is done automatically and can exhibit a certain error. Read more.

Find in this Page

Furthermore, SW14 associates with [CLB2](#) protein and is a substrate for the CLB2-associated CDC28 kinase in vitro.

Furthermore, the [Cks1](#) protein was shown to be physically **associated** with active forms of the Cdc28 [protein kinase](#).

The cyclin-dependent kinase Cdc28p **associates** with the cyclin [Clb2p](#) to induce mitosis in the yeast Saccharomyces cerevisiae.

We find that G1 arrest in the cdc37-1 mutant is accompanied by a decrease in the Cdc28 activity **associated** with the G1 [cyclin Cln2](#).

We found that [Hct1](#) was **phosphorylated** in vivo at multiple CDK consensus sites during cell cycle stages when activity of the cyclin-dependent kinase Cdc28 is high and APC activity is low.

It is likely, therefore, that [Cks1](#) mediates a more specialized **function** of the Cdc28 kinase such as its ability to form specific multimeric complexes or to localize properly in cellular compartments.

[Cdc37](#) **promotes** the stability of protein kinases Cdc28 and [Cak1](#).

**UMB 2006**  
Fortaleza, Brazil  
August 6-10, 2006

195

## Summary

- Information retrieval
  - Ad hoc IR methods are more flexible than text categorization methods
  - Text categorization methods can generally provide better performance than ad hoc IR
- Entity recognition
  - It is not sufficient to recognize names – the entities should also be identified
  - The best methods use curated synonyms lists

**UMB 2006**  
Fortaleza, Brazil  
August 6-10, 2006

196

## Information Extraction and Text/Data Mining

*Lars Juhl Jensen,*  
EMBL, Germany  
*jensen@embl.de*

## Overview

- Information extraction (IE)
  - Simple statistical co-occurrence methods
  - Combining co-occurrence and categorization
  - Natural Language Processing (NLP)
- Text/data mining
  - Making discoveries from text alone
  - Augmenting text mining with other data types
  - Annotation of high-throughput data

## IE by co-occurrence

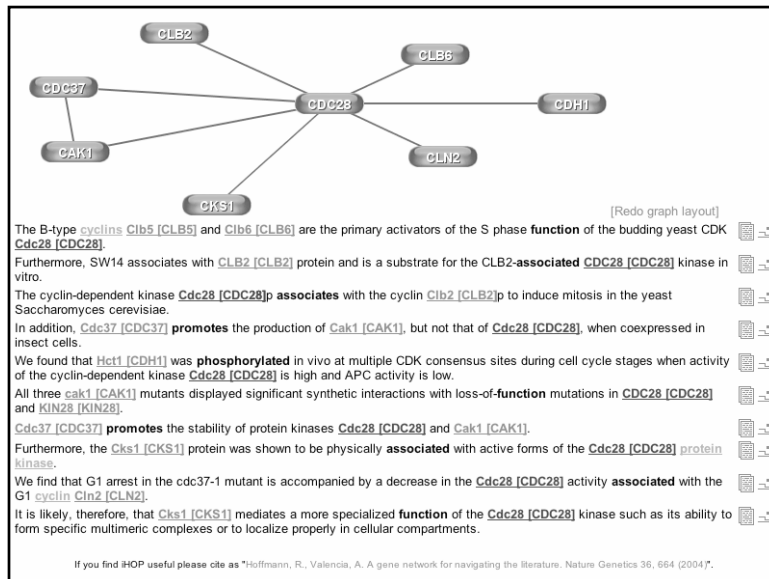
- Limitations of co-occurrence methods
  - Relations are always symmetric
  - The type of relation is not given
- Scoring the relations
  - More co-occurrences  $\Rightarrow$  more significant
  - Ubiquitous entities  $\Rightarrow$  less significant
- Simple, good recall, poor precision

## Example

Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homolog) directly phosphorylated Swe1 and this modification served as a priming step to promote subsequent Cdc5-dependent Swe1 hyperphosphorylation and degradation

### Relations extracted

Clb2–Cdc28, Clb2–Swe1, Cdc28–Swe1,  
Cdc5–Swe1, Clb2–Cdc5, and Cdc28–Cdc5



## Categorization

- Extracting specific types of relations
  - Text categorization can be used to identify sentences that mention a certain type of relations
- Well suited for database curation
  - Text categorization can be reused
  - Recall is most important since curators can correct the false positives



202

Summary of all potential interactors					
The list below shows all other proteins that co-occur in the literature with your query protein. The number of co-occurrence papers are listed under the column "View supporting papers". Clicking on this number will take you to a more detailed view of these co-occurrences.					
	name	short description	Is this interactor real?	View supporting papers	more info
Site map	CLB2	Involved in mitotic induction	Probably	34	SeqHound PreBIND
	CDC6	Protein involved in initiation of DNA replication	Probably	13	SeqHound PreBIND
Search PreBIND	CSR1	chs5 spa2 rescue; isolated as a multicopy suppressor of the lethality of chs5 spa2 double mutant at 37 degrees.	Probably	1	SeqHound PreBIND
	CDC37	cell cycle protein necessary for passage through START	Probably	3	SeqHound PreBIND
About PreBIND	HSL1	Negative regulator of swe1 kinase (which regulates cdc28)	Probably	7	SeqHound PreBIND
	SW14	Involved in cell cycle dependent gene expression	Probably	14	SeqHound PreBIND
Help	CLN2	role in cell cycle START	Probably	55	SeqHound PreBIND
	CLN3	role in cell cycle START; involved in G(sub)1 size control	Probably	34	SeqHound PreBIND
Credits	SWE1	Protein kinase that inhibits G2/M transition. S. pombe wee1+ homolog	Probably	12	SeqHound PreBIND
	FAR1	Factor arrest protein	Probably	10	SeqHound PreBIND
Go to BIND	CDH1	CDC20 homolog 1	Probably	6	SeqHound PreBIND
	DIB1	S. pombe dim1+ in budding yeast	Probably	4	SeqHound PreBIND
	SIC1	P40 inhibitor of Cdc28p-Clb5 protein kinase complex	Probably	16	SeqHound PreBIND
	SCS2	Likely to be involved in regulating INO1 expression, suppressor of a dominant nuclear mutation that is inositol-dependent in the presence of choline	Probably	16	SeqHound PreBIND
	CLN1	role in cell cycle START essential for mitotic and meiotic DNA synthesis.	Probably	37	SeqHound PreBIND
	CDC2	dispensable for meiotic spindle pole body duplication, but required for synaptonemal complexes and full intragenic recombination, spindle pole body separation and spindle formation	Probably	27	SeqHound PreBIND

## NLP

- Information is extracted based on parsing and interpreting phrases or full sentences
  - Good at extracting specific types of relations
  - Handles directed relations
- Complex, good precision, poor recall



204

## Example

Mitotic cyclin (Clb2)-bound Cdc28 (Cdk1 homolog) directly phosphorylated Swe1 and this modification served as a priming step to promote subsequent Cdc5-dependent Swe1 hyperphosphorylation and degradation

### Relations:

- Complex: Clb2–Cdc28
- Phosphorylation: Clb2→Swe1, Cdc28→Swe1, and Cdc5→Swe1

## An NLP architecture

- Tokenization
  - Entity recognition with synonyms list
  - Detection of multi words and sentence boundaries
- Part-of-speech tagging
  - TreeTagger trained on GENIA
- Semantic labeling
  - Dictionary of regular expressions
- Entity and relation chunking
  - Rule-based system implemented in CASS

### Semantic labeling

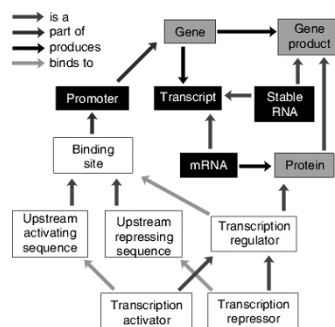
Gene and protein names  
Words for entity recognition  
Words for relation extraction

### Named entity chunking

[<sub>nxgene</sub> The GAL4 gene]

### Relation chunking

[<sub>nxexpr</sub> The expression of  
[<sub>nxgene</sub> the cytochrome genes  
[<sub>nxpg</sub> CYC1 and CYC7]]]  
is controlled by  
[<sub>nxpg</sub> HAP1]



[<sub>phosphorylation\_active</sub>  
Lyn, [negation but not Jak2]  
phosphorylates  
CrkL]

[<sub>expression\_repression\_active</sub>  
Btk  
regulates  
the IL-2 gene]

[<sub>phosphorylation\_active</sub>  
Lyn also participates in  
[<sub>phosphorylation</sub> the tyrosine phosphorylation  
and activation of syk]]

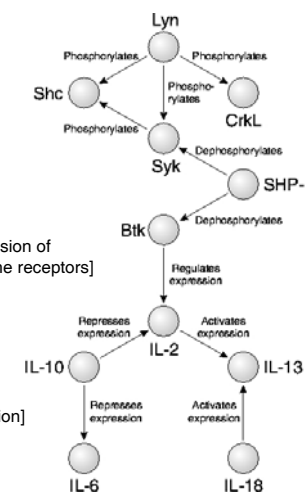
[<sub>phosphorylation\_nominal</sub>  
the phosphorylation of  
the adapter protein SHC  
by the Src-related kinase Lyn]

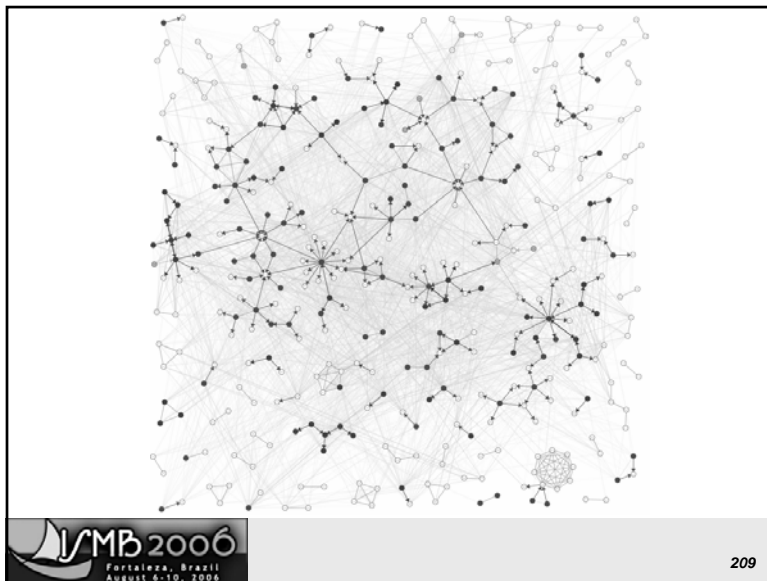
[<sub>expression\_repression\_active</sub>  
IL-10  
also decreased  
[<sub>expression</sub> mRNA expression of  
IL-2 and IL-18 cytokine receptors]

[<sub>phosphorylation\_nominal</sub>  
phosphorylation of Shc by  
the hematopoietic cell-specific  
tyrosine kinase Syk]

[<sub>dephosphorylation\_nominal</sub>  
Dephosphorylation of  
Syk and Btk  
mediated by  
SHP-1]

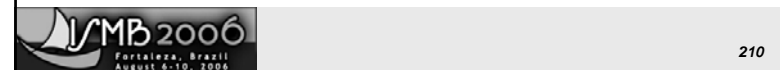
[<sub>expression\_activation\_passive</sub>  
IL-13 expression]  
[<sub>expression</sub> induced by  
IL-2 + IL-18]





## Mining text for nuggets


- Inferring new relations from old ones
  - This can lead to actual discoveries if no one knows all the facts required for the inference
  - Combining facts from disconnected literatures
- Swanson's pioneering work
  - Fish oil and Reynaud's disease
  - Magnesium and migraine



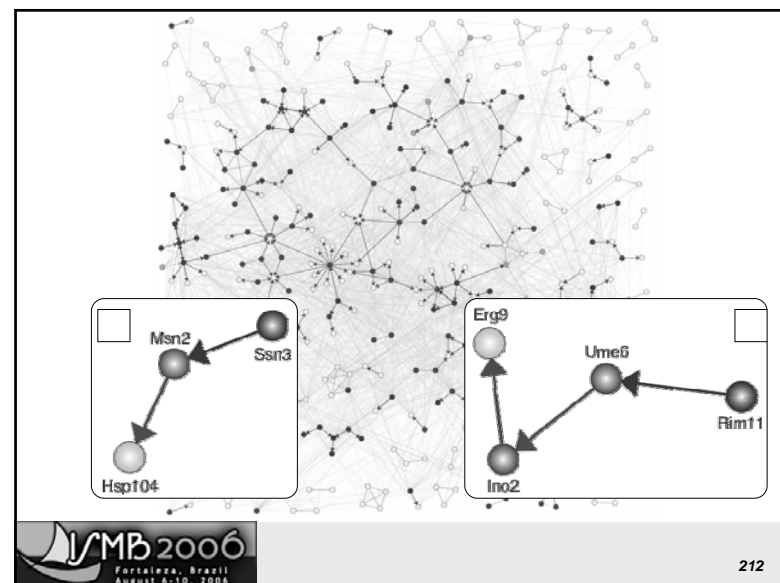
**ARROWSMITH LINKING**  
DOCUMENTS, DISCIPLINES, INVESTIGATORS AND DATABASES

WELCOME: This site is open during construction.

Ready to use Arrowsmith? <a href="#">Start</a>	References <a href="#">Start</a>	Corner for Collaborative Informatics <a href="#">Go to Website</a>
<b>Author-ity:</b> A tool for identifying Medline articles written by a particular author <a href="#">Start</a>	Compendium of Biomedical Text Mining Tools <a href="#">Start</a>	<b>Anne O'Tate:</b> A tool for summarizing the results of a PubMed query. <a href="#">Start</a>


 Fortaleza, Brazil  
 August 6-10, 2006

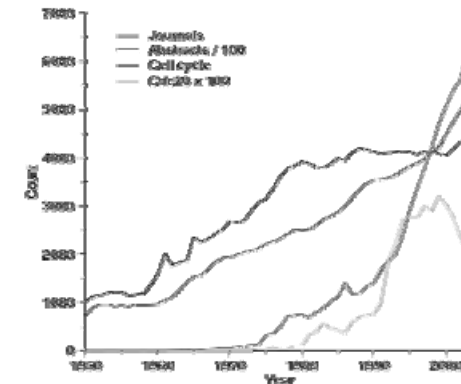
211



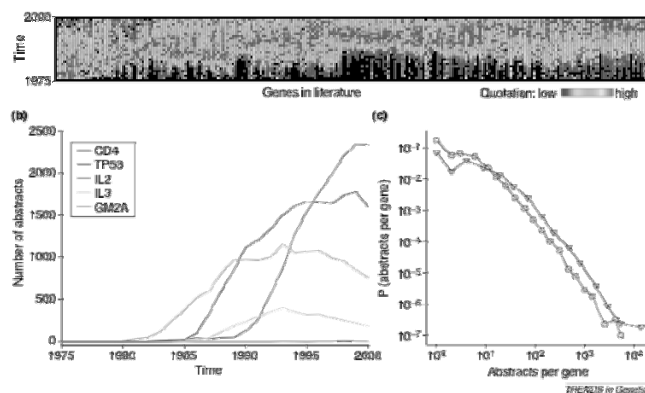
## Trends

- Similar to existing data mining approaches
  - Although all the detailed data is in the text, people may have missed the big picture
- Temporal trends
  - Historical summaries, forecasting
- Correlations
  - Customers who bought this item also bought

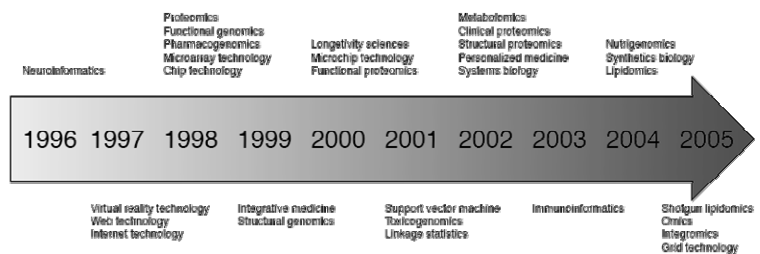
## Time



## Successful genes



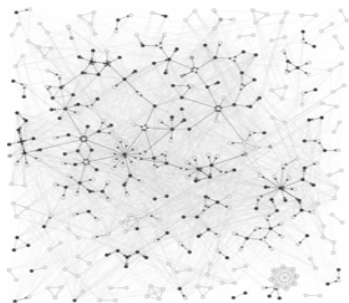
## Buzzwords



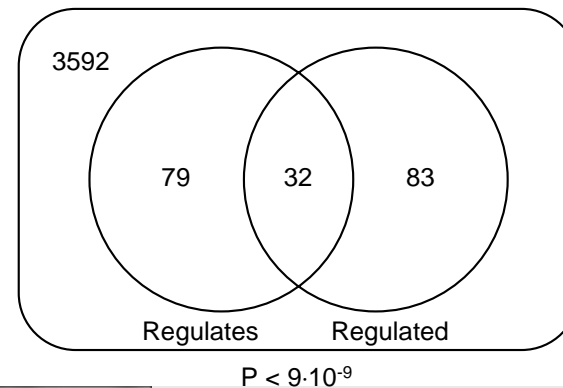


## Correlations

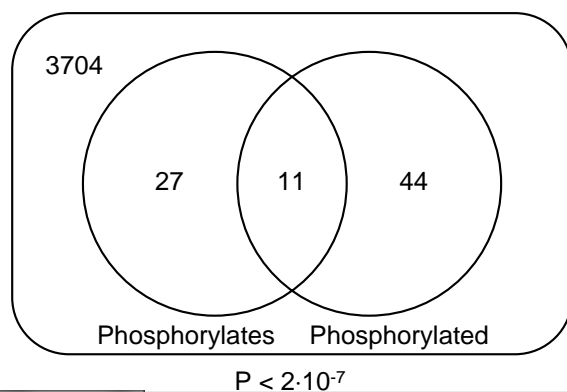
- “Customers who bought this item also bought ...”
- Protein networks
  - “Proteins that regulate expression ...”
  - “Proteins that control phosphorylation ...”
  - “Proteins that are phosphorylated ...”



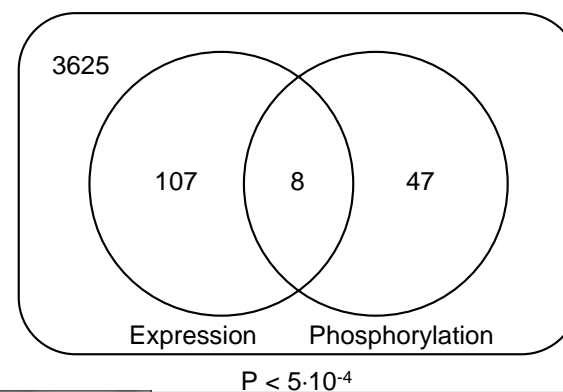
## Transcriptional networks



## Signaling pathways



## Multiple regulation



# Integration

- Annotation of high-throughput data
  - Loads of fairly trivial methods
- Protein interaction networks
  - Can unify many types of interaction data
- More creative strategies
  - Identification of candidate disease genes
  - Linking genotype to phenotype

Home · Download · Help/Info

**STRING**

**STRING - Search Tool for the Retrieval of Interacting Genes/Proteins**

**Enter your gene/protein of interest ...**

Identifier:  e.g. 'trpB', 'ANP1\_YEAST', ...  
you may also upload a [list](#)

alternatively, paste an amino-acid sequence:

Interactors wanted:

**What it does ...**

STRING is a database of known and predicted protein-protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources:


Genomic Context   High-throughput Experiments   (Conserved) Coexpression   Previous Knowledge


STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database currently contains 736429 proteins in 179 species.

**References / Info ...**

STRING uses orthology information from the excellent [COG database](#) (Ref). Up-to-date genomes and proteins are maintained at [SWISSPROT](#) and [ENSEMBL](#). STRING references: [von Mering et al. 2005](#) / [von Mering et al. 2003](#) / [Snel et al. 2000](#). Miscellaneous: [Access Statistics](#), [Robot Access Guide](#), [Supported Browsers](#).

**What's New?** You are looking at release 6.2 of STRING - latest additions are the 'HPRD' and 'Reactome' databases. **Previous Releases:** Trying to reproduce an earlier finding? Confused? Try our old releases: [version 6.0](#), [version 5.1](#)

 EMBL • BIOCOMPUTING • BORK GROUP


 IUMB 2006  
 Fortaleza, Brazil  
 August 6-10, 2006

Home · Download · Help/Info


**STRING**

**Your Input:**

☒ CDC20 Cell division control protein 20 (EC 2.7.1.37) (290 aa)

**Predicted Functional Associations:**

Association	Neighborhood	Gene Fusion	Co-occurrence	Conservation	Experimental Data	Text Mining	Homology	Score
<input checked="" type="checkbox"/> CK1L Serine/threonine-protein kinase CK1L (EC 2.7.1.37) (CDK-activating kin [...])	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.999
<input checked="" type="checkbox"/> CLN1 G1/S-specific cyclin CLN1 (546 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.999
<input checked="" type="checkbox"/> CLN2 G1/S-specific cyclin CLN2 (545 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.999
<input checked="" type="checkbox"/> CLN3 G1/S-specific cyclin CLN3 (580 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.999
<input checked="" type="checkbox"/> CLB1 G2/mitotic-specific cyclin 1 (471 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.999
<input checked="" type="checkbox"/> CLB2 G2/mitotic-specific cyclin 2 (491 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.999
<input checked="" type="checkbox"/> CLB3 G2/mitotic-specific cyclin 3 (427 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.999
<input checked="" type="checkbox"/> CLB4 G2/mitotic-specific cyclin 4 (460 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.999
<input checked="" type="checkbox"/> CLB5 S-phase entry cyclin 5 (435 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.999
<input checked="" type="checkbox"/> CLB6 S-phase entry cyclin 6 (380 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.999
<input checked="" type="checkbox"/> CKS1 Cyclin-dependent kinases regulatory subunit (Cell division control pro [...])	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.999
<input checked="" type="checkbox"/> SIC1 SIC1 protein (CDK inhibitor p40) (204 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.999
<input checked="" type="checkbox"/> SML1 Mitosis inhibitor protein kinase SML1 (EC 2.7.1.-) (619 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.990
<input checked="" type="checkbox"/> HCT1 Hypothetical 62.8 kDa Trp-Rsp repeats containing protein in PHC1-ITG2 [...]	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.996
<input checked="" type="checkbox"/> CDC6 Cell division control protein 6 (513 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.900
<input checked="" type="checkbox"/> PDS1 Securin (372 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.960
<input checked="" type="checkbox"/> PHO5 Negative regulator of the PHO system (EC 2.7.1.37) (Serine/threonine-p [...])	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.965
<input checked="" type="checkbox"/> ELM1 Protein kinase ELM1 (EC 2.7.1.-) (640 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.956
<input checked="" type="checkbox"/> CDC15 Cell division control protein 15 (EC 2.7.1.-) (594 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.955
<input checked="" type="checkbox"/> CDC7 Cell division control protein 7 (EC 2.7.1.37) (507 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.955
<input checked="" type="checkbox"/> SWI6 Regulatory protein SWI6 (Cell-cycle box factor, chain SWI6) (Trans-act [...])	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.954
<input checked="" type="checkbox"/> WHI5 ORF YOR003W (295 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.952
<input checked="" type="checkbox"/> CDC37 Hsp90 co-chaperone CDC37 (Hsp90 chaperone protein kinase-targeting sub [...])	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.951
<input checked="" type="checkbox"/> FAR1 Cyclin-dependent kinase inhibitor FAR1 (CKI FAR1) (factor arrest prote [...])	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.950
<input checked="" type="checkbox"/> GRK1 Ubiquitin ligase complex F-box protein GRK1 (1151 aa)	•••••	•••••	•••••	•••••	•••••	•••••	•••••	0.950


 IUMB 2006  
 Fortaleza, Brazil  
 August 6-10, 2006

Home · Download · Help/Info

**STRING**

**Relevant abstracts mentioning your query species (Saccharomyces cerevisiae):**

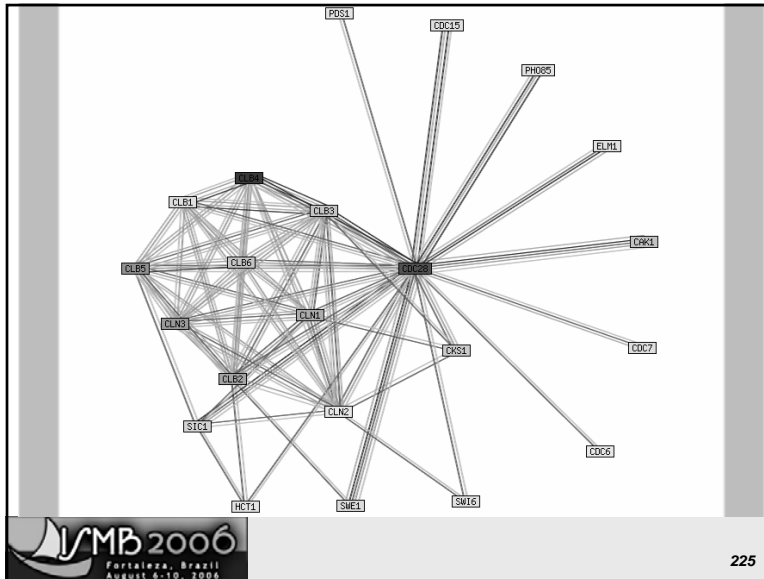
**CLB5 (■) and CLB6 (■), a new pair of B cyclins involved in DNA replication in Saccharomyces cerevisiae.** 

The functions of the Cdc28 (■) protein kinase in DNA replication and mitosis in *Saccharomyces cerevisiae* are thought to be determined by the type of cyclin subunit with which it is associated. G1-specific cyclins encoded by *CLN1* (■), *CLN2* (■), and *CLN3* (■) are required for entry into the cell cycle (Start) and thereby for S phase, whereas G2-specific B-type cyclins encoded by *CLB1* (■), *CLB2* (■), *CLB3* (■), and *CLB4* (■) are required for mitosis. We describe a new family of B-type cyclin genes, *CLB5* (■) and *CLB6* (■), whose transcripts appear in late G1 along with those of *CLN1* (■), *CLN2* (■), and many genes required for DNA replication. Deletion of *CLB6* (■) has little or no effect, but deletion of *CLB5* (■) greatly extends S phase, and deleting both genes prevents the timely initiation of DNA replication. Transcription of *CLB5* (■) and *CLB6* (■) is normally dependent on Cln activity, but ectopic *CLB5* (■) expression allows cells to proliferate in the absence of Cln cyclins. Thus, the kinase activity associated with *Clb5* (■) and not with Cln cyclins may be responsible for S-phase entry. *Clb5* (■) also has a function, along with *Clb3* (■) and *Clb4* (■), in the formation of mitotic spindles. Our observation that *CLB5* (■) is involved in the initiation of both S phase and mitosis suggests that a single primordial B-type cyclin might have been sufficient for regulating the cell cycle of the common ancestor of many, if not all, eukaryotes.

**G2 cyclins are required for the degradation of G1 cyclins in yeast.** 

Progression of the eukaryotic cell cycle is controlled by cyclin-dependent kinases (CDKs). Cdc28 (■), the budding yeast homologue of Cdc2 (Cdk1 (■)), is required for both the G1/S and G2/M transitions of the cell cycle. The functional specificity of the Cdc28 (■) kinase is determined by its association with G1 or G2 cyclins. Alteration of cell cycle phases is thus mainly due to mechanisms that ensure that one cyclin family succeeds another. Here we show that the G2 cyclins *Clb1* (■), *Clb2* (■), *Clb3* (■) and *Clb4* (■) are required for the proteolysis of the G1 cyclins *Cln1* (■) and *Cln2* (■), providing a mechanism for coupling synthesis of G2 cyclins with the disappearance of G1 cyclins. Our data indicate that this pathway involves the Ubr9 ubiquitin-conjugating enzyme. The Cdc34 ubiquitin-conjugating activity may function redundantly with Ubr9, or it may only be involved in Cln2 turnover through its role in promoting the degradation of Sic1 (■), a specific inhibitor of Cdc28 (■)-Clb complexes.

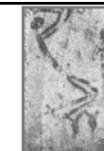
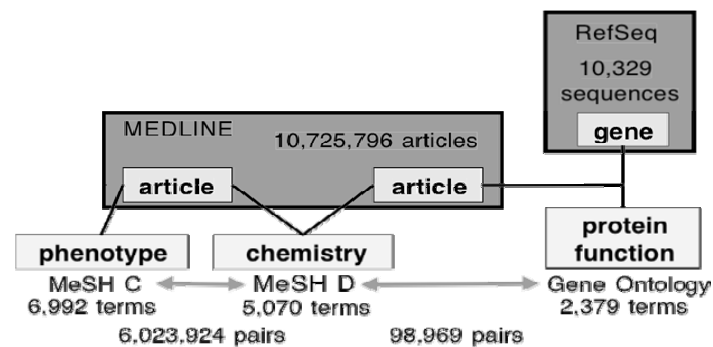

 IUMB 2006  
 Fortaleza, Brazil  
 August 6-10, 2006



## Disease candidate genes

- Rank the genes within a chromosomal region to which a disease has been mapped
- BITOLA
  - Gene→Words→Disease (similar to ARROWSMITH)
- G2D
  - Gene→Function→Chemical→Phenotype→Disease
  - Uses MEDLINE but not the text

## G2D



## G2D

### Candidate Genes to Inherited Diseases

Welcome to the G2D web server (version 2.0). Here you can use our algorithm to scan a human genomic region for genes related to an inherited disease. You can also access a database of pre-computed results for mapped monogenic human diseases and for asthma, a complex disease.

Candidate priorities are automatically established by a data mining algorithm that extracts putative genes in the chromosomal region where the disease is mapped, and evaluates their possible relation to the disease based on the phenotype of the disorder. [To know more about g2d.](#)

Use G2D

### CHOOSE YOUR G2D COMBO

**PHENOTYPE BOX**  
Type one OMIM ID defining the disease picked from [this list](#) (e.g. 131244)

**LOCATION BOX**  
Where do we look for the responsible gene?  
Select  In chromosome

**Position** start stop (e.g. 63950000 73950000)

**Cytogenetic band(s)** band1 band2 (e.g. q13.2)

**Cytogenetic marker(s)** marker1 marker2 (e.g. D9S201 D9S298)

(Maximum 50 Mb)

**Analysis of epilepsy, childhood absence 1**

GO TO: REASONS FOR ASSOCIATION OVERVIEW OF CANDIDATES BEST 10 CANDIDATES BEST 100 CANDIDATES

**CHROMOSOMAL REGION**  
 Disease mapped: epilepsy, childhood absence 1  
 Method: Md  
 Chromosome: 8  
 Genomic position start-stop: 117700000-146274826  
 length: 28574827  
 Band: 8q24

**LINKS**  
 EntrezGene: [50966]  
 MEDLINE: [1095568]  
 [9758624]  
 OMIM: [600131] XploreMed

**REASONS FOR ASSOCIATION**

**MEDLINE QUERY**  
 A set of 93 papers related to this disease was derived from MEDLINE using the query:  
 epilepsy [tw] AND childhood [tw] AND absence [tw] AND 1 [tw]

**MESH-C TERMS**  
 The MeSH-C terms associated to the derived papers were collected.  
 Here you can see the terms ordered by the number of papers where they were found (number in brackets).  
 Epilepsy [42]  
 Epilepsy, Absence [39]  
 Epilepsias, Myoclonic [15]  
 Epilepsias, Partial [13]

**GO terms**  
 We compiled the GO terms associated to the MeSH-C terms selected for this disease.  
 Here you can see those with higher score of association.  
 0.002425 GABA-B receptor (function)  
 0.001051 isolate transport (process)  
 0.000886 GABA-A receptor (function)  
 0.000646 gamma-amino butyric acid signaling pathway (process)  
 0.000598 N-methyl-D-aspartate

**IMB 2006**  
 Fortaleza, Brazil  
 August 6-10, 2006

229

**CANDIDATE 2**

**DNA (Note)**  
 R-score = 0.002301; GO-score = 0.000301  
 117700000 bp Chromosome 8 146274826 bp

1 aa NP\_003605 368 aa

**Similarity found to protein:**  
 [8484]NP\_003605 | Homo sapiens | galanin receptor 3 | length=368 aa

**GO annotation:**  
 0.000000 plasma membrane (component)  
 0.000000 feeding behavior (process)  
 0.000000 galanin receptor (function)  
 0.000232 learning and/or memory (process)  
 0.000152 synaptic transmission (process)  
 0.000000 integral membrane protein (component)  
 0.000520 negative regulation of adenylate cyclase activity (process)  
 0.000000 neuropeptide signaling pathway (process)

**BLASTX hits:**  
 E-value 1e-11 prot=18..330 DNA=142429946..142429071 fr=94/326 per=(28%) [EST] [U]

**IMB 2006**  
 Fortaleza, Brazil  
 August 6-10, 2006

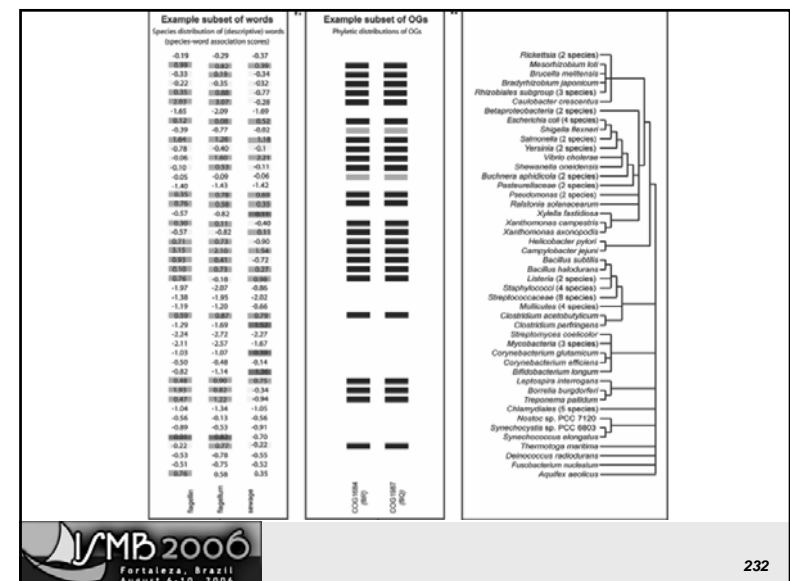
230

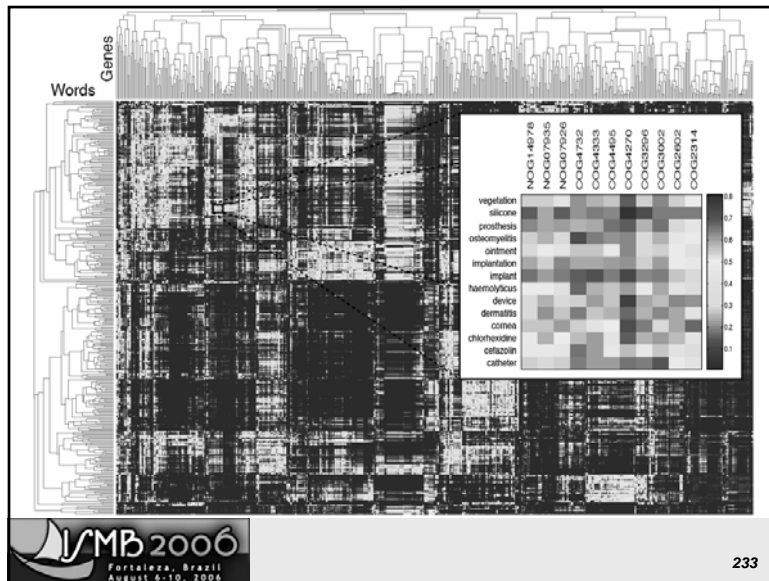
## Genotype-phenotype

- Genes and traits can be linked through similar phylogenetic profiles
  - Mainly works for prokaryotes so far
  - Traits are represented by keywords
- Finding the phylogenetic profiles
  - Gene profiles stem from sequence similarity
  - Keyword profiles are based co-occurrence with the species name in MEDLINE

**IMB 2006**  
 Fortaleza, Brazil  
 August 6-10, 2006

231





## Annotation

- Finding keywords for a group of genes
  - ER is used to find associated abstracts
  - The frequency of each word is counted
  - Background frequencies are recorded
  - A statistical test is used to rank the words
- The same strategy can be used to find MeSH terms related to a gene cluster



234

## Summary

- Information extraction
  - Co-occurrence methods generally give better recall but worse accuracy than NLP methods
  - Only NLP can handle directed interactions
- Text/data mining
  - New relations can be found from text alone
  - Methods that combine text and other data types have much better discovery potential



235



## Outlook

Lars Juhl Jensen  
EMBL, Germany  
jensen@embl.de

IAMB 2006 - Fortaleza, Brazil - August 6-10, 2006

## Necessity

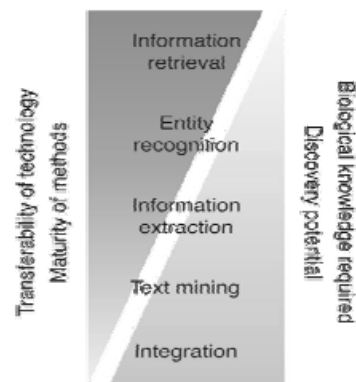
- Literature mining will remain important
  - Repositories are always made too late
  - There will always be new types of relations
  - Semantically tagged XML may replace ER
  - But no one will ever tag everything!
- Specific IE problems will become obsolete
  - Protein function and physical interactions

## Permission

- Open access
  - Literature mining methods cannot work on text unless it is accessible
  - Restricted access is now the limiting factor
- Standard formats
  - Getting the text out of a PDF file is not trivial
- Where do I get all the patent text?!

## Innovation

- The tools are in place for IR, ER, and IE
- Text- and data-mining
  - Biologists are needed
  - Work with linguists
- Lack of innovation
  - Combine text and data



## Acknowledgments

### EML Research

- Jasmin Saric
- Isabel Rojas

### EMBL Heidelberg

- Peer Bork
- Miguel Andrade
- Rossitza Ouzounova
- Michael Kuhn
- Jan Korb
- Tobias Doerks