cnio
Centro Nacional de Investigaciones Oncológicas

BIOINFORMÁTICA Y BIOLOGÍA COMPUTACIONAL

Universidad Complutense de Madrid, Escuela de Verano

# Text mining and Information Extraction in Biomedicine

http://zope.bioinfo.cnio.es/teaching

We have here much data, and we must proceed to lay out our campaign",

Van Helsing in Bram Stockers Dracula

Martin Krallinger

# Lecture Overview

1- BACKGROUND

2- BIOMEDICAL LITERATURE

3- TEXT MINING & NATURAL LANGUAGE PROCESSING (NLP)

4- MAIN NLP TASKS

[ Break: 15 minutes ]

5- NLP IN BIOLOGY AND BIOINFORMATICS

6- EXISTING NLP APPLICATIONS

7- RESOURCES FOR BIO-NLP

8- EVALUATION OF NLP IN BIOLOGY

9- CONCLUSIONS & OUTLOOK

(ALSO SOME PRACTICAL TASKS)

# TEXT MINING AND BIOINFORMATICS

- Comparative Genomics
- Databases and Data Integration
- Evolution and Phylogeny
- Biological and medical Ontologies
- Proteomics and Genomics bioinformatics
- Sequence Analysis
- Structural Bioinformatics
- Systems Biology, networks & pathways
- Text mining, Information Extraction & Retrieval

# BIOLOGICAL DATA TYPES

- Sequence data (e.g. DNA, RNA, proteins)
- Structural data (e.g. 3D coordinates, EM )
- Trees, hierarchies and graphs (e.g. phylogeny, ontologies, PPI networks).
- Other: e.g. Microarray data, proteomics experiments,..
- **Natural language texts: annotation records, keywords, concepts in ontologies and literature.**

# TEXT MINING IMPORTANCE

- Humans exchange information using natural language
- Knowledge in Biology mainly in free text
- Most Functional information/annotations used in bioinformatics were directly or indirectly derived from the literature.
- Crucial for efficient information access of biologists and biological annotation databases
- Useful to assist in drug discovery and target selection, adverse drugs effect descriptions,..
- Rapid literature data accumulation

# PROTEIN FUNCTION DESCRIPTIONS

- Heterogeneous information types.
- Direct and indirect links to functional descriptions.
- Functional descriptions: structured annotation database records often lack contextual information.
- Free text (literature): additional contextual information, e.g. time, space, experimental conditions, in vitro/in vivo,…
- Associations of gene/proteins to Controlled vocabularies (concepts).

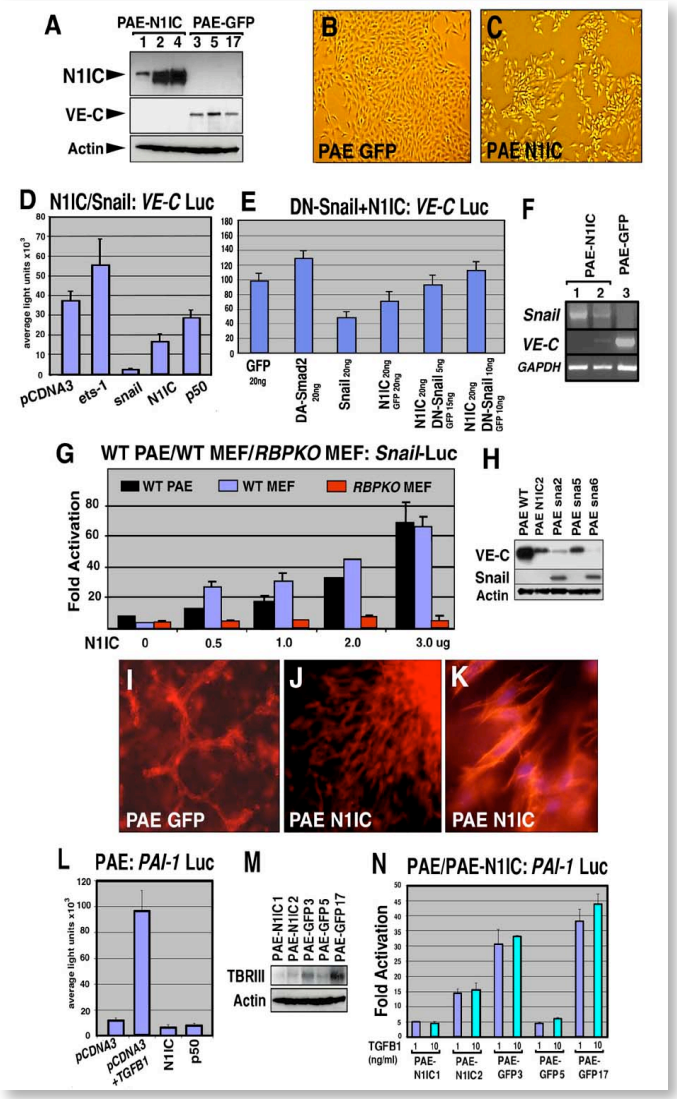
# BIOMEDICAL LITERATURE

- Base for **communicating** scientific discoveries.
- Important for competitive intelligence (CI)
- Information in form of human **natural language**.
- Base for **annotation** databases records.
- **Contextual** information of experimental results.
- Both peer-**reviewed** & community reviewed info.
- **Heavy use** of literature databases.
- Rapid **growth** of information.
- Both broad and specific **reader** communities.

# From experiments to literature (1)

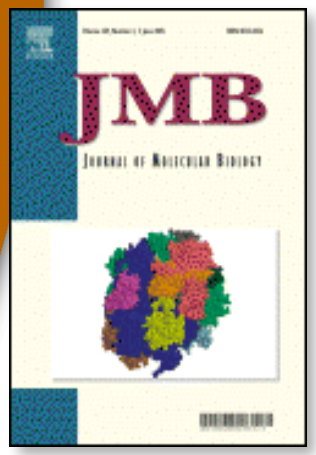
- First step: literature study
- Used for experimental planning.
- Information of experimental conditions
- Used for actual target selection.

# From experiments to literature (2)



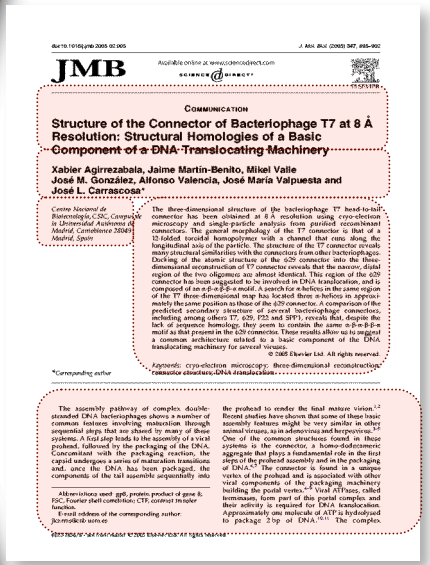

- Used for result **interpretation**.
- Used for **comparison** to other results.
- Used for **hypothesis** generation.
- **Communicate** results to scientific community.

# Data in scientific articles




## Considerations of Journal-specific characteristics:

- Journal/article Format (important for pre-processing)
- Paper structure (section types)
- Article type (review, clinical study, etc.)
- Target audience of journal/article.

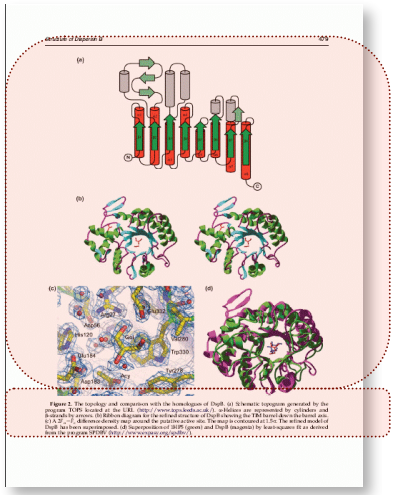



## Full text:

- Title
- Authors
- Abstract
- Text Body
- References

## Tables & table legends

## Figures & figure legends



# General characteristics of biomedical language (1)

- **Heavy use of domain specific terminology (12% biochemistry related technical terms), examples:** *chemoattractant, fibroblasts, angiogenesis*

- **Polysemic words (word sense disambiguation), examples**: *APC stands for both argon plasma coagulation and activated protein C; or teashirt can refer to a type of cloth and to a gene (tsh).*

- **Heavy use of acronyms, examples:** *Activated protein C (APC) , or vascular endothelial growth factor (VEGF)*

- **Most words with low frequency (data sparseness)**

- **Molecular Biology domain is very dynamic & poorly formalized nomenclature and terminology**

# General characteristics of biomedical language (2)

- **New names and terms created (novelty), example:**

'*This disorder maps to chromosome 7q11-21, and this locus was named* ***CLAM****.* ' *[PMID:12771259 ]*

- **Typographical variants (e.g. in writing gene names), example:** *TNF-alpha, TNF alpha, TNFalpha, TNF-a (without hyphen).*

- **Different writing styles (native languages): syntactic and semantic and word usage implications.**

- **Heavy use of referring expressions (anaphora, cataphora and ellipsis) and inference, example:**

**Glycogenin** is a *glycosyltransferase*.
**It** functions as the *autocatalytic initiator for the synthesis of glycogen in eukaryotic organisms*.

# Word usage in scientific English (1)

**Table 2** | Most frequently used words in various countries

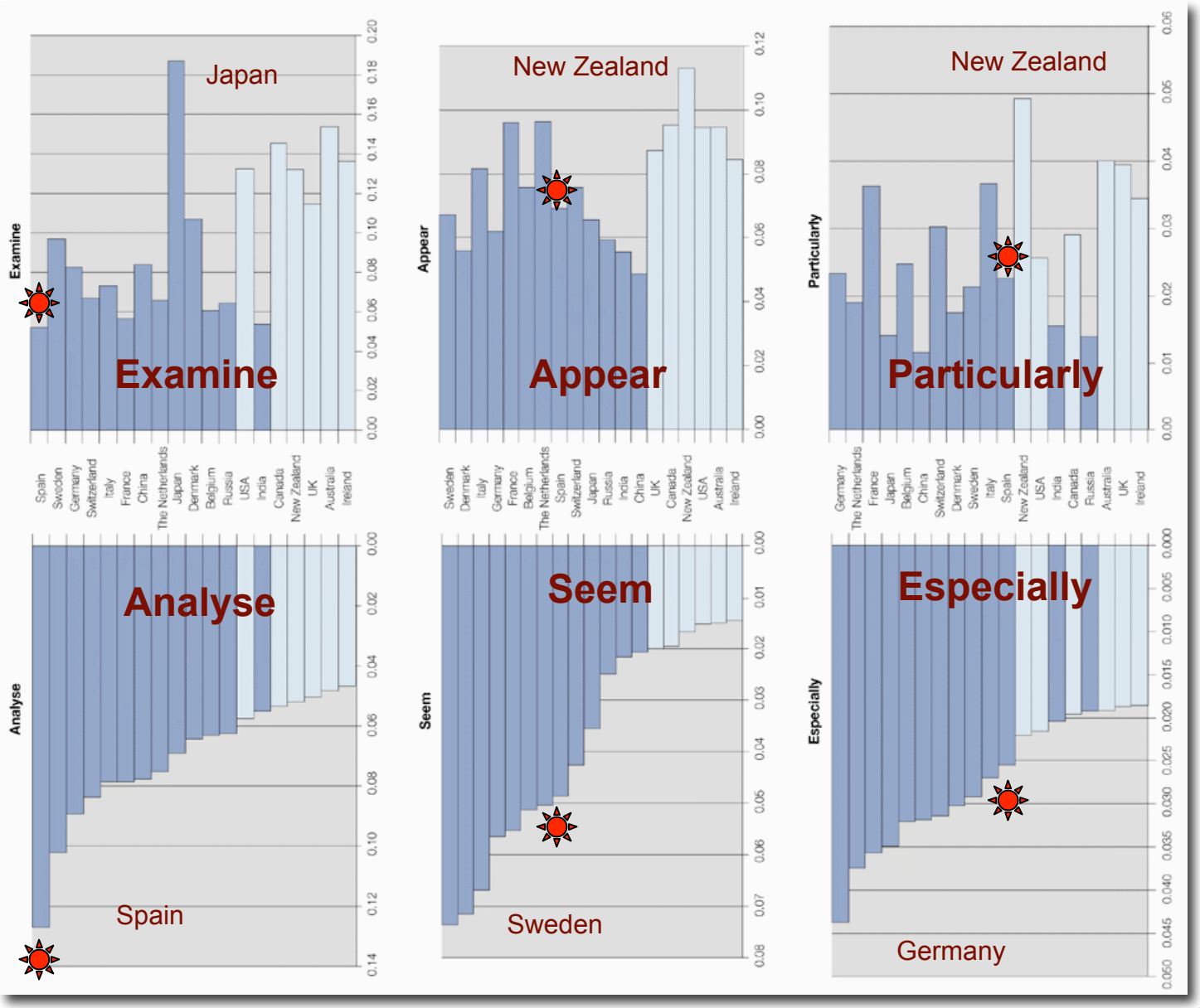
| Country | Adjectives | Nouns | Verbs | Adverbs | Example sentence | PMID ref |
|---|---|---|---|---|---|---|
| Spain | Infrequent, bibliographic | Repercussion, evolution, existence, **sunflower**, **olive**, **wine** | – | Basically | Prevalence of CYP2D6 gene duplication and its **repercussion** on the oxidative phenotype in a white population. | 7697944 |
| Japan | Useful | **Bullfrog**, **shadow** (in radiography) | Clarify | Faintly, next, suddenly, scarcely | MDR-1 protein was **faintly** expressed in one of four chemoresistant patients, but Bcl-2 were [**sic**] clearly detected in four patients. | 12538495 |
| UK | Unsuitable, unlinked, unfamiliar | Marmoset, **consultant**, **questionnaire** | Lie, mirror, arise, tackle | Wholly, principally, particularly | The morphology of these projection neurons was revealed in great detail and confirmed that the projection **arises wholly** from pyramidal cells. | 11602231 |
| Russia | **Gravitational** | (Space) **mission**, **quantum**, **hibernate**, peculiarity, regularity, realization | – | **Thermo-dynamically** | The article is devoted to the question of **peculiarity** of bronchopulmonary system's pathology in the workers of the animal fodder production [**sic**]. | 10341521 |
| India | **Malarial**, -wise (as in stepwise), ascorbic | Malaria, buffalo, peanut, garlic, catfish, | Impart (convey) | Appreciable | Hydroxypropylmethylcellulose (HPMC) was used to **impart** strength and sphericity to the agglomerates. | 12476867 |
| France | Exceptional, digestive | Trouble | Envisage (imagine) | Successively (sequentially), essentially, sometimes | These 2 cells [**sic**] lines being able to clone, it is hard to **envisage** clonogenic assays. | 3051563 |
| China | **Medicinal**, **radiant** (heat), **noxious** (heat) | **Acupuncture**, **coal**, **tea** | **Burn**, replenish, alleviate | Obviously, meanwhile | Because only a catalytic amount of ERK2/pTpY is required, this method **alleviates** the need for large quantities of phospho-ERK2. | 12056917 |
| Germany | Satisfying practicable, unremarkable | Hint, precondition multitude | – | Additionally, exactly, | In clinically presumed spontaneous spinal cord infarction and **unremarkable** signaling of the spinal cord during sequential MRI investigations vertebral body infarction may serve as the only confirmatory sign of spinal cord ischemic stroke. | 11987007 |
| US | **Federal**, investigational, supplemental | Residency, **cocaine**, **payment**, **veteran**, **reimbursement**, physician, care, plan, **noncompliance**, effort, **profit** | Sponsor, mandate | – | Loss of revenue, mainly from **noncompliance** with charge capture resulted in the hospital billing only US$386,794.32 with a total **reimbursement** of US$165,779.86. | 12488156 |

Words in bold typeface have specific meanings and are probably related to local research rather than to local language usage. The bold and underlined words in the example sentences indicate the most abundant country-specific terms. The words shown were found to be more common in the abstracts of the corresponding country than in the abstracts of any other of the 19 representative countries (as in Fig. 2). Note that most of the sentences are grammatically correct, but the usage of the marked (bold and underlined) words is unusual. PMID ref, PubMed reference number.



Netzel R, Perez-Iratxeta C, Bork P, Andrade MA. The way we write. EMBO Rep. 2003 May;4(5):446-51

# Word usage in scientific English (2)






Netzel R, Perez-Iratxeta C, Bork P, Andrade MA. The way we write. EMBO Rep. 2003 May;4(5):446-51

# Overview of Biomedical literature databases

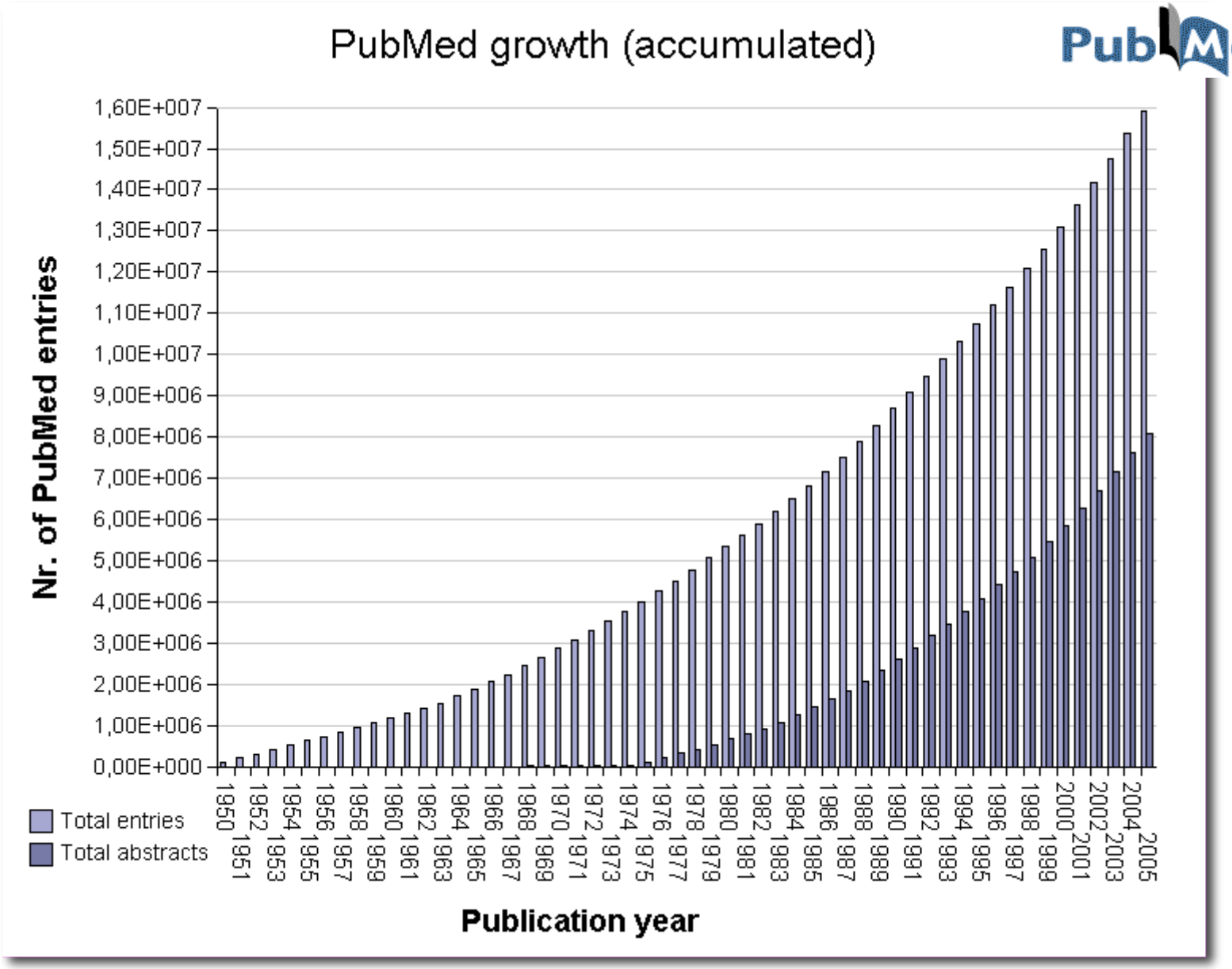
- NLP: need electronically (digitalized) accessible texts
- Main scientific textual data types: e-books and e-articles and the Web (online reports, etc).
- e-Books: e.g. NCBI bookshelf.
- Biomedical article citations (abstracts): PubMed
- Full text articles: PubMed Central (PMC)
- Repositories such as HighWire Press, BioMed Central

# The PubMed database (1)

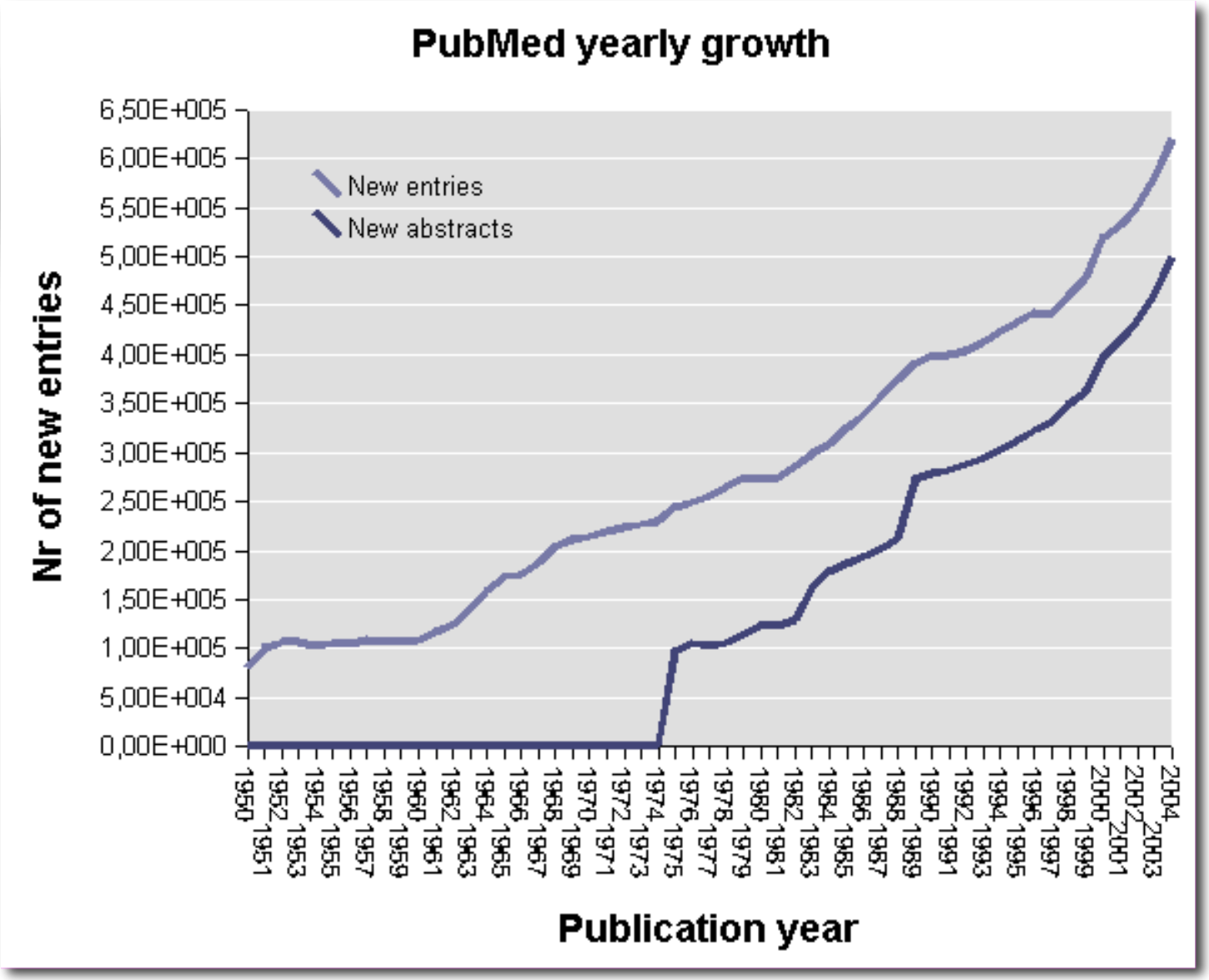
- Scientific articles: new scientific discoveries.
- Citation entries of scientific articles of all biomedical sciences, nursing, biochemistry, engineering, chemistry, environmental sciences, psychology, etc,...
- Developed at the **NCBI** (NIH).
- Digital library contains more than **16 million citations**
- From over **4,800 biomedical journals**
- Most articles (over 12,000,000) in English.
- Each entry is characterized by a unique identifier, the PMID.
- More than half of them (over 7,000,000) have abstracts
- Often links to the full text articles are displayed.

# The PubMed database (2)

- Approx. one million entries (with abstracts) refer to gene descriptions.
- Author, journal and title information of the publication.
- Some records with gene symbols and molecular sequence databank numbers
- Indexed with Medical Subject Headings (MeSH)
- Accessed online through a text-based search query system called Entrez
- Offers additional programming utilities, the Entrez
- Programming Utilities (eUtils)
- NLM also leases the content of the PubMed/ Medline database on a yearly basis



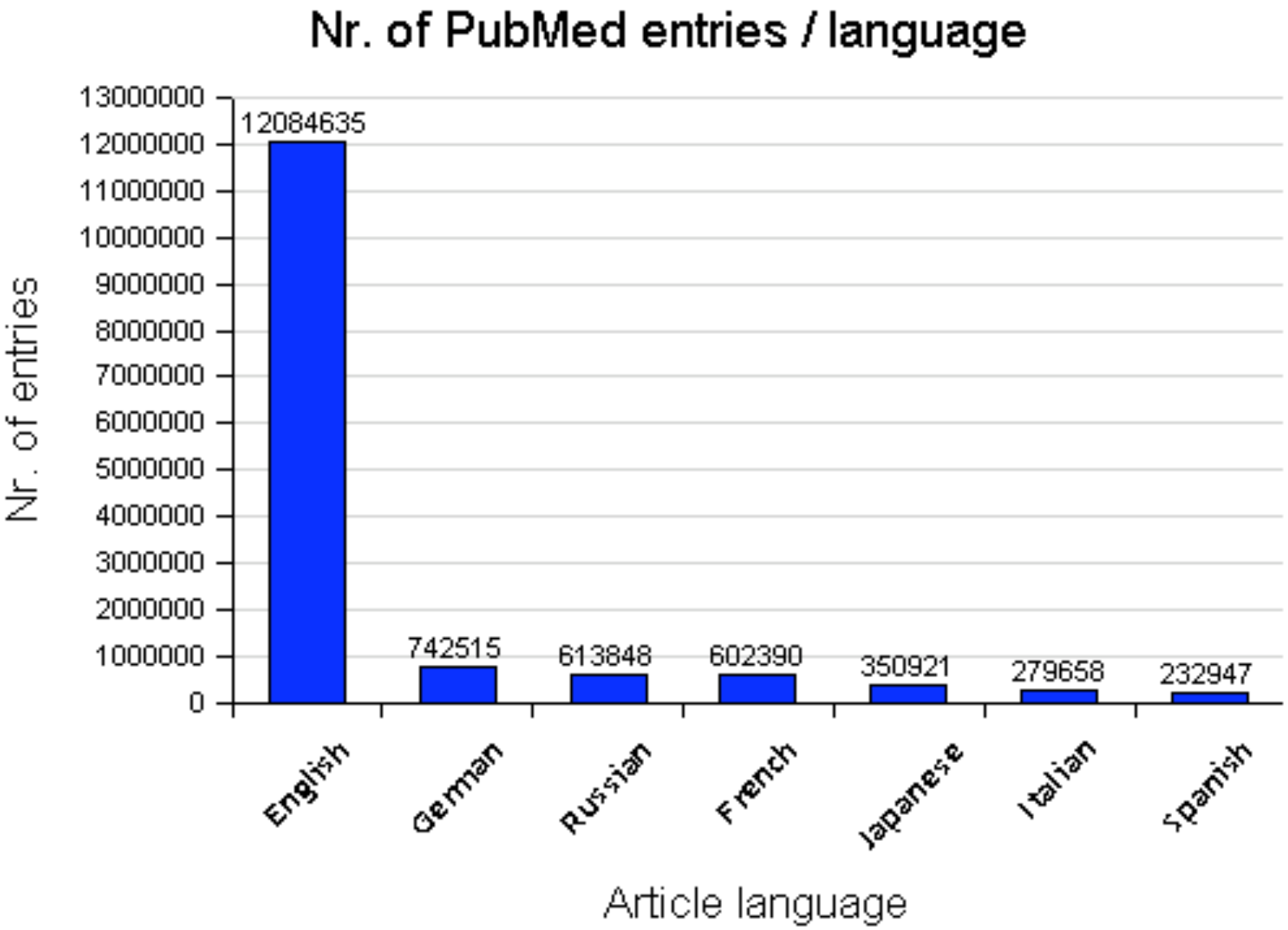

**PubMed has over 16 million entries, most of the recent ones with abstracts**




**PubMed is accumulating over 600,000 new entries every year**

cnio
Centro Nacional de Investigaciones Oncológicas

PubMed



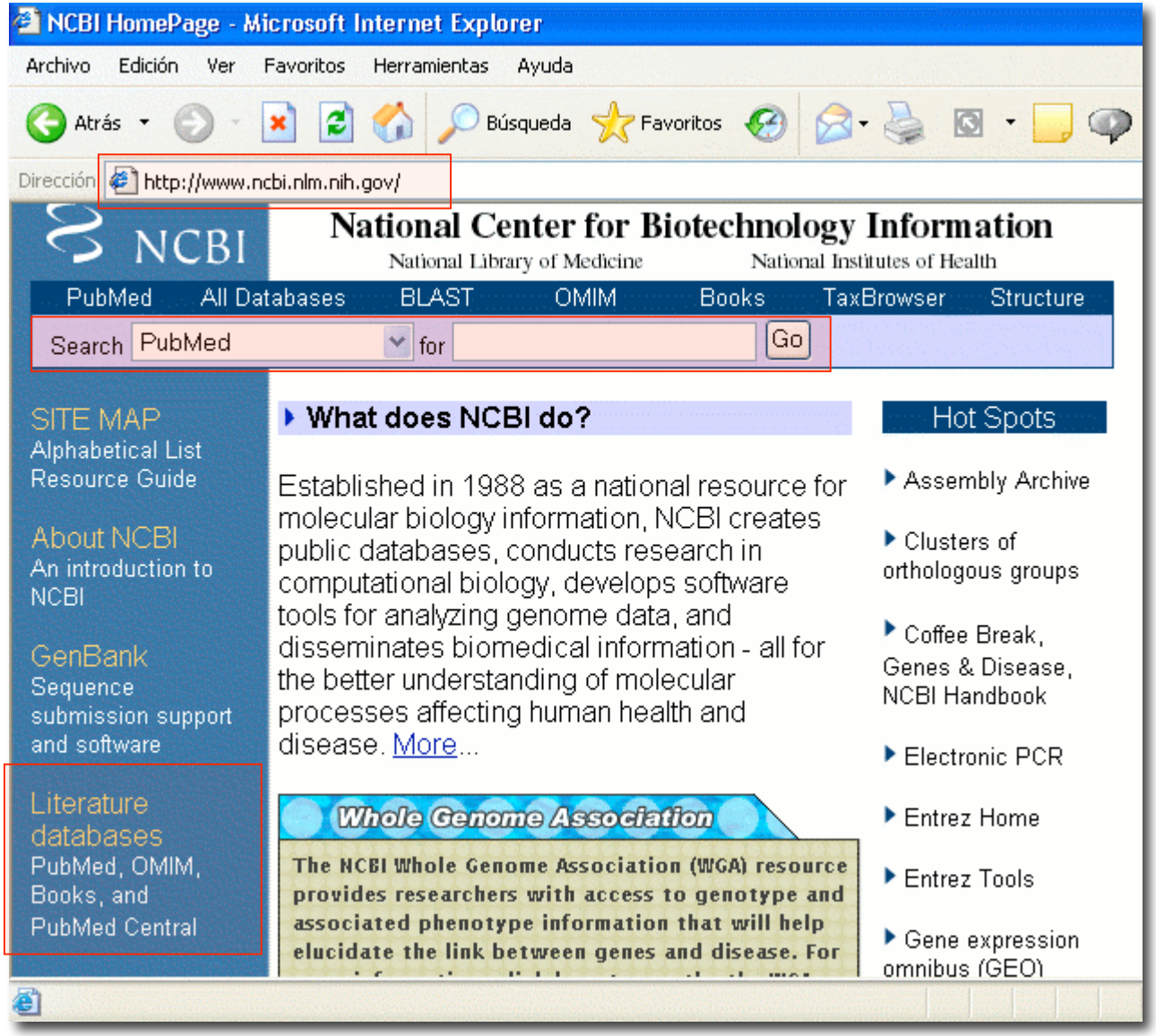

MK

cnio
Centro Nacional de Investigaciones Oncológicas

**NCBI URL**

**NCBI Entrez Query field**

**NCBI links To PMC and E-Books**




**PubMed receives over 70 million queries every month**

MK

# The PubMed search

Entrez PubMed - Microsoft Internet Explorer

Archivo Edición Ver Favoritos Herramientas Ayuda

Atrás Búsqueda Favoritos

Dirección http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=search&term=glycogenin Ir Vínculos

NCBI PubMed www.pubmed.gov

A service of the National Library of Medicine and the National Institutes of Health

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search PubMed for glycogenin **Query term** Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Send to

All: 106 Review: 7

Items 1 - 20 of 106 Page 1 of 6 Next

**Ranked list of hits**

1: Katz A. Related Articles, Links

Glycogenin, proglycogen, and glycogen biogenesis: what's the story?
Am J Physiol Endocrinol Metab. 2006 Apr;290(4):E757-8; author reply E758-9. No abstract available.
PMID: 16533952 [PubMed - indexed for MEDLINE]

2: Torija MJ, Novo M, Lemassu A, Wilson W, Roach PJ, Francois J, Parrou JL. Related Articles, Links

Glycogen synthesis in the absence of glycogenin in the yeast Saccharomyces cerevisiae.
FEBS Lett. 2005 Jul 18;579(18):3999-4004.
PMID: 16004992 [PubMed - indexed for MEDLINE]

3: Bazan S, Curtino JA. Related Articles, Links

About Entrez
NCBI Toolbar

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation

Internet

# The PubMed search result

Entrez PubMed - Microsoft Internet Explorer

Archivo Edición Ver Favoritos Herramientas Ayuda

Atrás Búsqueda Favoritos

Dirección http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15811343&query_ Ir Vínculos

www.pubmed.gov

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search PubMed for Go Clear

Limits Preview/Index History Clipboard Details

Display Abstract Show 20 Sort by Send to

All: 1 Review: 0

About Entrez
NCBI Toolbar

Text Version

Entrez PubMed
Overview
Help | FAQ
Tutorials
New/Noteworthy
E-Utilities

PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Special Queries

1: FEBS Lett. 2005 Apr 11;579(10):2208-14. Related Articles, Links

ELSEVIER FULL-TEXT ARTICLE **Link to full text**

**Biochemical characterization of Neurospora crassa glycogenin (GNN), the self-glucosylating initiator of glycogen synthesis.**

de Paula RM, Wilson WA, Roach PJ, Terenzi HF, Bertolini MC.

Instituto de Quimica, UNESP, Departamento de Bioquimica e Tecnologia Quimica, R. Professor Francisco Degni, s/n, 14800-900 Araraquara, SP, Brazil.

Glycogenin acts in the initiation step of glycogen biosynthesis by catalyzing a self-glucosylation reaction. In a previous work [de Paula et al., Arch. Biochem. Biophys. 435 (2005) 112-124], we described the isolation of the cDNA gnn, which encodes the protein glycogenin (GNN) in

Internet

Entrez PubMed - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Display&DB=pubmed

1: Katz A. Glycogenin, proglycogen, and ...[PMID: 16533952]


```
<PubmedArticle>
    <MedlineCitation Owner="NLM" Status="In-Data-Review">
        <PMID>16533952</PMID>
        <DateCreated>
            <Year>2006</Year>
            <Month>03</Month>
            <Day>14</Day>
        </DateCreated>
        <Article PubModel="Print">
            <Journal>
                <ISSN IssnType="Print">0193-1849</ISSN>
                <JournalIssue CitedMedium="Print">
                    <Volume>290</Volume>
                    <Issue>4</Issue>
                    <PubDate>
                        <Year>2006</Year>
                        <Month>Apr</Month>
                    </PubDate>
                </JournalIssue>
                <Title>American journal of physiology. Endocrinology and metabolism.</Title>
                <ISOAbbreviation>Am. J. Physiol. Endocrinol. Metab.</ISOAbbreviation>
            </Journal>
            <ArticleTitle>Glycogenin, proglycogen, and glycogen biogenesis: what's the story
```


## PubMed XML-format entry:

http://www.nlm.nih.gov/bsd/licensee/data_elements_doc.html

# PubMed Central (PMC)

- Digital archive of **full text** life science journals.
- Articles have a unique PMCID.
- Allows Boolean query search.
- Offers free full text articles
- Journal Publishing XML DTD, but also other widely used DTD in life science
- http://www.pubmedcentral.nih.gov/index.html

cnio
Centro Nacional de Investigaciones Oncológicas

PubMed Central Homepage - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://www.pubmedcentral.nih.gov/ Search

Home Bookmarks

PubMed Central

A free archive of life sciences journals

About PMC Journal List Search Utilities

**PubMed Central (PMC)** is the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life sciences journal literature.

Find Articles Advanced search

**Browse PMC journals:** [A-B] [C-H] [I-M] [N-S] [T-Z] [Full List] [New Journals]

Add your name to the **PMC News list** to get email notification of new PMC journals and other significant updates.

All the articles in PMC are free (sometimes on a delayed basis). Some journals go beyond free, to **Open Access**. Find out what that means.

PMC's **utilities** include an OAI service that provides XML of the full-text of some articles, functions for scripting PMC searches and linking to specific PMC articles from your site, and more ...

Looking for a modern journal article DTD? Take a look at NLM's **Journal Publishing XML DTD and schema**.

It's about preservation and access: **digitizing the complete run of back issues** of many of the

The **PMC journal list** comprises journals that deposit material in PMC on a routine basis and generally make all their published articles available here. Find out how to **include your journal** in PMC.

PMC also has the **author manuscripts** of articles published by NIH-funded researchers in various non-PMC journals. Increasing free access to these articles is the goal of the **NIH Public Access** policy. Similar manuscripts from researchers funded by the Wellcome Trust are available in PMC as well.

Eligible researchers should use the **NIH Manuscript Submission** system to deposit manuscripts.

Get answers to other questions about PubMed

MK

# NCBI Bookshelf

- Collection of electronic **biomedical text books.**
- Allows Boolean query search.
- Offers free full text articles
- Direct searching the books or from PubMed abstracts
- http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=books

cnio
Centro Nacional de Investigaciones Oncológicas

Bookshelf - Online books - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=books Search

Home Bookmarks

NCBI **Bookshelf**

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books

Search Books for Go Clear

Limits Preview/Index History Clipboard Details

About Entrez

Books

Overview

Using the books

Information for authors and publishers

Contact us

Mailing list

Project background

FAQ

My NCBI

Privacy Policy

The **Bookshelf** is a growing collection of biomedical books that can be searched directly by typing a concept into the textbox above and selecting "Go". Try one of these searches:

- cell cycle control
- immunodeficiency
- protein evolution

Books are also linked to terms in PubMed abstracts: when viewing an abstract, select the "Books" link to see phrases within the abstract hyperlinked to book sections.

## New on the Bookshelf:

### Disease Control Priorities in Developing Countries 2nd ed.

Dean T. Jamison, Joel G. Breman, Anthony R. Measham, George Alleyne, Mariam Claeson, David B. Evans, Prabhat Jha, Anne Mills, Philip Musgrove, editors
Washington (DC): IBRD/The World Bank and Oxford University Press; 2006

### Global Burden of Disease and Risk Factors

Alan D. Lopez, Colin D. Mathers, Majid Ezzati, Dean T. Jamison, Christopher J. L. Murray, editors
Washington (DC): IBRD/The World Bank and Oxford University Press; 2006

### Priorities in Health

Dean T. Jamison, Joel G. Breman, Anthony R. Measham, George Alleyne, Mariam Claeson, David B. Evans, Prabhat

MK

# Biomedical corpora and text collections

- Medtag corpus, includes the Abgene, MedPost and GENETAG corpora
- TREC Genomics Track collections
- BioCreative corpora
- GENIA corpus
- Yapex corpus
- Others, e.g. LL05 dataset, BioText Data, PennBioIE, OHSUMED text collection, Medstract corpus,...

Centro Nacional de Investigaciones Oncológicas

United States
National Library of Medicine
National Institutes of Health

Search NLM Web Site
Go
NLM Home | Contact NLM | Site Map | FAQs

Home > Bibliographic Services Division > Distance Education Resources

# PubMed® Online Training

Return to PubMed

## • PubMed Tutorial

The PubMed Tutorial is based on the NLM's one-day PubMed training course.

## • Quick Tours

The following are brief animated tutorials with audio for using PubMed. Running times are rounded to the nearest minute. Click on the link to launch the tour.

### Searching PubMed Quick Tour

- Search PubMed for an Author (3 min., June 2005)
- Searching PubMed by Author and Subject (1 min., June 2005)
- PubMed Simple Subject Search Example (1 min., June 2005)
- Search for a Journal (5 min, February 2006)
- Retrieving Citations from a Journal Issue (1 min., December 2005)

### My NCBI Quick Tour

- Getting Started with My NCBI (approx. 5 min., revised April 2006)
  How to register, sign in and out, change your password, and what to do if you've forgotten your password.
- Saving Searches (approx. 4 min., revised June 2005)
  How to save a PubMed search, to run later or to have results sent to your e-mail account.
- Filters (approx. 7 min., revised July 2005)
  How to create filters to group your search results.
- LinkOut Filters (approx. 6 min., revised September 2005)

MK

# Text mining and Natural language processing (NLP)

- Techniques that analyze, understand and generate language (free text, speech).
- Multidisciplinary field: information technology, computational linguistics, AI, statistics, psychology, language studies, etc,.
- Strongly language dependent (Bio-NLP mostly English).
- Create computational models of language.
- Learn statistical properties of language.
- Methods: statistical analysis, machine learning, rule-based, pattern-matching, AI, etc...
- Explore the grammatical, morphological, syntactical and semantic features of well-structured language
- The statistical analysis of these features in large text collections is generally the basic approach used by NLP techniques.
- Often combinations of these inter-related features are explored by NLP strategies.

# Grammatical features

- Grammar: rules governing a particular language.
- Rules for correct formulation of a specific language
- Grammatical features in NLP, e.g. part of speech (POS)
- POS of a word depends on sentence context
- Examples: noun, verb, adjective, adverb or preposition.
- Programs label words with POS: POS taggers.
- Example:

***Caspase-3*** Proper noun, sing. ***was*** Verb, past tense ***partially*** Adverb ***activated*** Verb, past part. ***by*** Prep. or subord. Conjunction ***IFN-gamma*** Proper noun, sing. [PMID 12700631].

- POS taggers are usually based on machine learning
- Trained with a set of manually POS-tagged sentences.
- POS useful for gene name identification and protein interactions detection from text,
- MedPost {Smith, 2004} a POS for biomedical domain
- MedPost: 97% accuracy in PubMed abstracts (86.8% gen. POS tagger)

# Example Online generic POS tagger

POS Tagging - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://cpk.auc.dk/~tb/tagger/tagit.php

Search

Home Bookmarks Yahoo Google MK Homepage ORF Zope on http://... PubMed

kom.aau.dk/~tb

Top
Publications
Teaching
Projects
NLPSuite
Resources
Demos
Online
Statistics

tom.brondsted.dk

@ email

## POS Tagging

[UP] [ABOUT THE TAGGER] [MORE ONLINE]

Copyright Tom Brøndsted March 2004, Test version 1.0

**NB!** *Netscape and Mozilla users can toggle between this view and a view without tags by selecting "view - use style - variant 3".*

**Tagged text:**

Glycogenin<Noun, sing. or mass> is<Verb, 3rd person sing. present> the<Determiner> self-glycosylating<Adjective> protein<Noun, sing. or mass> primer<Noun, sing. or mass> that<Prep. or subord. Conjunction> initiates<Verb, 3rd person sing. present> glycogen<Verb sing. present, non-3rd.> granule<Noun, sing. or mass> formation.<Noun, sing. or mass>

Output text POS-labele

Input (English) text for POS-tagging

```
Glycogenin is the self-glycosylating protein primer that initiates
glycogen granule formation.
```

Input text

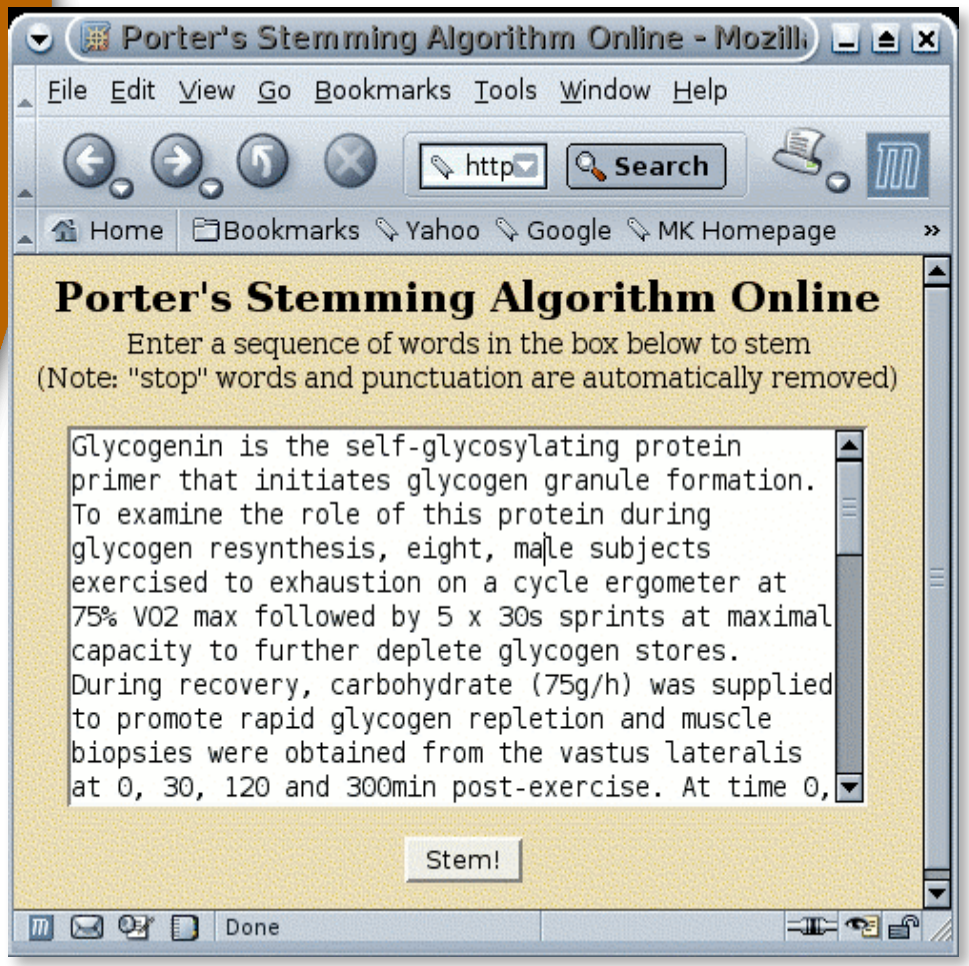
submit text

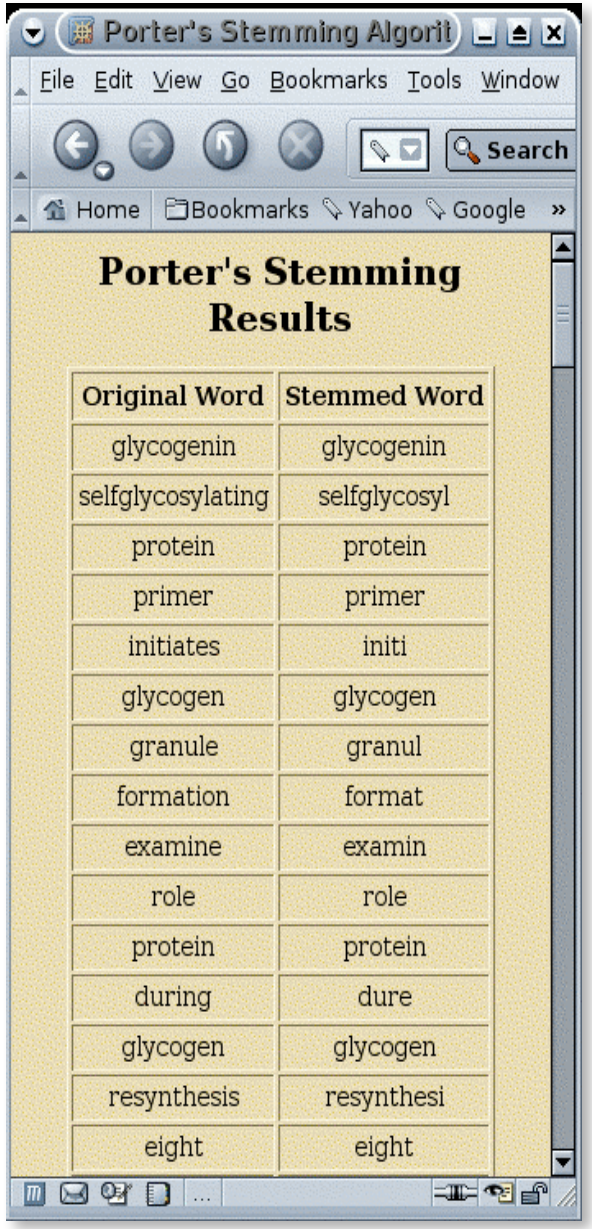
CHECK

# Morphological features

- **Word structure analysis**
- **Rules of how words relate to each other.**
- **Example 1: <u>plural formation rules</u>, e.g.:**
  *gene* and *genes* or *caspase* and *caspases*
- **Example 2: <u>verb inflection rules</u>, e.g.**
  *phosphorylate*, *phosphorylates* and *phosphorylating*
  all have the same verb stem, word root.
- **<u>Stemmer algorithms</u> to standardize word forms to a common stem**
- **Linking different words to the same entity.**
- **Different algorithms, e.g. Porter stemmer {Porter, 1980}**
- **Problem: collapse two semantically different words, e.g.:**
  gallery and gall.

# Online Porter Stemmer




| Original Word | Stemmed Word |
|---|---|
| glycogenin | glycogenin |
| selfglycosylating | selfglycosyl |
| protein | protein |
| primer | primer |
| initiates | initi |
| glycogen | glycogen |
| granule | granul |
| formation | format |
| examine | examin |
| role | role |
| protein | protein |
| during | dure |
| glycogen | glycogen |
| resynthesis | resynthesi |
| eight | eight |

http://maya.cs.depaul.edu/~classes/ds575/porter.htm

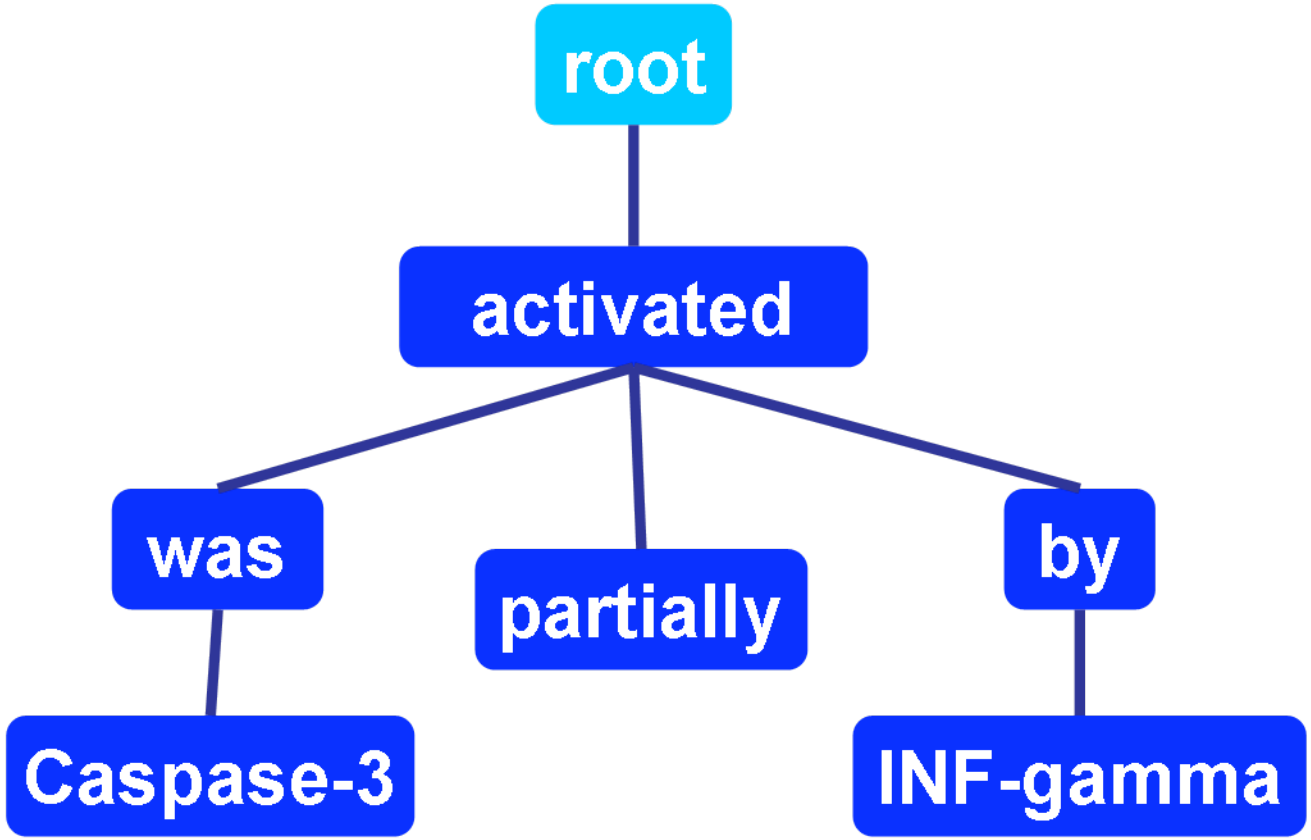
# SYNTACTICAL FEATURES (1)

- **Relationships between words in a sentence: syntactic structur**
- **Shallow parsers analyze such relations at a coarse level, identification of phrases (groups of words which function as a syntactic unit), e.g. noun phrase or verbal phrase.**

- **Example:** Connexor shallow parser output:

| | |
|---|---|
| *Caspase-3* | <: nominal head, noun, single-word noun phrase,> |
| *was,* | <auxiliary verb, indicative past> *partially* <adverbial head, adverb |
| *activated* | <main verb, past participle, perfect> |
| *by* | <preposed marker, preposition> |
| *IFN-* | <premodifier, noun, noun phrase begins,> |
| *gamma* | <nominal head, noun, noun phrase ends>. |

- **Word labeled to corresponding phrase.**
- **Noun phrases (head is a noun, NP) e.g. '**Caspase-3**' and** 'INF-gamma' **and verbal phrases (head is a verb, VP).**




*Caspase-3* <: nominal head, noun, single-word noun phrase,>
*was,* <auxiliary verb, indicative past> *partially* <adverbial head, adverb>
*activated*<main verb, past participle, perfect>
*by* <preposed marker, preposition>
*IFN-* <premodifier, noun, noun phrase begins,>
*gamma* <nominal head, noun, noun phrase ends>.

# SYNTACTICAL FEATURES (2)

- **Other features: identification of subject-object relationships**
- **{Koike,.ea: 2005}.**
- **E.g. for NP-VP-NP :**

'Smith and Mitchell (1989) found that [overexpression of <gene> **IMEI**</gene> ] induced [an <GO> **early meiotic event (recombination)** </GO> in rich medium], but later meiotic events did not occur (i.e., they detected [no spore formation])'.

- **In this case the subject is represented by the 'IMEI' gene and the**
- **object is the Gene Ontology term 'early meiotic event'.**

# SEMANTIC FEATURES

- **Associations of words with their corresponding <u>meaning in a given context.</u>**
- **Semantics (meanings) of a word -> understand meaning sentence.**
- **<u>Dictionaries</u> and thesauri provide such associations**
- **Gene Ontology (GO) provides concepts for biological aspects of genes**
- **Gene names and symbols contained in SwissProt**
- **Example:**

  ***Caspase-3*** /GENE PRODUCT **was partially** ***activated*** /INTERACTION VERB **by** ***IFN-gamma*** /GENE PRODUCT.
- **Caspase-3 and INF-gamma are identified as gene products**
- **The verb 'activated' refers in this context to a certain type of interaction**

# CONTEXTUAL FEATURES

- Words occurrence in textual context - association.
- Co-occurrence of Caspase-3 and INF-gamma in the same sentence indicates some relationship between them.
- Determine contextual similarity of proteins documents.
- Use for instance: list of words (bag of words)
- The statistical analysis of word frequencies or patterns
- Features are interrelated

Centro Nacional de Investigaciones Oncológicas

# MAIN TASKS in NLP

- Information Retrieval (IR).
- Information extraction/Text mining (IE).
- Question Answering (QA).
- Natural Language Generation (NLG).
- Automatic summarization.
- Machine translation (MT).
- Text proofing.
- Anaphora resolution
- Text zoning
- Speech recognition.
- Document clustering
- Document categorization
- Optical character recognition (OCR).

MK

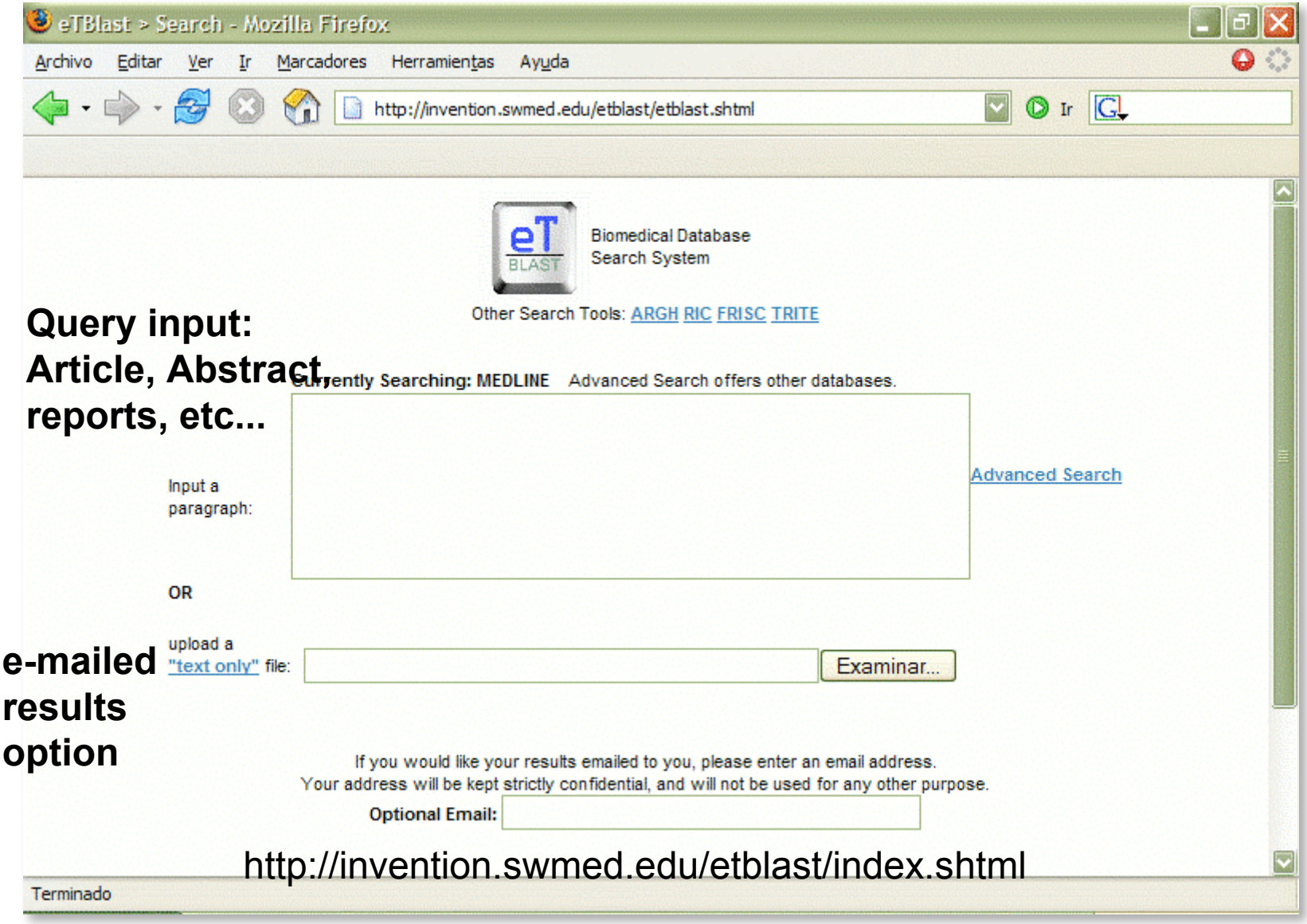
# INFORMATION RETRIEVAL (IR)

- IR: process of recovery of those documents from a collection of documents which satisfy a given information demand.
- Information demand often posed in form of a search query.
- Example: retrieval of web-pages using search engines, e.g. Google.
- Important steps for indexing document collection:
  - Tokenization
  - Case folding
  - Stemming
  - Stop word removal
- Efficient indexing to reduce vocabulary of terms and query formulations.
- Example: *'Glycogenin AND binding'* and 'glycogenin *AND* bind'.
- Query types: Boolean query and Vector Space Model based query.

# SELECTIVE DISSEMINATION OF INFORMATION (SDI)

- Service provided by a library or data repository institution which periodically alerts users of new publications.
- New publications can be associated to certain subjects or information demands
- Often based on automated iterative/periodical IR queries.
- Advantages: new publications are automatically announced (e.g. using e-mail alerts)
- Disadvantages: implicit to IR based on Boolean queries, ofter not-relevant articles.
- Free SDI services based on PubMed / Biomedical literature:
  - Cubby (NCBI)
  - PubCrawler
  - BioMail

# eTBLAST




**Query input: Article, Abstract, reports, etc...**

**e-mailed results option**

http://invention.swmed.edu/etblast/index.shtml

# eTBLAST

eTBlast > Success! - Mozilla Firefox

Archivo Editar Ver Ir Marcadores Herramientas Ayuda

http://invention.swmed.edu/cgi-bin/etblast/etblast_parser_new

eTBLAST Search FRISC TRITE RIC Help Disclaimer Contact

**Submission Successful!**

Your results will be available at http://invention.swmed.edu/etblast/user/user-1117008679/results.html in just a few minutes. The time it takes to process your query depends on the length of your query, the number of people using our system at this moment, and whether or not you selected features from our Advanced Search.

Your results will be accessible at this location for several weeks. You may want to bookmark it so it does not get lost.

While you wait, you may want to check out our search utilities.

- RIC allows you to build a short profile, and upload a query which we will automatically re-run weekly.
- TRITE is a set of general interest topics we re-run weekly so that you can easily find the latest research.
- FRISC is a set of profiles we built to keep the faculty members of our department at UTSouthwestern up to date on the latest research.

You can also view search terms generated by your input paragraph.

**NOTE:** Some users report that spam-filtering softare intercepts etBlast output. If you chose to have your results emailed to you, and you do not recieve them, check to see whether your filter has intercepted them.

If for any reason you should fail to recieve results, send us an email referencing your user number: **user-1117008679**

http://invention.swmed.edu/etblast/user/user-1117008679/results.html

# eTBLAST results

Results of the search - Mozilla Firefox

Archivo Editar Ver Ir Marcadores Herramientas Ayuda

http://invention.swmed.edu/etblast/user/user-1117008679/results.html

Submitted paragraph | Statistics Summary | Disclaimer

**Closest Matches:**

1. Glycogen metabolism in quail embryo muscle. The role of the glycogenin primer and the intermediate proglycogen.
   J Lomako ... W J Whelan
   Eur J Biochem 1995 Nov;234(1);343-9. Score: 61.496
2. A new look at the biogenesis of glycogen.
   M D Alonso ... W J Whelan
   FASEB J 1995 Sep;9(12);1126-37. Score: 54.209
3. Glycogen synthesis in the astrocyte: from glycogenin to proglycogen to glycogen.
   J Lomako ... M D Norenberg
   FASEB J 1993 Nov;7(14);1386-93. Score: 50.494
4. Further studies on the role of glycogenin in glycogen biosynthesis.
   C Smythe ... P Cohen
   Eur J Biochem 1990 Apr;189(1);199-204. Score: 48.524
5. Glycogenin-dependent organization of Ascaris suum muscle glycogen.

Terminado

**Similarity ranked document list**

# eTBLAST results: high scoring words

PMID: 8529663 - Mozilla Firefox

Archivo Editar Ver Ir Marcadores Herramientas Ayuda

http://invention.swmed.edu/cgi-bin/etblast/abstract_local?pmid=8529663&use

Ir

eT BLAST

Eur J Biochem 1995 Nov;234(1);343-9.

**Glycogen metabolism in quail embryo muscle. The role of the glycogenin primer and the intermediate proglycogen.**

J Lomako
W M Lomako
W J Whelan

**Terms with high weight**

Department of Biochemistry and Molecular Biology, University of Miami School of Medicine, FL 33101, USA.

Cultured quail embryo muscle has proven to be an excellent model system for studying the synthesis of macromolecular glycogen from, and its degradation to, glycogenin, the autocatalytic, self-glucosylating primer for glycogen synthesis. We recently demonstrated that proglycogen, a low-M(r) form of glycogen, is an intermediate in the synthesis. Here we show that proglycogen also functions as an intermediate in macroglycogen degradation and, in one set of circumstances, represents an arrest point in glycogen breakdown, which does not continue to glycogenin. We suggest that in the nutritionally dependent turnover of glycogen in tissues, the molecules cycle between proglycogen and macromolecular glycogen and are not normally degraded to glycogenin. Nevertheless, when this does happen, the released glycogenin is active, capable of re-initiating glycogen synthesis. Under culture conditions where the conversion of proglycogen into glycogenin does take place, the intermediates lying between form a discrete rather than a continuous series, suggestive of a cluster structure for proglycogen and indicating that breakdown is stepwise. Evidence of post-translational modification of glycogenin was obtained by the finding that, in glycogen from cultured muscle, glycogenin is phosphorylated.

MedlineID: 0
PMID: 8529663

# IR EVALUATION

- **Precision**: fraction of relevant documents retrieved divided by the total returned documents
- **Recall**: proportion of relevant documents returned divided by the total number of relevant documents
- **F-score**: the harmonic mean of precision and recall
- **Precision-recall curves**

Centro Nacional de Investigaciones Oncológicas

# 15 MINUTES BREAK

MK

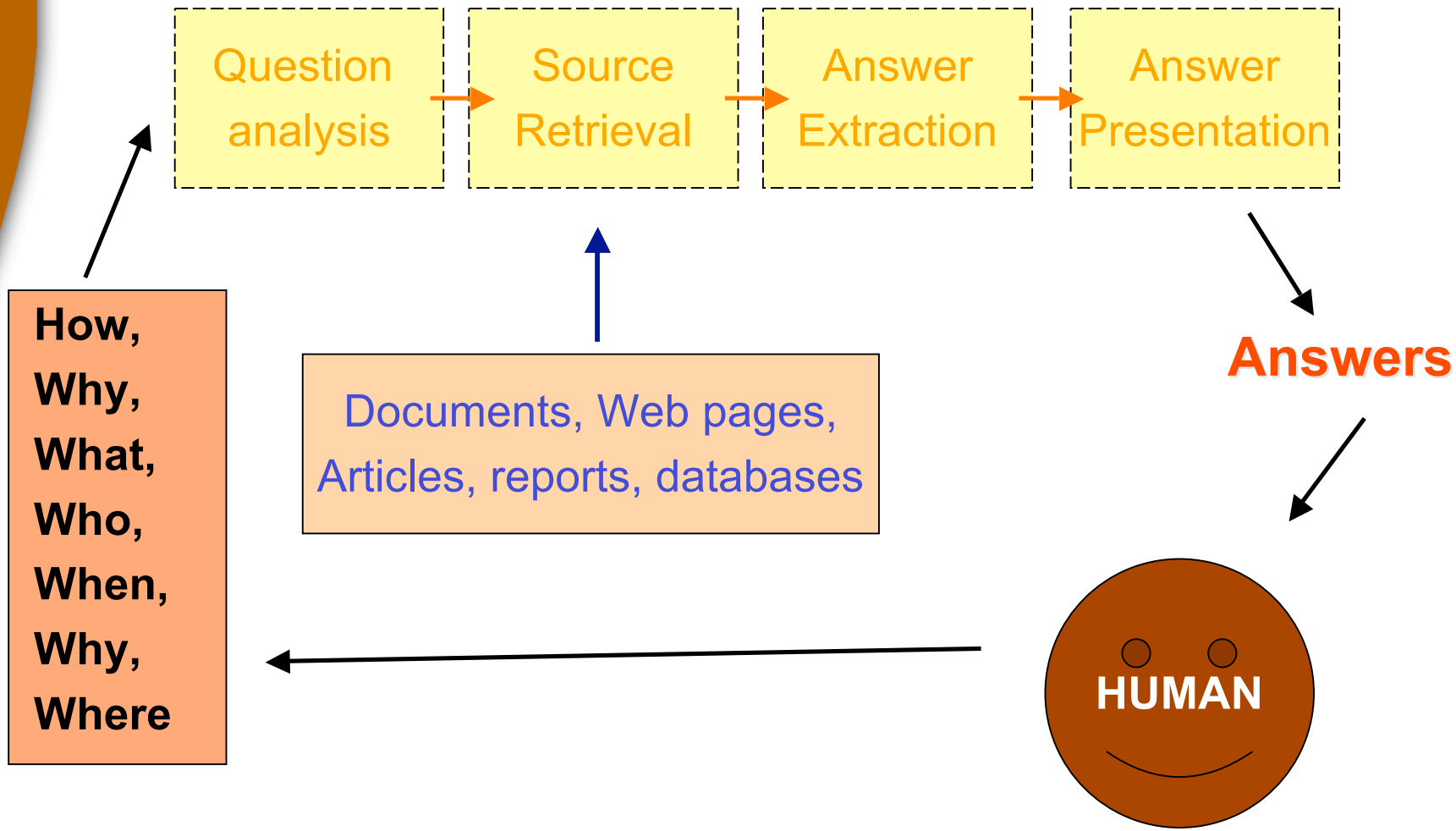
# INFORMATION EXTRACTION (IE) & TEXT MINING

- Identification of semantic structures within free text.
- Use of syntactic and Part of Speech (POS) information.
- Integration of domain specific knowledge (e.g. ontologies).
- Identification of textual patterns.
- Extraction of predefined entities (NER), relations, facts.
- Entities like: companies, places, dates
- Bio-entities like: proteins, genes, chemical compounds.
- Relations like: protein interactions
- Methods: heuristics, rule-based systems, machine learning and statistical techniques, regular expressions,.

# QUESTION ANSWERING

- Humans formulate questions using natural language.
- Example: *What are the molecular functions of Glycogenin?*
- QA: automatic generation of answers to queries in form NL expressions from document collections.
- Most systems limited to generic literature or newswire.
- QA difficult: heterogeneous, poorly formalized domain, new scientific terms
- Ad hoc retrieval task of the TREC Genomics Track 2005.
- NL query system example: Ask Jeeves
- Galitsky system (semantic skeletons (SSK), logical programming).

# QUESTION ANSWERING ARCHITECTURE




New directions in Question Answering, Mark Maybury

# NATURAL LANGUAGE GENERATION

- NLG: constructing automatically natural language texts.
- Display the content of databases: reports, error messages.
- Based on semantic input, providing computer-internal representation of the information.
- Different degrees of complexity.
- Biology: modeling the domain language difficult.
- Simpathica/XSSYS trace analysis tool.

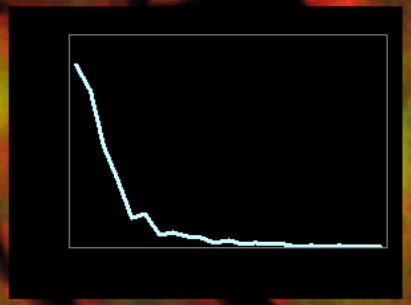
# AUTOMATIC TEXT SUMMARIZATION

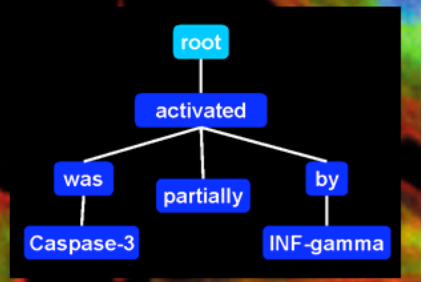
- Process of distilling the most important information from a source to <u>produce a short version</u>.
- Single document vs. multi-document summarization
- Techniques to produce customized summarizations (e.g. 10% from the original).
- Mainly word-for-word sentences extracted from the given documents.
- Selecting the right sentences (sentence clustering/ranking)
- Use of time, sentence length and position, terms,..
- Technically similar to clustering & categorization, select a representative of the cluster.
- Resulting summary: set of sentences

Advances in automatic summarization, Many & Maybury & Text Mining, Weiss et al

cnio
Centro Nacional de Investigaciones Oncológicas

# Bio-NLP

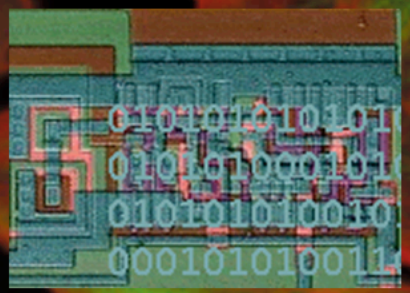




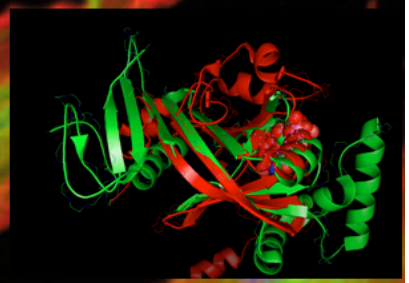

MATHEMATICS & STATISTICS

COMPUTATIONAL LINGUISTICS & NLP

IT & COMPUTER SCIENCE



BIOINFORMATICS



LIFE SCIENCES & BIOMEDICINE

MK

# Extracting functional Annotations

(1) <u>Manual:</u>

- Annotation extraction by database curators.
- Scientific literature analysis.
- Time-consuming & labor-intensive.
- Accurate and usage of human inference
- Example: Gene Ontology annotation (GOA).

(2) <u>Text mining:</u>

- To assist annotation extraction
- Identification of annotation relevant sentences.
- Identification of protein-term associations.

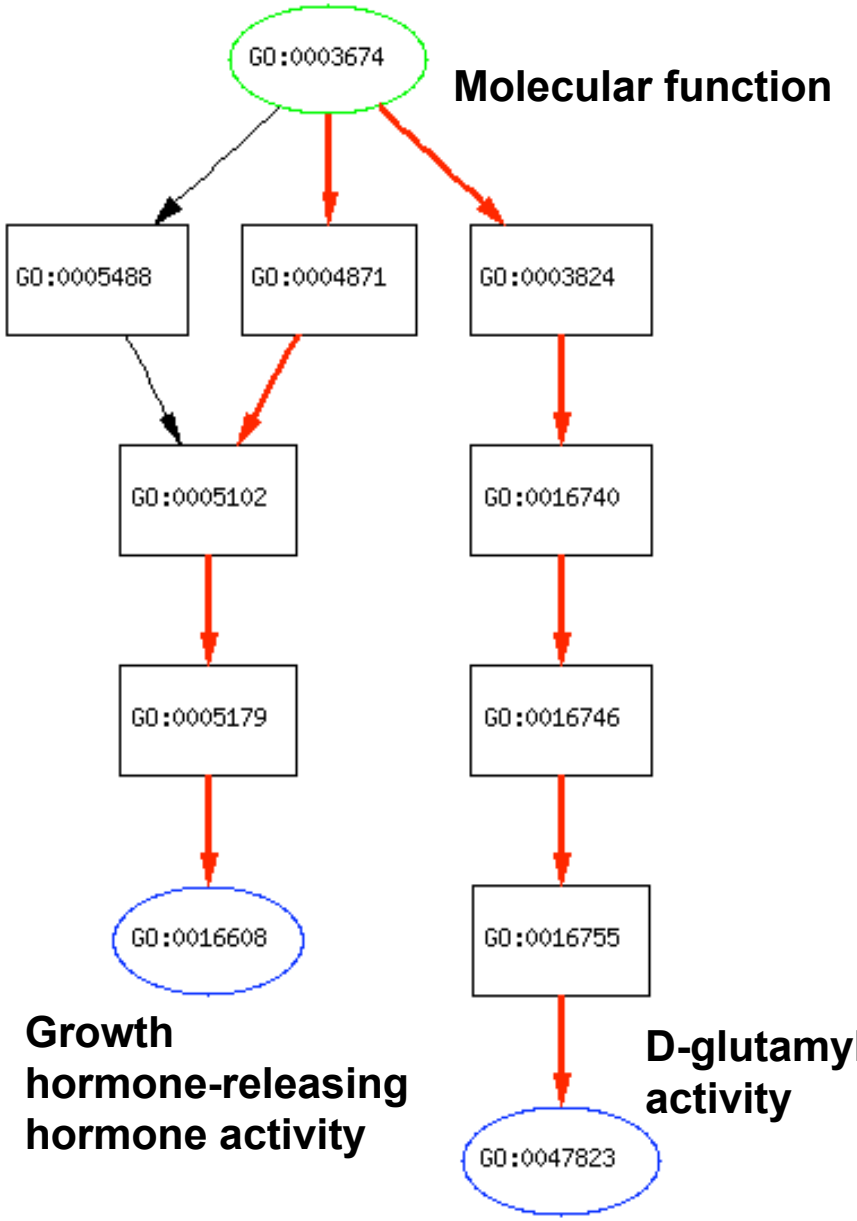
# Citations in Annotation Databases

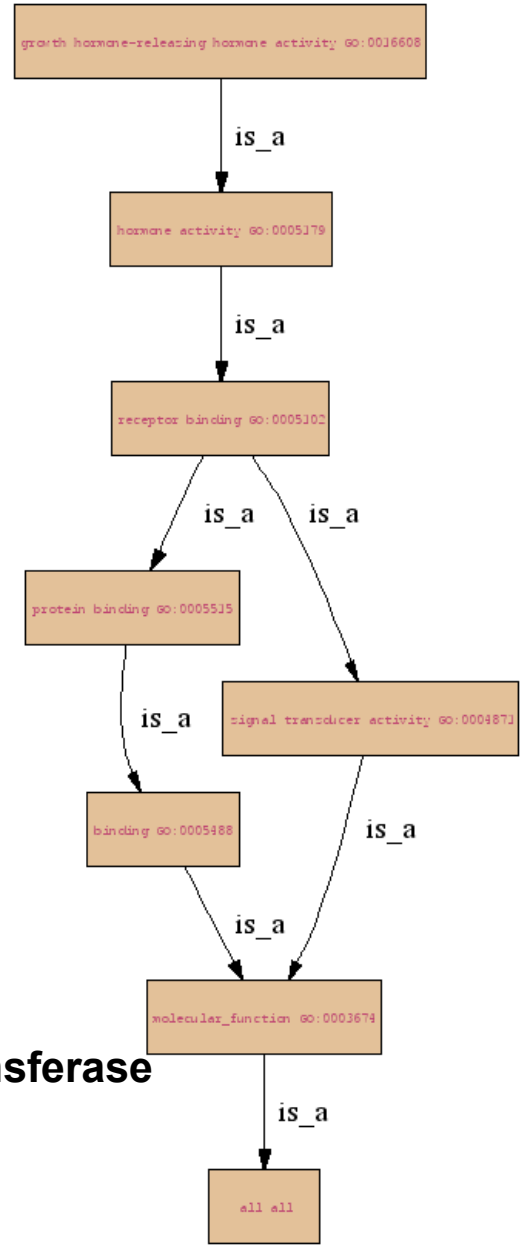
| DB | GOA | GeneRif | UniProt | OMIM | PDB |
|---|---|---|---|---|---|
| GOA | **29,248** | 3,972 | 15,409 | 9,465 | 135 |
| GeneRif | 3,972 | **84,380** | 4,890 | 6,637 | 620 |
| UniProt | 15,409 | 4,890 | **112,476** | 19,859 | 5,061 |
| OMIM | 9,465 | 6,637 | 19,859 | **88,766** | 295 |
| PDB | 135 | 620 | 5,061 | 296 | **11,790** |

# Gene Ontology (GO)

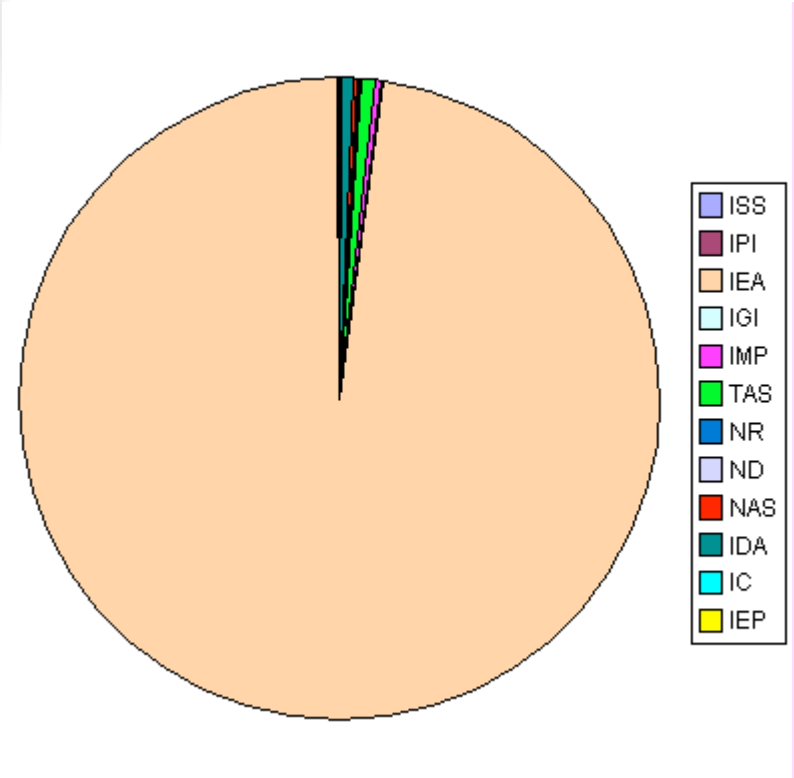
- Ontology deacyclic graph structure.
- Controlled vocabulary of concepts.
- Three main categories:
  - Molecular Function
  - Cellular Component
  - Biological Process
- Describe relevant biological aspects of gene products
- Synonyms, links to external keywords.
- Currently most important source annotation terms.
- IS A and PART OF relations.

Centro Nacional de Investigaciones Oncológicas



# Gene Ontology annotations




| Ev.C. | Annot | Perc. |
|---|---|---|
| **IEA** | **6421817** | **0.97529** |
| ISS | 19576 | 0.00297 |
| NR | 2191 | 0.00033 |
| ND | 4433 | 0.00067 |
| IPI | 7130 | 0.00108 |
| IGI | 3014 | 0.00046 |
| IMP | 19072 | 0.00290 |
| IDA | 38862 | 0.00590 |
| IEP | 1495 | 0.00023 |
| IC | 831 | 0.00013 |
| TAS | 49630 | 0.00754 |
| NAS | 16456 | 0.00250 |

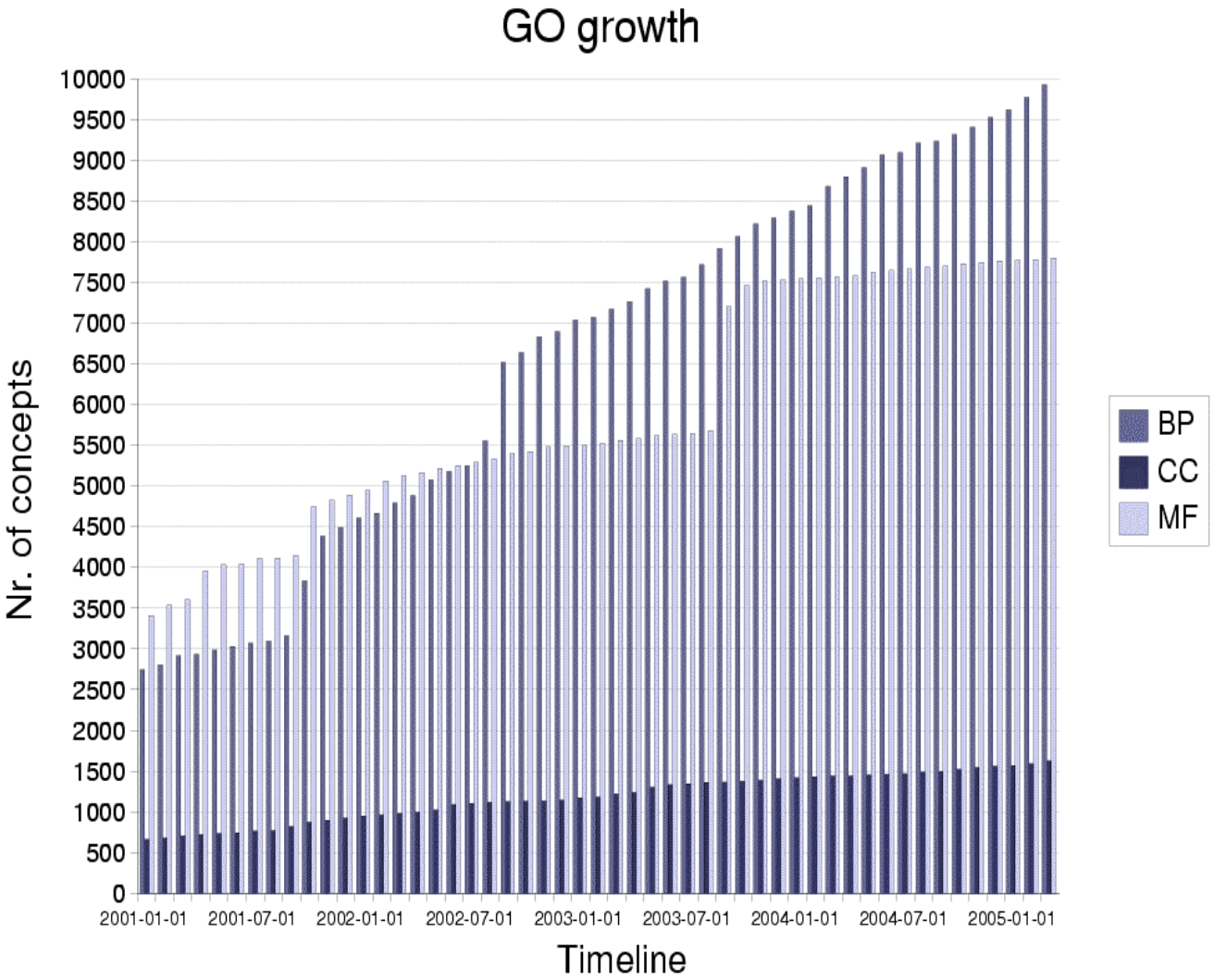
TAS: Traceable Author Statement; IDA: Inferred by direct assay; IC: Inferred by curator ; ND:No data;

IMP: Inferred from mutant phenotype; IGI: Inferred from genetic interaction; 3.8) IPI :Inferred from physical interaction;

ISS: Inferred from sequence similarity; IEP: Inferred from expression pattern; NAS: Non traceable author statement;

IEA: Inferred by electronic annotation; NR: Not recorded;

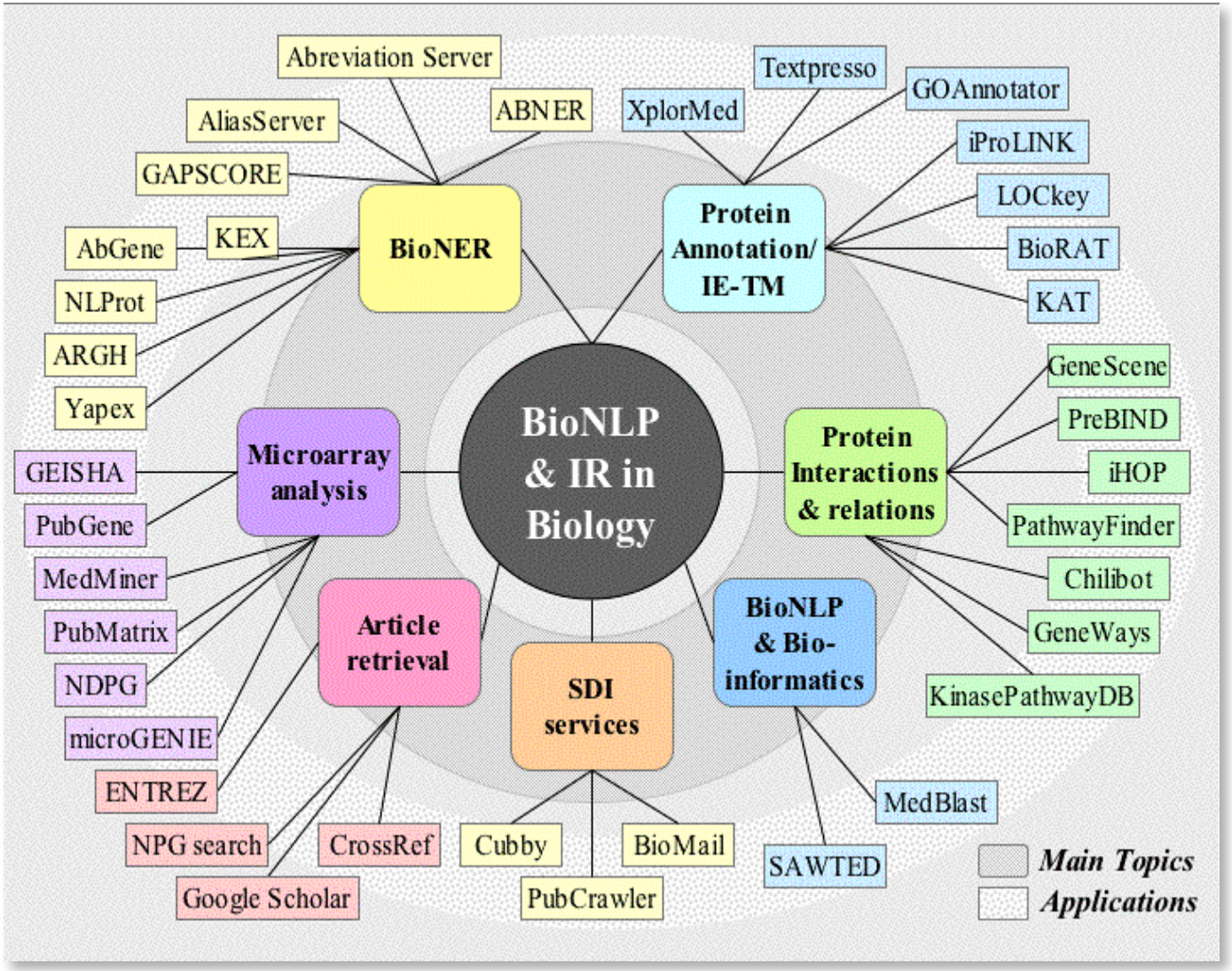
# Gene Ontology concept growth



# Ontologies in Biology

- Cell type
- Human Disease
- Mammalian phenotype
- Multiple alignment
- Pathway ontology
- Protein domain
- Protein-protein interactions
- Systems Biology
- Protein modification
- Mouse pathology
- Mouse adult gross anatomy, …. and a growing number of many more .

# Bio-NLP/Text mining applications



# APPLICATIONS AND USER COMMUNITIES

Pathway extraction

Enzyme Kinetics parameters

Mutation extraction

Gene prioritization

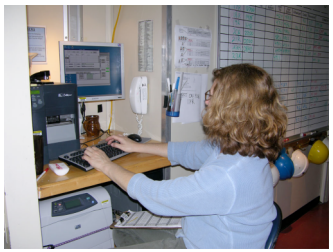
Improve Sequence search

Gene cluster analysis

**Biologists**

BioNLP & IR in Biology

**Database curators**



Protein Interaction

Protein-term association

Protein/gene normalization

Bio-entity tagging

Term extraction

Gene regulation

# Bio-NLP/Text mining applications

- NER: tagging biological entities (e.g. protein names).
- Automatic annotation extraction: associating proteins to functional descriptions/terms.
- Protein interactions: extracting interactions of proteins, genes and drugs.
- Gene cluster analysis: providing biological context through literature mining (microarrays)
- Protein sub-cellular localizations
- Improving sequence-based homology detection
- Other: kinetic parameters, sequence patterns, chromosome aberrations, …

# NER: TAGGING BIO-ENTIRIES IN TEXT

- Aim: Identify biological entities in articles and to link them to entries in biological databases.
- Generic NER: corporate names and places (0.9 f-score), Message Understanding Conferences (MUC) .
- Biology NER: more complex (synonyms, disambiguation, typographical variants, official symbols not used,..).
- Bioinformatics vs. NLP approach.
- Performance organism dependent.
- Methods: POS tagging, rule-based, flexible matching, statistics, ML (naïve Bayes, ME, SVM, CRF, HMM).
- Important for down-stream text mining.

# DIFFICULTIES OF GENE TAGGING

- Authors often do not use the official gene symbols
- Genes have often synonyms.
- Use of full gene names and/or gene symbols/acronyms
- Gene names - medical terms ambiguity
- Gene names - common English words ambiguity ( fly)
- Alternative typographical variants
- 14% of genes display inter-species ambiguity {Chen, 2005}.
- Ambiguity between protein names and their protein family names
- Identification of new gene names (novel genes)

# SOME TRICKY CASES OF GENE TAGGING

(1) The nightcap mutation caused severe defects in these cells [PMID:12399306]

(2) In the present investigation, we have discovered that Piccolo, a CAZ (cytoskeletal matrix associated with the active zone) protein in neurons that is structurally related to Rim2, [PMID:12401793]

(3) The Drosophila takeout gene is regulated by the somatic sex-determination pathway and affects male courtship behavior. [PMID:12435630]

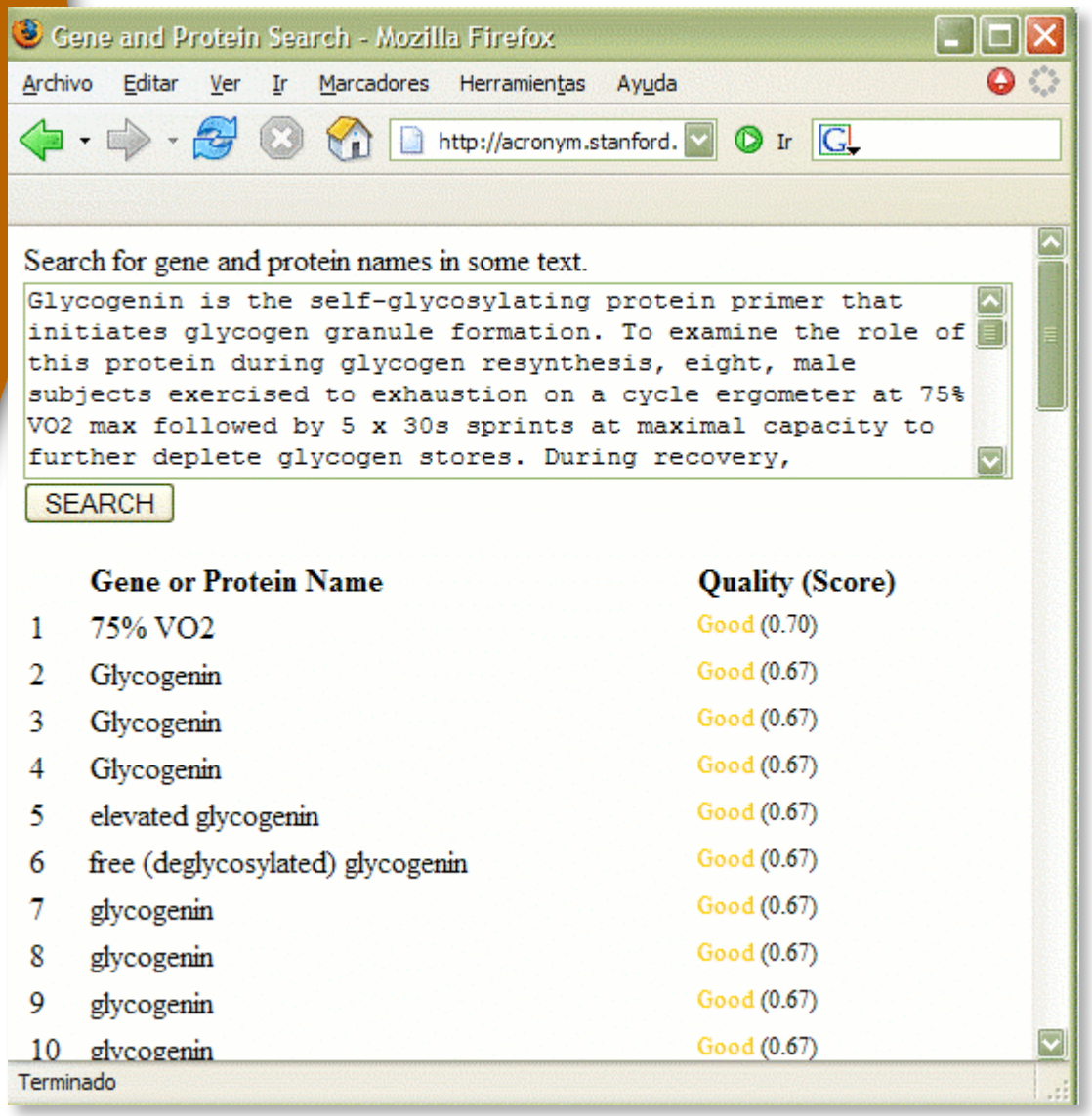
(4) This function is independent of Chico, the Drosophila insulin receptor substrate (IRS) homolog [PMID:12702880].

(5) A new longevity gene, Indy (for I'm not dead yet), which doubles the average …. [PMID:12391301]

(6) The Drosophila peanut gene is required for cytokinesis and encodes a protei similar to yeast putative bud neck filament proteins [PMID 8181057].

(6) Ambiguity of PKC: Protein kinase C and Pollution kerato-conjunctivitis

# GAPSCORE: GENE TAGGING




- **Scores words based**
- **on a statistical model of gene names**
- **Quantifies:**
  - **Appearance**
  - **Morphology**
  - **Context.**
- **Online.**

Chang JT, Schütze H, and Altman RB. GAPSCORE: Finding Gene and Protein Names One Word at a Time. *Bioinformatics*. 2004 Jan 22;20(2):216-25.

http://bionlp.stanford.edu/gapscore

# NLProt: GENE TAGGING

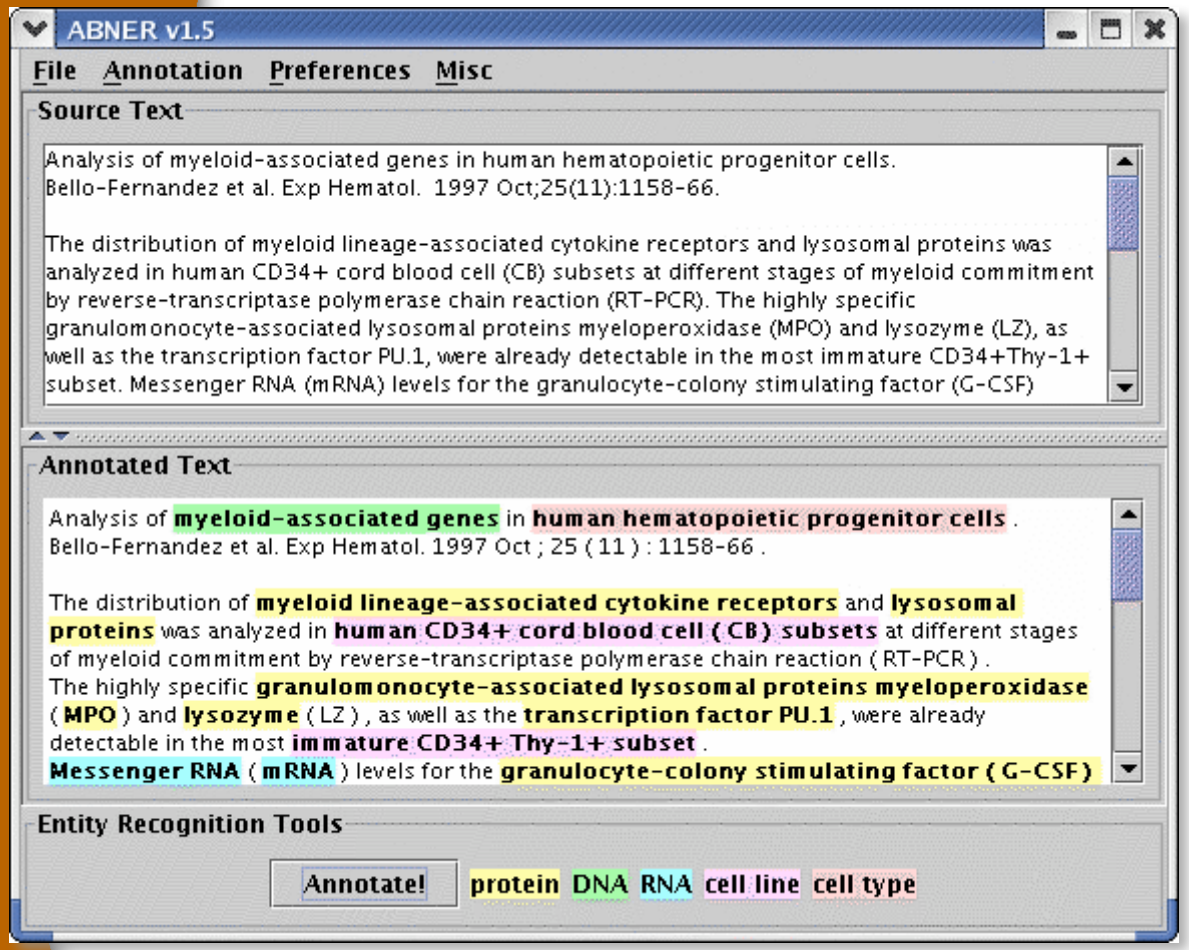
| NAME | ORGANISM | TXT-POS | SCORE | METHOD | DB-ID(S) |
|---|---|---|---|---|---|
| Glycogenin | homo sapiens | 1 | 1.040 | SVM | GYG2 HUMAN (86%) |
| glycogenin | homo sapiens | 96 | 0.856 | SVM | GYG2 HUMAN (91%) |
| glycogenin | homo sapiens | 103 | 1.040 | SVM | GYG2 HUMAN (91%) |
| Glycogenin | homo sapiens | 109 | 0.871 | SVM | GYG2 HUMAN (86%) |
| glycogenin | homo sapiens | 138 | 0.980 | SVM | GYG2 HUMAN (91%) |
| Glycogenin | homo sapiens | 157 | 0.971 | SVM | GYG2 HUMAN (86%) |
| glycogenin | homo sapiens | 161 | 0.311 | SVM | GYG2 HUMAN (91%) |
| glycogenin | homo sapiens | 214 | 0.819 | SVM | GYG2 HUMAN (91%) |
| glycogenin | homo sapiens | 234 | 0.747 | SVM | GYG2 HUMAN (91%) |

- **Online (e-mail alert).**
- **Downloadable.**
- **SVM-based**
- **Pre-processing dictionary**
- **Rule-based filtering step**
- **PubMed words.**
- **Precision of 75%**
- **Recall of 76%**
- **Provides reliability score**

http://cubic.bioc.columbia.edu/services/nlprot/

Mika, S, Rost, B NLProt: extracting protein names and sequences from papers. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W634-7.

# ABNER: GENE TAGGING




- Downloadable.
- Conditional Random Fields (CRF)-based
- Trained on BioCreative and GENIA
- Orthographic and contextual features
- Can be trained on
- new corpora
- Genes, proteins, cell lines
- Java-based

Burr Settles. "ABNER: A Biomedical Named Entity Recognizer." http://www.cs.wisc.edu/~bsettles/abner/. 2004.

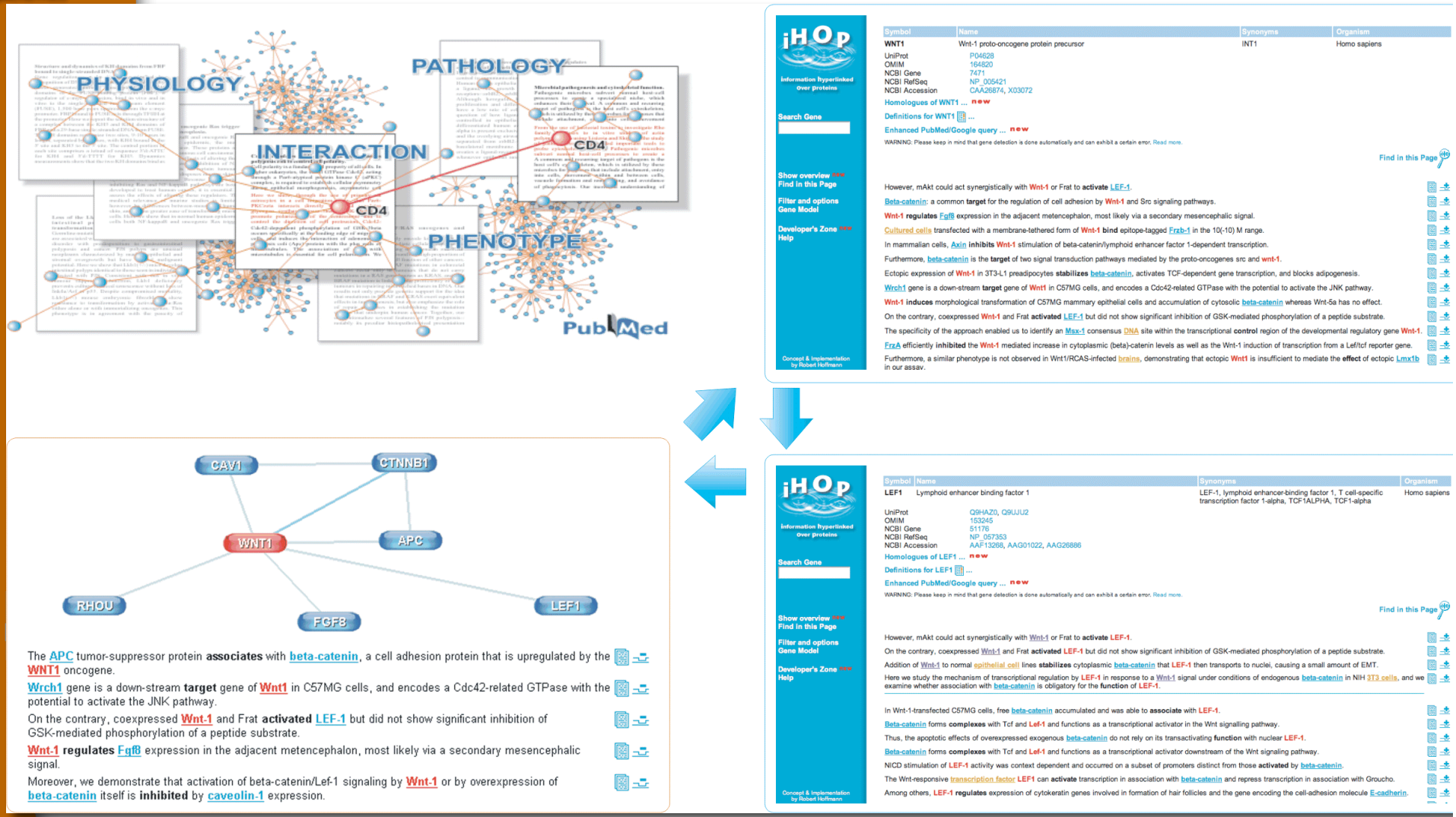
# Biomedical Named Entity Recognizer

Activated protein C ( APC ) is a serine protease that plays a central role in physiological anticoagulation , and has more recently been shown to be a potent anti inflammatory mediator . We show here that APC upregulates the angiogenic promoters , vascular endothelial growth factor ( VEGF ) , monocyte chemoattractant protein 1 ( MCP 1 ) , interleukin 8 ( IL 8 ) or matrix metalloproteinase 2 ( MMP 2 ) in cultured human skin fibroblasts ( HF ) , keratinocytes ( HK ) or umbilical vein endothelial cells ( HUVE ) . In the chick embryo chorio allantoic membrane assay , APC promoted angiogenesis. In a full thickness rat skin healing model , a single topical application of APC enhanced wound healing compared to saline control . In summary , our results demonstrate that APC promotes cutaneous wound healing at least partly by stimulation of angiogenesis.

- **Based on Machine learning**
- **Good results in the COLING Bio-NER contest (Geneva)**
- **Many classes (entity types), including** Virus, Tissue, RNA, Protein, Polynucleotide, Peptide, Organism, Nucleotide, Lipid, DNA, Cell Type, Cell Line, Cell Component, Carbohydrate, Body Part Atom and Amino Acid Monomer

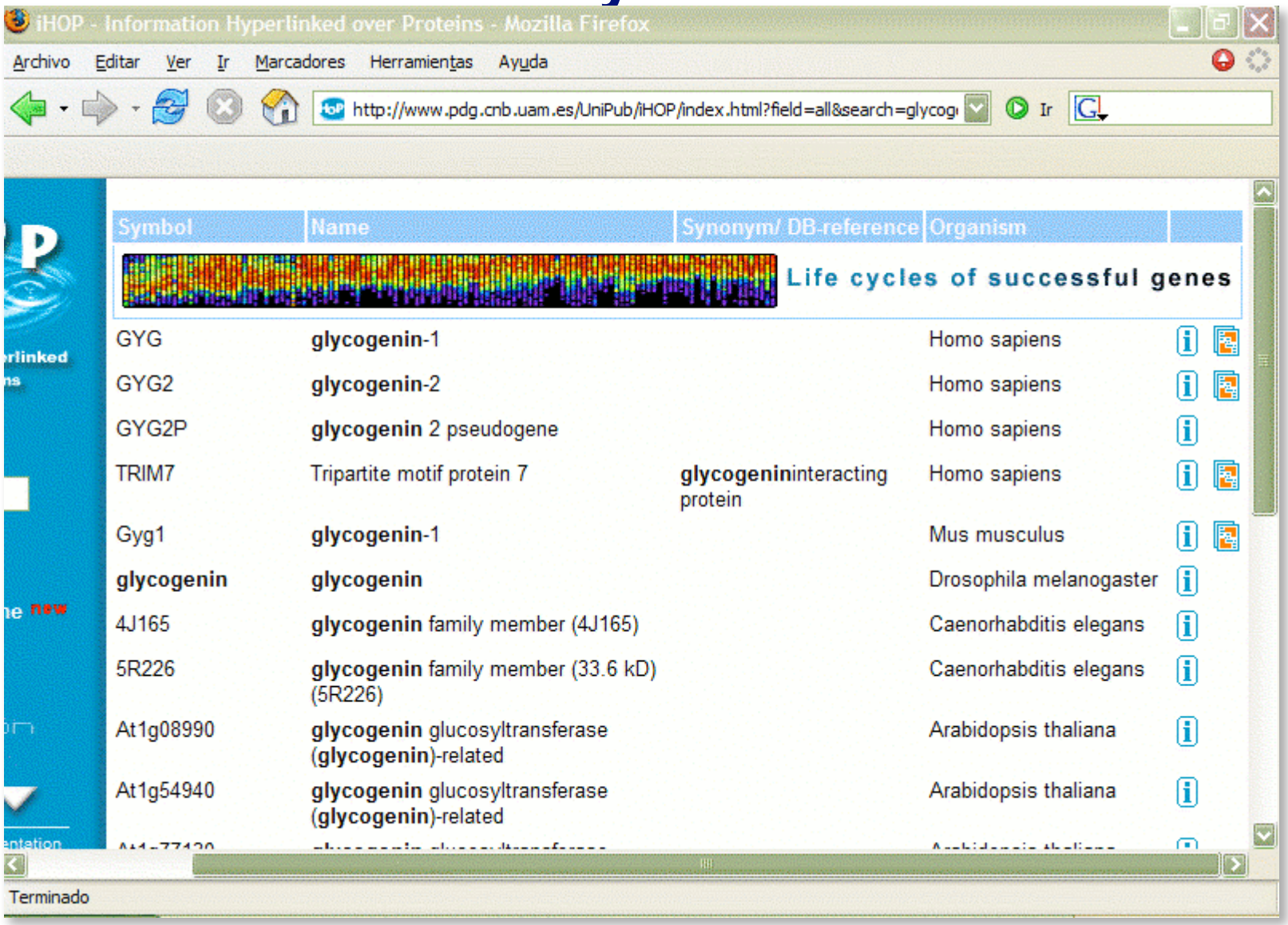
http://textmining.i2r.a-star.edu.sg/NLS/webdemo/bioner.html

# iHOP system




Hoffmann R, Valencia A. A gene network for navigating the literature *Nat Genet*. 2004 Jul;36(7):664.

# iHOP system




Hoffmann R, Valencia A. A gene network for navigating the literature *Nat Genet*. 2004 Jul;36(7):664.

# iHOP system

iHOP - Information Hyperlinked over Proteins [ GYG ] - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://www.pdg.cnb.uam.es/UniPub/iHOP/gs/88913.html?IN=1

Search

Home | Bookmarks | Yahoo | Google | MK Homepage | ORF | Zope on http://... | PubMed | Python | Zope | PyTut | OEAW | GeneDic | biocreative | GenomeNet

iHOP

information hyperlinked over proteins

Search Gene

Show overview
Find in this Page

Filter and options
Gene Model

Developer's Zone

Help

Concept & Implementation by Robert Hoffmann

| Symbol | Name | Synonyms | Organism |
|---|---|---|---|
| GYG | Glycogenin-1 | glycogenin, GYG1 | Homo sapiens |

| | |
|---|---|
| UniProt | P46976, Q9UNV0 |
| OMIM | 603942 |
| NCBI Gene | 2992 |
| NCBI RefSeq | NP_004121 |
| NCBI Accession | AAB00114, AAB09752, AAD31084 |

Homologues of GYG ... new

Definitions for GYG ...

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

Mutation of Tyr-196 in glycogenin-2 to a Phe residue abolished the ability of glycogenin-2 to self-glucosylate but not to **interact** with glycogenin-1.

Mutational analysis of the coding regions of the genes encoding protein kinase B-alpha and -beta, phosphoinositide-dependent protein kinase-1, phosphatase targeting to glycogen, protein phosphatase inhibitor-1, and glycogenin: lessons from a search for genetic variability of the insulin-**stimulated** glycogen synthesis pathway of skeletal muscle in NIDDM patients.

Effects of exercise on GLUT-4 and glycogenin gene expression in human skeletal muscle.

The third cDNA encoded a polypeptide of unknown function and was designated GNIP (glycogenin interacting protein).

GNIP, a novel protein that binds and activates glycogenin, the self-glucosylating initiator of glycogen biosynthesis.

Overall, GN-2 has 40-45% identity to muscle glycogenin but is 72% identical over a 200-residue segment thought to contain the catalytic domain.

Glycogenin-1 and glycogenin-2 interact with one another, based on in vitro interactions and co-immunoprecipitation from liver and cell extracts.

Mouse glycogenin-1 has a predicted molecular mass of 37¿ omitted¿399 Da, and the deduced amino acid sequence exhibited 87% homology with human glycogenin-1.

For the first time, we report that a single bout of exercise is sufficient to cause upregulation of GLUT-4 and glycogenin gene expression in human skeletal muscle.

Fasting plasma insulin concentrations, muscle creatine, glycogen and GLUT-4 protein content as well as GLUT-4, glycogen synthase-1 (GS-1) and glycogenin-1 (Gln-1) mRNA expression were determined.

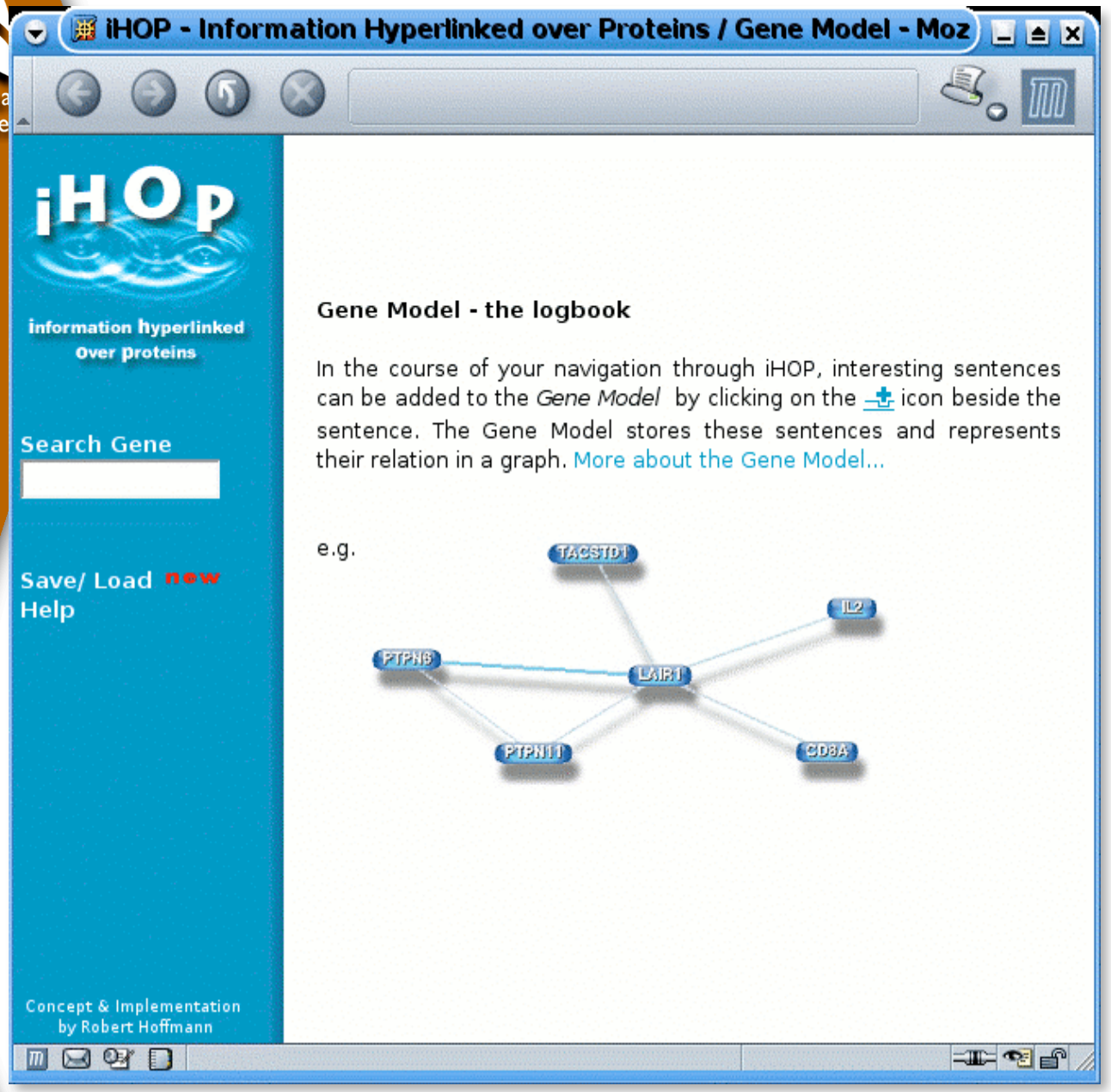
In conclusion, the co-expression of glycogenin with GLUT3 might enable glycogen-storing cells to exchange glucose quite effectively according to prevailing metabolic demands of glycogen synthesis or degradation.

The discovery of a second human gene, GYG2, encoding a liver-specific isoform of glycogenin, the self-glucosylating initiator of glycogen biosynthesis, raised the possibility for differential controls of this protein in liver and muscle.

The present study investigated the expression of glycogenin, the protein primer for glycogen synthesis, and the high affinity glucose transporter isoform GLUT3 as a further potential regulator of cellular glycogen metabolism, in first trimester and term human placenta using immunohistochemistry and

Transferring data from www.pdg.cnb.uam.es...

# iHOP system




## iHOP: Visualization of protein interactions using network graphs

Hoffmann R, Valencia A.
A gene network for navigating the literature *Nat Genet*. 2004 Jul;36(7):664.

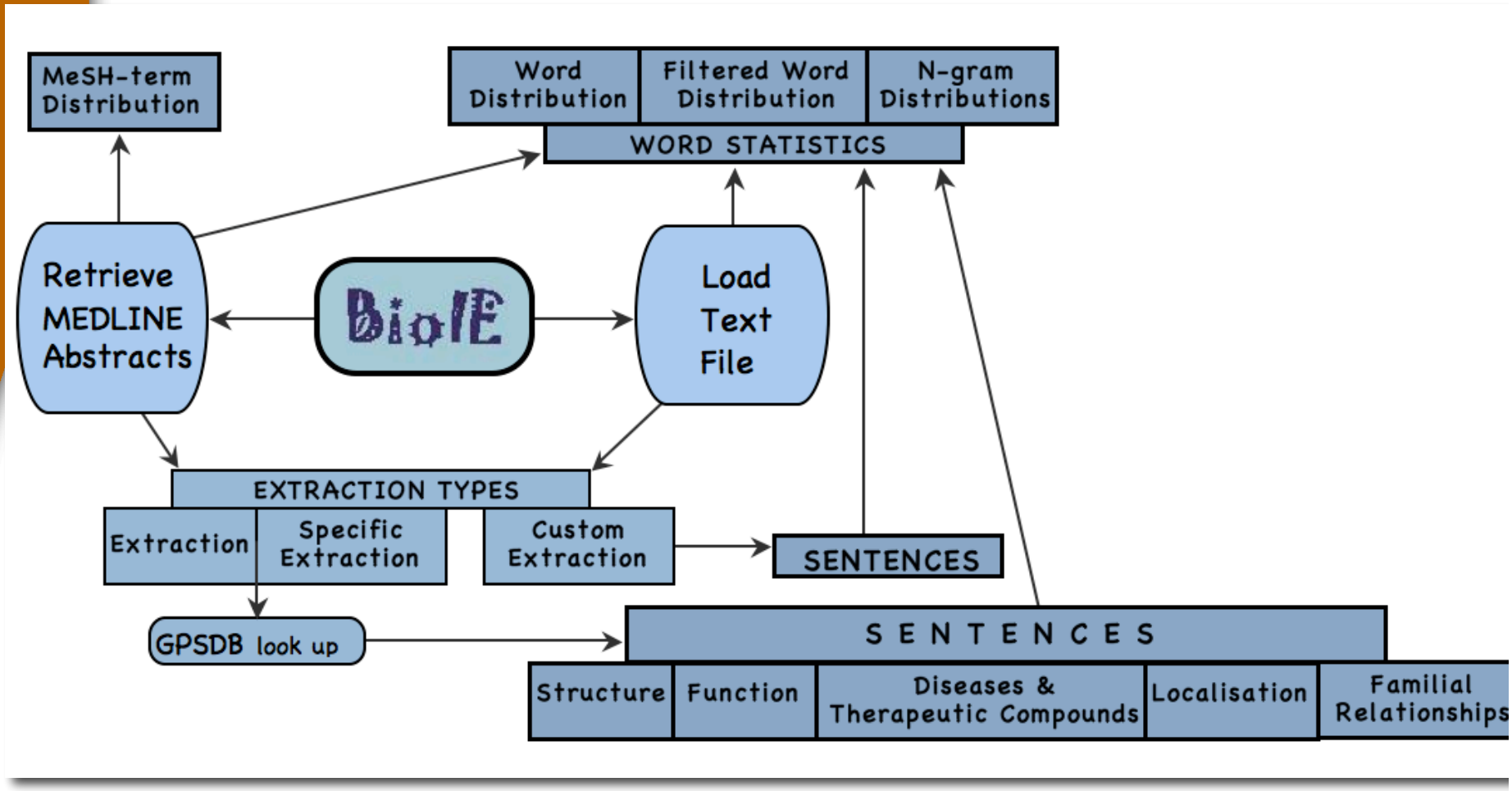
# METIS and BioIE: Text mining annotations

- Rule-based method for identifying informative sentences.
- For a given query term (e.g. gene names) it returns a list of sentences that match a set of manually defined templates and rules.
- The identified templates are highlighted within these sentences
- Information on word distributions is provided.
- This system (available as an on-line server)
- Focuses on predefined categories, namely structure, function, diseases and therapeutic compounds, location and family relationships

http://umber.sbs.man.ac.uk/dbbrowser/bioie/

Divoli, A. Atwood TK, BioIE: extracting informative sentences from the biomedical literature. Bioinformatics. 2005 May 1;21(9):2138-9. Epub 2005 Feb 2.

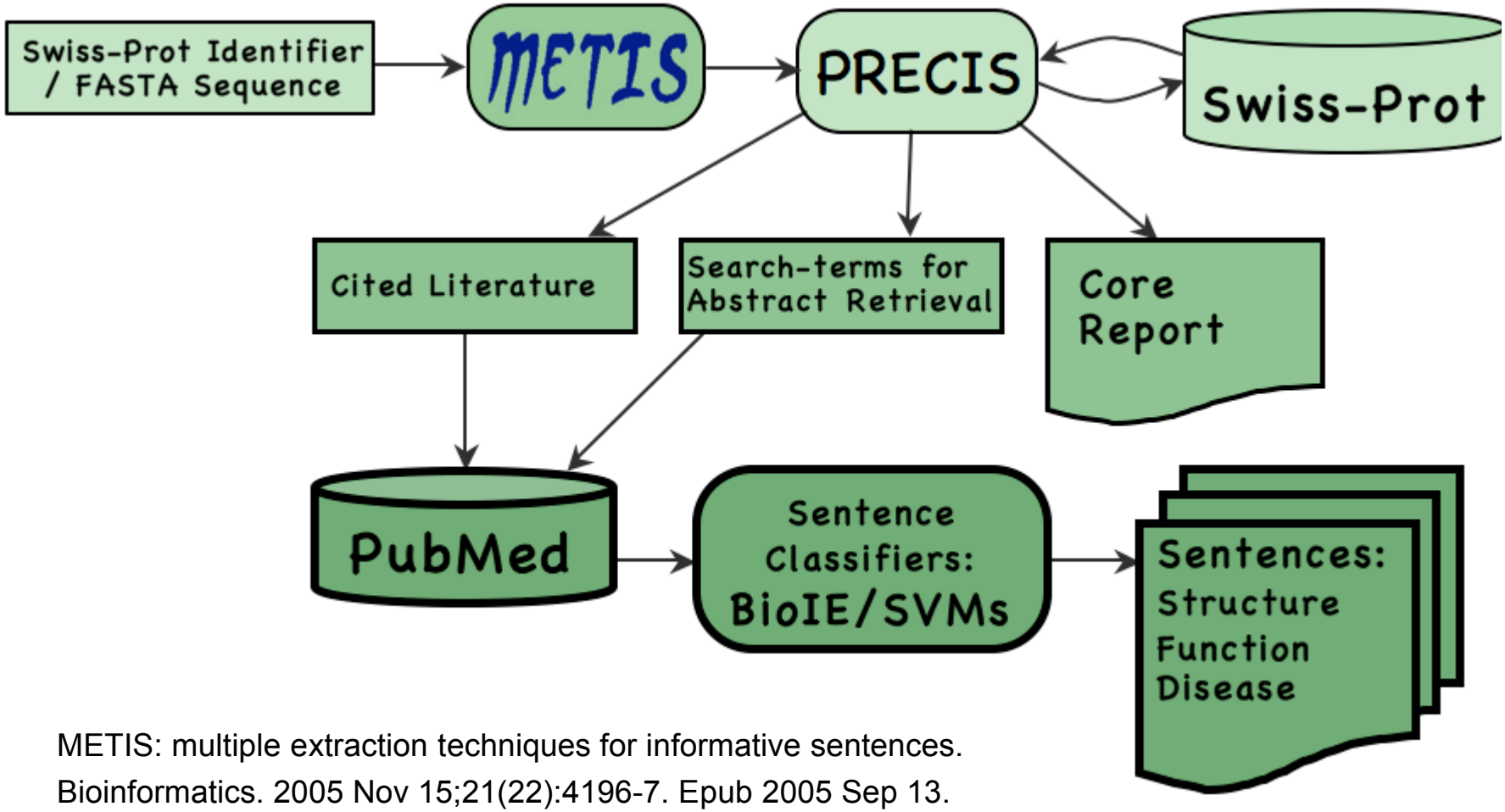
# BioIE: Text mining annotations




http://umber.sbs.man.ac.uk/dbbrowser/bioie/

Divoli, A. Atwood TK, BioIE: extracting informative sentences from the biomedical literature. Bioinformatics. 2005 May 1;21(9):2138-9. Epub 2005 Feb 2.

# METIS and BioIE: Text mining annotations




METIS: multiple extraction techniques for informative sentences. Bioinformatics. 2005 Nov 15;21(22):4196-7. Epub 2005 Sep 13.

http://umber.sbs.man.ac.uk/cgi-bin/dbbrowser/precis/metis_precis.cgi

# METIS and BioIE: Text mining annotations

1. Blast search in the Swiss-Prot database.
2. Structured report is generated
3. Informative sentences are extracted from PubMed
4. Use BioIE as well as a SVM-based sentence classifier.

METIS: multiple extraction techniques for informative sentences.
Bioinformatics. 2005 Nov 15;21(22):4196-7. Epub 2005 Sep 13.

http://umber.sbs.man.ac.uk/cgi-bin/dbbrowser/precis/metis_precis.cgi

# Textpresso

Textpresso - Mozilla Firefox

Archivo Editar Ver Ir Marcadores Herramientas Ayuda

http://www.textpresso.org/

Home

Simple Retrieval

Simple Retrieval (sorted by year)

Advanced Retrieval

Ontology

DTD

User Guide

## Simple Retrieval

Query should be met in a sentence publication.

Type in keywords to be searched for, separated by **white spaces:**

Exact match

Specify categories that should also be met (optional):

none , none .

Search in

Abstract Author Full Text Title Year

Search!

## News and Messages

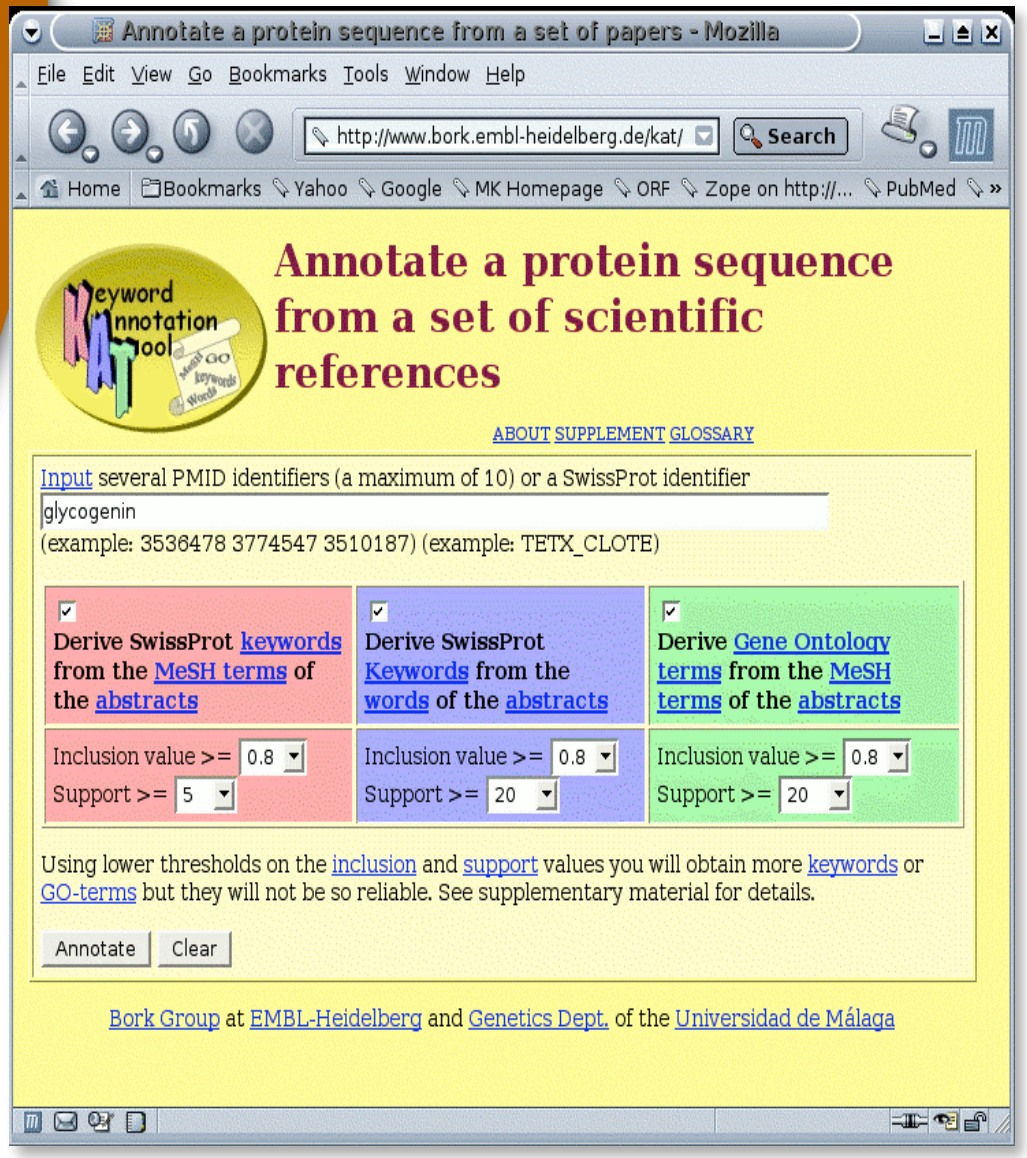
**Welcome to *Textpresso* ! The *Textpresso* search engine for *C. elegans* abstracts and fulltexts was developed at Wormbase to service the *C. elegans* community, and is being expanded to serve other communities.**

**March 21st, 2005:** A new build that contains over 1,000 new full text papers and an updated Textpresso ontology (version 1.1) is released. The new ontology version has four new categories; Reporter Gene, Restriction Enzyme, Second Messenger and Vector. In addition, new terms have been added to the Drugs and Small Molecules and Organism categories. Details of the updated ontology can be found by following the "Ontology" link on the menu to the left of this page.

**March 1st, 2005:** We are current testing a new build that contains 4 new categories and over 1,000 new full text papers on the Development Site. This site will be under a heavy barrage of testing and may experience

Terminado

# KEYWORD ANNOTATION TOOL (KAT)




- **Extraction of mappings between related terms using a model of fuzzy associations**
- **Mesh terms/SwissProt keywords/GO terms**

Perez AJ, Perez-Iratxeta C, Bork P, Thode G
Andrade MA.
Gene annotation from scientific literature usi
mappings between keyword systems.
Bioinformatics . 2004 Sep 1;20(13):2084-91.
Epub 2004 Apr 1.

# GOPUBMED


http://www.gopubmed.org/

# PROTEIN INTERACTIONS

- Advances in experimental large scale protein interaction analysis
- Exp. Methods for protein interaction characterization:
  - protein arrays
  - mRNA expression microarrays
  - Yeast two-hybrid
  - Affinity purification with MS
  - X-ray, NMRFRET, chemical cross-linking,..
- Bioinformatics methods for protein characterization:
- Genome-based
- Sequence-based

# PROTEIN INTERACTION DATABASES

| Database Name | Reference | URL |
|---|---|---|
| BIND | (Bader *et al.*, 2003) | http://bind.ca |
| DIP | (Xenarios *et al.*, 2002) | http://dip.doe-mbi.ucla.edu |
| GRID | (Breitkreutz *et al.* 2003) | http://biodata.mshri.on.ca/grid |
| HPID | (Han *et al.*, 2004) | http://www.hpid.org |
| HPRD | (Peri *et al.*, 2004) | http://www.hprd.org |
| IntAct | (Hermjakob *et al.*, 2004) | http:/www.ebi.ac.uk/intact |
| MINT | (Zanzoni *et al.*, 2002) | http://cbm.bio.uniroma2.it/mint |
| STRING | (vonMering et al., 2003) | http://string.embl.de |
| ECID | (Juan *et al.*, 2004) | http://www.pdg.cnb.uam.es/ECID |

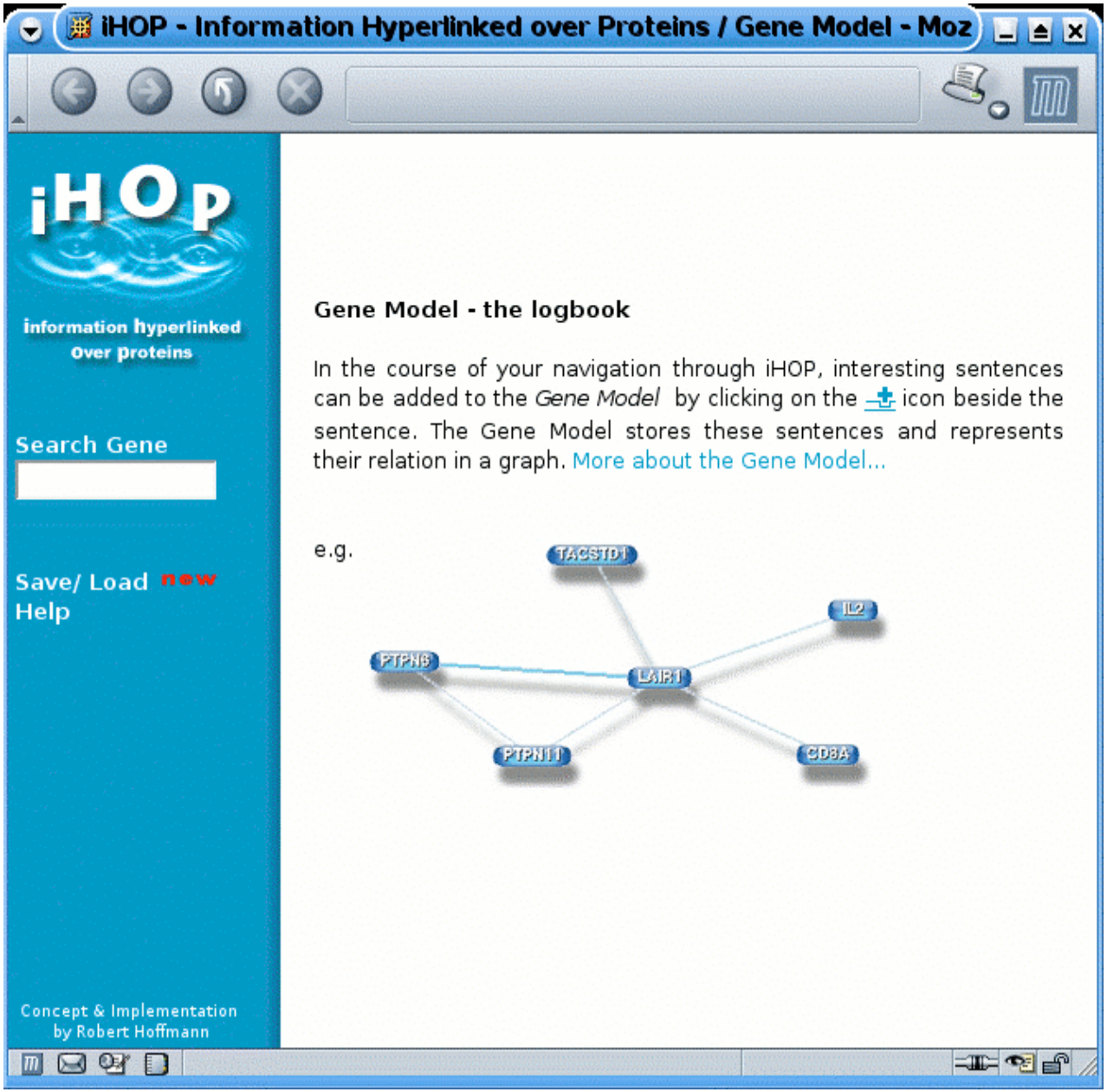
# PubGene: PROTEIN INTERACTION

- Use the co-occurrence of protein and gene names.
- Assumption: co-occurrence imply biological relationship
- Indexing PubMed abstracts and titles with human proteins.
- Construction of interaction networks.
- Build upon binary interactions between co-occurring proteins

http://www.pubgene.org/

Jenssen TK, Laegreid A, Komorowski J, Hovig E.
A literature network of human genes for high-throughput analysis of gene expression.Nat Genet. 2001 May;28(1):21-8.

cnio
Centro Nacional de Investigaciones Oncológicas

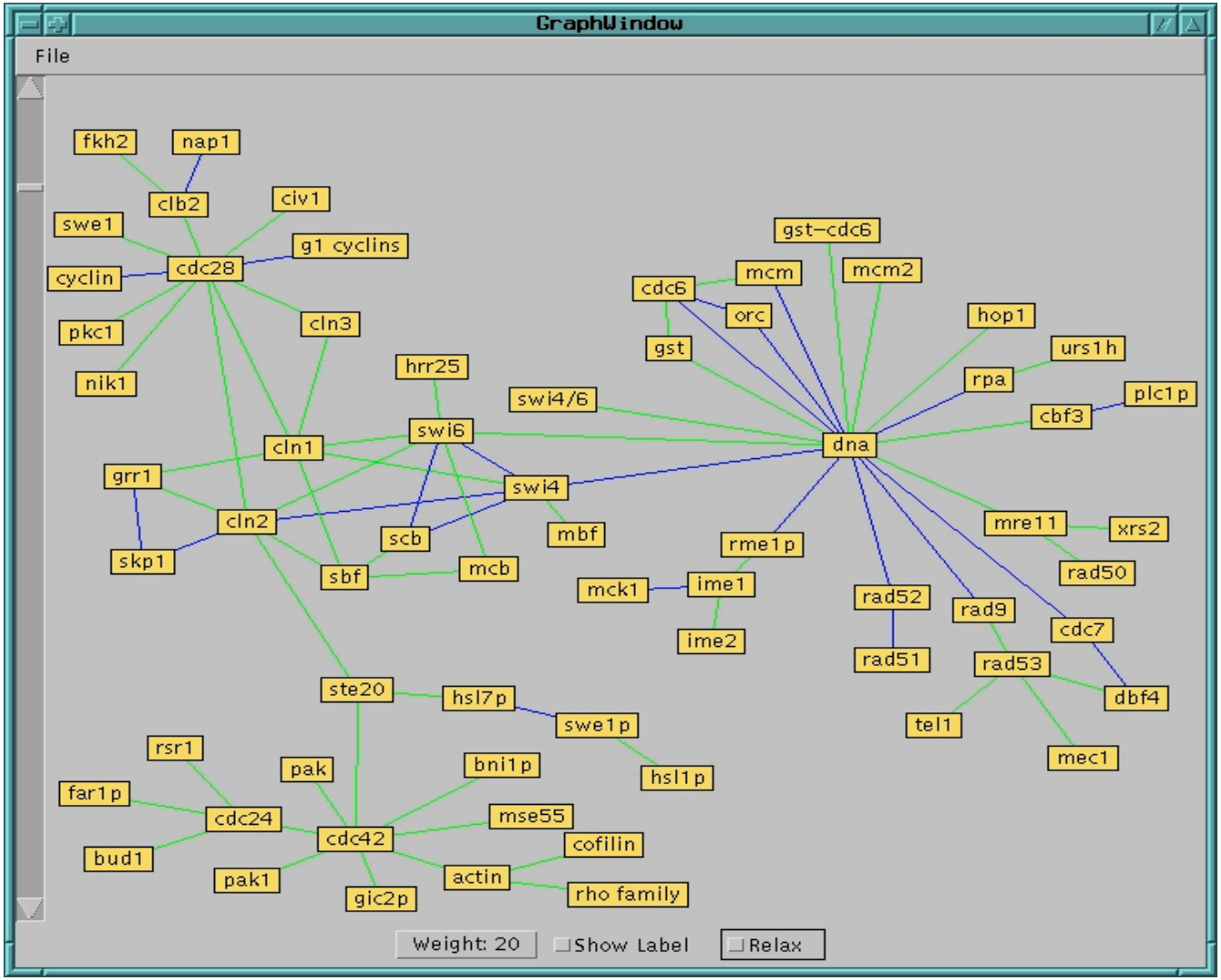



MK

# SUISEKI

- Relationship between the co-occurring proteins using frames
- Frames: textual patterns used to express interactions
- Initial set of 14 interaction words based on domain knowledge.
- Examples: *activate*, *bind*, *suppress*
- Analyzed the order of protein names within sentences.
- Take into account distance (off-set) between protein names.
- System effective for simple interaction types.
- Difficult cases: long sentences with complex grammatical structures

# SUISEKI (2)


GraphWindow
File
fkh2
nap1
clb2
civ1
swe1
g1 cyclins
cyclin
cdc28
pkc1
cln3
nik1
hrr25
swi4/6
swi6
cln1
grr1
swi4
cln2
scb
mbf
skp1
sbf
mcb
gst-cdc6
mcm
mcm2
cdc6
orc
gst
hop1
urs1h
rpa
plc1p
cbf3
dna
mre11
xrs2
rad50
rme1p
mck1
ime1
ime2
rad52
rad9
cdc7
rad51
rad53
dbf4
tel1
mec1
ste20
hsl7p
swe1p
hsl1p
rsr1
pak
bni1p
far1p
cdc24
mse55
cdc42
cofilin
bud1
pak1
actin
rho family
gic2p
Weight: 20
Show Label
Relax

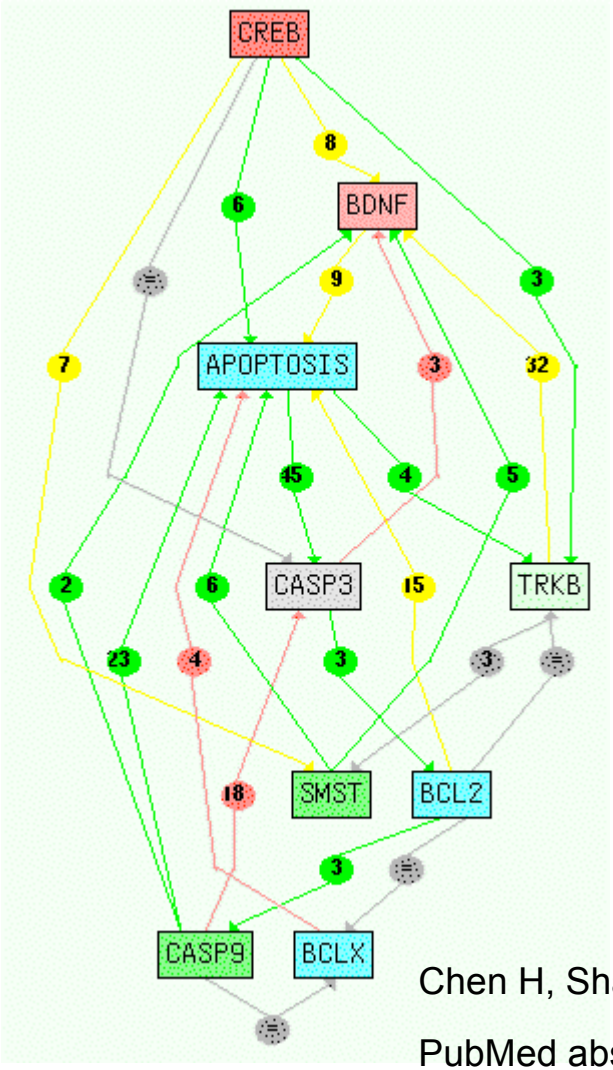
# CHILIBOT (1)

- NLP-based text mining approach.
- Content-rich relationship networks among biological
- Concepts, genes, proteins or drugs.
- Nature of the relationship: inhibitory, stimulative, neutral and simple co-occurrence.
- Internet-based application with graphical visualization
- Sentence as unit, POS tagging, shallow parsing and rules

http://www.chilibot.net/

Chen H, Sharp BM.Content-rich biological network constructed by mining PubMed abstracts.BMC Bioinformatics. 2004 Oct 8;5(1):147.

# CHILIBOT (2)




- **Need registration.**
- **Hypothesis generation.**

http://www.chilibot.net/

Chen H, Sharp BM.Content-rich biological network constructed by mining PubMed abstracts.BMC Bioinformatics. 2004 Oct 8;5(1):147.

# Microarray data analysis

- Coordinated expression of genes.
- Functional co-regulation within biological processes.
- Mine micro array data using the associated biomedical literature.
- Characterize groups of genes extracting functional keywords.
- Score the coherence of gene clusters.
- Group genes based on their associated literature and functional descriptions.

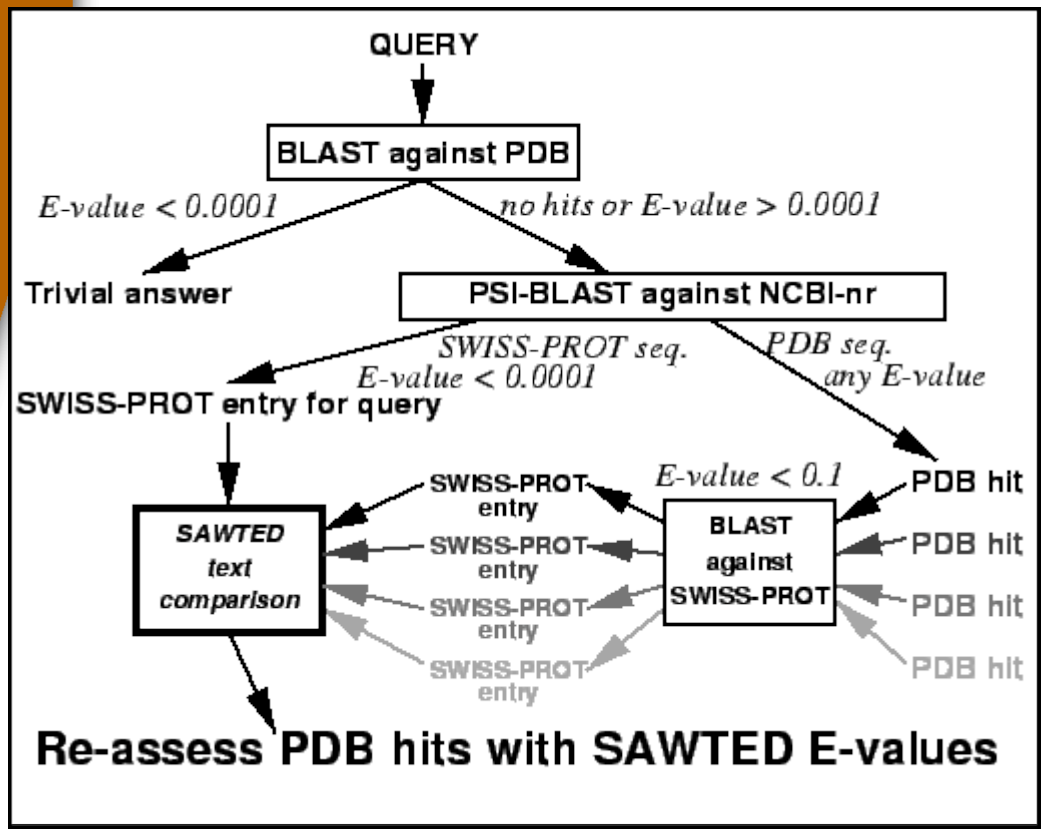
# PROTEIN SUBCELLULULAR LOCATION

- Protein activity -> specific cellular environments.
- Localization determination:
- Experimental techniques.
- Bioinformatics techniques (PSORT).
- Text mining.
- Nair and Rost: lexical information in annotation database records.
- Stapley et al: Use SVM to classify proteins according to their subcellular localization, extracted from PubMed abstracts.

# NLP AND SEQUENCE ANALYSIS: MEDBLAST

- Use NLP techniques to retrieve the related articles for a given sequence (online).
- Related articles:
  - those describing the query sequence (protein) or
  - Its redundant sequences and close homologues
- Direct search with the sequence.
- Indirect search with gene symbols.
- Use Blast against GenBank.
- Use Eutilities toolset to retrieve documents

# NLP AND SEQUENCE ANALYSIS: SAWTED




http://www.bmm.icnet.uk/~sawted/

**Structure Assignment Witl Text**

**Sequence similarity the base for identifying structure templates for query sequence**

**Structure Assignment With Text Description**

**Document comparison Algorithms**

**Used within 3D-PSSM**

# Resources for Bio-NLP

http://biocreative.sourceforge.net/index.html

- Literature databases : PubMed and PubMed Central.
- Annotated text corpora: GENIA corpus {Kim, 2003}, BioCreative corpus, Yapex corpus, the Genic Interaction Extraction Challenge provided both a training and test set {Nedellec, 2005}.
- General NLP tools: for statistical text analysis, the Bow Toolkit is very useful {McCallum Libbow},
- NLTK , CCG, the Porter stemmer {Porter, 1980}.
- Dictionaries and ontologies: Gene Ontology and The Unified Medical Language System and MeSH
- Biomedical domain NLP systems

# Why community assessments?

Compare different methods and strategies

Reproduce performance of systems on common data

Provide useful data collections: Gold Standard data

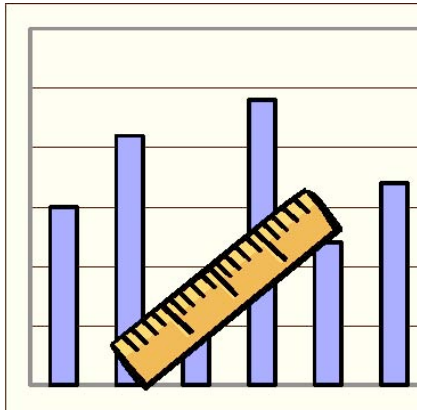
Explore meaningful evaluation strategies and tools

Determine the state of the art

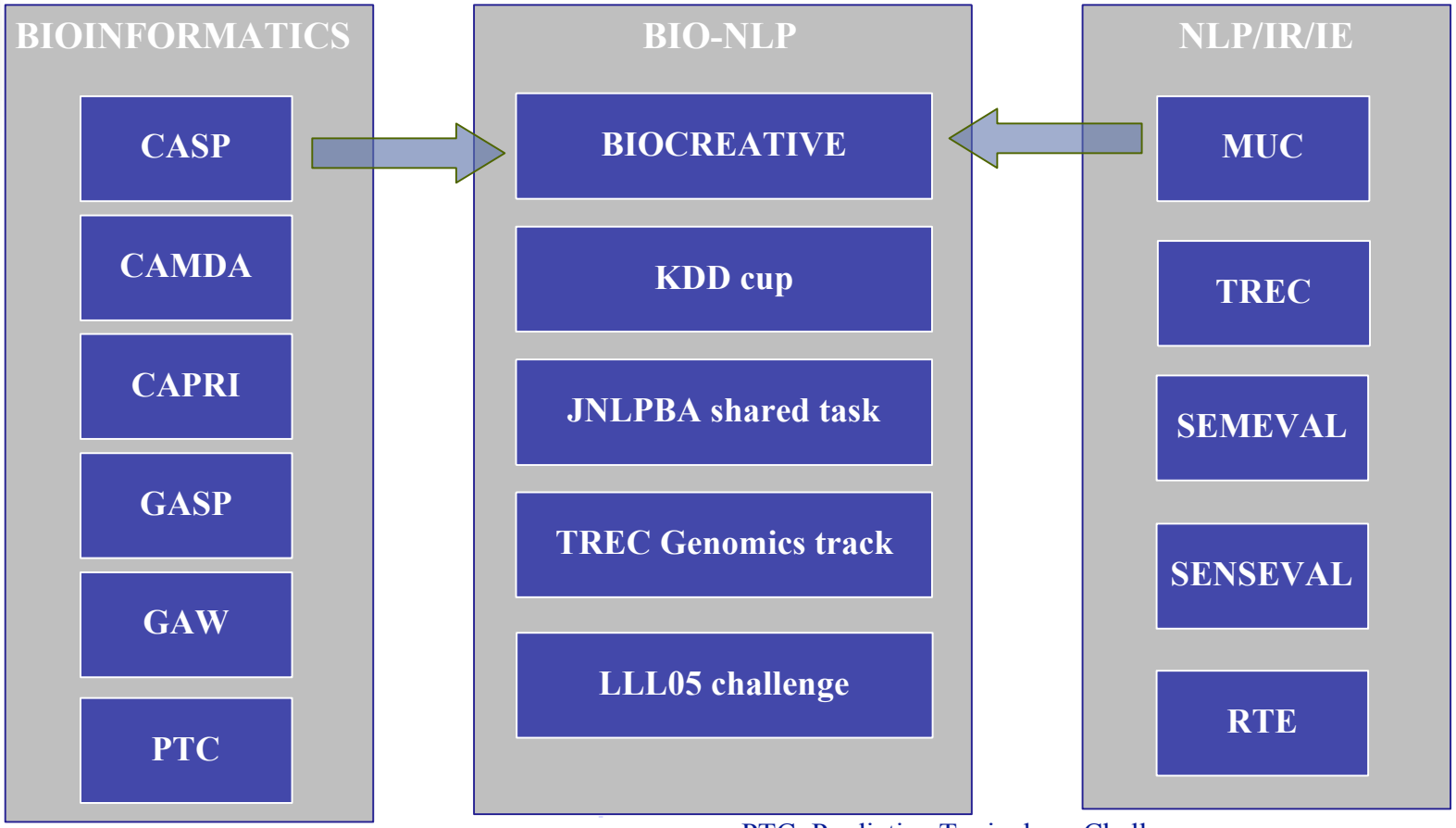
Monitor improvements in the field

Point out needs of the user community

Promote collaborative efforts



# Community evaluations



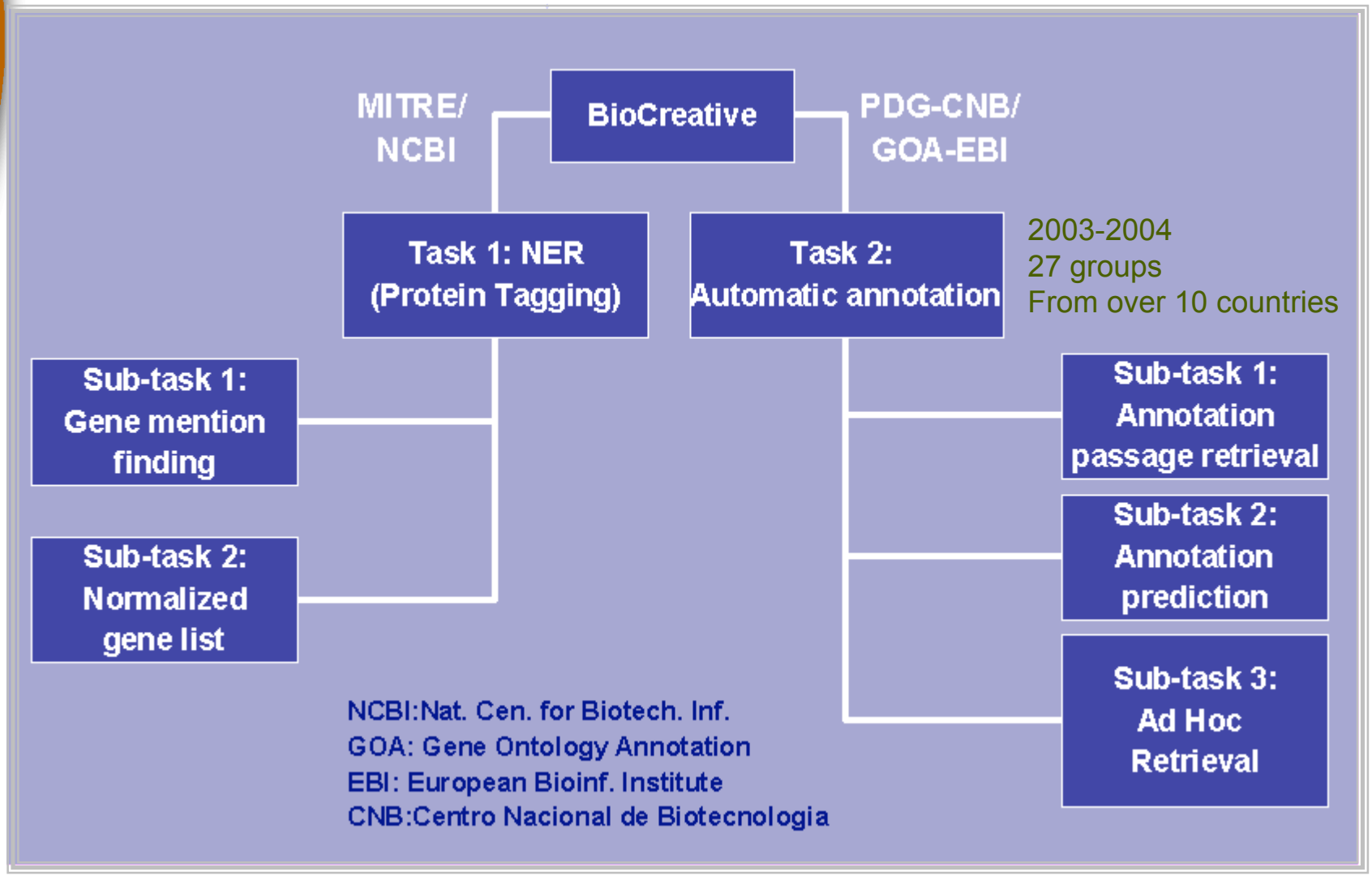

CASP: Critical assessment of Protein Structure Prediction
CAMDA: Critical Assessment of Microarray Data Analysis
CAPRI: Critical Assessment of Prediction of Interactions
GASP: Genome Annotation Assessment Project
GAW: Genome Access Workshop

PTC: Predictive Toxicology Challenge
KDD: Knowledge Discovery and Data mining
JNLPBA: Joint workshop on Natural Language Processing in Biomedi
TREC: Text Retrieval conference
MUC: Message Understanding conference
LLL05: Genic interaction extraction challenge
RTE: Textual Entailment challenge

# The BioCreative Challenge

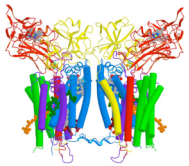
- Critical Assessment of Information Extraction systems in Biology
- Community challenge evaluation a community-wide effort for evaluating text mining and information extraction systems applied to the biological domain.
- Increasing nr. of groups working in the area of text mining, new systems, publications.
- Need of common standards or shared evaluation criteria to enable comparison
- Avoid the limitations of using private data sets: One system = one evaluation data set
- Promote development of systems which scale to real applications
- Community assessment of scientific progress: Monitor improvements
- Involve domain experts (end users) and biological database curators and domain experts
- Extraction of biologically relevant and useful information from the literature.

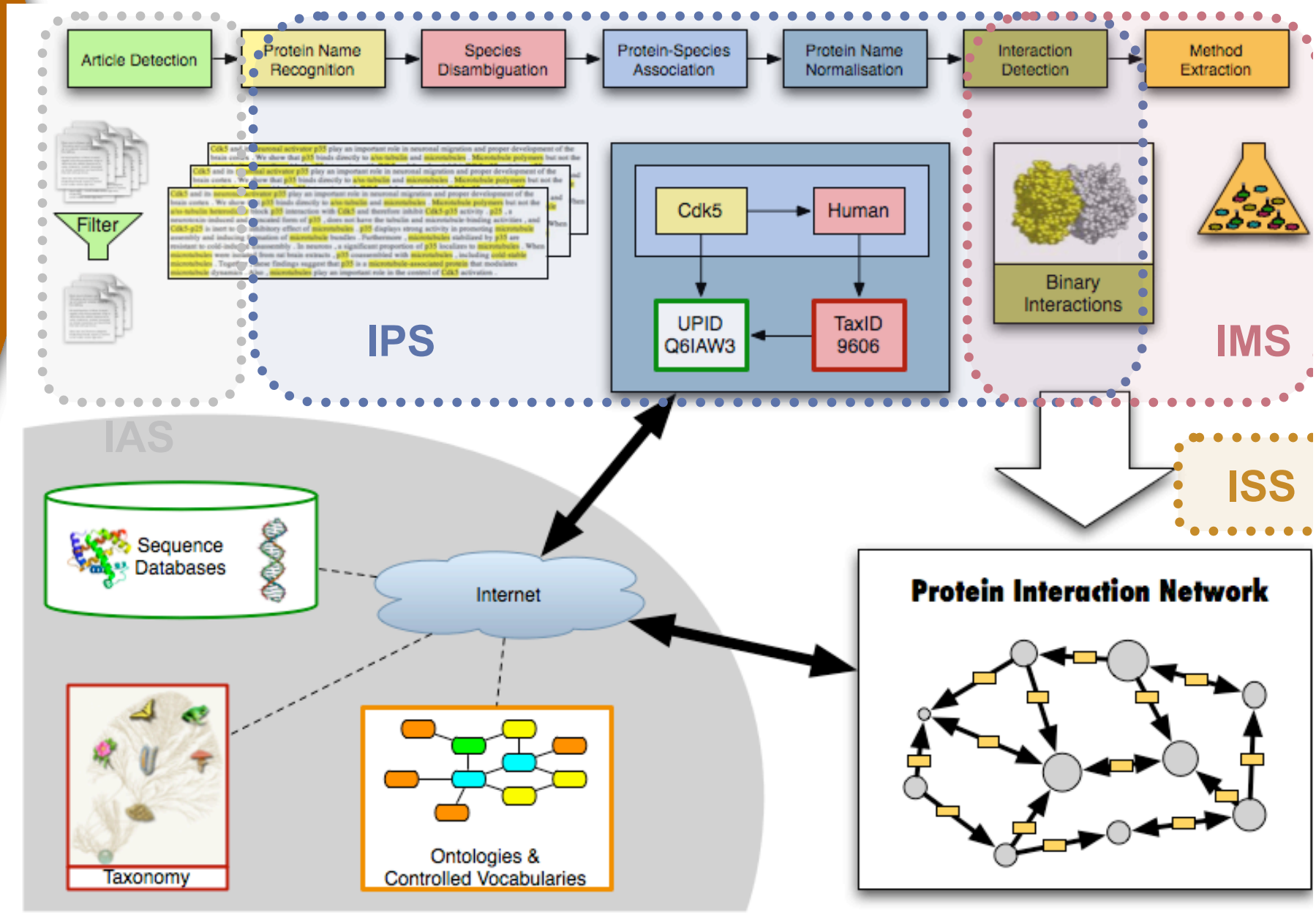
# BioCreative I



# Intro: Protein-Protein Interactions

- Protein interactions crucial for functional role and biological processes.
- High throughput yeast two-hybrid screening or affinity purification coupled with mass spectroscopy
- IntAct and MINT: interaction information in well structured database records in standard formats (PSI-MI, MI-ontology)
- IE and text mining techniques to automatically extract interaction information from free texts
- Rapid growth of literature databases and so of PPI publications
- No evaluation of PPI extracting systems from full text articles
- Annotations vs. statements: experimental characterizations
- Manual curation is time consuming and requires trained domain expert curators




Article Detection
Protein Name Recognition
Species Disambiguation
Protein-Species Association
Protein Name Normalisation
Interaction Detection
Method Extraction
Filter
IPS
Cdk5
Human
UPID Q6IAW3
TaxID 9606
Binary Interactions
IMS
IAS
ISS
Sequence Databases
Internet
Protein Interaction Network
Taxonomy
Ontologies & Controlled Vocabularies

# Conclusions and outlook

- There is an increasing interest in text mining applied to the biomedical and biology domain
- Important to assist in database curation
- Important for more efficient information access in biology
- Need of corpora and training data
- Need of community-wide evaluations
- Increasing diversification of applications
- Integration of bioinformatics and text mining

# Useful links, reviews and articles

R. Hoffmann, M. Krallinger, E. Andres, J. Tamames, C. Blaschke and A. Valencia. Text Mining for Metabolic Pathways, Signaling Cascades, and Protein Networks. Science STKE 283, pe21 (2005).

M. Krallinger, R. Alonso-Allende Erhadt and A. Valencia. Text-mining approaches in molecular biology and biomedicine. Drug Discovery Today 10, 439-445 (2005).

M. Krallinger and A. Valencia. Applications of Text Mining in Molecular Biology, from name recognition to Protein interaction maps. In Data Analysis and Visualization in Genomics and Proteomics, chapter 4, Wiley.