Review

Text-mining and information-retrieval services for molecular biology

Martin Krallinger and Alfonso Valencia

Address: Protein Design Group, National Center of Biotechnology, CNB-CSIC, Cantoblanco, E-28049 Madrid, Spain.

Correspondence: Martin Krallinger. E-mail: martink@cnb.uam.es. Alfonso Valencia. E-mail: valencia@cnb.uam.es

Published: 28 June 2005

Genome ***Biology*** 2005, **6:**224 (doi:10.1186/gb-2005-6-7-224)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/7/224

Abstract

Text-mining in molecular biology - defined as the automatic extraction of information about genes, proteins and their functional relationships from text documents - has emerged as a hybrid discipline on the edges of the fields of information science, bioinformatics and computational linguistics. A range of text-mining applications have been developed recently that will improve access to knowledge for biologists and database annotators.

The use of large-scale experimental techniques and bioinformatic tools has increased the pace at which biologists produce relevant information. This also promotes the growth of the scientific literature, which contains information on those experimental results in the form of free text that is structured in a way that makes it straightforward for humans to read but more difficult for computers to interpret automatically. As a consequence, there is increasing interest in methods that can handle collections of biological texts. Such methods include systems that efficiently retrieve and classify documents in response to complex user queries, and beyond this, systems that carry out a deeper analysis of the literature to extract specific associations, such as protein-protein interactions and protein functions. This deeper analysis is called text-mining. The complex and concise nature of the scientific literature means that the use of text-mining tools developed for generic texts is often impractical; a set of freely available text-mining applications adapted to the needs of biology have been developed, however, and some of them are now available for practical use. In parallel, a number of strategies for evaluating text-mining applications have appeared, with the goal of assessing and improving the field by providing datasets that can be used for training and testing applications.

Finding relevant articles

Throughout the last decade, the amount of electronically accessible textual material has been growing exponentially. Internet-based technologies exploit the availability of these large collections of documents for the development of information-retrieval systems. Currently, biologists and bioinformaticians take advantage of those tools, not only when searching generic documents such as news articles using search engines such as Alta Vista [1] and Google [2], but especially when querying publications specific to biomedicine, for example those stored in PubMed [3,4]. The range of community-wide genome projects, for which Internet-based information exchange is crucial, together with the heavy use of biology databases through web-based tools, means that natural language processing (NLP) techniques could be useful. NLP is based on the use of computers to process language, and it includes techniques developed to provide the basic methodology required for automatically extracting relevant functional information from unstructured data, such as scientific publications. Information retrieval and NLP systems are soon likely to become important not only for extracting information but also for assisting in various aspects of research such as the discovery of new facts, the interpretation of findings, and the design of experiments.

One of the first steps when handling textual data is the extraction of relevant documents from a large collection. This process is commonly known as information retrieval. In the case of indexed web pages, powerful search engines such as Google [2] return a ranked list of documents relevant to a given user search. There are two basic search strategies: query-based and document-based searches. In query-based searches, documents are returned that contain certain user-specified combinations of keywords. As some words - 'stop words' such as 'and', 'if' and 'the' - are found at a high frequency within most documents and thus display a low information content, they are often excluded during the retrieval process. Keywords may be combined by Boolean operators, such as AND, OR and NOT. The second type of retrieval, document-based searching, aims to return a ranked list of documents similar to a given query document as a whole, rather than to a combination of a few keywords. The most widely used retrieval tool in molecular biology is Entrez [3,4], the PubMed information retrieval system provided at the US National Center for Biotechnology Information (NCBI) [5]. It supports basic keyword and Boolean query-based searches, as well as document-based searches to return all abstracts that are similar to a given document. The popular search engine Google [2] has recently incorporated a search tool specific to the academic literature, Google Scholar [6,7], for the retrieval of scientific articles, reports and books. The ranking of the returned hits is mainly based on the extent to which documents are connected by citations and web links. Other scientific literature databases and search engines include Crossref Search [8], which enables searches of the full content provided by a set of publishers, and the Nature Publishing Group search engine [9], which allows advanced search strategies.

Although these tools are useful for many tasks, it is time-consuming to use them for efficient searches and article selection, and such functions must be repeated periodically to keep knowledge up-to-date. As PubMed already contains over 15 million citations of biomedical articles [4] and is steadily growing (more than 450,000 articles are added every year [10]), services that periodically retrieve relevant articles and automatically alert the user have been implemented. Among those systems, known as selective dissemination of information (SDI) services, are My NCBI (formerly PubMed Cubby) [4,11], BioMail [12] and PubCrawler [13,14] (these and other services described in this article are listed in Table 1). These, together with some commercial tools, have been evaluated independently [15], showing that the combined use of different SDI systems results in useful automated searching.

The first step in text-mining: identification of biological entities

Biological research is name-centered: proteins are referred to in free text by their names or symbols rather than using the unambiguous identifiers provided by annotation databases (such as SwissProt accession numbers [16]). Identifying mentions of proteins and genes unambiguously within free text is a fundamental step for the later extraction of functional attributes of these entities. Unfortunately this is a difficult process, partly because of the complex nature and usage of gene and protein names. Genes and proteins may be referred to in free text in a range of different ways: as full names (for example, porin), as symbols (the *Saccharomyces cerevisiae* gene *POR1*), and also through typographical variants (*POR-1*). Many genes also have several synonyms (such as *OMP2* for *POR1*), or the gene name may be ambiguous [17] and refer to words that also have a different meanings depending on the context (for example, *big brain*, the full name for the *Drosophila melanogaster* gene *bib*, could also be an anatomical description). Furthermore, it has been suggested that errors in gene names might be introduced automatically by certain applications in bioinformatics [18].

In the NLP field, the identification of entities in free text is known as named-entity recognition (NER). To identify biological entities such as genes, proteins and drugs automatically and unambiguously within free text, over 50 information-extraction and text-mining tools have recently been implemented, and two community-wide evaluations have been carried out [19,20]. The top left of Figure 1 shows nine existing NER applications for biology that are provided via an online server or are directly downloadable. Note that the average recovery of biological entities from free text by 15 NER tools was 80%, and the results had an accuracy of 80% [21]; these figures are significantly lower than in the case of entities found in documents from fields such as economics, which demonstrates the complex nature of protein names.

Proteins and genes are characterized within biological databases through unique identifiers; each identifier is associated with its corresponding protein or nucleotide sequence and functional descriptions. The automatic recognition of entities such as genes and proteins in free text is insufficient if it is not linked to the corresponding database identifiers. Distinguishing between the use of protein names and protein-family names constitutes a serious obstacle in the task of highlighting protein entities in free text, as text passages sometimes refer to the general properties of protein families and at other times to the properties of individual proteins.

Different research communities have addressed the issue of named-entity recognition in biology in different ways. The NLP community has typically tried to identify names by analyzing the syntactic structure of sentences, making use of information about parts of speech in a sentence and the syntactic roles of words, whereas bioinformaticians have instead explored the identification of variants of the names contained in databases, even adapting standard bioinformatics algorithms such as BLAST to the problem of protein-name identification [22]. Neither of these two strategies seems to

Table 1

Biomedical text-mining resources, servers and programs

Name	Description	URL	Published reference or URL*
Abbreviation Server	Biomedical abbreviation server	http://bionlp.stanford.edu/abbreviation/	[35]
AbGene	Protein name tagger	ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe	[29]
ABNER	Protein/Gene/DNA/RNA/cell tagger	http://www.cs.wisc.edu/~bsettles/abner/	[31]
AliasServer	Protein alias handler	http://cbi.labri.fr/outils/alias/index.php	[37]
ARGH	Biomedical acronym resolver	http://invention.swmed.edu/argh/	[88,89]
ARROWSMITH	Extended MEDLINE search tool	http://kiwi.uchicago.edu/	[84]
BioMail	PubMed updating and alerting service	http://biomail.sourceforge.net/biomail/	[12]
BioRAT	Biology information extraction tool	http://bioinf.cs.ucl.ac.uk/biorat/	[81]
BITOLA	Literature-based biomedical discovery system	http://www.mf.uni-lj.si/bitola/	[86]
Chilibot	Relationship extraction	http://www.chilibot.net	[57]
CrossRef Search	Full content search engine	http://www.crossref.org/crossrefsearch.html	[8]
GAPSCORE	Protein name tagger	http://bionlp.stanford.edu/gapscore	[23]
Geisha	Text-mining tool to assist microarray analysis	http://www.pdg.cnb.uam.es/blaschke/cgi-bin/geisha	[67]
GeneScene	Information extraction for regulatory pathways	http://genescene.arizona.edu/index.html	[59]
GOAnnotator	Annotation extraction from literature	http://xldb.fc.ul.pt/rebil/tools/goa/	[51]
Google Scholar	Scholar literature search engine	http://scholar.google.com/	[6]
iHOP	Information on hyperlinked proteins	http://www.pdg.cnb.uam.es/UniPub/iHOP/	[40]
iProLINK	Protein annotation and tagging	http://pir.georgetown.edu/iprolink	[55]
KAT	Annotate proteins from scientific references	http://www.bork.embl-heidelberg.de/kat/	[52]
KeX	Protein name tagger	http://www.hgc.jp/service/tooldoc/KeX	[33]
KinasePathway database	Tool for extraction of protein, gene and compound interactions from text	http://kinasedb.ontology.ims.u-tokyo.ac.jp	[46]
MedBlast	Document retrieval for sequences	http://medblast.sibsnet.org/	[63]
MedMiner	Extraction of sentences relevant to genes	http://discover.nci.nih.gov/textmining/main.jsp	[69]
microGENIE	Text-mining for microarrays	http://www.cs.vu.nl/microgenie	[76]
My NCBI	PubMed updating and alerting service	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed	[11]
NDPG	Scores the literature based coherence of gene clusters	None	[66]
NLProt	Protein name tagger	http://cubic.bioc.columbia.edu/services/nlprot/	[25]
NPG search engine	Nature Publishing Group search engine	http://search.nature.com/search/?sp_a=sp1001702d&sp_t =advanced&sp_x_1=ujournal&sp-p=all&sp	[9]
PreBIND	Classifier of protein interaction documents	http://bind.ca/	[44]
PubCrawler	PubMed updating and alerting service	http://pubcrawler.gen.tcd.ie/	[13]
PubGene	Text-mining tool for microarrays	http://www.pubgene.org/	[72]
PubMatrix	Multiplex literature mining tool	http://pubmatrix.grc.nia.nih.gov/secure-bin/index.pl	[74]
PubMed Entrez	Biomedical citation retrieval system	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed	[3]
Relationship Extractor	Biomedical relationship extractor	http://www-personal.engin.umich.edu/~murthyr/ Relationship_Extractor.html	[90]
SAWTED	Text-enhanced remote homolog detector	http://www.sbg.bio.ic.ac.uk/~sawted/	[61]
Scopus	Scientific literature database and search	http://www.scopus.com/scopus/home.url	[93]
Textpresso	*C. elegans* literature information retrieval and extraction tool	http://www.textpresso.org/	[48]
XplorMed	Explores bibliographic MEDLINE searches	http://www.bork.embl-heidelberg.de/xplormed	[91]
Yapex	Protein name tagger	http://ellis.sics.se:8080/cgi-bin/Yapex/yapex.cgi	[27]

An overview of some of the available text-mining, information-extraction, information-retrieval and selective dissemination of information services currently available. *References to articles describing each tool are given; where no article has been published, the reference is to the URL.

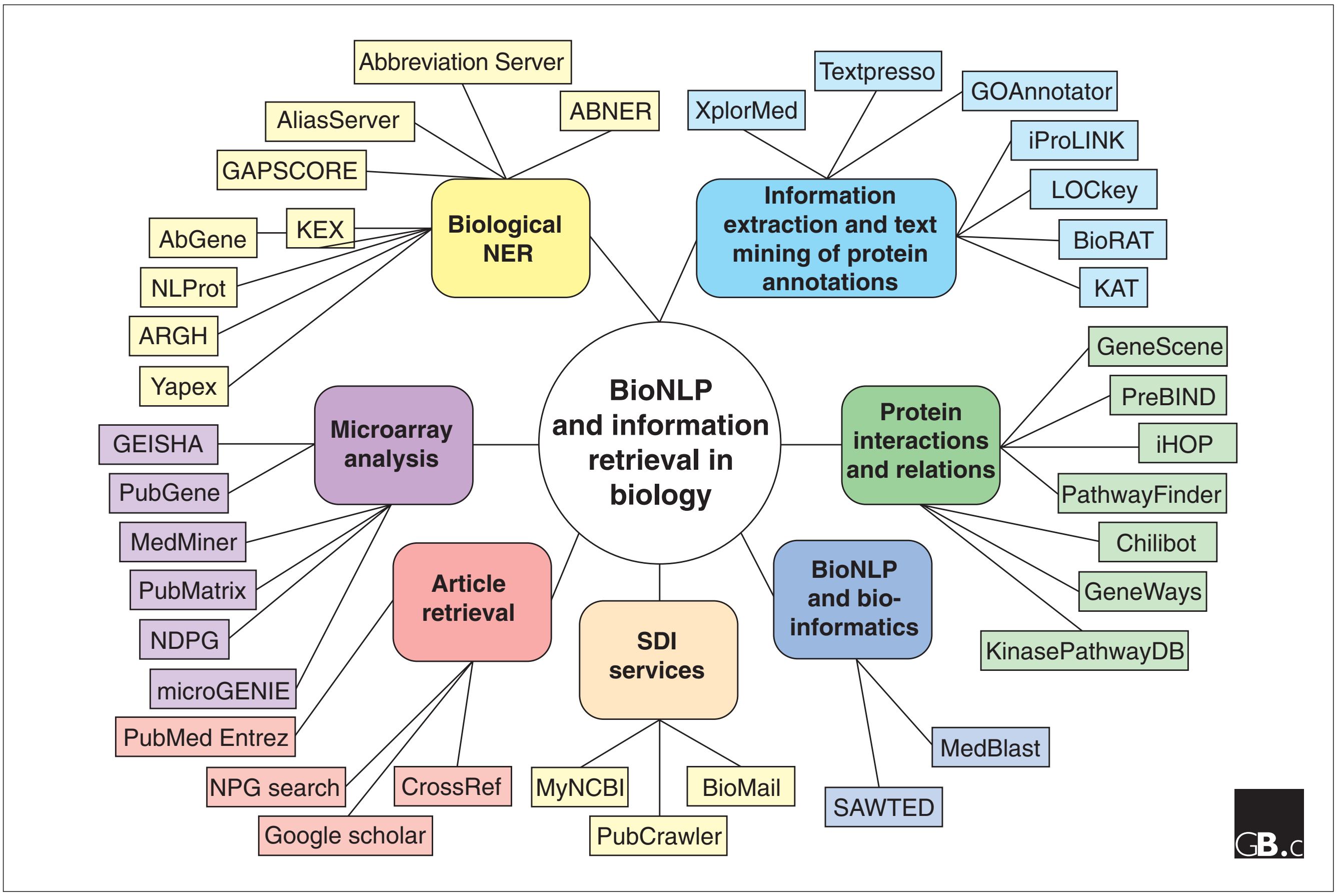


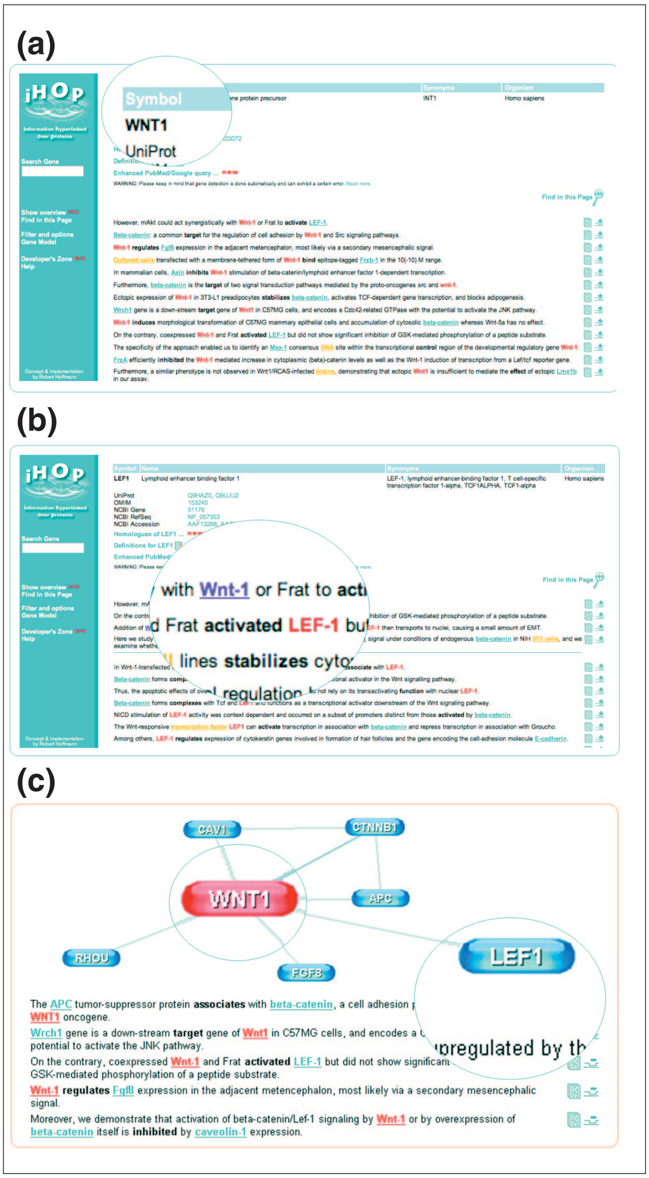
Figure 1
An overview of biological natural language processing (BioNLP) and text-mining applications for biology. The major topics are represented by the inner circle of seven approaches, and the corresponding applications are given in the outer layers of boxes. Most of the tools are available online or for download. Some applications could be classified into multiple topics; they are shown here associated with one of their most significant topics. For instance, most of the text-mining applications (that is, the applications that are not simply for article retrieval) have integrated modules for named entity recognition (NER), and selective dissemination of information (SDI) services often use automated Boolean queries for article retrieval. References and URLs for each application, where available, are given in Table 1.

be efficient by itself, and many intermediate combinations are therefore appearing, including the following examples. GAPSCORE [23,24] is an easy-to-use online tool for detecting protein and gene names within free text (a 'protein tagger'). The text to be analyzed can be pasted into an online form and submitted to the server, which returns a list of the words observed in the document and a statistical quality score that indicates how probable it is that the each word represents a gene or protein name. Another online protein tagger is NLProt, developed at Columbia University [25,26]. NLProt is based on a machine learning technique called support vector machines (SVMs) and allows protein identification either in a submitted text or in the text corresponding to a list of submitted PubMed article identifiers. Additional protein taggers include Yapex [27,28], also available online, and three downloadable tools, AbGene [29,30], ABNER [31,32] and KEX [33,34]. Abbreviations or acronyms are often used as a shorter form to refer to gene names in articles; the Abbreviation Server [35,36] developed at Stanford University allows a similar search strategy to that used by GAPSCORE to be applied to biomedical abbreviations such as gene symbols. Finally, the AliasServer [37,38] helps in linking the various aliases of a given gene through different biological databases for various species.

One of the main challenges when linking protein names to database entries is distinguising between proteins that have the same names but belong to different genomes - a process called inter-species gene disambiguation. This is especially cumbersome in the case of mouse and human genes; the same gene symbol is often used in both species and both names are often mentioned in the same textual passage. The complex nature of protein- and gene-name identification is reinforced further by the dynamic nature of gene-name usage and name creation, with official gene names being changed and new synonyms being created [39]; it is clear that static approaches and dictionaries will not be sufficient for solving the problem.

One step further: mining interactions and relations

Although the identification of biological entities is a crucial step, in practice it is the extraction of associations between proteins and their functional features that poses an interesting biological problem. Several systems have been constructed for extracting annotations of genes and proteins automatically and for detecting protein-protein interactions and regulatory pathways. Protein-protein interactions have attracted particular interest in the light of recent developments in high-throughput proteomics. One system that extracts annotations and detects interactions is the iHOP system that we have implemented at the Spanish National Biotechnology Center [40]. This facilitates the direct linking of information in the INTACT [41] protein-interaction database with corresponding bibliographic references (Figure 2). As well as highlighting direct associations between genes and functional descriptions, iHOP also includes advanced search modes for discovery and visualization of literature-based protein-interaction networks for a range of organisms, including human, mouse and yeast [42]. The basic approach followed by iHOP is protein-centric: it arranges relevant sentences from the literature around protein names, and the use of co-citation of protein names in each sentence facilitates navigation through the dispersed literature relevant to a particular protein. As a result, users can successively explore the functions of related proteins by building virtual protein-relation networks (Figure 2c). The iHOP system is based on the ideas previously developed for the SUISEKI knowledge-discovery system [43].



Symbol		Synonyms	Organism
WNT1	ne protein precursor	INT1	Homo sapiens

Definition
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

However, mAkt could act synergistically with Wnt-1 or Frat to **activate** LEF-1.

Beta-catenin: a common **target** for the regulation of cell adhesion by Wnt-1 and Src signaling pathways.

Wnt-1 **regulates** Fgf8 expression in the adjacent metencephalon, most likely via a secondary mesencephalic signal.

Cultured cells transfected with a membrane-tethered form of Wnt-1 **bind** epitope-tagged Frzb-1 in the 10(-10) M range.

In mammalian cells, Axin **inhibits** Wnt-1 stimulation of beta-catenin/lymphoid enhancer factor 1-dependent transcription.

Furthermore, beta-catenin is the **target** of two signal transduction pathways mediated by the proto-oncogenes src and wnt-1.

Ectopic expression of Wnt-1 in 3T3-L1 preadipocytes **stabilizes** beta-catenin, activates TCF-dependent gene transcription, and blocks adipogenesis.

Wrch1 gene is a down-stream **target** gene of Wnt1 in C57MG cells, and encodes a Cdc42-related GTPase with the potential to activate the JNK pathway.

Wnt-1 **induces** morphological transformation of C57MG mammary epithelial cells and accumulation of cytosolic beta-catenin whereas Wnt-5a has no effect.

On the contrary, coexpressed Wnt-1 and Frat **activated** LEF-1 but did not show significant inhibition of GSK-mediated phosphorylation of a peptide substrate.

The specificity of the approach enabled us to identify an Msx-1 consensus DNA site within the transcriptional **control** region of the developmental regulatory gene Wnt-1.

FrzA efficiently **inhibited** the Wnt-1 mediated increase in cytoplasmic (beta)-catenin levels as well as the Wnt-1 induction of transcription from a Lef/tcf reporter gene.

Furthermore, a similar phenotype is not observed in Wnt1/RCAS-infected brains, demonstrating that ectopic Wnt1 is insufficient to mediate the **effect** of ectopic Lmx1b in our assay.

Symbol	Name	Synonyms	Organism
LEF1	Lymphoid enhancer binding factor 1	LEF-1, lymphoid enhancer-binding factor 1, T cell-specific transcription factor 1-alpha, TCF1ALPHA, TCF1-alpha	Homo sapiens

UniProt	Q9HAZ0, Q9UJU2
OMIM	153245
NCBI Gene	51176
NCBI RefSeq	NP_057353
NCBI Accession	AAF13268, AA

Homologues of LEF1 ... new
Definitions for LEF1
Enhanced PubMed
WARNING: Please ke d more.

Find in this Page

with Wnt-1 or Frat to acti
d Frat activated LEF-1 but
ll lines **stabilizes** cyto
l regulation

However, mA

On the contr hibition of GSK-mediated phosphorylation of a peptide substrate.

Addition of w EF-1 then transports to nuclei, causing a small amount of EMT.

Here we study signal under conditions of endogenous beta-catenin in NIH 3T3 cells, and we examine whethe

In Wnt-1-transfected ssociate with LEF-1.

Beta-catenin forms comp ional activator in the Wnt signalling pathway.

Thus, the apoptotic effects of ove o not rely on its transactivating **function** with nuclear LEF-1.

Beta-catenin forms **complexes** with Tcf and and functions as a transcriptional activator downstream of the Wnt signaling pathway.

NICD stimulation of LEF-1 activity was context dependent and occurred on a subset of promoters distinct from those **activated** by beta-catenin.

The Wnt-responsive transcription factor LEF1 can **activate** transcription in association with beta-catenin and repress transcription in association with Groucho.

Among others, LEF-1 **regulates** expression of cytokeratin genes involved in formation of hair follicles and the gene encoding the cell-adhesion molecule E-cadherin.

The APC tumor-suppressor protein **associates** with beta-catenin, a cell adhesion p WNT1 oncogene.

Wrch1 gene is a down-stream **target** gene of Wnt1 in C57MG cells, and encodes a potential to activate the JNK pathway.

upregulated by th

On the contrary, coexpressed Wnt-1 and Frat **activated** LEF-1 but did not show significant GSK-mediated phosphorylation of a peptide substrate.

Wnt-1 **regulates** Fgf8 expression in the adjacent metencephalon, most likely via a secondary mesencephalic signal.

Moreover, we demonstrate that activation of beta-catenin/Lef-1 signaling by Wnt-1 or by overexpression of beta-catenin itself is **inhibited** by caveolin-1 expression.

Figure 2
Basic steps in the use of the iHOP text-mining tool [40], illustrated with screenshots [42]. For a given query (for example, the protein symbols **(a)** Wnt-1 or **(b)** LEF-1), all the sentences mentioning the name are retrieved from PubMed. These sentences also contain mentions of other proteins, which are highlighted and which might show associations with the query protein (see the magnified area in (b)). Functional terms (such as 'target' and 'complexes' and interaction verbs (such as 'activated' and 'stabilizes') are in bold. **(c)** By clicking on the 'Gene model' link in the left panel in (a,b), interaction networks of proteins that co-occur in sentences with the query proteins can be displayed.

Some other text-mining applications include PreBIND [44,45], developed to assist in the extraction of protein-protein interactions; the KinasePathway database text-mining system, which extracts interactions between proteins, genes and compounds [46,47]; and Textpresso [48,49], an information-retrieval and extraction tool developed for the *Caenorhabditis elegans* literature in the context of the model-organism database WormBase [50]. Textpresso defines 33 categories of word describing entities or relationships - such as genes, pathways, or regulation - and integrates this 'Textpresso Ontology' with a text-mining system for searching the *C. elegans* literature. Among the text-mining services available online that focus on automatic annotation extraction are GOAnnotator, which provides associations between protein names and Gene Ontology terms [51]; KAT [52,53], a system for deriving terms relevant to annotations such as SwissProt keywords and Gene Ontology terms [54] from PubMed abstracts for a given query protein; and the iProLINK tool [55,56], which performs automated extraction of annotations for given protein names and provides information related to the organisms in which proteins are found and the protein families of which they are members. Figure 1 and Table 1 provide an overview of the different systems currently available.

A system with a special focus on the extraction of relationships between genes, proteins and other information is Chilibot ([57,58]; user registration is required before running queries); it allows searches using gene symbols and keywords, and the color-coded output provides information about gene-expression levels when available. The extraction of complex relationships can be handled by GeneScene [59,60], a toolkit that provides visualization and navigation facilities for exploring regulatory networks; the tool currently provides information only on the literature on yeast and on the p53 tumor suppressor and the AP1 transcription factor.

Some attempts have been made to merge text-mining methods and bioinformatic methods involving sequence analysis into a single system. The integration of functional information extracted by NLP algorithms with standard bioinformatic methods such as sequence-comparison techniques has been exploited by the Structure Assignment With Text Description (SAWTED) system [61,62], which can be tested online. It combines a document-comparison algorithm called a 'vector-cosine model' with the PSI-BLAST sequence retrieval method, which is especially useful for detecting sequences that are distantly related. Another strategy that makes use of sequence information and free text is MedBlast [63,64]; using the web-based interface of MedBlast, for a given query sequence and optional additional keywords the system returns articles related to the protein corresponding to the query sequence.

Text-mining and large gene collections

Technical advances in molecular biology mean that large collections of genes are nowadays often studied simultaneously using genomic approaches. Using conventional information retrieval to link these genes with the associated literature is not efficient, and a large list of irrelevant documents can be returned. For example, microarray experiments result in groups of genes with particular expression patterns; to interpret these groups in terms of the underlying biological meaning, information is needed not only on each individual gene but also on commonalities among the whole group. The functional information is commonly extracted from databases such as SwissProt [16] or GO [65], which in turn are nourished by extracting relevant functional features from the literature.

A number of text-mining methods have been developed for linking groups of genes found in microarrays and other experiments directly and automatically with information contained in biomedical article databases. The neighbor divergence per gene (NDPG) approach [66] uses the literature to score the functional coherence of gene clusters. GEISHA [67,68] automatically mines the literature for functional terms associated with gene groups and carries out a statistical analysis of the significance of those terms. Among the available online tools for assisting in interpreting microarray data are MedMiner [69,70], which can be used to filter and organize information from free text obtained from automatic PubMed [4] and GeneCard [71] searches and PubGene [72,73] which has additional visualization capabilities for displaying network information and pathway mapping. The analysis of frequency matrices of term co-occurrences of two lists of keywords is the basis of the PubMatrix system [74,75], which can be used online after registering. Finally, microGENIE [76] enables semi-automatic queries of very large collections of genes (UniGene and SwissProt gene names and GenBank accession numbers) in PubMed to speed up the retrieval of relevant articles. It is important to realize that existing text-mining technologies in biology are focused on identification and linking of functional information of proteins in free text, they are currently not providing automatically generated summaries of biologically relevant information.

Towards knowledge discovery

The field of 'BioNLP' - text-mining and information extraction for molecular biology - is very recent, but the existing applications are improving steadily. This is partly because of newly available resources, such as collections of annotated documents suitable for training new systems (for example, the GENIA [77] corpus and the BioCreative [19] corpus). The improvement also reflects the effect of community-wide assessments such as the BioCreative contest [19] and the KDD challenge cup [78], which enable evaluation of the efficiency of different methodologies, and the genomics track of the Text Retrieval Conference (TREC) workshops [79,80], a forum for developing solutions to information-retrieval and document-classification tasks in biology. The development of controlled, computer-readable vocabularies (ontologies), dictionaries, and functional keywords (Gene Ontology concepts [54] and SwissProt keywords [16]) defining relevant biological aspects of proteins have also been valuable for text-mining tools. Because of the restricted availability of full-text articles most of the existing text-mining systems for biology are centered on the analysis of abstracts, but changes in publishing policy and increasing access to repositories of whole articles make mining of full text a likely development in the near future. Some initiatives in this direction have been started already, for example the BioRAT system [81,82], which processes full-text articles so as to identify target facts.

Perhaps the most likely future developments will be the construction of networks and interactions for discovering new relationships through intermediate entities, followed by the proposal of new functions - this process is referred to as 'knowledge discovery'. Several exploratory attempts have been made to develop knowledge-discovery systems, but they are not yet of general practical use. Our SUISEKI system [83], for instance, extracts indirect relationships between proteins through associations with intermediate proteins in text. Two online tools that directly address the difficulty of making knowledge-discovery practical are ARROWSMITH [84,85] and BITOLA [86,87]. ARROWSMITH [84,85] aims to discover indirect relations between two entities that are not directly connected in the literature; the indirect relationship can be a substance or disease condition. BITOLA [86,87] is a biomedical discovery-support system with a focus on the discovery of disease candidate genes, taking advantage of Medical Subject Heading (MeSH) terms.

Undoubtedly, the development of text-mining applications specific for biology is the only way to cope with the increasing amount of free textual data produced in this field. The increasing interest of users in efficiently retrieving and extracting relevant information, the need to keep up with new discoveries described in the literature or in biological databases, and the demands posed by the analysis of high-throughput experiments, are the underlying forces motivating the development of text-mining applications in molecular biology. Those technologies should provide the foundation for future knowledge-discovery tools able to identify previously undiscovered associations, something that will assist in the formulation of models of biological systems.

Acknowledgements

The work of our group was supported by grants from the European Commission (ORIEL IST-2001-32688, TEMBLOR QLRT-2001-00015, Biosapiens LSHC-CT-2003-505265). We thank Robert Hoffmann for providing Figure 2 and Christian Blaschke, as well as all the members of the group, for interesting discussions.

References

1. **Altavista** [http://www.altavista.com]
2. **Google** [http://www.google.com]
3. Schuler G, Epstein J, Ohkawa H, Kans J: **Entrez: molecular biology database and retrieval system.** *Methods Enzymol* 1996, **266:**141-162.
4. **Entrez PubMed** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed]
5. Wheeler D, Church D, Federhen S, Lash A, Madden T, Pontius J, Schuler G, Schriml L, Sequeira E, Tatusova T, Wagner L: **Database resources of the National Center for Biotechnology.** *Nucleic Acids Res* 2003, **31:**28-33.
6. Editorial: **The ultimate search engine?** *Nat Cell Biol* 2005, **7:**1.
7. **Google Scholar** [http://scholar.google.com]
8. **CrossRef Search, publisher pilot for full-text scholarly research** [http://www.crossref.org/crossrefsearch.html]
9. **Nature Publishing Group search engine** [http://search.nature.com/search/?sp_a=sp1001702d&sp_t=advanced&sp_x_1=ujournal&sp-p=all&sp]
10. Staab S, Blaschke C, Nedellec C, Park J, Schatz B, Valencia A, Bernardi L, Ratsch E, Kania R, Saric J, Rojas I, Staab S: **Mining information for functional genomics.** *IEEE Intelligent Systems* 2002, **17:**66-80.
11. Knecht L, Shooshan S: **Internet Grateful Med to be retired; reminder of NLM Gateway availability.** *NLM Tech Bull* 2001, **318:**e3.
12. **Biomail** [http://biomail.sourceforge.net/biomail]
13. Hokamp K, Wolfe K: **PubCrawler: keeping up comfortably with PubMed and GenBank.** *Nucleic Acids Res* 2004, **32:**W16-W19.
14. **PubCrawler** [http://pubcrawler.gen.tcd.ie/]
15. Shultz M, DeGroote S: **MEDLINE SDI services: how do they compare?** *J Med Libr Assoc* 2003, **91:**460-467.
16. **Expasy - SwissProt and TrEMBL** [http://us.expasy.org/sprot]
17. Chen L, Liu H, Friedman C: **Gene name ambiguity of eukaryotic nomenclatures.** *Bioinformatics* 2005, **21:**248-256.
18. Zeeberg B, Riss J, Kane D, Bussey K, Uchio E, Linehan W, Barrett J, Weinstein J: **Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics.** *BMC Bioinformatics* 2004, **5:**80.
19. Hirschman L, Yeh A, Blaschke C, Valencia A: **Overview of BioCreAtIvE: critical assessment of information extraction for biology.** *BMC Bioinformatics* 2005, **6(Suppl 1):**S1.
20. Kim J, Ohta T, Tsuruoka Y, Tateisi Y: **Introduction to the bio-entity recognition task at JNLPBA.** In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications* 28-29 August 2004; Geneva. 70-76. [http://www.genisis.ch/~natlang/JNLPBA04/JNLPBA.final.pdf]
21. Yeh A, Morgan A, Colosimo M, Hirschman L: **BioCreAtIvE task 1A: gene mention finding evaluation.** *BMC Bioinformatics* 2005, **6(Suppl 1):**S2.
22. Krauthammer M, Rzhetsky A, Morozov P, Friedman C: **Using BLAST for identifying gene and protein names in journal articles.** *Gene* 2000, **259:**245-252.
23. Chang J, Schutze H, Altman R: **GAPSCORE: finding gene and protein names one word at a time.** *Bioinformatics.* 2004, **20:**216-225.
24. **Gene and Protein Name Server** [http://bionlp.stanford.edu/gapscore]
25. Mika S, Rost B: **NLProt: extracting protein names and sequences from papers.** *Nucleic Acids Res* 2004, **32:**W634-W637.
26. **CUBIC: NLProt/Index** [http://cubic.bioc.columbia.edu/services/nlprot]
27. Franzen K, Eriksson G, Olsson F, Asker L, Liden P, Coster J: **Protein names and how to find them.** *Int J Med Inform* 2002, **67:**49-61.
28. **Yapex** [http://ellis.sics.se:8080/cgi-bin/Yapex/yapex.cgi]
29. Tanabe L, Wilbur W: **Tagging gene and protein names in biomedical text.** *Bioinformatics* 2002, **18:**1124-1132.
30. **AbGene** [ftp://ftp.ncbi.nlm.nih.gov/pub/tanabe]
31. Settles B: **Biomedical named entity recognition using conditional random fields and rich feature sets.** *Proc NLPBA/COLING 2004.* 2004.
32. **ABNER: a biomedical named entity recognizer** [http://www.cs.wisc.edu/~bsettles/abner]
33. Fukuda K, Tsunoda T, Tamura A, Takagi T: **Toward information extraction: identifying protein names from biological papers.** *Pac Symp Biocomput* 1998, **3:**707-718.
34. **KeX** [http://www.hgc.jp/service/tooldoc/KeX]
35. Chang J, Schuetze H, Altman R: **Creating an online dictionary of abbreviations from MEDLINE.** *J Am Med Inform Assoc* 2002, **9:**612-620.
36. **Biomedical Abbreviation Server** [http://bionlp.stanford.edu/abbreviation]
37. Iragne F, Barre A, Goffard N, DeDaruvar A: **AliasServer: a web server to handle multiple aliases used to refer to proteins.** *Bioinformatics* 2004, **20:**2331-2332.
38. **AliasServer** [http://cbi.labri.fr/outils/alias/index.php]
39. Hoffmann R, Valencia A: **Life cycles of successful genes.** *Trends Genet* 2003, **19:**79-81.
40. Hoffmann R, Valencia A: **A gene network for navigating the literature.** *Nat Genet.* 2004, **36:**664.
41. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, *et al.*: **IntAct: an open source molecular interaction database.** *Nucleic Acids Res* 2004, **32(Database issue):**D452-D455.
42. **Information hyperlinked over proteins (iHOP)** [http://www.pdg.cnb.uam.es/UniPub/iHOP]
43. Blaschke C, Valencia A: **The frame-based module of the Suiseki information extraction system.** *IEEE Intelligent Systems* 2002, **17:**14-20.
44. Donaldson I, Martin J, deBruijn B, Wolting C, Lay V, Tuekam B, Zhang S, Baskin B, Bader G, Michalickova K, *et al.*: **PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine.** *BMC Bioinformatics* 2003, **4:**11.
45. **BIND - The Biomolecular Interaction Network** [http://bind.ca]
46. Koike A, Kobayashi Y, Takagi T: **Kinase pathway database: an integrated protein-kinase and NLP-based protein-interaction resource.** *Genome Res* 2003, **13:**1231-1243.
47. **Kinase Pathway database** [http://kinasedb.ontology.ims.u-tokyo.ac.jp]
48. Muller H, Kenny E, Sternberg P: **Textpresso: an ontology-based information retrieval and extraction system for biological literature.** *PLoS Biol* 2004, **2:**e309.
49. **Textpresso** [http://www.textpresso.org]

50. **Wormbase** [http://www.wormbase.org]
51. **GOAnnotator** [http://xldb.fc.ul.pt/rebil/tools/goa]
52. Perez A, Perez-Iratxeta C, Bork P, Thode G, Andrade M: **Gene annotation from scientific literature using mappings between keyword systems.** *Bioinformatics* 2004, **20:**2084-2091.
53. **KAT** [http://www.bork.embl-heidelberg.de/kat]
54. **An Introduction to the Gene Ontology** [http://www.geneontology.org/GO.doc.shtml]
55. Hu Z, Mani I, Hermoso V, Liu H, Wu C: **iProLINK: an integrated protein resource for literature mining.** *Comput Biol Chem* 2004, **28:**409-416.
56. **iProLINK** [http://pir.georgetown.edu/iprolink]
57. Che H, Sharp B: **Content-rich biological network constructed by mining PubMed abstracts.** *BMC Bioinformatics* 2004, **5:**147.
58. **Chilibot** [http://www.chilibot.net]
59. Leroy G, Chen H: **Filling preposition-based templates to capture information from medical abstracts.** *Pac Symp Biocomput* 2002, **7:**350-361.
60. **GeneScene** [http://genescene.arizona.edu/index.html]
61. MacCallum R, Kelley L, Sternberg M: **SAWTED: structure assignment with text description-enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons.** *Bioinformatics* 2000, **16:**125-129.
62. **SAWTED** [http://www.sbg.bio.ic.ac.uk/~sawted]
63. Tu Q, Tang H, Ding D: **MedBlast: searching articles related to a biological sequence.** *Bioinformatics* 2004, **20:**75-77.
64. **MedBlast** [http://medblast.sibsnet.org]
65. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.** *Bioinformatics* 2004, **20:**578-580.
66. Raychaudhuri S, Altman R: **A literature-based method for assessing the functional coherence of a gene group.** *Bioinformatics* 2003, **19:**396-401.
67. Oliveros J, Blaschke C, Herrero J, Dopazo J, Valencia A: **Expression profiles and biological function.** *Genome Inform Ser Workshop Genome Inform* 2000, **11:**106-117.
68. **DNA Array Analysis with Geisha** [http://www.pdg.cnb.uam.es/blaschke/cgi-bin/geisha]
69. Tanabe L, Scherf U, Smith L, Lee J, Hunter L, Weinstein J: **MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling.** *Biotechniques* 1999, **27:**1210-1217.
70. **MedMiner** [http://discover.nci.nih.gov/textmining/main.jsp]
71. **GeneCards** [http://bioinformatics.weizmann.ac.il/cards]
72. Jenssen T, Laegreid A, Komorowski J, Hovig E: **A literature network of human genes for high-throughput analysis of gene expression.** *Nat Genet* 2001, **28:**21-28.
73. **PubGene** [http://www.pubgene.org]
74. Becker K, Hosack D, Dennis G, Lempicki R, Bright T, Cheadle C, Engel J: **PubMatrix: a tool for multiplex literature mining.** *BMC Bioinformatics* 2003, **4:**61.
75. **PubMatrix** [http://pubmatrix.grc.nia.nih.gov/secure-bin/index.pl]
76. **MicroGENIE** [http://www.cs.vu.nl/microgenie]
77. Kim JD, Ohta T, Tateisi Y, Tsujii J: **GENIA corpus - semantically annotated corpus for bio-textmining.** *Bioinformatics* 2003, **19:**i180-i182.
78. Yeh A, Hirschman L, Morgan A: **Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup.** *Bioinformatics* 2003, **19(Supp 11):**i331-i339.
79. Hersh W, Bhupatiraju R: **TREC GENOMICS track overview.** In *Proceedings of the Twelfth Text Retrieval Conference* 18-21 November 2003, Gaithersburg. Edited by Voorhees EM, Buckland LP. Gaithersburg: National Institute of Standards and Technology; 2003: 14-24.
80. **TREC Genomics Trach** [http://ir.ohsu.edu/genomics]
81. Corney D, Buxton BF, Langdon W, Jones D: **BioRAT: extracting biological information from full-length papers.** *Bioinformatics.* 2004, **20:**3206-3213.
82. **BioRAT** [http://bioinf.cs.ucl.ac.uk/biorat]
83. Blaschke C, Valencia A: **The potential use of SUISEKI as a protein interaction discovery tool.** *Genome Inform Ser Workshop Genome Inform.* 2001, **12:**123-134.
84. Smalheiser N, Swanson D: **Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses.** *Comput Methods Programs Biomed* 1998, **57:**149-153.
85. **ARROWSMITH** [http://kiwi.uchicago.edu/]
86. Hristovski D, Peterlin B: **Literature-based disease candidate gene discovery.** *Proceedings of Medinfo 2004*. Edited by Fieschi M. Bethesda: American Medical Informatics Association; 2004:1649.
87. **BITOLA - Biomedical Discovery Support System** [http://www.mf.uni-lj.si/bitola]
88. Wren J, Garner H: **Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries.** *Methods Inf Med* 2002, **41:**426-434.
89. **ARGH - Biomedical Acronym Resolver** [http://invention.swmed.edu/argh]
90. **Relationship Extractor** [http://www-personal.engin.umich.edu/~murthyr/Relationship_Extractor.html]
91. Perez-Iratxeta C, Bork P, Andrade M: **XplorMed: a tool for exploring MEDLINE abstracts.** *Trends Biochem Sci* 2001, **26:**573-575.
92. **XplorMed** [http://www.bork.embl-heidelberg.de/xplormed]
93. **Scopus** [http://www.scopus.com/scopus/home.url]