Text mining the Biomedical Literature

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

Van Helsing in Bram Stockers Dracula

Talk overview:

- The Biomedical literature
- Natural language processing (NLP)
- NLP in the Molecular Biology domain
- Text mining applications
- Evaluation of Text mining tools
- Conclusions and outlook
- Useful links, reviews and articles

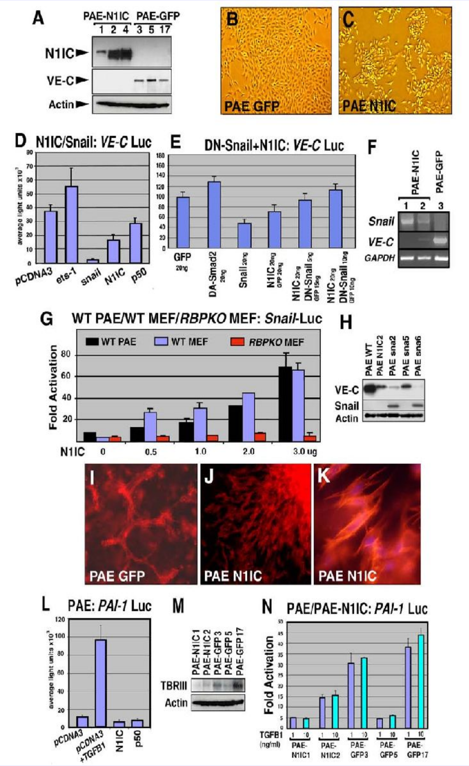
From experiments to scientific publications

1- Experiments

Planning and carrying out experiments (lab work)

2- Results

Processing and interpretation of obtained results

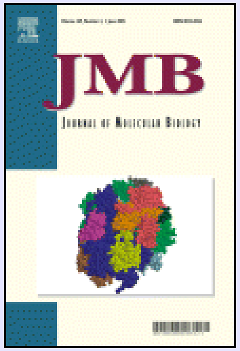


3- Scientific Peer-reviewed articles

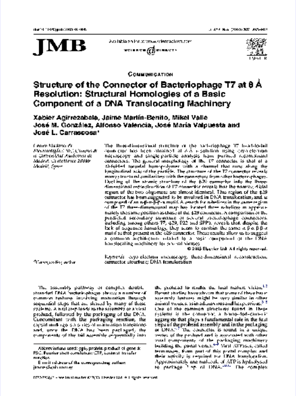
'Relevant' results are published in scientific journals

Data in scientific articles

Scientific Journals



Free Text


JMB

Communication

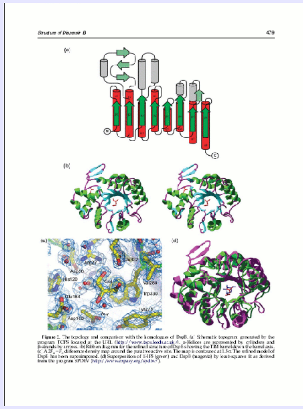
Structure of the Connector of Bacteriophage T7 at 8 Å Resolution: Structural Homologies of a Basic Component of a DNA Translocating Machinery

Xabier Agirrezabala, Jaime Martín-Benito, Mikel Valle José M. González, Alfonso Valencia, José María Valpuesta and José L. Carrascosa*

- Title
- Abstracts
- Keywords
- Text body
- References

Tables

Figures



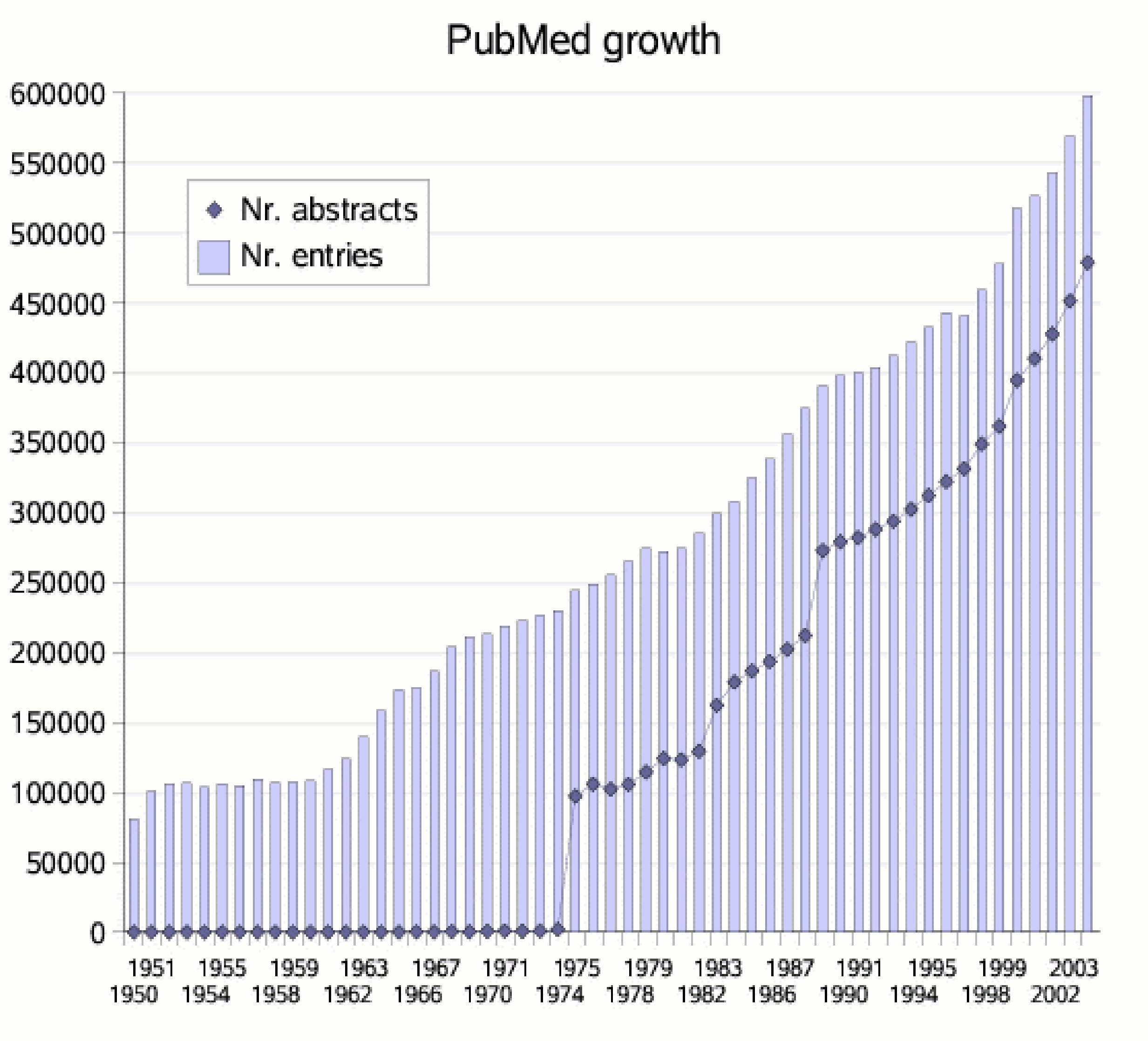
Journal-specific Information:

- Format
- Paper structure (sections)
- Article type

Biomedical literature characteristics

- Heavy use of domain specific terminology (12% biochemistry related technical terms).
- Polysemic words (word sense disambiguation).
- Most words with low frequency (data sparseness).
- New names and terms created.
- Typographical variants
- Different writing styles (native languages)

PubMed/Medline database at NCBI



PubMed

- Developed at the National Center for Biotechnology Information (NCBI).

- The core 'Textome'.

- repository of citation entries of scientific articles.

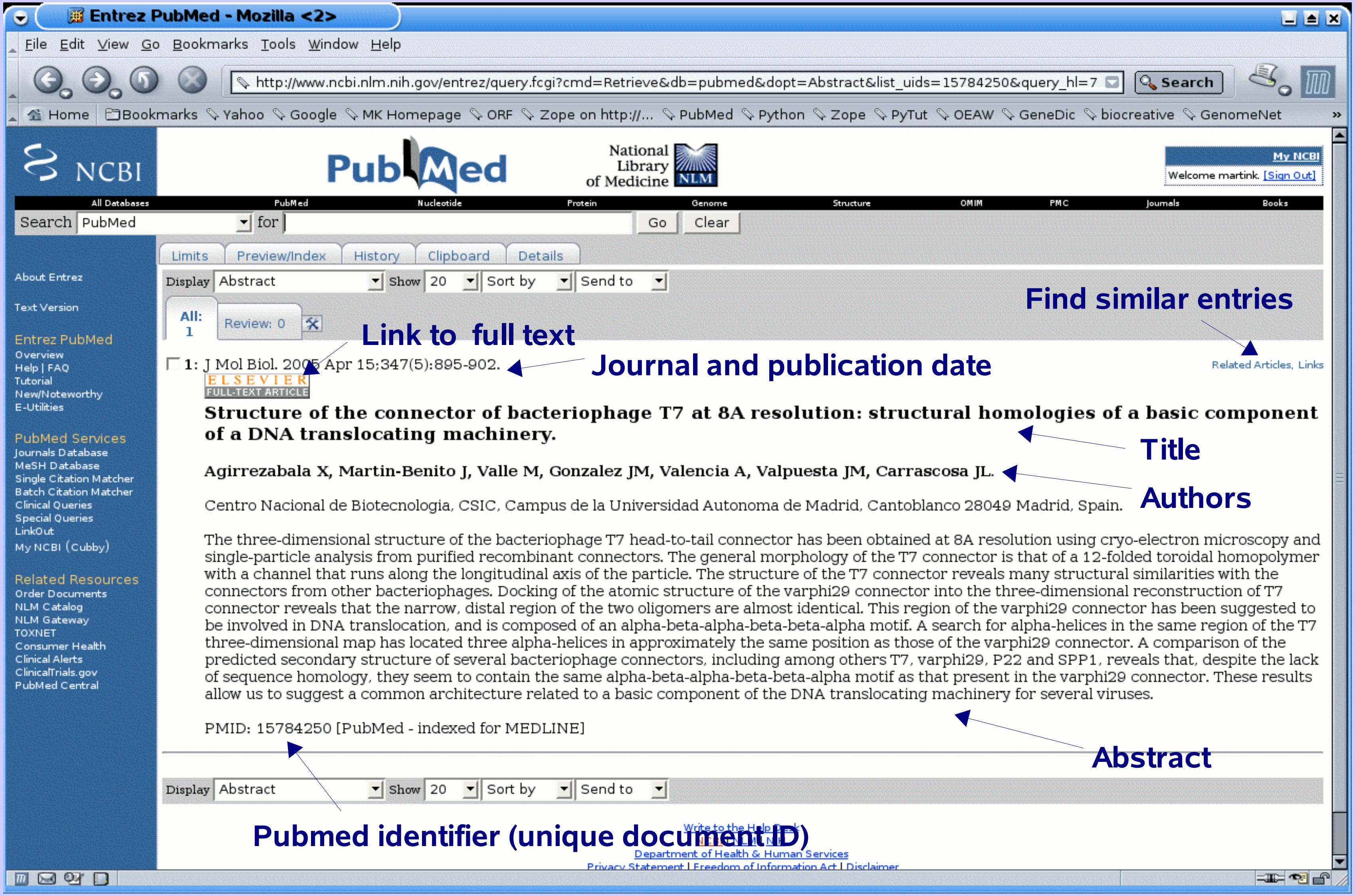
- PubMed titles and abstracts are primary data source for Bio-NLP.

- ~ 450,000 new abstracts/a

- > 4,800 biomedical journals

- ENTREZ search engine

PubMed online


Entrez PubMed - Mozilla <2>
File Edit View Go Bookmarks Tools Window Help
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15784250&query_hl=7
Search
Home Bookmarks Yahoo Google MK Homepage ORF Zope on http://... PubMed Python Zope PyTut OEAW GeneDic biocreative GenomeNet
NCBI
PubMed
National Library of Medicine NLM
My NCBI
Welcome martink [Sign Out]
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
Text Version
Entrez PubMed
Overview
Help | FAQ
Tutorial
New/Noteworthy
E-Utilities
PubMed Services
Journals Database
MeSH Database
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Special Queries
LinkOut
My NCBI (Cubby)
Related Resources
Order Documents
NLM Catalog
NLM Gateway
TOXNET
Consumer Health
Clinical Alerts
ClinicalTrials.gov
PubMed Central
Related Articles, Links
ELSEVIER FULL-TEXT ARTICLE
Display Abstract Show 20 Sort by Send to
Write to the Help Desk
NCBI | NLM | NIH
Department of Health & Human Services
Privacy Statement | Freedom of Information Act | Disclaimer

□ 1: J Mol Biol. 2005 Apr 15;347(5):895-902.

Structure of the connector of bacteriophage T7 at 8A resolution: structural homologies of a basic component of a DNA translocating machinery.

Agirrezabala X, Martin-Benito J, Valle M, Gonzalez JM, Valencia A, Valpuesta JM, Carrascosa JL.

Centro Nacional de Biotecnologia, CSIC, Campus de la Universidad Autonoma de Madrid, Cantoblanco 28049 Madrid, Spain.

The three-dimensional structure of the bacteriophage T7 head-to-tail connector has been obtained at 8A resolution using cryo-electron microscopy and single-particle analysis from purified recombinant connectors. The general morphology of the T7 connector is that of a 12-folded toroidal homopolymer with a channel that runs along the longitudinal axis of the particle. The structure of the T7 connector reveals many structural similarities with the connectors from other bacteriophages. Docking of the atomic structure of the varphi29 connector into the three-dimensional reconstruction of T7 connector reveals that the narrow, distal region of the two oligomers are almost identical. This region of the varphi29 connector has been suggested to be involved in DNA translocation, and is composed of an alpha-beta-alpha-beta-beta-alpha motif. A search for alpha-helices in the same region of the T7 three-dimensional map has located three alpha-helices in approximately the same position as those of the varphi29 connector. A comparison of the predicted secondary structure of several bacteriophage connectors, including among others T7, varphi29, P22 and SPP1, reveals that, despite the lack of sequence homology, they seem to contain the same alpha-beta-alpha-beta-beta-alpha motif as that present in the varphi29 connector. These results allow us to suggest a common architecture related to a basic component of the DNA translocating machinery for several viruses.

PMID: 15784250 [PubMed - indexed for MEDLINE]

Find similar entries

Link to full text

Journal and publication date

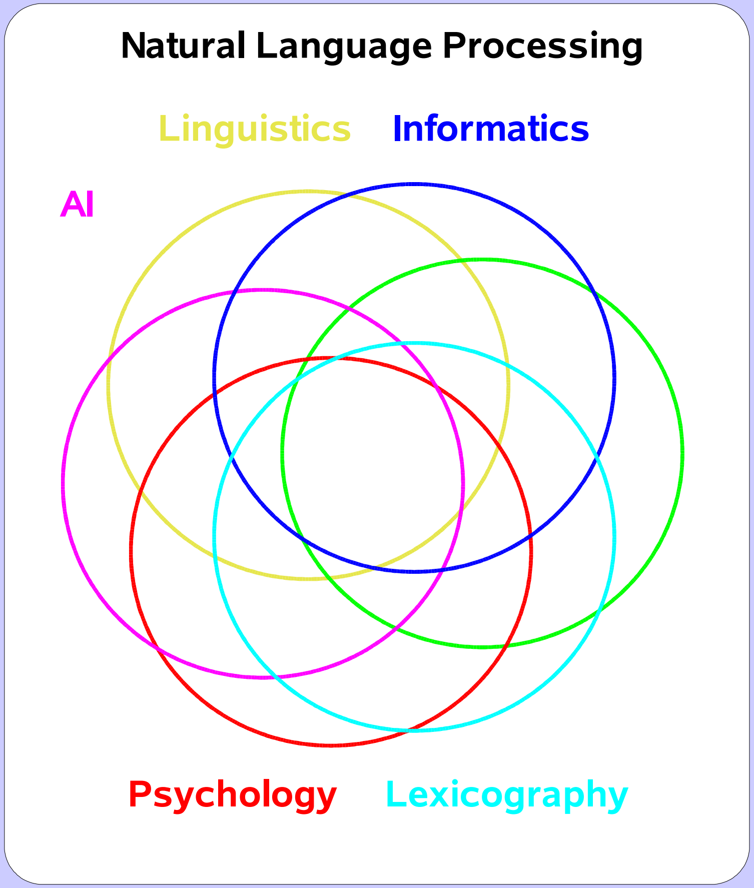
Title

Authors

Abstract

Pubmed identifier (unique document ID)

Natural Language Processing (NLP) basics



Natural Language Processing

Linguistics Informatics

AI

Psychology Lexicography

Domain, e.g. Biomedicine/ Molecular Biology

- **Techniques that analyse, understand and generate language** (free text, speech).
- Linguistic tools, e.g. syntactic analyser and semantic classification.
- Multidisciplinary field.
- Strongly language dependent.
- Create computational models of language.
- Learn statistical properties of language.
- Methods: statistical analysis, machine learning, rule-based, pattern-matching, AI, etc...
- Domain dependent (biomedical) vs generic NLP.

Major NLP tasks

- Information Retrieval (IR).
- Information extraction/Text mining (IE).
- Question Answering (QA).
- Natural Language Generation (NLG).

- Automatic summarisation.
- Machine translation.
- Text proofing.
- Speech recognition.
- Optical character recognition (OCR).

Information Retrieval (IR)

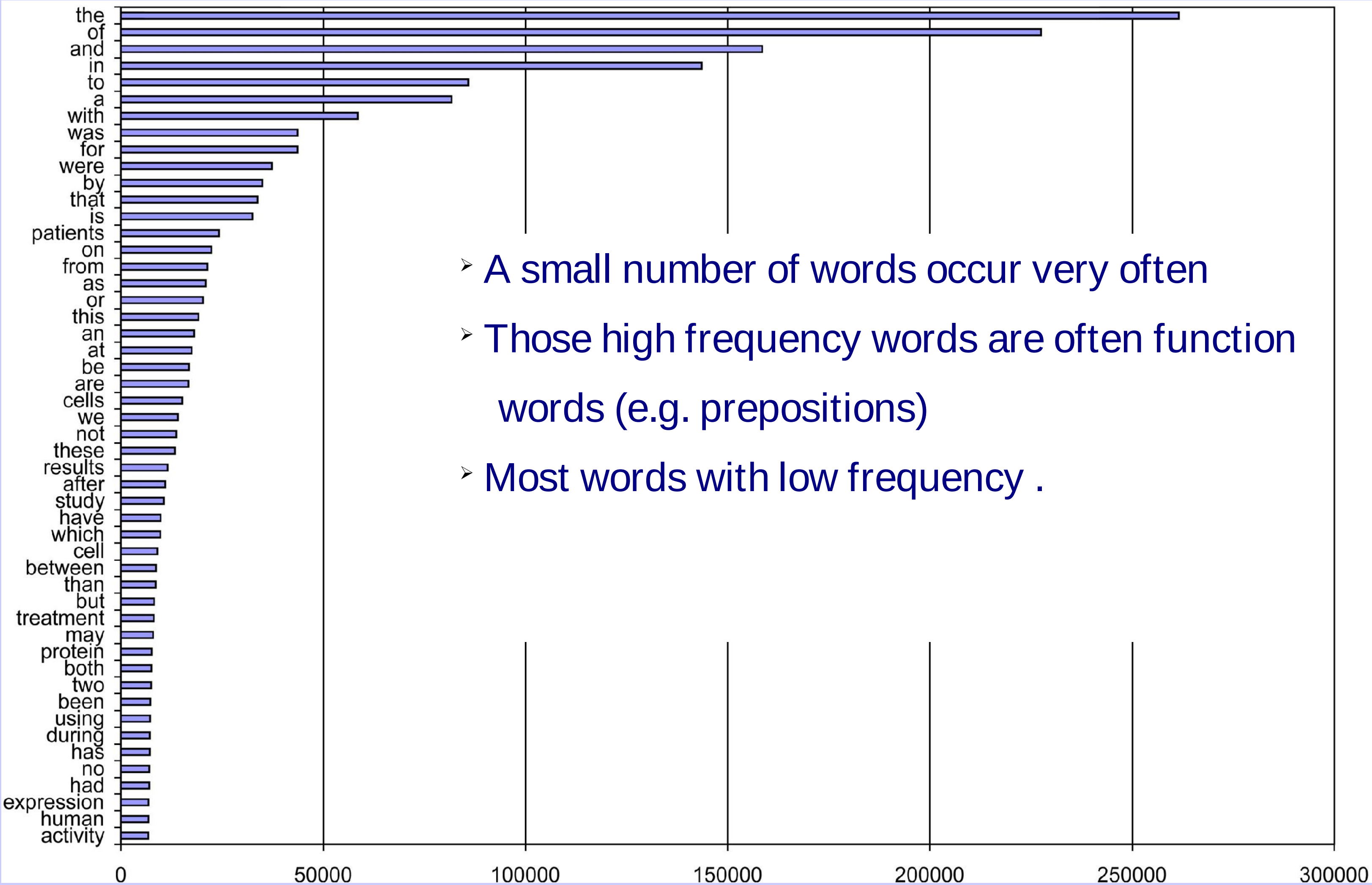
- IR: process of **recovery of those documents** from a collection of documents which satisfy a given information demand.
- Information demand often posed in form of a **search query**.
- Example: retrieval of web-pages using search engines, e.g. Google.
- First step: indexing document collection:
    - 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.

Boolean query

- Based on **combination of terms** using Boolean operators.
- Basic **Boolean operators**: AND, OR and NOT.
- Queries matched against the terms in the inverted index file.
- Entrez – Boolean search in PubMed.
- Fast, easy to implement.
- **Search engines**: often stop word removal and case folding.
- Stop word removal : space saving speed improvement.
- Return a **unranked list**.
- Return large list of documents, many not relevant.
- Terms have different information content ->

  better weighted query.

Zipf's law



- A small number of words occur very often
- Those high frequency words are often function words (e.g. prepositions)
- Most words with low frequency .

From: Rebholz-Schuhmann D, Kirsch H, Couto F (2005) Facts from Text—Is Text Mining Ready to Deliver? PLoS Biol 3(2): e65

Commonly excluded stop words

after	also	an	and
as	at	be	because
before	between	but	before
for	however	from	if
in	into	of	or
other	out	since	such
than	that	the	these
there	this	those	to
under	upon	when	where
whether	which	with	within
without	.	.	.

Vector space model

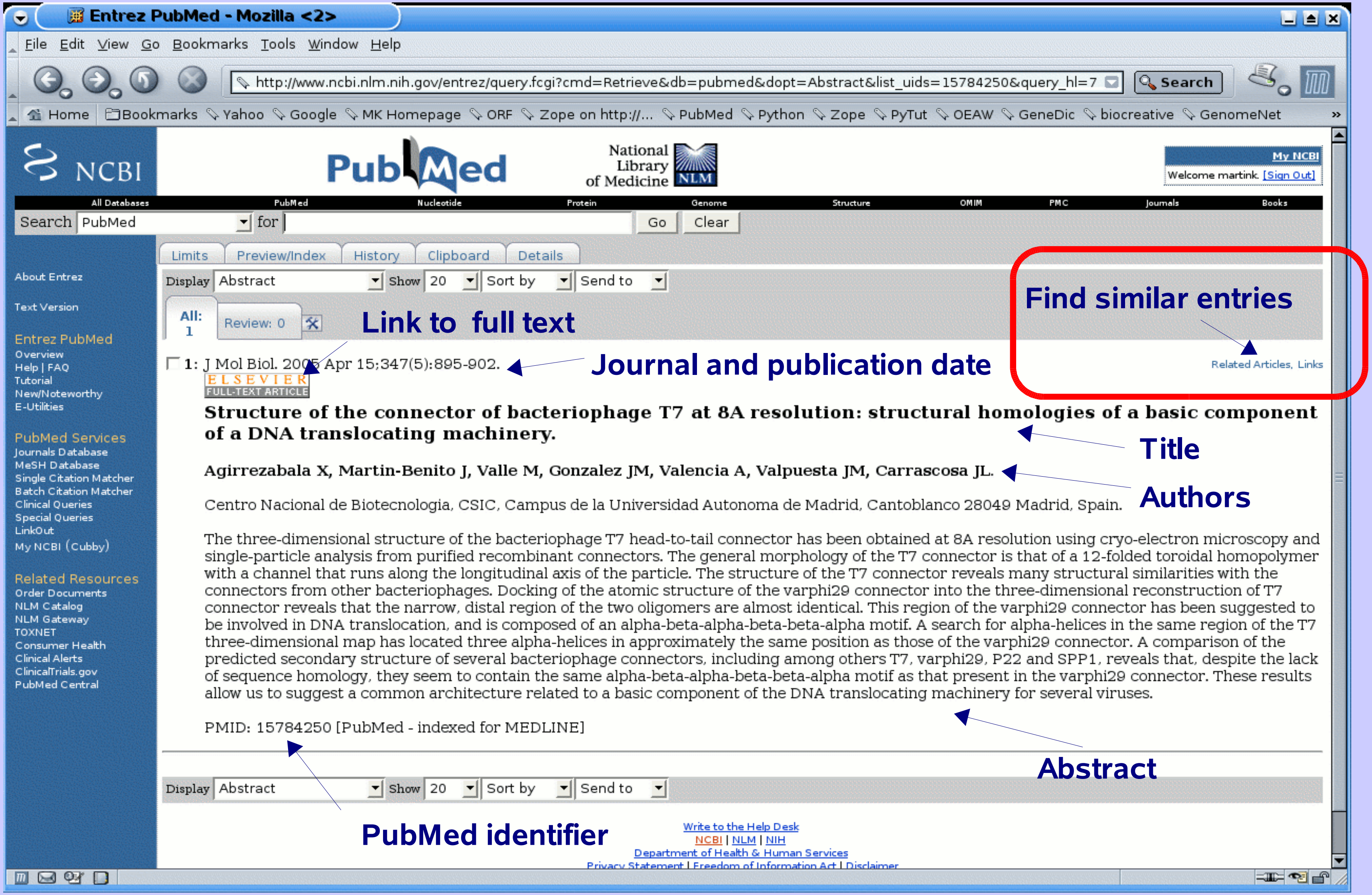
- Measure **similarity** between query and documents.
- Query can be a list of terms or whole documents.
- Documents and queries as **vectors of terms**.
- **Term weighting** according to their frequency:
  - within the document
  - within the document collection
- Widespread term weighting: tf x idf.
- Calculate similarity between those vectors.
- Cosine similarity.
- Return a ranked list.
- Example: related article search in PubMed

$$w_{i,j} = tf_{i,j} \times idf_j$$

$$idf_{i,j} = \log\left(\frac{N}{df_j}\right)$$

$$sim(Q,D) = \frac{\sum_{j=1}^{V} w_{Q,j} \times w_{i,j}}{\sqrt{\sum_{j=1}^{V} w_{Q,j} \times \sum_{j=1}^{V} w_{i,j}^2}}$$

PubMed online



http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15784250&query_hl=7

Welcome martink. [Sign Out]

Search PubMed for

Display Abstract Show 20 Sort by Send to

□ 1: J Mol Biol. 2005 Apr 15;347(5):895-902.

Structure of the connector of bacteriophage T7 at 8A resolution: structural homologies of a basic component of a DNA translocating machinery.

Agirrezabala X, Martin-Benito J, Valle M, Gonzalez JM, Valencia A, Valpuesta JM, Carrascosa JL.

Centro Nacional de Biotecnologia, CSIC, Campus de la Universidad Autonoma de Madrid, Cantoblanco 28049 Madrid, Spain.

The three-dimensional structure of the bacteriophage T7 head-to-tail connector has been obtained at 8A resolution using cryo-electron microscopy and single-particle analysis from purified recombinant connectors. The general morphology of the T7 connector is that of a 12-folded toroidal homopolymer with a channel that runs along the longitudinal axis of the particle. The structure of the T7 connector reveals many structural similarities with the connectors from other bacteriophages. Docking of the atomic structure of the varphi29 connector into the three-dimensional reconstruction of T7 connector reveals that the narrow, distal region of the two oligomers are almost identical. This region of the varphi29 connector has been suggested to be involved in DNA translocation, and is composed of an alpha-beta-alpha-beta-beta-alpha motif. A search for alpha-helices in the same region of the T7 three-dimensional map has located three alpha-helices in approximately the same position as those of the varphi29 connector. A comparison of the predicted secondary structure of several bacteriophage connectors, including among others T7, varphi29, P22 and SPP1, reveals that, despite the lack of sequence homology, they seem to contain the same alpha-beta-alpha-beta-beta-alpha motif as that present in the varphi29 connector. These results allow us to suggest a common architecture related to a basic component of the DNA translocating machinery for several viruses.

PMID: 15784250 [PubMed - indexed for MEDLINE]

Display Abstract Show 20 Sort by Send to

Find similar entries

Link to full text

Journal and publication date

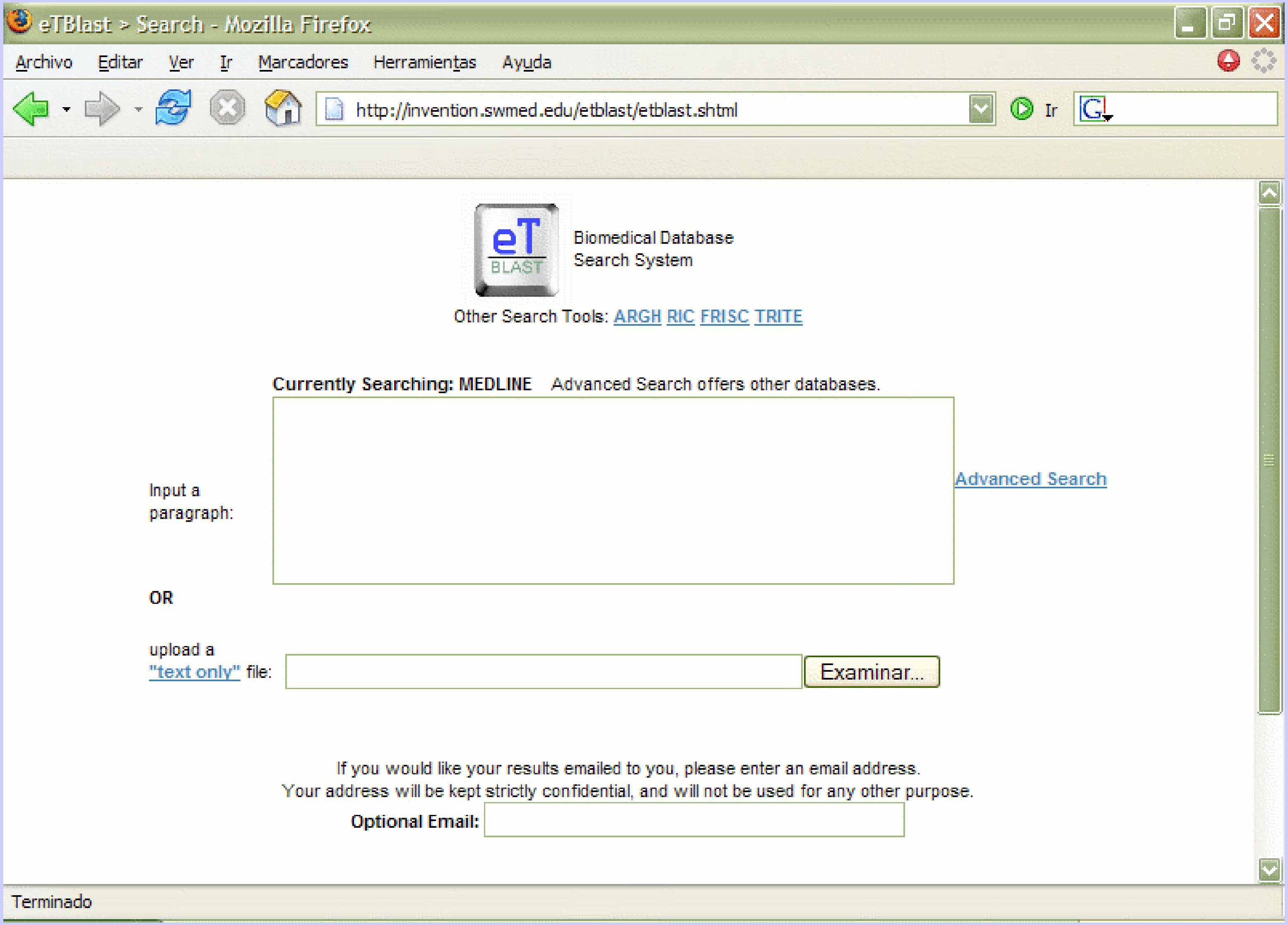
Title

Authors

Abstract

PubMed identifier

eTBlast (1)



eTBlast > Search - Mozilla Firefox
Archivo Editar Ver Ir Marcadores Herramientas Ayuda
Ir
eT BLAST
Examinar...
Terminado

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

Biomedical Database Search System

Other Search Tools: ARGH RIC FRISC TRITE

Currently Searching: MEDLINE Advanced Search offers other databases.

Input a paragraph:

Advanced Search

OR

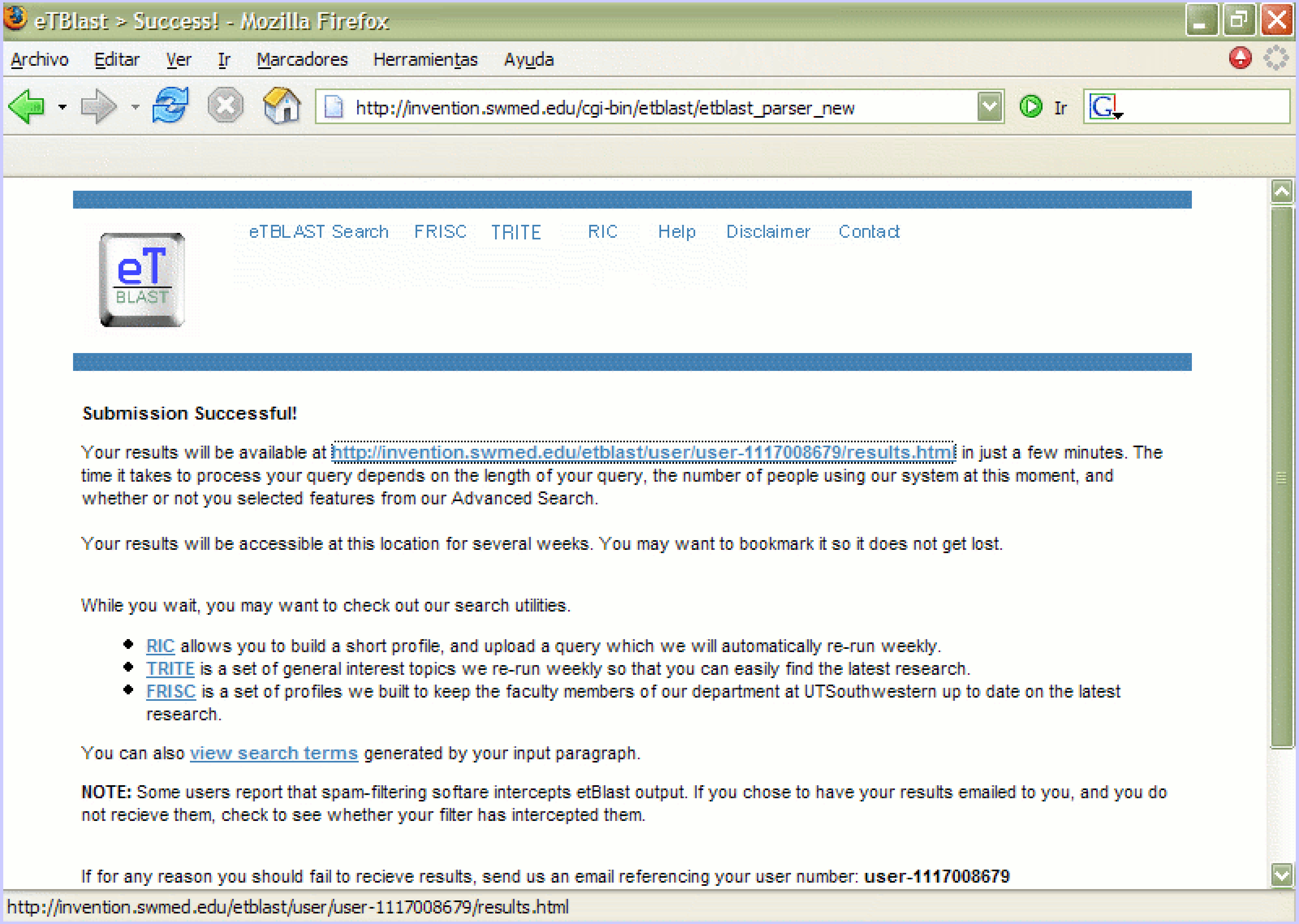
upload a "text only" file:

If you would like your results emailed to you, please enter an email address.
Your address will be kept strictly confidential, and will not be used for any other purpose.

Optional Email:

Science, May 14, 2004 issue. Under NetWatch, see the topic, "TOOLS: Just the Right Words".

eTBlast (2)


eTBlast > Success! - Mozilla Firefox
Archivo Editar Ver Ir Marcadores Herramientas Ayuda
http://invention.swmed.edu/cgi-bin/etblast/etblast_parser_new
Ir
eTBLAST Search FRISC TRITE RIC Help Disclaimer Contact
eT BLAST

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.

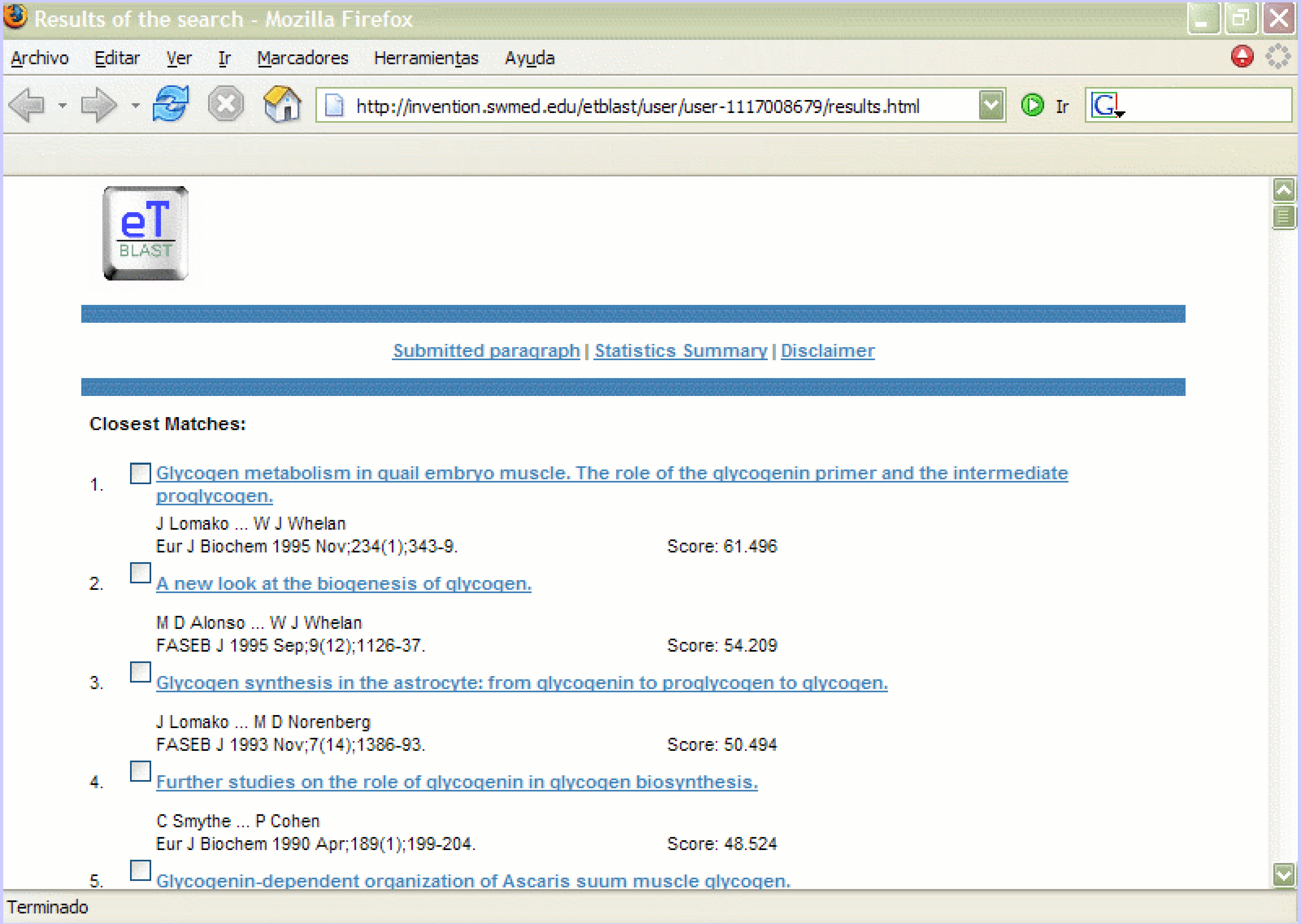
You can also view search terms generated by your input paragraph.

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 (3)


Results of the search - Mozilla Firefox
Archivo Editar Ver Ir Marcadores Herramientas Ayuda
http://invention.swmed.edu/etblast/user/user-1117008679/results.html
Ir
eT BLAST

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

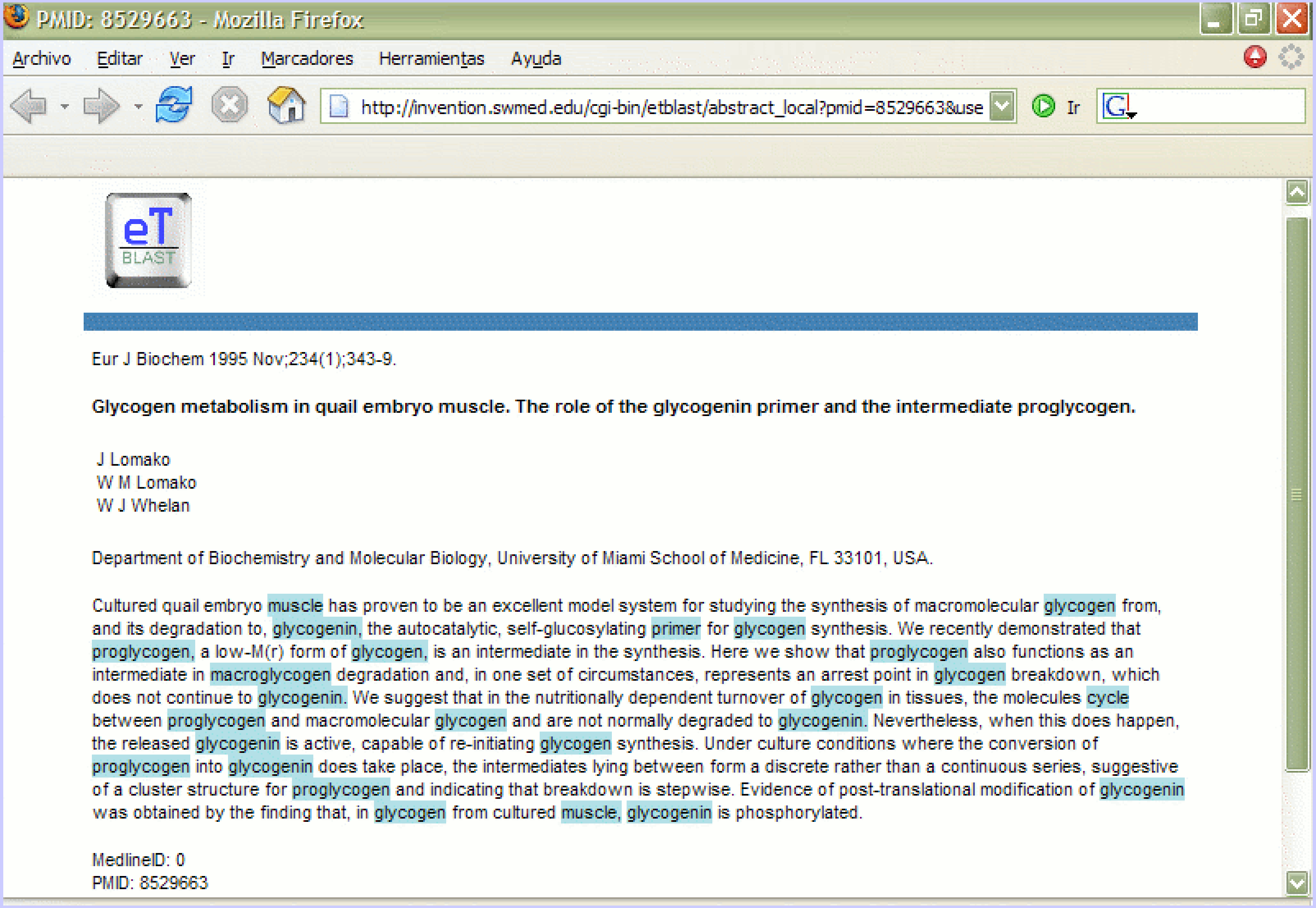
2. □ A new look at the biogenesis of glycogen.
M D Alonso ... W J Whelan
FASEB J 1995 Sep;9(12):1126-37. Score: 54.209

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.

eTBlast (4)


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

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 performance

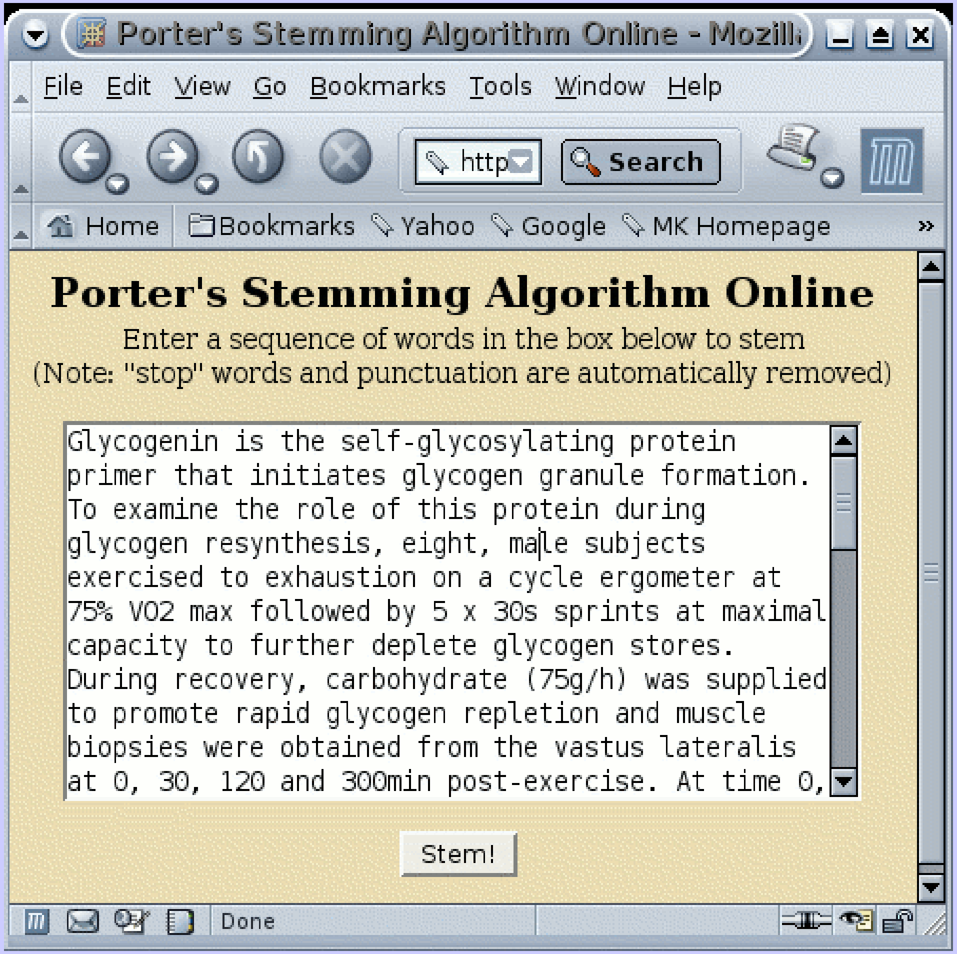
- **Precision**: fraction of relevant documents retrieved divided by the total returned documents
- **Recall**: proportion of relevant documents returned divided by the total number of relevant documents
- **F-score**: the harmonic mean of precision and recall
- Precision-recall curves

Information Extraction and 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 or proteins, drugs.
- Relations like: protein interactions
- Methods: heuristics, rule-based systems, machine learning and statistical techniques, regular expressions,..

Stemming

Process of removing affixes of words transforming them to their corresponding morphological base form or root.

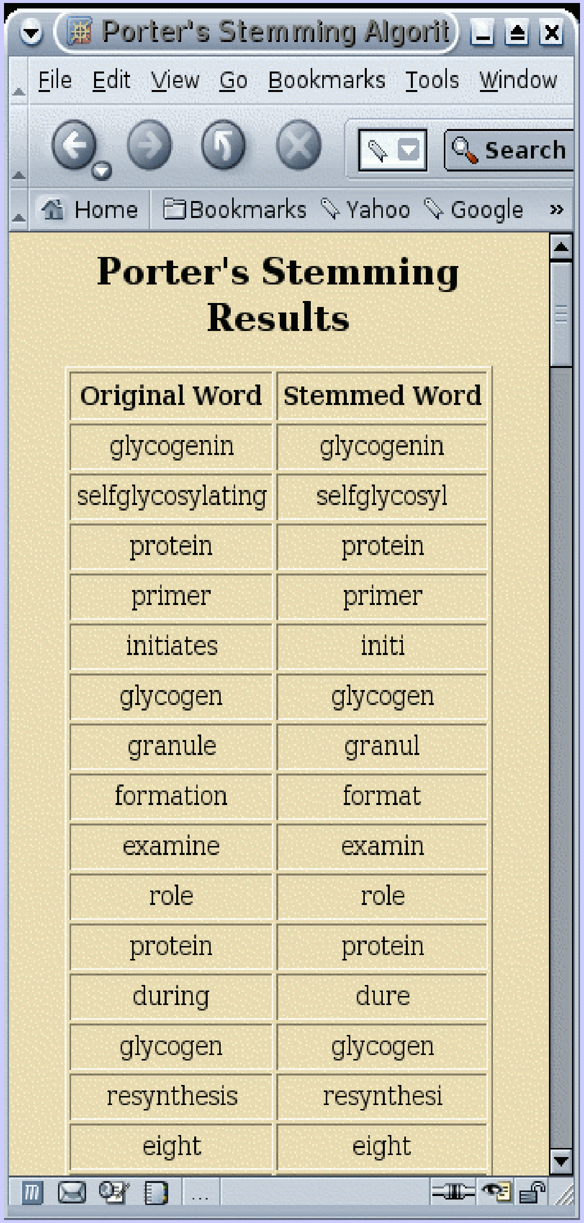

Porter's Stemming Algorithm Online - Mozill
File Edit View Go Bookmarks Tools Window Help
http
Search
Home
Bookmarks
Yahoo
Google
MK Homepage
Stem!
Done

Porter's Stemming Algorithm Online

Enter a sequence of words in the box below to stem
(Note: "stop" words and punctuation are automatically removed)

```
Glycogenin is the self-glycosylating protein
primer that initiates glycogen granule formation.
To examine the role of this protein during
glycogen resynthesis, eight, male subjects
exercised to exhaustion on a cycle ergometer at
75% VO2 max followed by 5 x 30s sprints at maximal
capacity to further deplete glycogen stores.
During recovery, carbohydrate (75g/h) was supplied
to promote rapid glycogen repletion and muscle
biopsies were obtained from the vastus lateralis
at 0, 30, 120 and 300min post-exercise. At time 0,
```

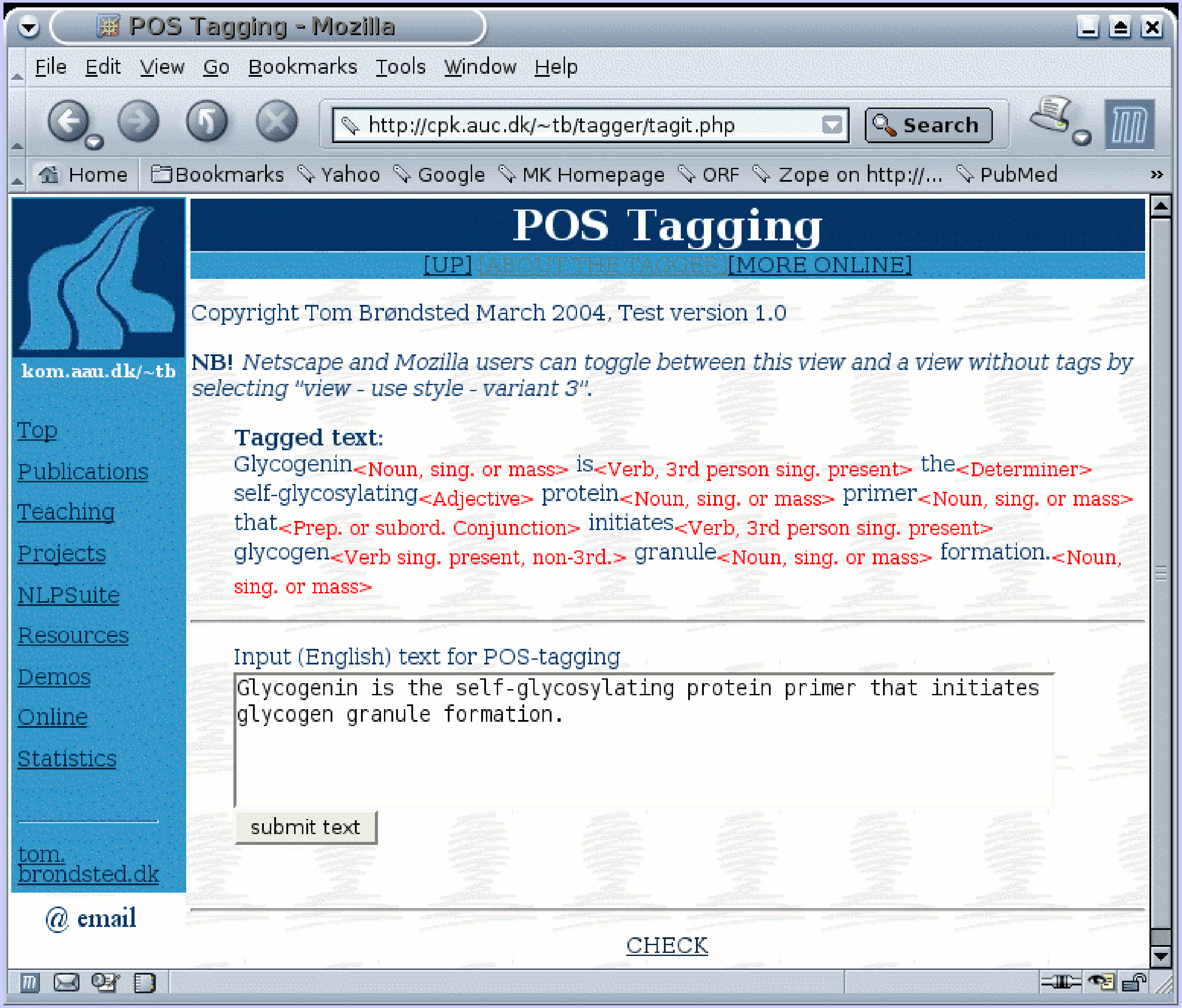
http://maya.cs.depaul.edu/~classes/ds575/porter.html


Porter's Stemming Algorit
File Edit View Go Bookmarks Tools Window
Search
Home
Bookmarks
Yahoo
Google

Porter's Stemming Results

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

POS–tagging



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>

Input (English) text for POS-tagging

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

submit text

CHECK

Providing each word given a sentence with its corresponding part of speech label , e.g. whether it is a noun, verb, preposition, article, etc.

Question Answering (QA)

- 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 formalised domain, new scientific terms
- Ad hoc retrieval task of the TREC Genomics Track 2005.
- Galitsky system (semantic skeletons (SSK), logical programming).

Natural Language Generation (NLG)

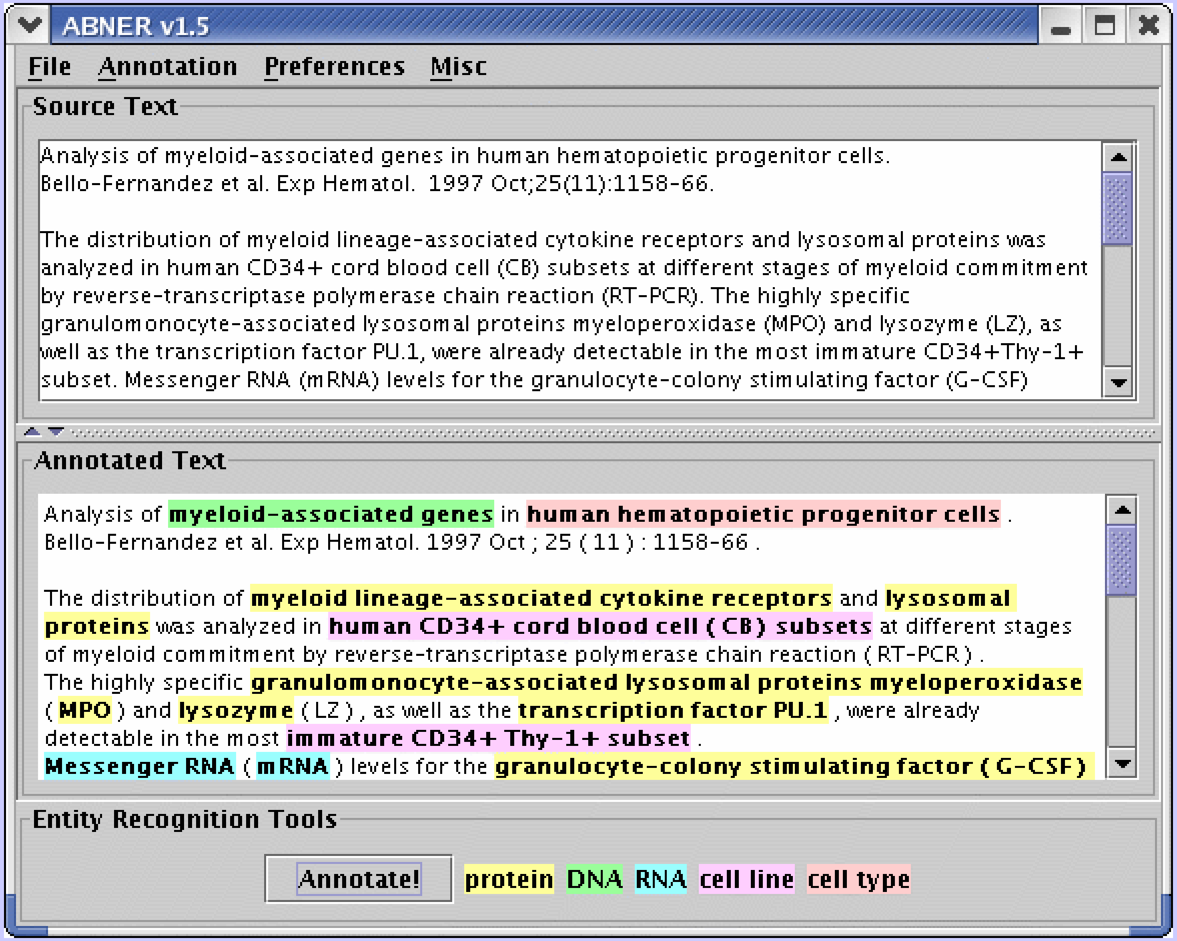
- NLG: constructing automatically natural language texts.
- Display the content of databases: reports, error messages.
- Based on semantic input, providing computer-internal representation of the information.
- Different degrees of complexity.
- Biology: modelling the domain language difficult.
- Simpathica/XSSYS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

Named entity recognition (NER)

- **Identification of entity types** in textual data.
- Semantic tagging.
- Example identification of company names and places
- Mainly identification of proper nouns.
- NER in Molecular Biology: identification of genes,

  proteins, chemical compounds, diseases,...
- Methods: ad-hoc rule based systems,
- ML techniques (HMM,SVM,...), statistical tools.
- Tools: GAPSCORE, ABNER, AbGene, NLProt

ABNER



Source Text

Analysis of myeloid-associated genes in human hematopoietic progenitor cells.
Bello-Fernandez et al. Exp Hematol. 1997 Oct;25(11):1158-66.

The distribution of myeloid lineage-associated cytokine receptors and lysosomal proteins was analyzed in human CD34+ cord blood cell (CB) subsets at different stages of myeloid commitment by reverse-transcriptase polymerase chain reaction (RT-PCR). The highly specific granulomonocyte-associated lysosomal proteins myeloperoxidase (MPO) and lysozyme (LZ), as well as the transcription factor PU.1, were already detectable in the most immature CD34+Thy-1+ subset. Messenger RNA (mRNA) levels for the granulocyte-colony stimulating factor (G-CSF)

Annotated Text

Analysis of **myeloid-associated genes** in **human hematopoietic progenitor cells** .
Bello-Fernandez et al. Exp Hematol. 1997 Oct ; 25 (11) : 1158-66 .

The distribution of **myeloid lineage-associated cytokine receptors** and **lysosomal proteins** was analyzed in **human CD34+ cord blood cell (CB) subsets** at different stages of myeloid commitment by reverse-transcriptase polymerase chain reaction (RT-PCR) .
The highly specific **granulomonocyte-associated lysosomal proteins myeloperoxidase** (**MPO**) and **lysozyme** (LZ) , as well as the **transcription factor PU.1** , were already detectable in the most **immature CD34+ Thy-1+ subset** .
Messenger RNA (**mRNA**) levels for the **granulocyte-colony stimulating factor (G-CSF)**

Entity Recognition Tools

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

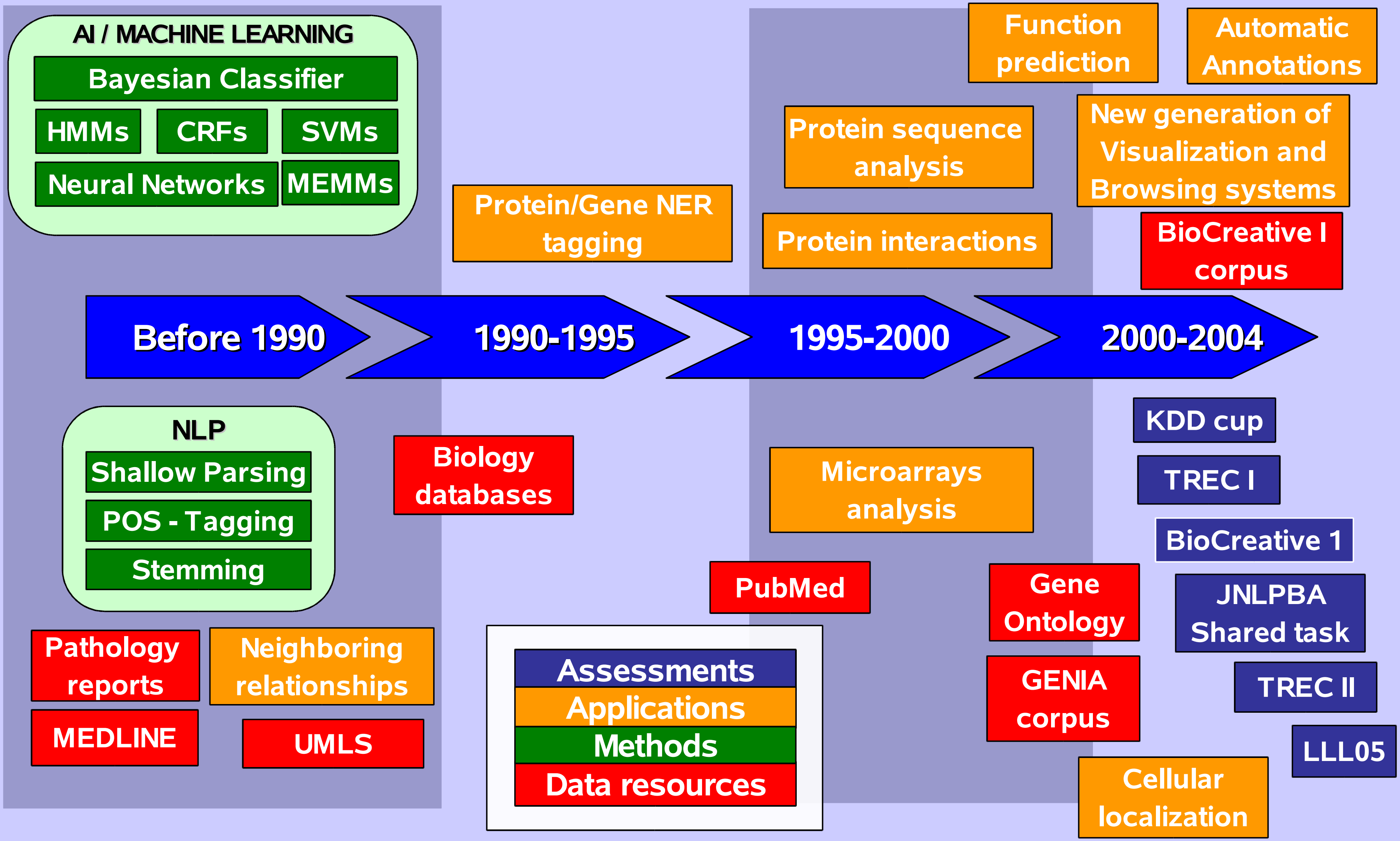
Basic NLP terms

Corpus: collection of documents.

POS tagging: labeling each word in a sentence with its part of speech (verb,noun,..) depending on its context.

Word sense disambiguation: assigning the semantic class (meaning) to a given word depending on its context.

NLP in Biomedicine – Timeline



Before 1990

1990-1995

1995-2000

2000-2004

AI / MACHINE LEARNING

- Bayesian Classifier
- HMMs
- CRFs
- SVMs
- Neural Networks
- MEMMs

NLP

- Shallow Parsing
- POS - Tagging
- Stemming

Pathology reports

Neighboring relationships

MEDLINE

UMLS

Protein/Gene NER tagging

Biology databases

Protein sequence analysis

Protein interactions

Microarrays analysis

PubMed

Function prediction

Automatic Annotations

New generation of Visualization and Browsing systems

BioCreative I corpus

KDD cup

TREC I

BioCreative 1

JNLPBA Shared task

TREC II

LLL05

Gene Ontology

GENIA corpus

Cellular localization

Assessments

Applications

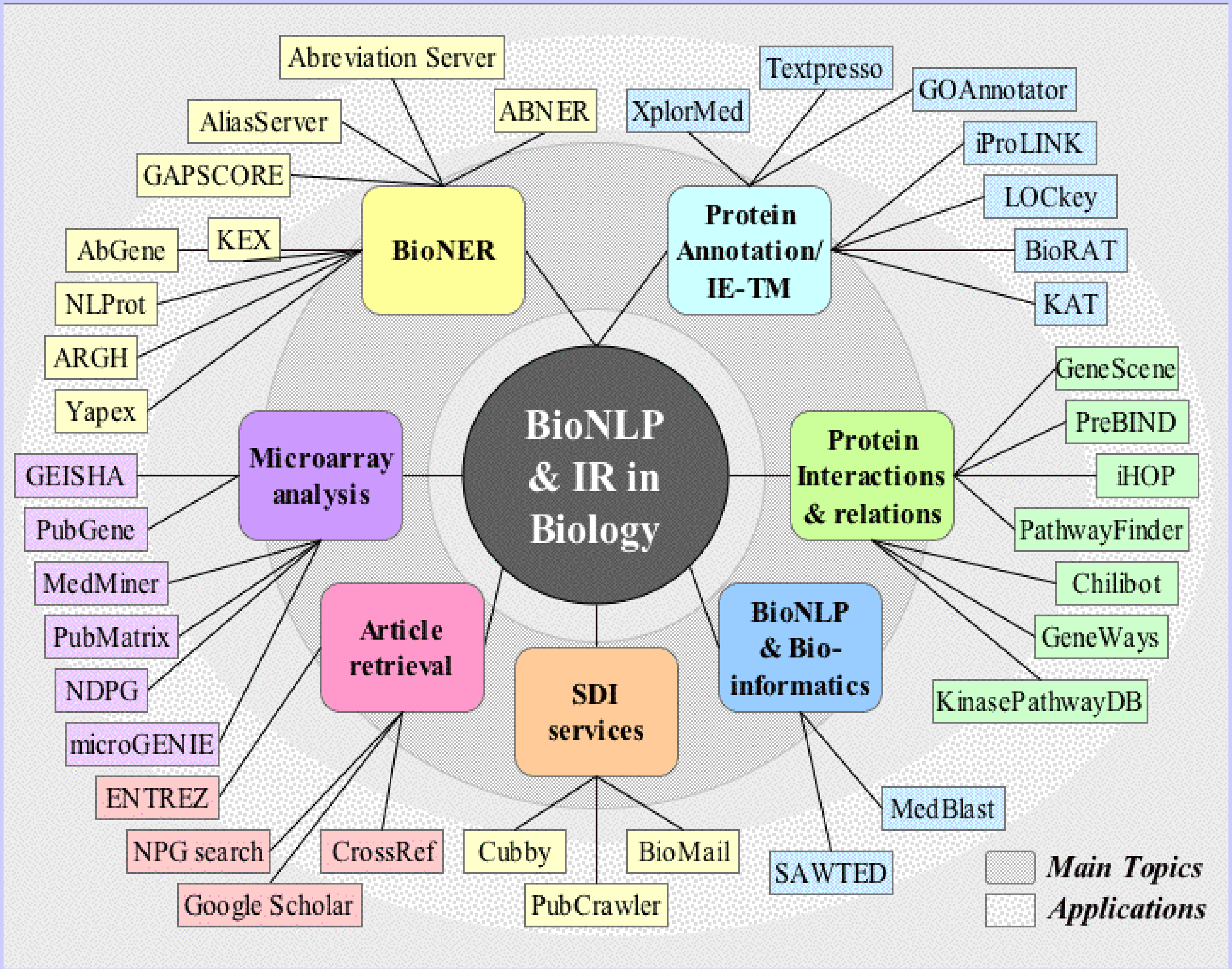
Methods

Data resources

Text mining applications in biology

- NER: tagging biological entities.
- Automatic annotation: associating proteins to functional descriptions.
- Protein interactions: extracting interactions of proteins, genes and drugs.
- Microarray analysis: providing biological context through literature mining
- Protein localisation
- Improving sequence-based homology detection.

Text mining applications in biology



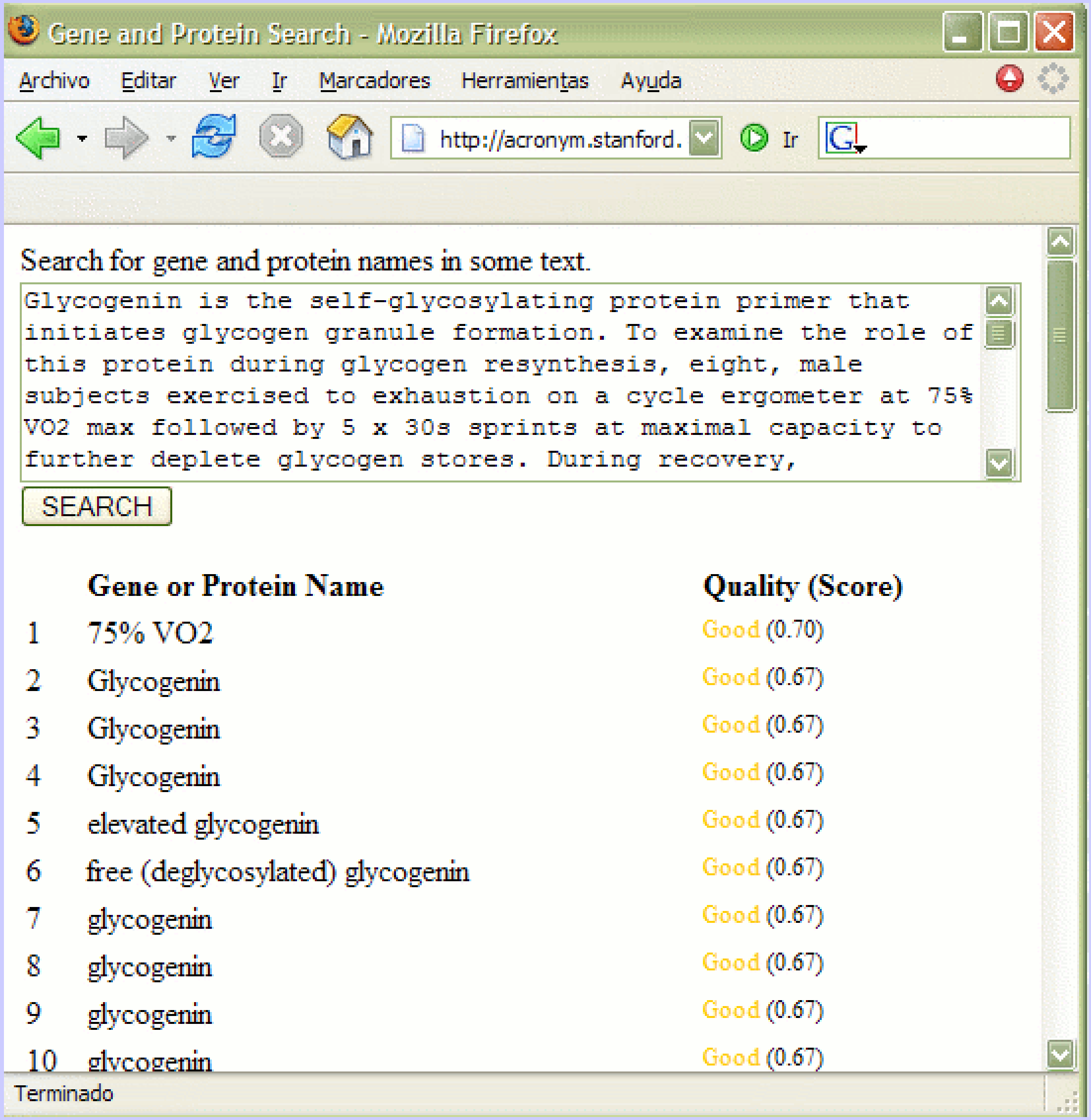
Text mining biomedical literature (2005)

Tagging Biological entities

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).
- 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).

GAPSCORE (1)



Search for gene and protein names in some text.

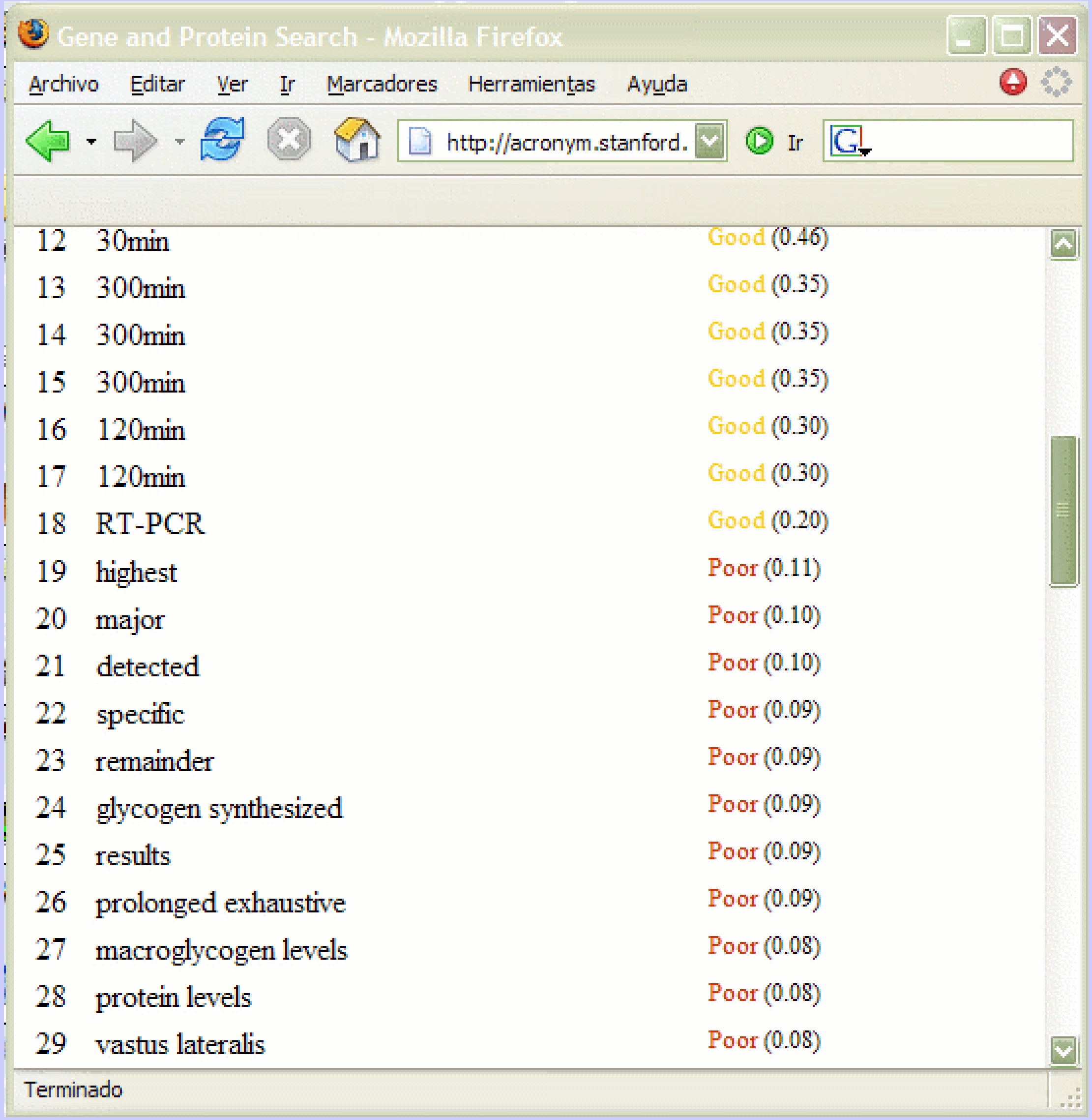
```
Glycogenin is the self-glycosylating protein primer that
initiates glycogen granule formation. To examine the role of
this protein during glycogen resynthesis, eight, male
subjects exercised to exhaustion on a cycle ergometer at 75%
VO2 max followed by 5 x 30s sprints at maximal capacity to
further deplete glycogen stores. During recovery,
```

	Gene or Protein Name	Quality (Score)
1	75% VO2	Good (0.70)
2	Glycogenin	Good (0.67)
3	Glycogenin	Good (0.67)
4	Glycogenin	Good (0.67)
5	elevated glycogenin	Good (0.67)
6	free (deglycosylated) glycogenin	Good (0.67)
7	glycogenin	Good (0.67)
8	glycogenin	Good (0.67)
9	glycogenin	Good (0.67)
10	glycogenin	Good (0.67)

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

http://bionlp.stanford.edu/gapscore/

GAPSCORE (2)

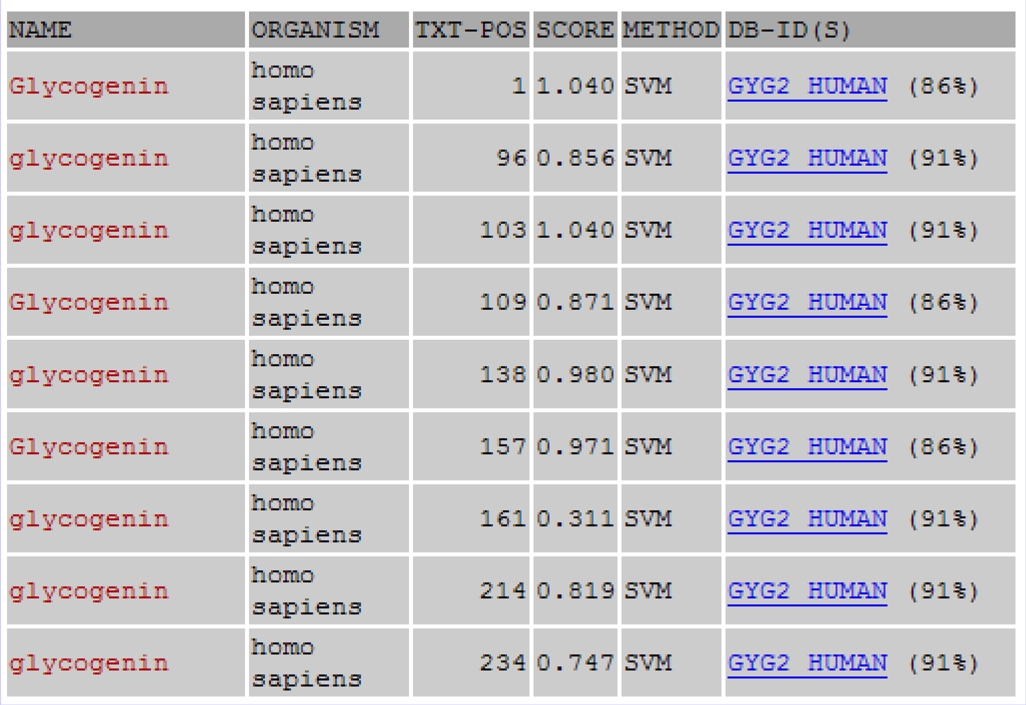

Gene and Protein Search - Mozilla Firefox
Archivo Editar Ver Ir Marcadores Herramientas Ayuda
http://acronym.stanford.
Ir
Terminado

12	30min	Good (0.46)
13	300min	Good (0.35)
14	300min	Good (0.35)
15	300min	Good (0.35)
16	120min	Good (0.30)
17	120min	Good (0.30)
18	RT-PCR	Good (0.20)
19	highest	Poor (0.11)
20	major	Poor (0.10)
21	detected	Poor (0.10)
22	specific	Poor (0.09)
23	remainder	Poor (0.09)
24	glycogen synthesized	Poor (0.09)
25	results	Poor (0.09)
26	prolonged exhaustive	Poor (0.09)
27	macroglycogen levels	Poor (0.08)
28	protein levels	Poor (0.08)
29	vastus lateralis	Poor (0.08)

- Choose cut-offs.
- Online.
- Based on Medline analysis
- Score new words using SVM
- Statistical analysis of PubMed words.

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.

NLProt



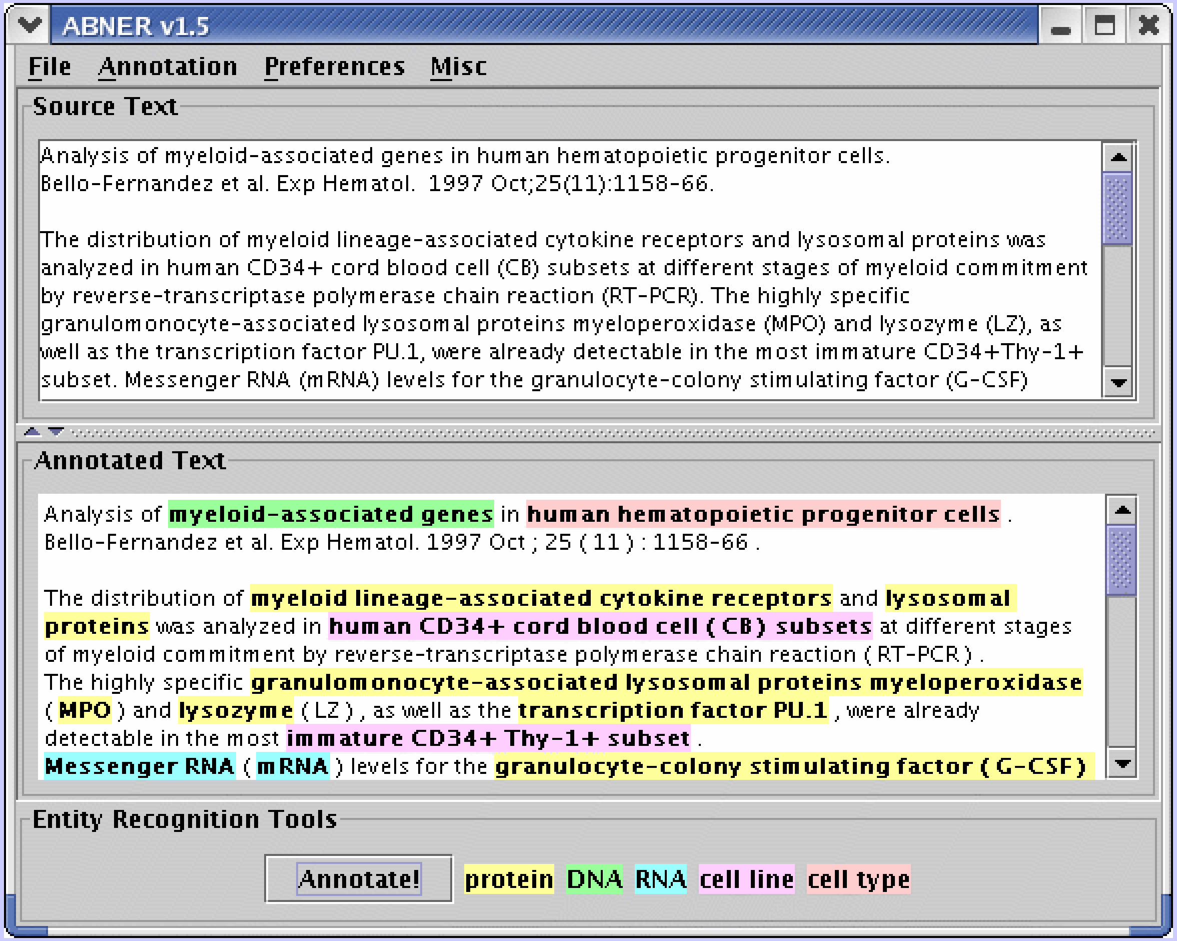
NAME	ORGANISM	TXT-POS	SCORE	METHOD	DB-ID(S)
Glycogenin	homo sapiens	1	1.040	SVM	GYG2_HUMAN (86%)
glycogenin	homo sapiens	96	0.856	SVM	GYG2_HUMAN (91%)
glycogenin	homo sapiens	103	1.040	SVM	GYG2_HUMAN (91%)
Glycogenin	homo sapiens	109	0.871	SVM	GYG2_HUMAN (86%)
glycogenin	homo sapiens	138	0.980	SVM	GYG2_HUMAN (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%

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

Chang JT, Schutze H, Altman RB. GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics*. 2004 Jan 22;20(2):216-25.

ABNER



Analysis of myeloid-associated genes in human hematopoietic progenitor cells.
Bello-Fernandez et al. Exp Hematol. 1997 Oct;25(11):1158-66.

The distribution of myeloid lineage-associated cytokine receptors and lysosomal proteins was analyzed in human CD34+ cord blood cell (CB) subsets at different stages of myeloid commitment by reverse-transcriptase polymerase chain reaction (RT-PCR). The highly specific granulomonocyte-associated lysosomal proteins myeloperoxidase (MPO) and lysozyme (LZ), as well as the transcription factor PU.1, were already detectable in the most immature CD34+Thy-1+ subset. Messenger RNA (mRNA) levels for the granulocyte-colony stimulating factor (G-CSF)

Analysis of **myeloid-associated genes** in **human hematopoietic progenitor cells** .
Bello-Fernandez et al. Exp Hematol. 1997 Oct ; 25 (11) : 1158-66 .

The distribution of **myeloid lineage-associated cytokine receptors** and **lysosomal proteins** was analyzed in **human CD34+ cord blood cell (CB) subsets** at different stages of myeloid commitment by reverse-transcriptase polymerase chain reaction (RT-PCR) .
The highly specific **granulomonocyte-associated lysosomal proteins myeloperoxidase** (**MPO**) and **lysozyme** (LZ) , as well as the **transcription factor PU.1** , were already detectable in the most **immature CD34+ Thy-1+ subset** .
Messenger RNA (**mRNA**) levels for the **granulocyte-colony stimulating factor (G-CSF)**

- A Biomedical Named Entity Recogniser
- Downloadable.
- CRF-based
- Trained on BioCreative and GENIA
- orthographic and contextual features
- Can be trained on new corpora

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

Extracting functional annotations

- **Manual annotation** extraction by database curators.
  - Scientific literature analysis.
  - Time-consuming & labour-intensive.
  - Example: Gene Ontology annotation (GOA).

- **Text mining** to assist annotation extraction:
  - Identification of annotation relevant sentences.
  - Identification of protein-term associations.

Function extraction – applications

- Andrade et. (1997)
- iHOP
- Textpresso system
- Gene Information System (GIS)
- Medical Knowledge Explorer (MeKE)
- GO engine,...

Andrade et al. (1997)

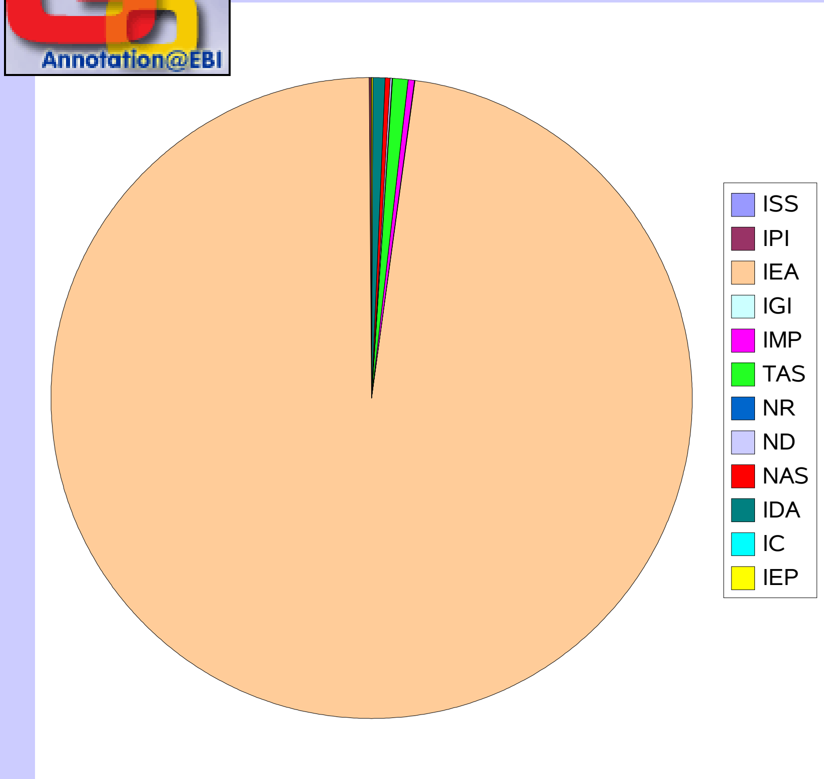
- Extracts sentences from PubMed which contain functional information.
- Statistical analysis of the word frequencies.
- Analysis in protein families.
- Background frequencies of those words.

GENE ONTOLOGY (GO)

- Ontology direct acyclic 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.

http://www.geneontology.org/

Gene Ontology Annotation

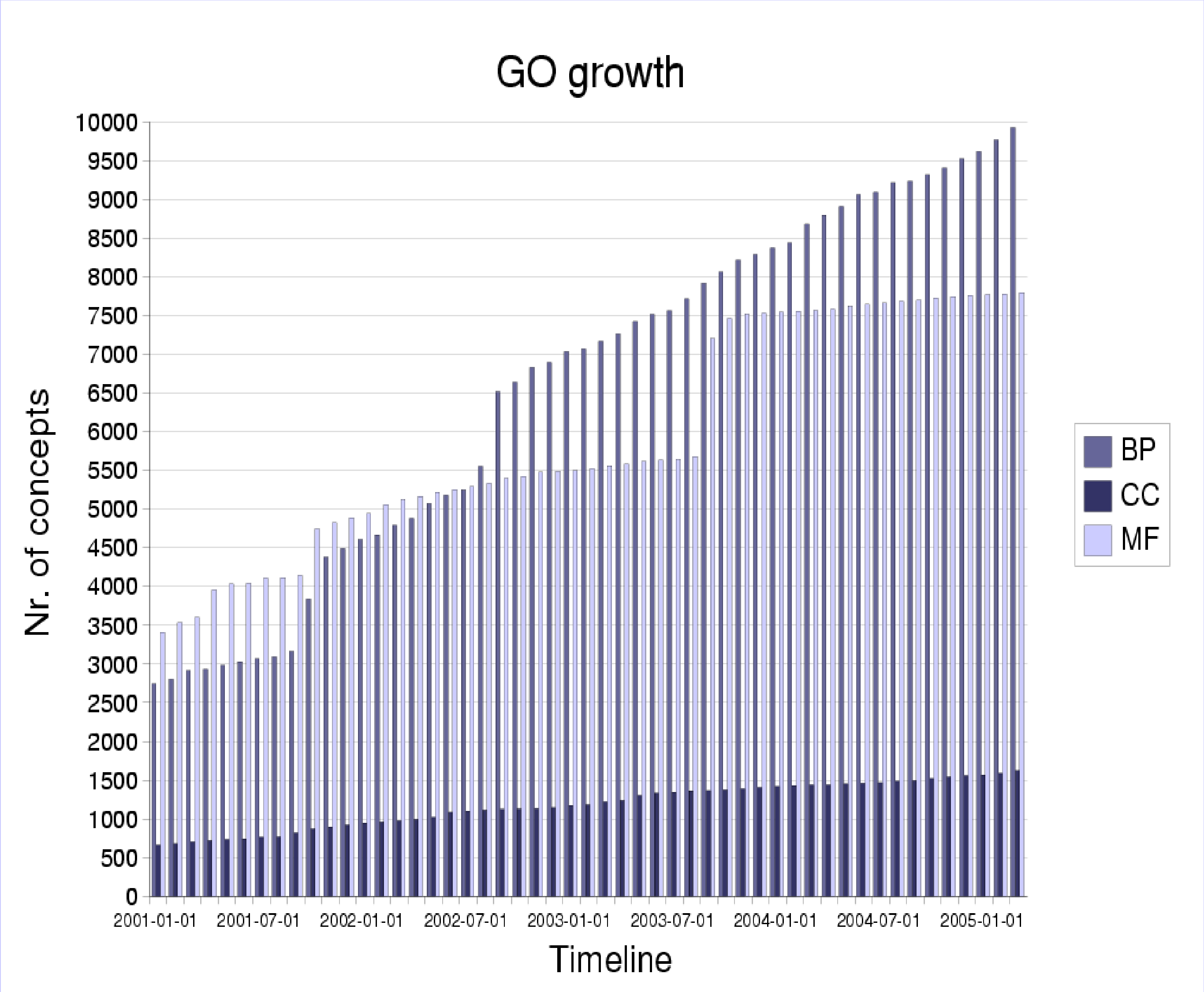


Ev.C.	Annot	Perc.	
IEA	**6421817**	**0.97529**	**Electronic/ sequence-based annotation**
ISS	19576	0.00297	
NR	2191	0.00033	
ND	4433	0.00067	
IPI	7130	0.00108	
IGI	3014	0.00046	**Experimental evidence**
IMP	19072	0.00290	
IDA	38862	0.00590	
IEP	1495	0.00023	
IC	831	0.00013	
TAS	49630	0.00754	**Curator knowledge**
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;

http://www.ebi.ac.uk/GOA/ 04/22/05

Gene Ontology Growth



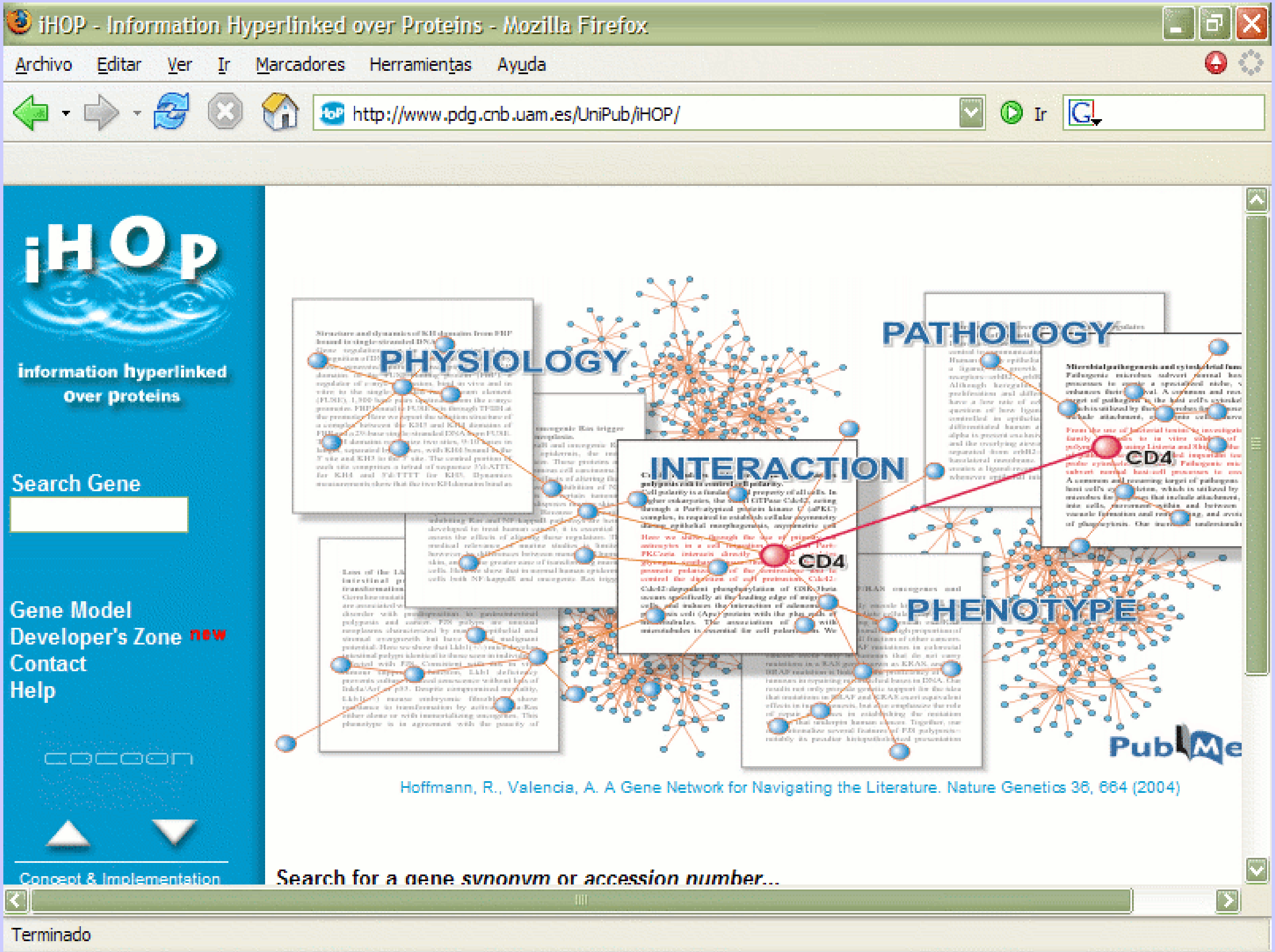
- MF:Molecular Function
- CC: Cellular Component
- BP: Biological Process

iHOP

- Protein centric: nucleates the literature around protein name.
- For a range of model organisms (e.g. Human, yeast,..)
- Hyperlinks proteins through co-occurrence
- Highlight direct associations between proteins and functional terms.
- Online, fast, easy to use.

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

iHOP



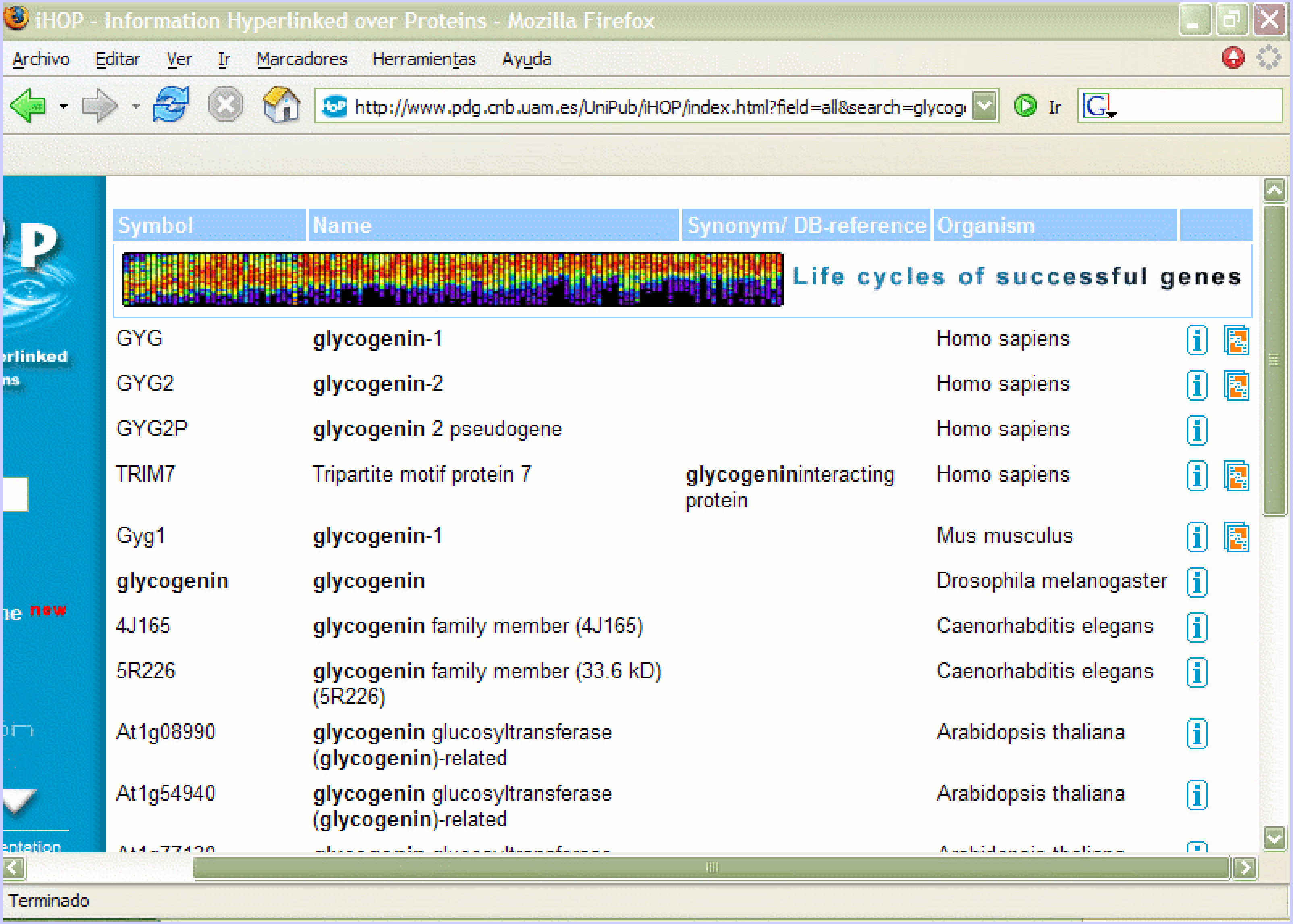
http://www.pdg.cnb.uam.es/UniPub/iHOP/

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

Search for a gene synonym or accession number...

Text mining biomedical literature (2005)

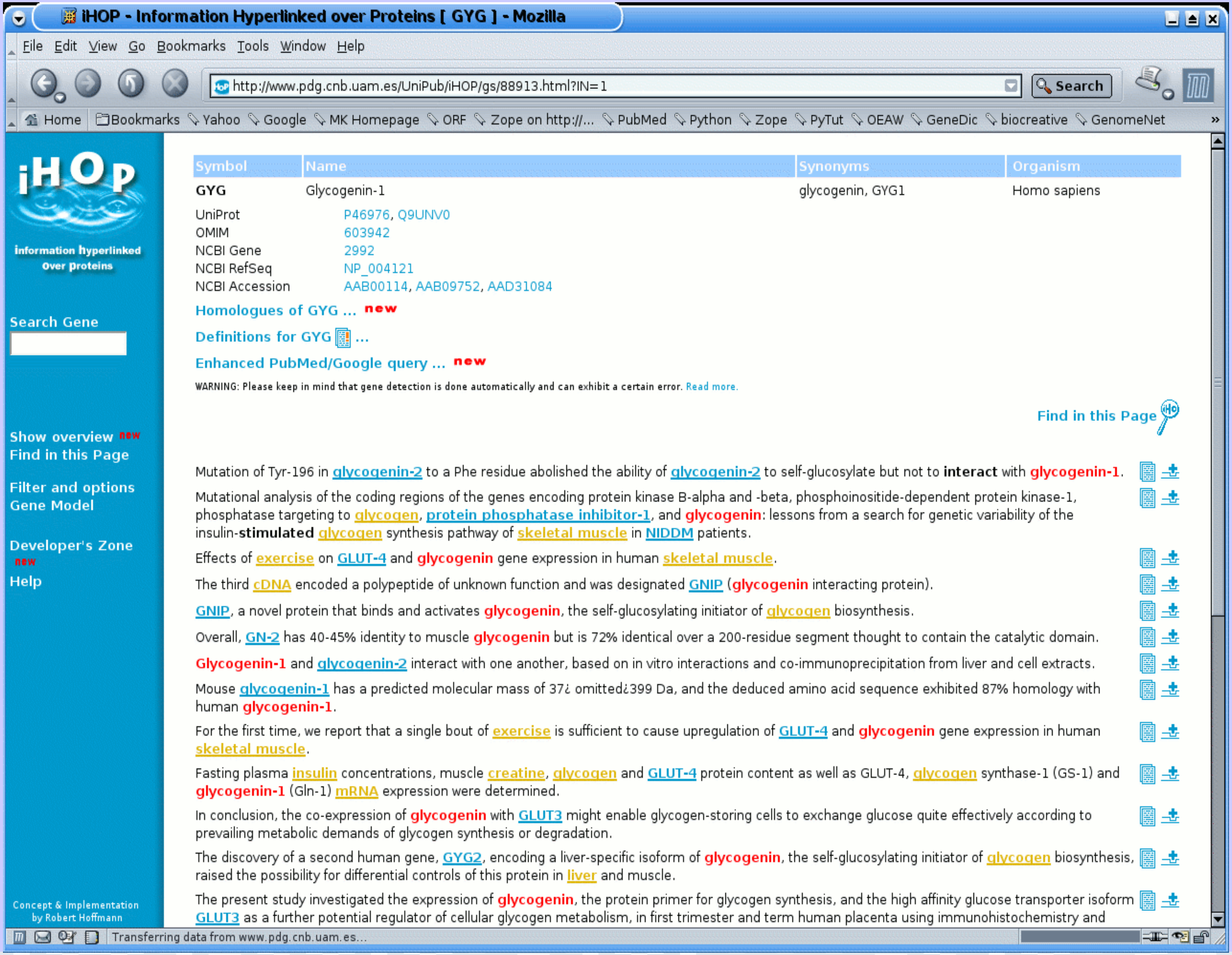
iHOP


iHOP - Information Hyperlinked over Proteins - Mozilla Firefox
Archivo
Editar
Ver
Ir
Marcadores
Herramientas
Ayuda
Ir
Life cycles of successful genes
rlinked
ns
ne new
entation
Terminado

http://www.pdg.cnb.uam.es/UniPub/iHOP/index.html?field=all&search=glycog

Symbol	Name	Synonym/ DB-reference	Organism	
GYG	**glycogenin**-1		Homo sapiens	
GYG2	**glycogenin**-2		Homo sapiens	
GYG2P	**glycogenin** 2 pseudogene		Homo sapiens	
TRIM7	Tripartite motif protein 7	**glycogenin**interacting protein	Homo sapiens	
Gyg1	**glycogenin**-1		Mus musculus	
glycogenin	**glycogenin**		Drosophila melanogaster	
4J165	**glycogenin** family member (4J165)		Caenorhabditis elegans	
5R226	**glycogenin** family member (33.6 kD) (5R226)		Caenorhabditis elegans	
At1g08990	**glycogenin** glucosyltransferase (**glycogenin**)-related		Arabidopsis thaliana	
At1g54940	**glycogenin** glucosyltransferase (**glycogenin**)-related		Arabidopsis thaliana	

Text mining biomedical literature (2005)


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 new
Find in this Page
Filter and options
Gene Model
Developer's Zone new
Help
Concept & Implementation by Robert Hoffmann
Transferring data from www.pdg.cnb.uam.es...

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.

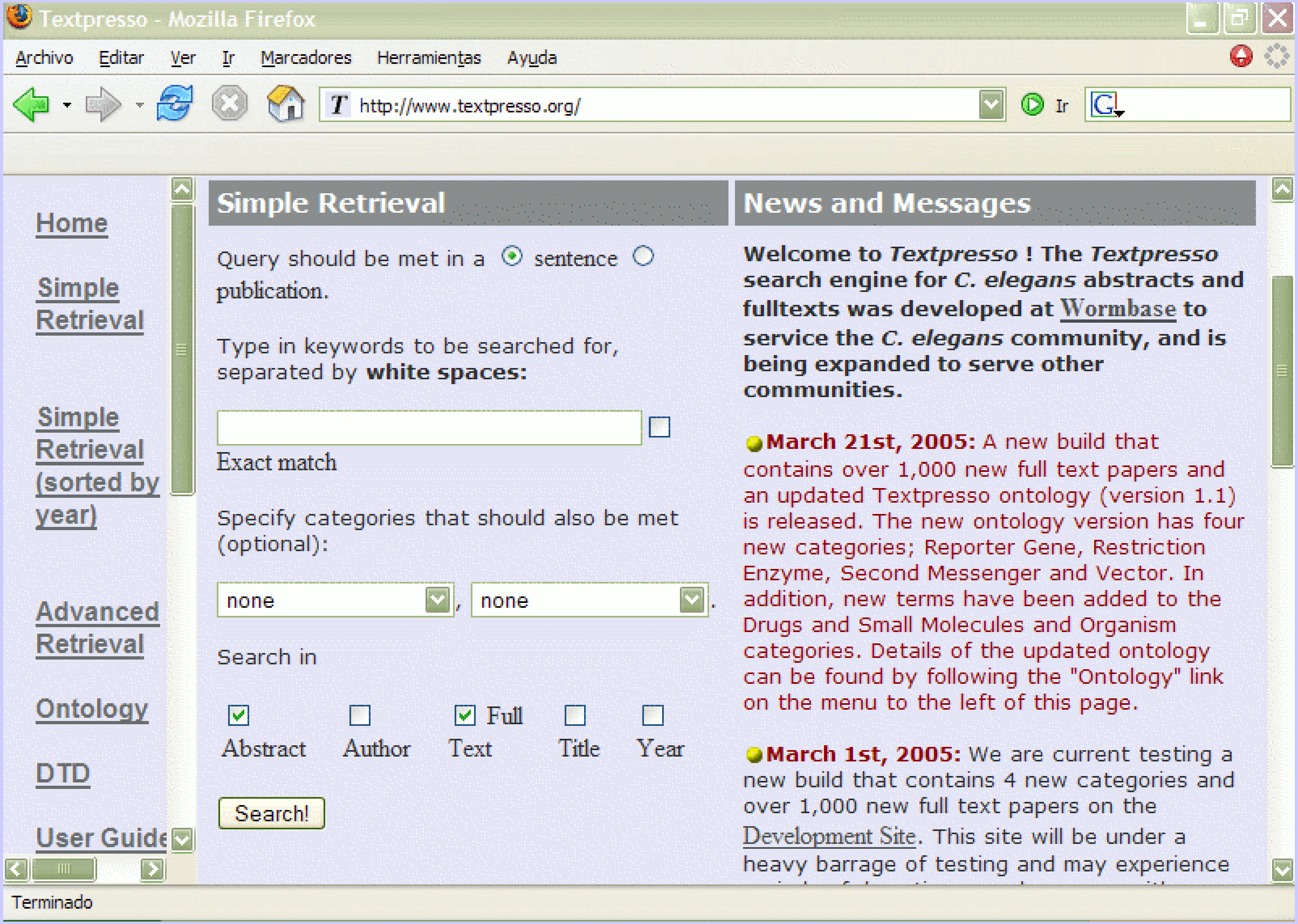
In conclusion, the co-expression of glycogenin with GLUT3 might enable glycogen-storing cells to exchange glucose quite effectively according to prevailing metabolic demands of glycogen synthesis or degradation.

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

Text mining biomedical literature (2005)

Textpresso


Textpresso - Mozilla Firefox
Archivo Editar Ver Ir Marcadores Herramientas Ayuda
http://www.textpresso.org/
Ir
Home
Simple Retrieval
Simple Retrieval (sorted by year)
Advanced Retrieval
Ontology
DTD
User Guide
Terminado

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

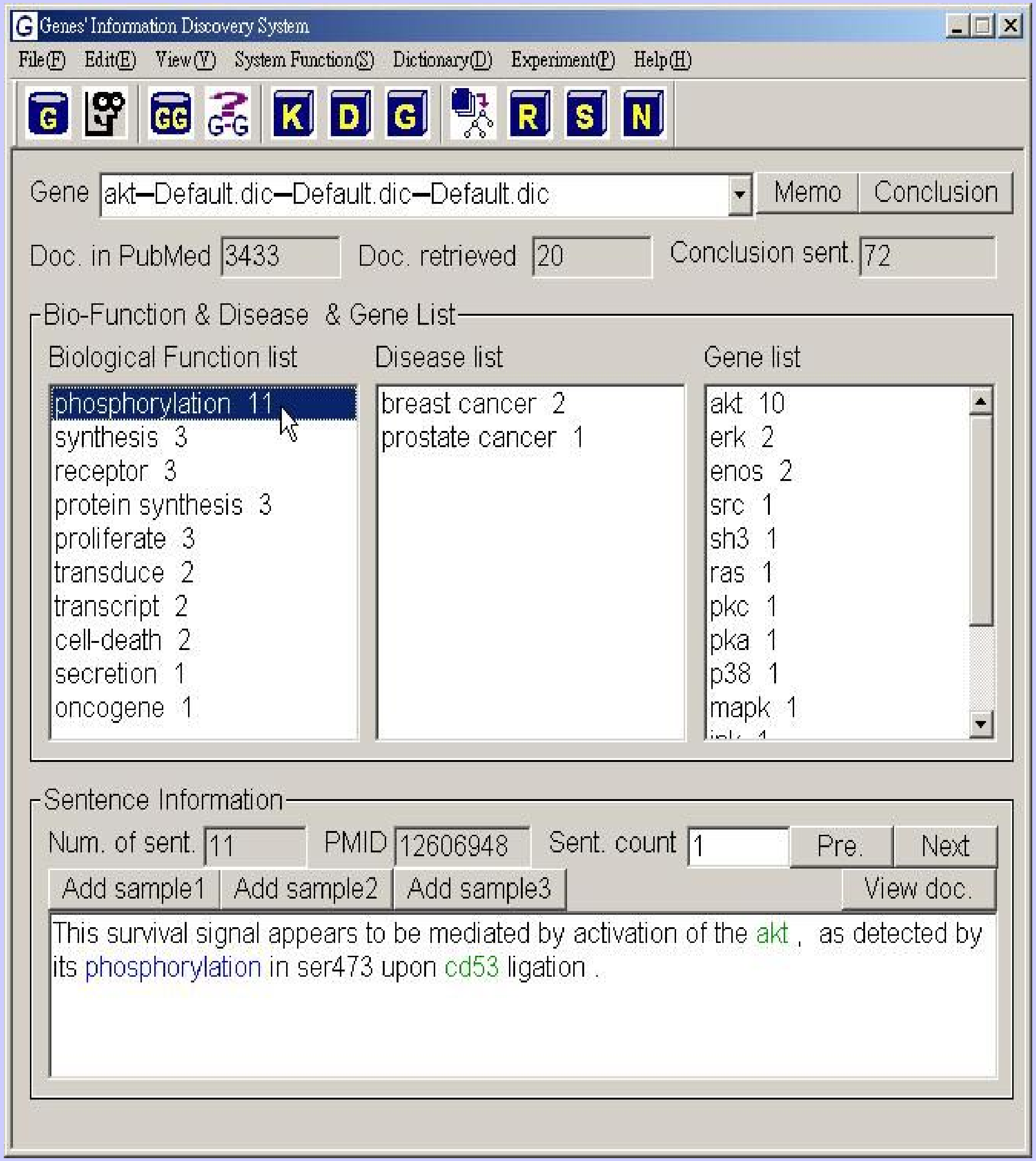
Text mining biomedical literature (2005)

Gene Information System (GIS)

- Focus on 4 types of gene-related info:
  - Biological function
  - Associated disease
  - Related genes
  - Gene-gene relations
- Gene information screening
- Gene-gene relation extraction.
- Downloadable

http://iir.csie.ncku.edu.tw/~yuhc/gis/

GIS


Genes' Information Discovery System
File(F) Edit(E) View(V) System Function(S) Dictionary(D) Experiment(P) Help(H)
G GG G-G K D G R S N
Memo
Conclusion
Pre.
Next
Add sample1
Add sample2
Add sample3
View doc.

Gene akt—Default.dic—Default.dic—Default.dic

Doc. in PubMed 3433 Doc. retrieved 20 Conclusion sent. 72

Bio-Function & Disease & Gene List

Biological Function list	Disease list	Gene list
phosphorylation 11	breast cancer 2	akt 10
synthesis 3	prostate cancer 1	erk 2
receptor 3		enos 2
protein synthesis 3		src 1
proliferate 3		sh3 1
transduce 2		ras 1
transcript 2		pkc 1
cell-death 2		pka 1
secretion 1		p38 1
oncogene 1		mapk 1

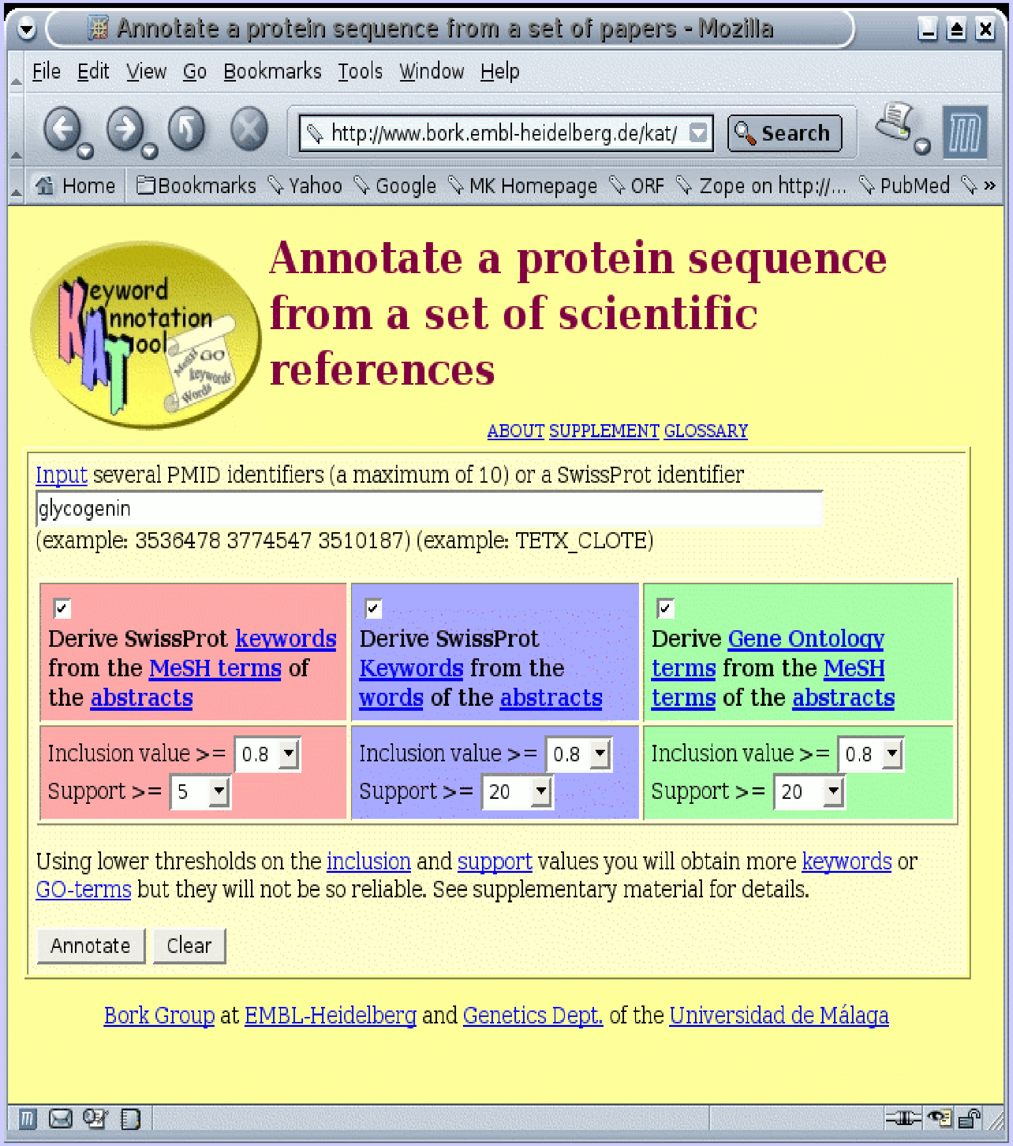
Sentence Information

Num. of sent. 11 PMID 12606948 Sent. count 1

This survival signal appears to be mediated by activation of the akt , as detected by its phosphorylation in ser473 upon cd53 ligation .

Chiang JH, Yu HC, Hsu HJ.GIS: a biomedical text-mining system for gene information discovery.*Bioinformatics*. 2004 Jan 1;20(1):120-1.

Keyword Annotation Tool (KAT)


Annotate a protein sequence from a set of papers - Mozilla
File Edit View Go Bookmarks Tools Window Help
http://www.bork.embl-heidelberg.de/kat/
Search
Home Bookmarks Yahoo Google MK Homepage ORF Zope on http://... PubMed »
Keyword Annotation Tool
KAT
GO
MeSH
keywords
Words

Annotate a protein sequence from a set of scientific references

ABOUT SUPPLEMENT GLOSSARY

Input several PMID identifiers (a maximum of 10) or a SwissProt identifier

glycogenin

(example: 3536478 3774547 3510187) (example: TETX_CLOTE)

☑ Derive SwissProt keywords from the MeSH terms of the abstracts	☑ Derive SwissProt Keywords from the words of the abstracts	☑ Derive Gene Ontology terms from the MeSH terms of the abstracts
Inclusion value >= 0.8 Support >= 5	Inclusion value >= 0.8 Support >= 20	Inclusion value >= 0.8 Support >= 20

Using lower thresholds on the inclusion and support values you will obtain more keywords or GO-terms but they will not be so reliable. See supplementary material for details.

Annotate Clear

Bork Group at EMBL-Heidelberg and Genetics Dept. of the Universidad de Málaga

- 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 using mappings between keyword systems. Bioinformatics. 2004 Sep 1;20(13):2084-91. Epub 2004 Apr 1.

Medical Knowledge Explorer (MeKE)

- Ontology-based text mining system.
- Methods of sentence alignment.
- Sentence classification methods.
- Flexible matching, stemming and indexing.
- Create new GO-term synonyms from text.
- Edit distance calculation
- Learn sentence motifs via sentence alignment
- Naïve Bayes sentence classifier

http://gen.csie.ncku.edu.tw/meke3/

Chiang JH, Yu HC.MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics*. 2003 Jul 22;19(11):1417-22

GO engine

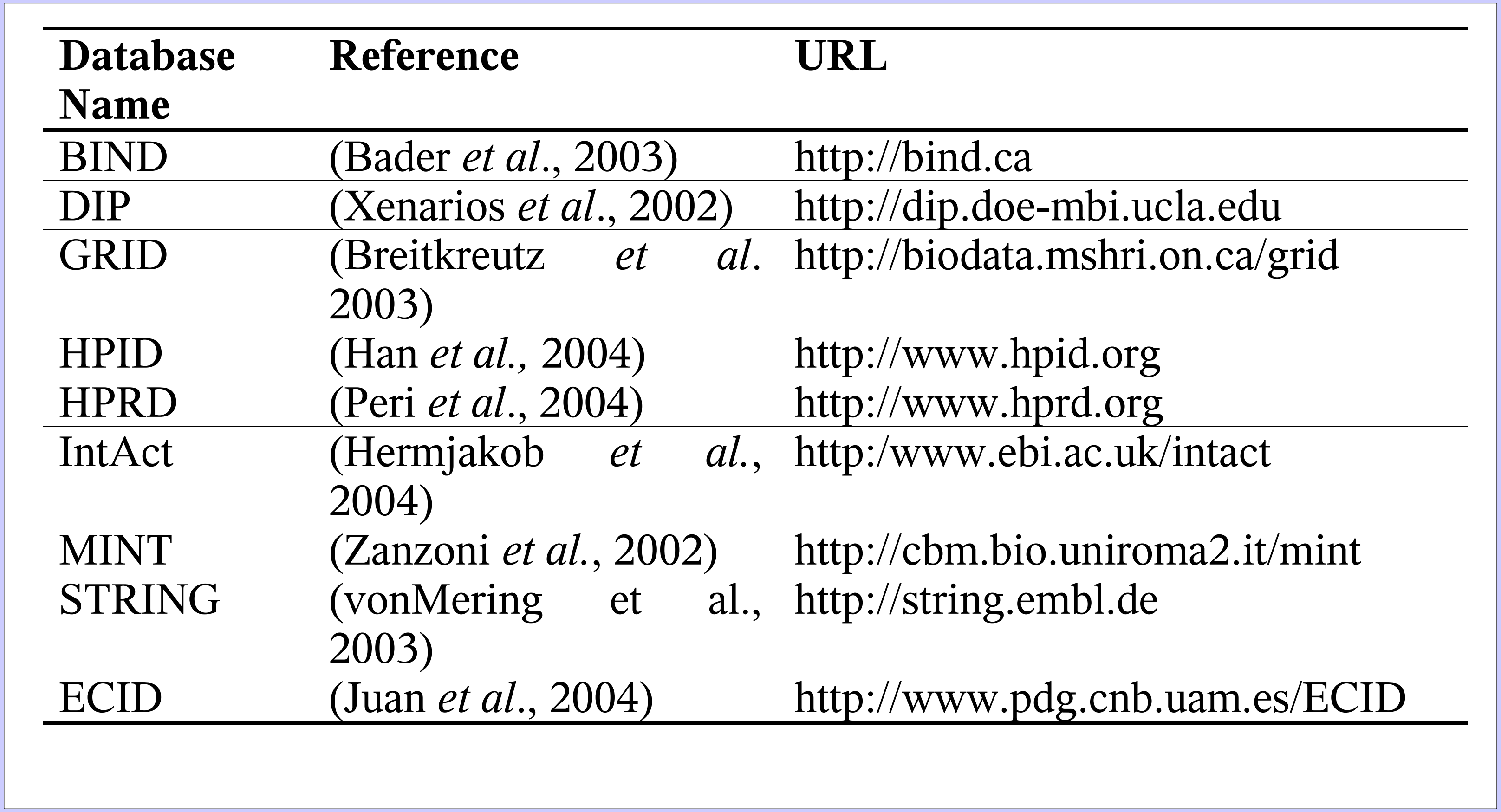
- Computational platform for GO annotation.
- Correlation of text info with specific GO nodes.
- Combines: homology info, protein clustering and text analysis.
- Calculate frequency of association of terms to GO nodes.

Xie H, Wasserman A, Levine Z, Novik A, Grebinskiy V, Shoshan A, Mintz L. Large-scale protein annotation through gene ontology. Genome Res. 2002 May;12(5):785-94.

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

Text mining and Protein interactions

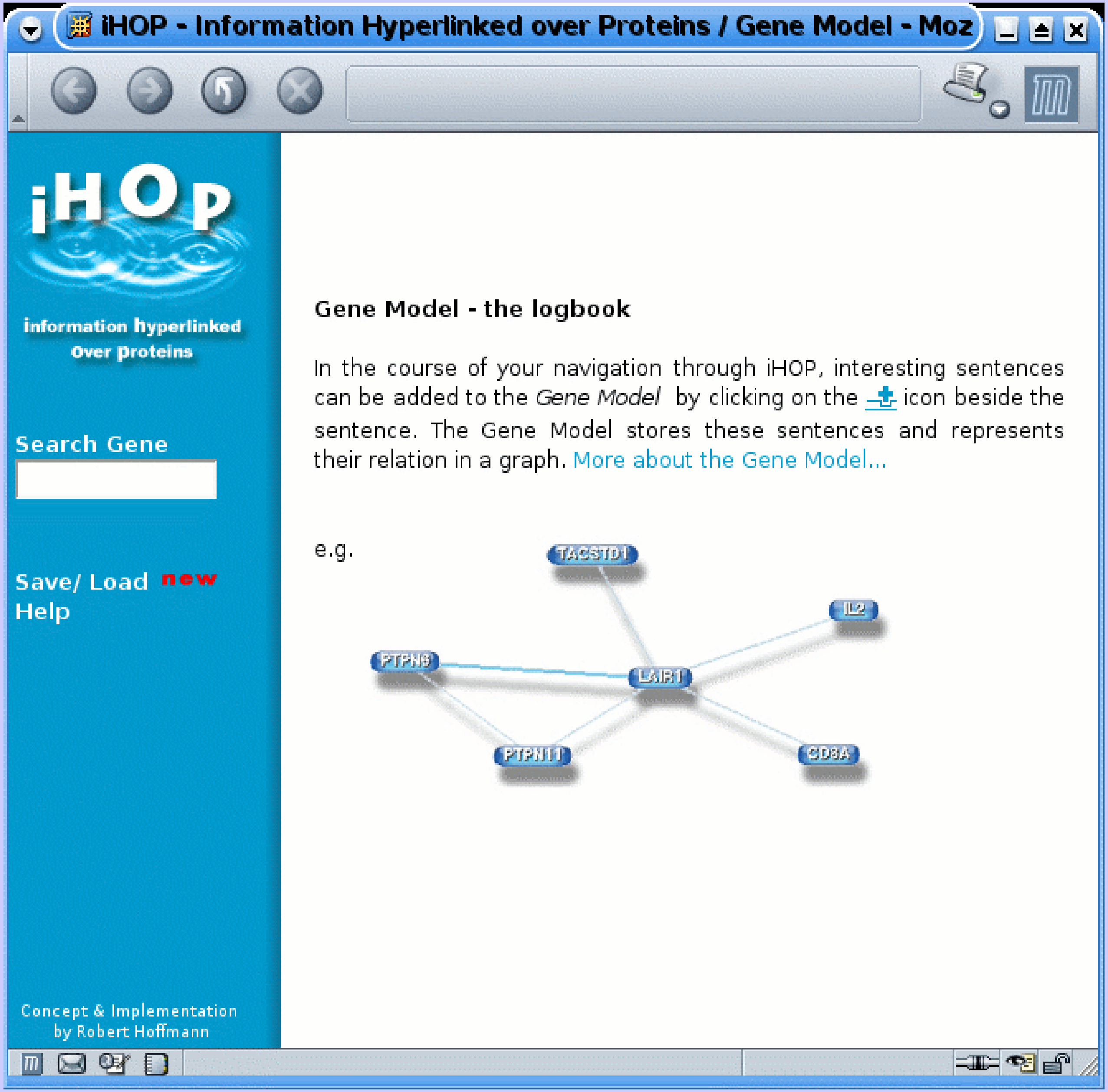
- Extract automatically those interactions from articles.
- NL used to characterise the nature of the interaction and its directionality.
- Literature-derived interaction networks:
  - power law distribution
  - scale free topology
- Visualised using network graphs.
- Methods range from: simple occurrence, expert derived word patterns (frames) to machine learning.

PubGene

- 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

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.

http://www.pubgene.org/


iHOP - Information Hyperlinked over Proteins / Gene Model - Moz
iHOP
information hyperlinked over proteins
Search Gene
Save/ Load new
Help
Concept & Implementation by Robert Hoffmann
TACSTD1
IL2
PTPN6
LAIR1
PTPN11
CD8A

Gene Model - the logbook

In the course of your navigation through iHOP, interesting sentences can be added to the *Gene Model* by clicking on the ✚ icon beside the sentence. The Gene Model stores these sentences and represents their relation in a graph. More about the Gene Model...

e.g.

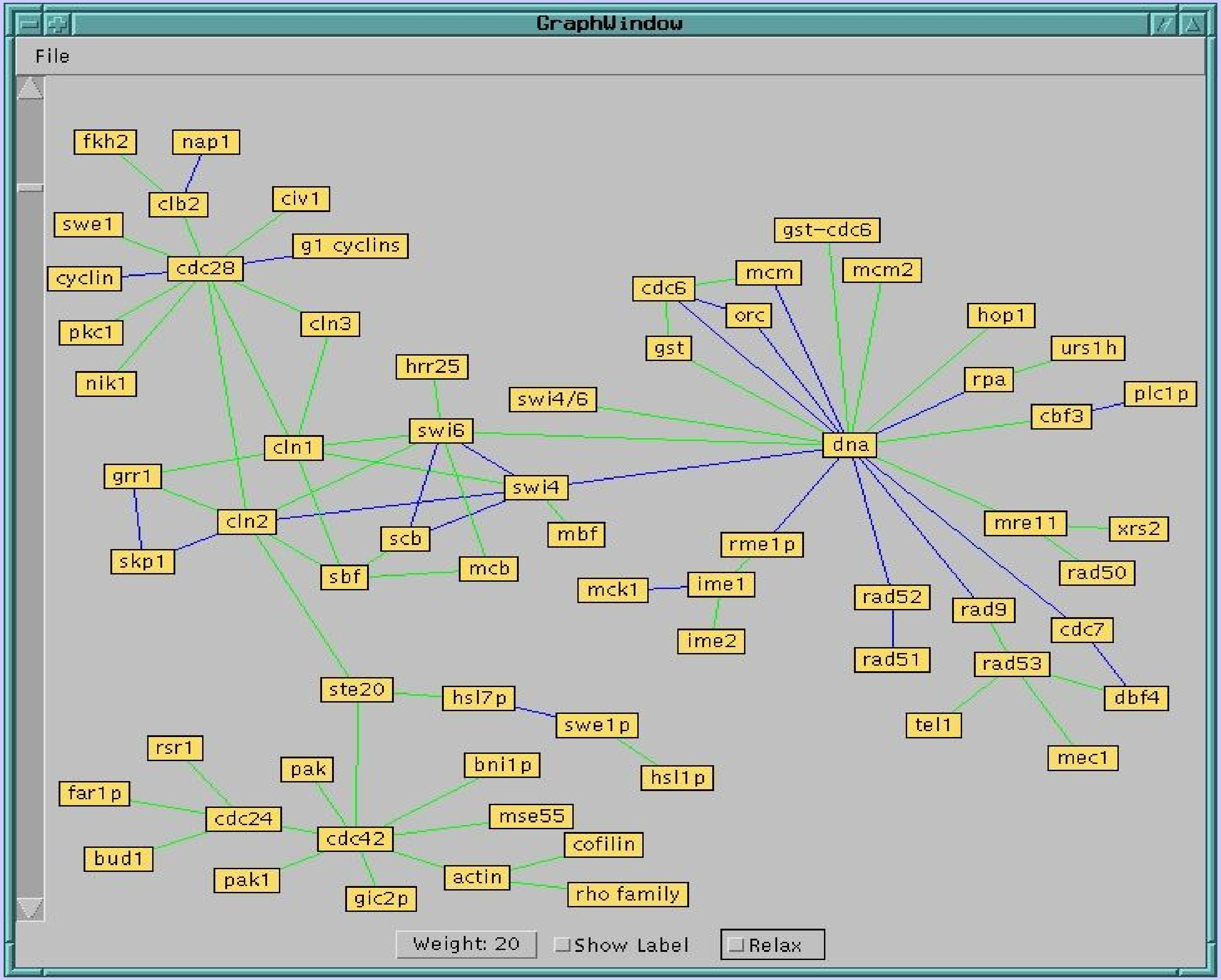
iHOP:

Visualization of protein interactions using network graphs

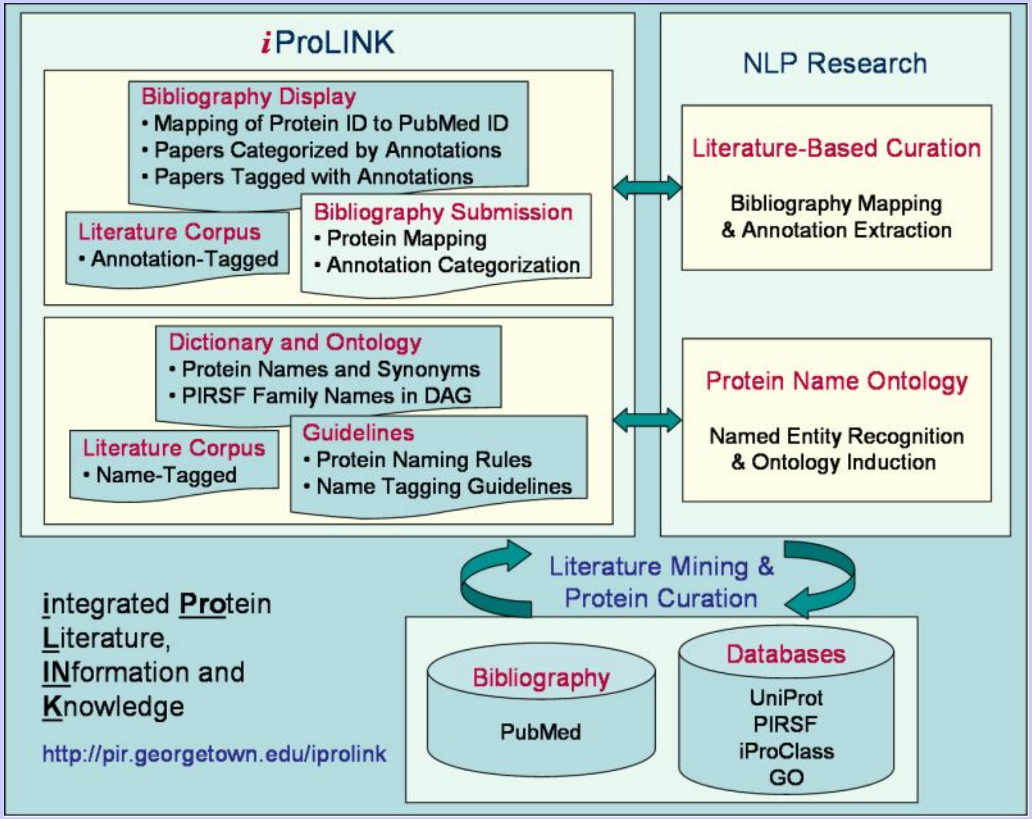
SUISEKI

- 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*
- Analysed 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 interaction network



iProLINK



- Mapped citations
- Annotation tagged literature corpora
- Data source for protein name ontology development

http://pir.georgetown.edu/iprolink/

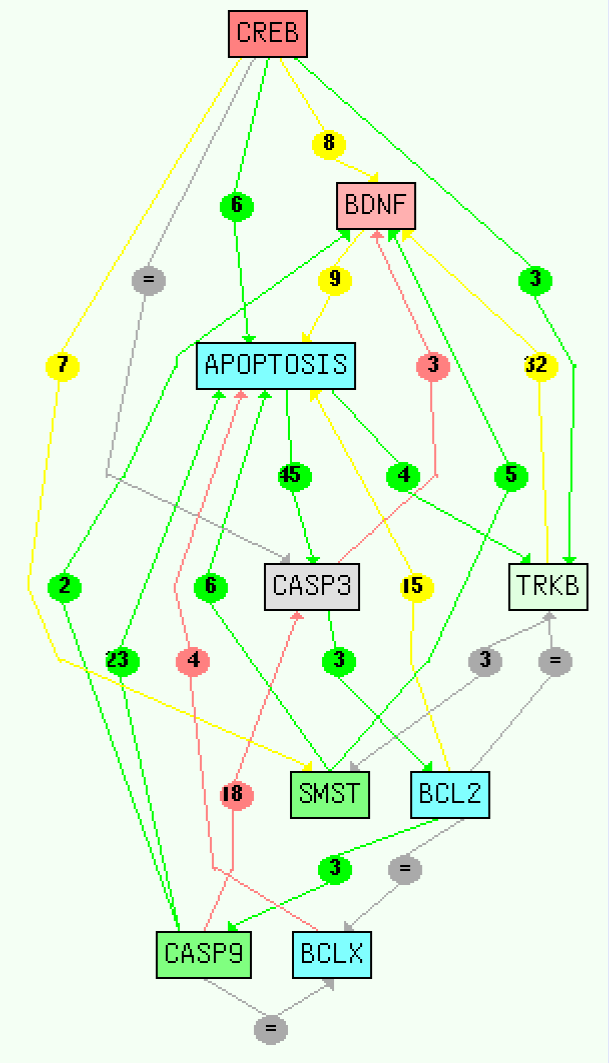
Hu ZZ, Narayanaswamy M, Ravikumar KE, Vijay-Shanker K, Wu CH.Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics*. 2005 Jun 1;21(11):2759-65. Epub 2005 Apr 6

Chilibot

- 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 visualisation
- Sentence as unit, POS tagging, shallow parsing and rules.

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

http://www.chilibot.net/



Chilibot (2)

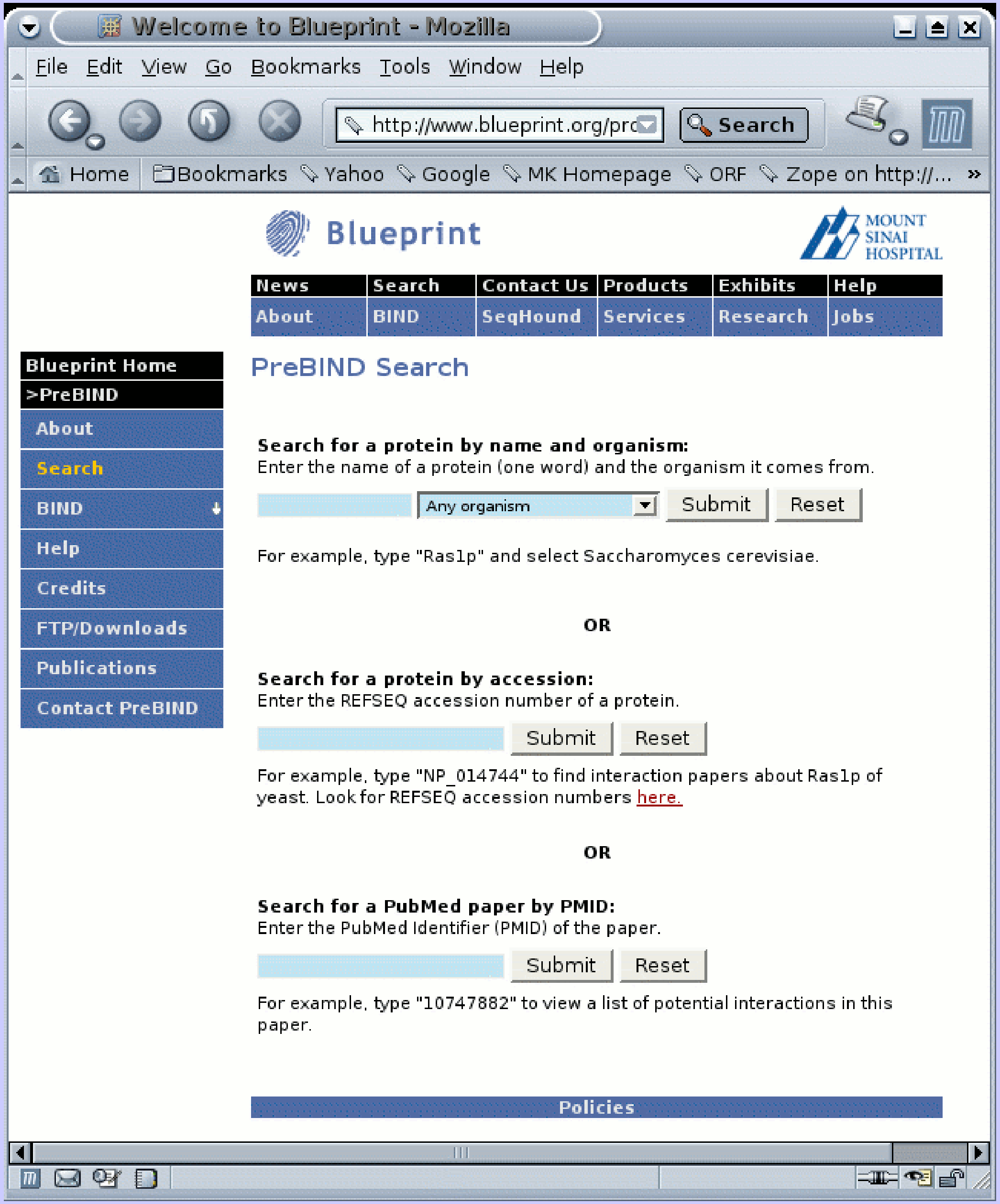
- Need registration.
- Hypothesis generation.

Chen H, Sharp BM.

Content-rich biological network constructed by mining PubMed abstracts.

BMC Bioinformatics. 2004 Oct 8;5(1):147.

http://www.chilibot.net/



PreBIND Search

Search for a protein by name and organism:
Enter the name of a protein (one word) and the organism it comes from.

Any organism ▾ Submit Reset

For example, type "Ras1p" and select Saccharomyces cerevisiae.

OR

Search for a protein by accession:
Enter the REFSEQ accession number of a protein.

Submit Reset

For example, type "NP_014744" to find interaction papers about Ras1p of yeast. Look for REFSEQ accession numbers here.

OR

Search for a PubMed paper by PMID:
Enter the PubMed Identifier (PMID) of the paper.

Submit Reset

For example, type "10747882" to view a list of potential interactions in this paper.

PreBIND

- Based on SVM.
- Query protein or accession number.
- Assist the Biomolecular Interaction Network Database (BIND)

Donaldson I, Martin J, de Bruijn B, Wolting C, Lay V, Tuekam B, Zhang S, Baskin B, Bader GD, Michalickova K, Pawson T, Hogue CW.PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine.*BMC Bioinformatics*. 2003 Mar 27;4(1):11.

http://www.blueprint.org/products/prebind

Text mining biomedical literature (2005)

Microarray data analysis

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

GEISHA

- Text mining tool for microarray analysis.
- Analyse the correlation between:
  - the increase of the level of expression patterns and
  - the significance of functional information derived from the literature.
- Extract functional information from the literature linked to the microarray genes.
- Calculates statistical significance of terms from documents associated to genes of each cluster.

Protein localization

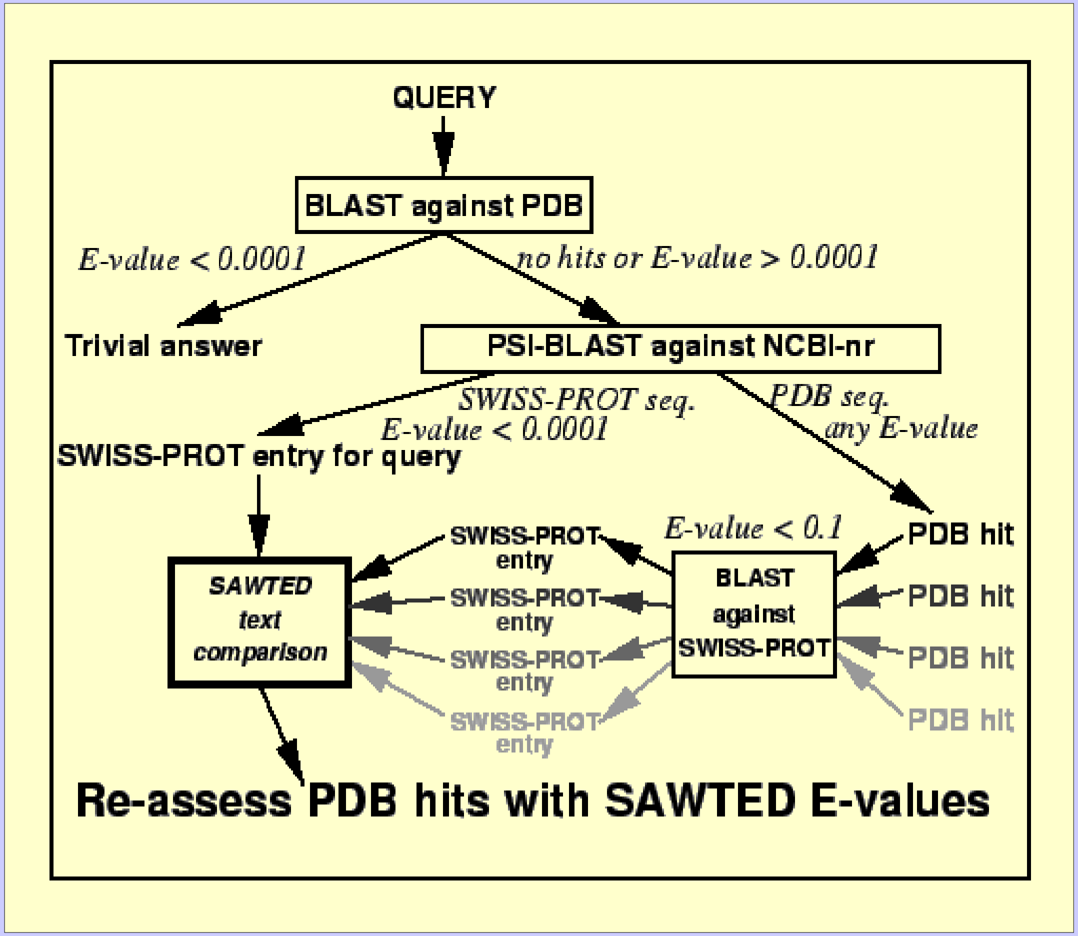
- Protein activity -> specific cellular environments.
- Localisation determination:
  - Experimental techniques.
  - Bioinformatics techniques (PSORT).
  - Text mining.
- Nair and Rost: lexical information in annotation database records.
- Stapley et al: Use SVM to classify proteins according to their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

http://medblast.sibsnet.org/

NLP and sequence analysis: SAWTED



QUERY

BLAST against PDB

E-value < 0.0001

Trivial answer

no hits or E-value > 0.0001

PSI-BLAST against NCBI-nr

SWISS-PROT seq. E-value < 0.0001

SWISS-PROT entry for query

PDB seq. any E-value

PDB hit

E-value < 0.1

BLAST against SWISS-PROT

PDB hit

PDB hit

PDB hit

PDB hit

SWISS-PROT entry

SWISS-PROT entry

SWISS-PROT entry

SWISS-PROT entry

SAWTED text comparison

Re-assess PDB hits with SAWTED E-values

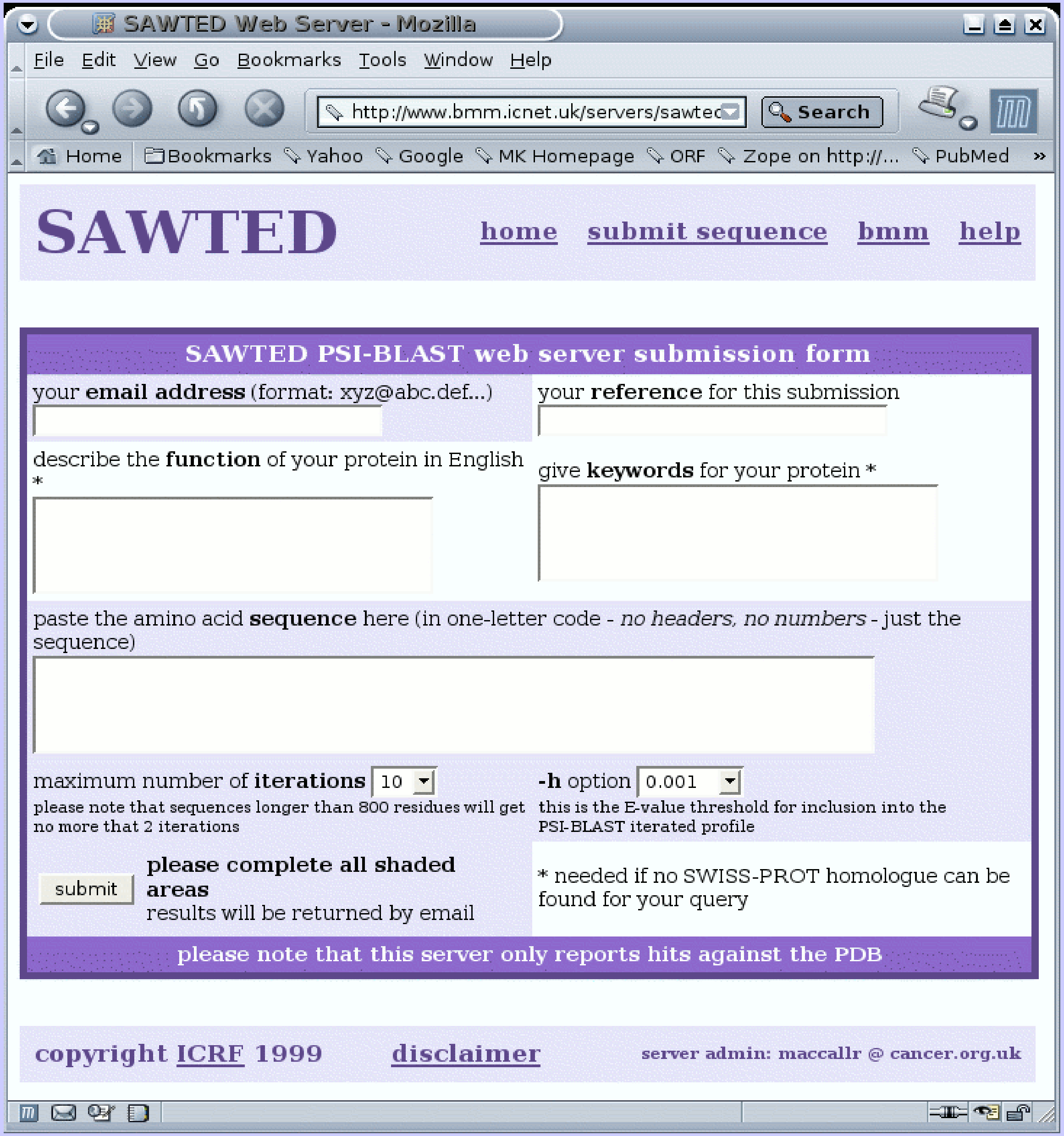
Sequence similarity

the base for identifying structure templates for query sequence

Structure Assignment With Text Description

Document comparison algorithms

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


SAWTED Web Server - Mozilla
File Edit View Go Bookmarks Tools Window Help
http://www.bmm.icnet.uk/servers/sawted
Search
Home
Bookmarks
Yahoo
Google
MK Homepage
ORF
Zope on http://...
PubMed
home
submit sequence
bmm
help
submit

SAWTED

SAWTED PSI-BLAST web server submission form

your **email address** (format: xyz@abc.def...)

your **reference** for this submission

describe the **function** of your protein in English *

give **keywords** for your protein *

paste the amino acid **sequence** here (in one-letter code - *no headers, no numbers* - just the sequence)

maximum number of **iterations** 10

please note that sequences longer than 800 residues will get no more that 2 iterations

-h option 0.001

this is the E-value threshold for inclusion into the PSI-BLAST iterated profile

please complete all shaded areas

results will be returned by email

* needed if no SWISS-PROT homologue can be found for your query

please note that this server only reports hits against the PDB

copyright ICRF 1999 disclaimer server admin: maccallr @ cancer.org.uk

Use information contained in text descriptions of SwissProt annotations

identification of remote homologues

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

Community wide evaluations

BIOINFORMATICS

- CASP
- CAMDA
- CAPRI
- GASP
- GAW
- PTC

BIO-NLP

- KDD cup
- BioCreative
- JNLPBA shared task
- TREC Genomics track
- LLL05 challenge

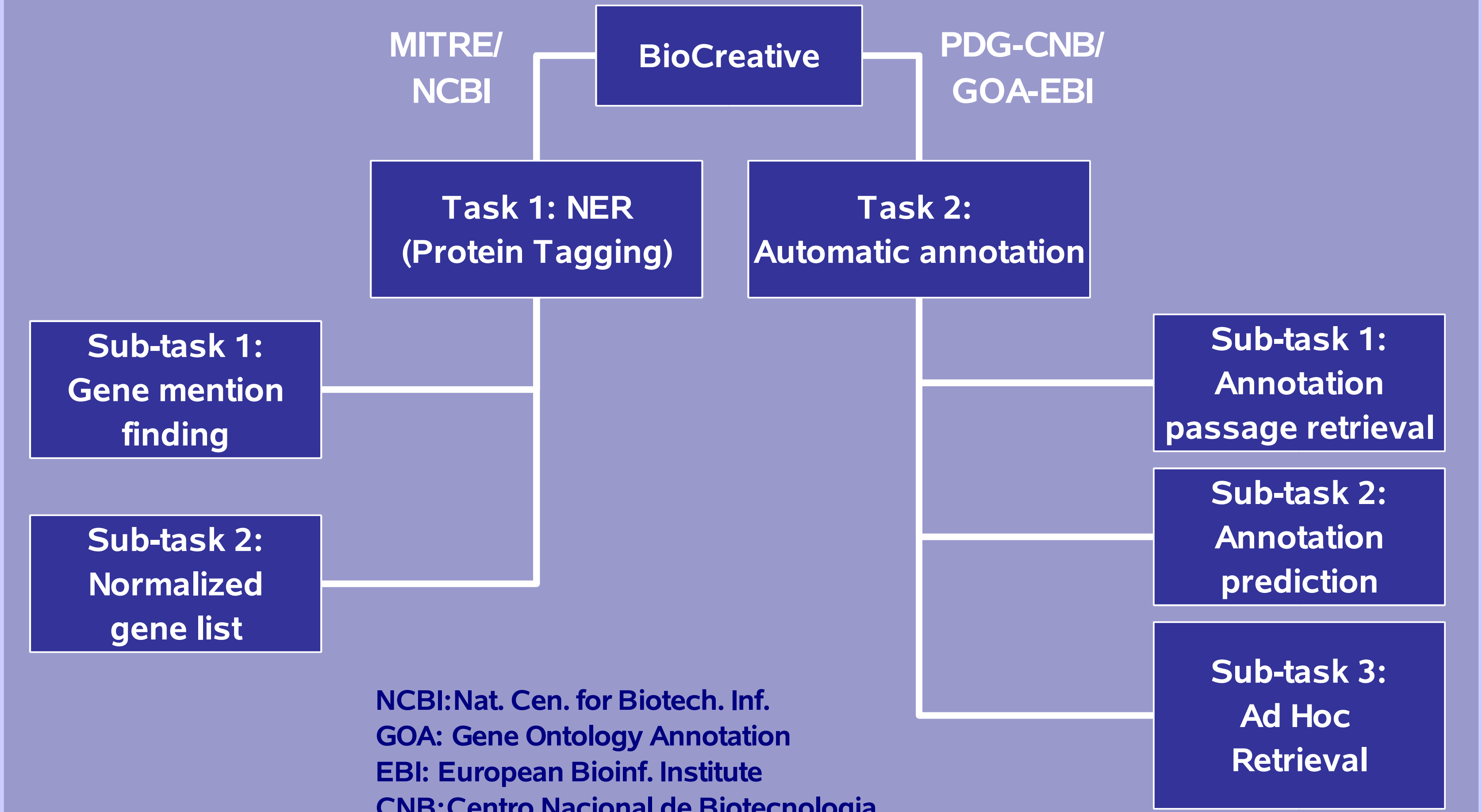
NLP/IR/IE

- MUC
- TREC

CASP: Critical asessment 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 Biomedicine
TREC: Text Retrieval conference
MUC: Message Understanding conference
LLL05: Genic interaction extraction challenge

Overview: BioCreative tasks


BioCreative

MITRE/ NCBI

PDG-CNB/ GOA-EBI

Task 1: NER (Protein Tagging)

Task 2: Automatic annotation

Sub-task 1: Gene mention finding

Sub-task 2: Normalized gene list

Sub-task 1: Annotation passage retrieval

Sub-task 2: Annotation prediction

Sub-task 3: Ad Hoc Retrieval

NCBI:Nat. Cen. for Biotech. Inf.
GOA: Gene Ontology Annotation
EBI: European Bioinf. Institute
CNB:Centro Nacional de Biotecnologia

BioCreative – Why?

- Open evaluation to determine the state of the art.
- Compare the performance of different methods.
- Produce a gold standard training set.
- Monitor improvements in the field.
- Produce useful evaluation tools/metrics.

BioCreative Task 1.1 summary

- Finding gene mentions in abstracts (NER).
- 15 teams, 3-4 submissions per team.
- Data and evaluation software provided by the NCBI.
- Performance: over 80% F-score (balanced precision and recall).
- Top scoring participants used some type of markov modelling (ME,HMM,CRF), SVM or manual rules.

BioCreative Task 1.2 summary

- Gene identifier list task.
- 8 teams, 3 submissions per team.
- Given an abstract from a specific model organism (Fly, Mouse, yeast) create the list of unique gene identifiers.
- F-score: yeast 0.92, fly 0.82 and mouse 0.79.
- Difficulties: ambiguity, complex names, distinguish between multiple identifiers.
- Methods: matching against lexical resources (e.g. exhaustive matching) or task 1.1 type systems.

BioCreative Task 2 description

- Automatic extraction and assignment of GO annotations for human proteins using full text articles.
- Based on triplets: protein – GO term – article passage.
- Task 2.1: Passage retrieval task, find the text passage which support a protein – GO term annotation.
- Task 2.2: text categorization task, predict protein – GO term associations and the corresponding text passage.
- Task 2.3*: ad hoc information retrieval, retrieve annotation relevant articles

Data sets and evaluation strategy

- GO: Gene Ontology: controlled vocabulary (concepts) within an ontology (DAG), 3 categories, MF: Molecular function, BP: Biological Process and CC: Cellular Component.
- GO concepts used for annotation purposes: GOA.
- Training set: 803 GOA derived full text articles from JBC journal.
- Test set: 113 articles for task 2.1 and 99 for 2.2 and triplets.
- Triplets: GO-term - protein – article -> return passage (task 2.1).
- Evaluation by GOA annotators from the EBI.
- Manually evaluation of the predicted passages within its context in the paper using a highlighting tool.
- Evaluation types: High:correct, Generally: OK but to general for practical use and Low:wrong.

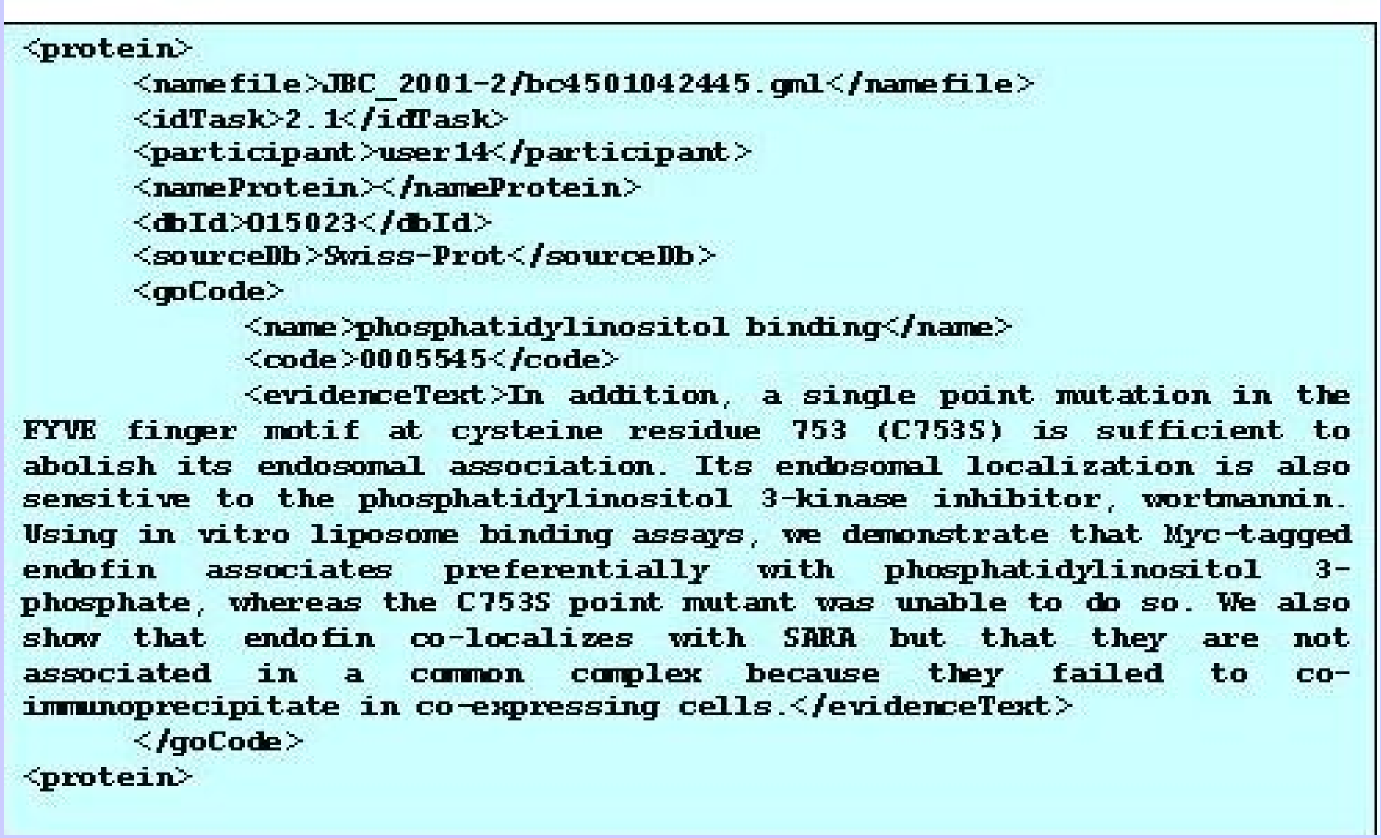
BioCreative Task 2 data set

Description	Training set	Test set 2.1	Test set 2.2
Full text articles	803	113	99
Total of GO annotation	2317	1076	1227
Nr of proteins in the GO annot	939	138	138
Nr GO terms used for annot	776	580	544
Average nr of annot/protein	2.467	7.797	8.891
Annotations of MF GO terms	709	330	356
Annotations of BP GO terms	1061	544	701
Annotations of CC GO terms	547	182	170
MF terms in the annotations	343	173	179
BP terms in the annotations	339	334	314
CC terms in the annotations	94	57	51

Data sets and evaluation strategy

- GO: Gene Ontology: controlled vocabulary (concepts) within an ontology (DAG), 3 categories, MF: Molecular function, BP: Biological Process and CC: Cellular Component.
- GO concepts used for annotation purposes: GOA.
- Training set: 803 GOA derived full text articles from JBC journal.
- Test set: 113 articles for task 2.1 and 99 for 2.2 and triplets.
- Triplets: GO-term - protein – article -> return passage (task 2.1).

BioCreative Task 2.1 sample submission



```
<protein>
        <namefile>JBC_2001-2/bc4501042445.gml</namefile>
        <idTask>2.1</idTask>
        <participant>user14</participant>
        <nameProtein></nameProtein>
        <dbId>015023</dbId>
        <sourceDb>Swiss-Prot</sourceDb>
        <goCode>
                <name>phosphatidylinositol binding</name>
                <code>0005545</code>
                <evidenceText>In addition, a single point mutation in the
FYVE finger motif at cysteine residue 753 (C753S) is sufficient to
abolish its endosomal association. Its endosomal localization is also
sensitive to the phosphatidylinositol 3-kinase inhibitor, wortmannin.
Using in vitro liposome binding assays, we demonstrate that Myc-tagged
endofin associates preferentially with phosphatidylinositol 3-
phosphate, whereas the C753S point mutant was unable to do so. We also
show that endofin co-localizes with SARA but that they are not
associated in a common complex because they failed to co-
immunoprecipitate in co-expressing cells.</evidenceText>
        </goCode>
<protein>
```

BioCreative Task 2 participating systems

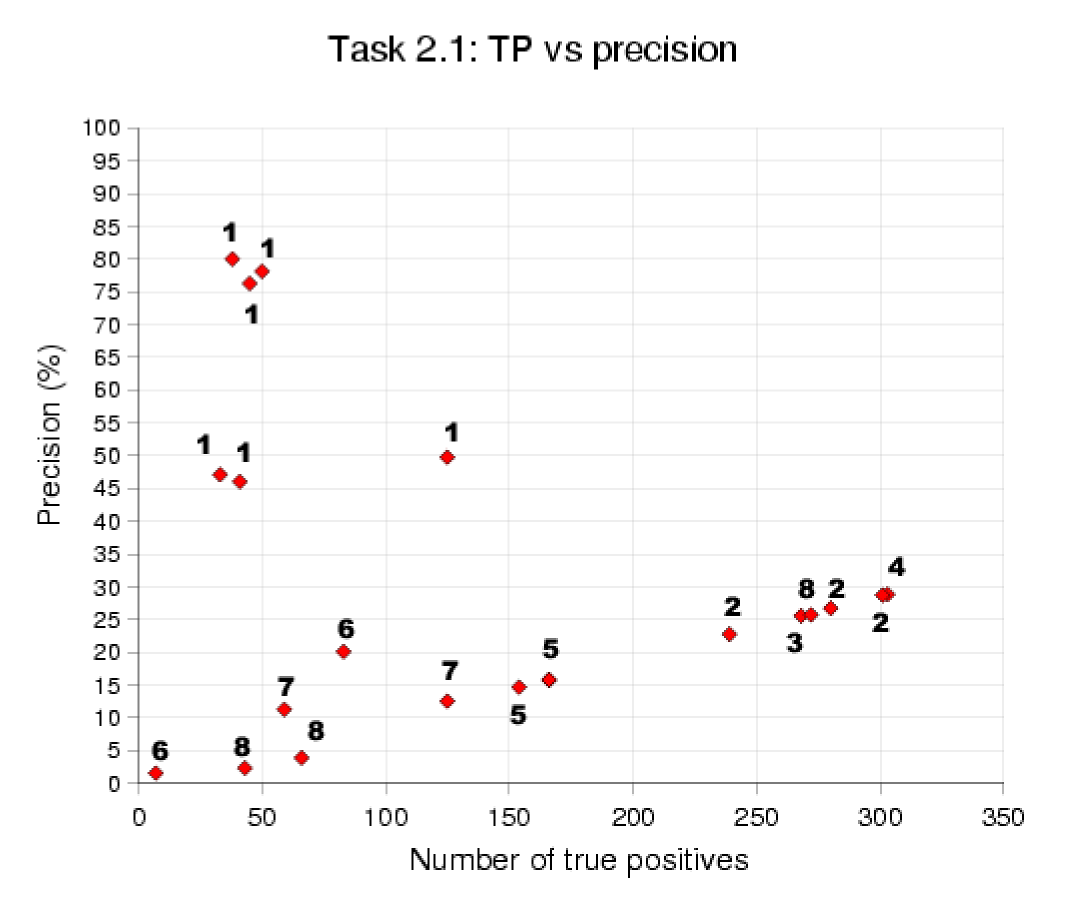
- 8 groups, max. 3 runs.
- Three strategies:

  (1) GO term centred, pattern matching, GO words (IC, word weight), recall centred

  (2) Machine learning techniques.

  (3) High precision, pattern matching and template extraction
- Tendency: sentence level, pattern matching, regular expressions and use of external resources (e.g. HuGO, UMLS), but:
- In general little overlap between the methods and the used resources.

BioCreative Task 2.1 results



TP: prediction evaluated as protein and GO terms correct

Precision: TP / Total nr. of evaluated submissions

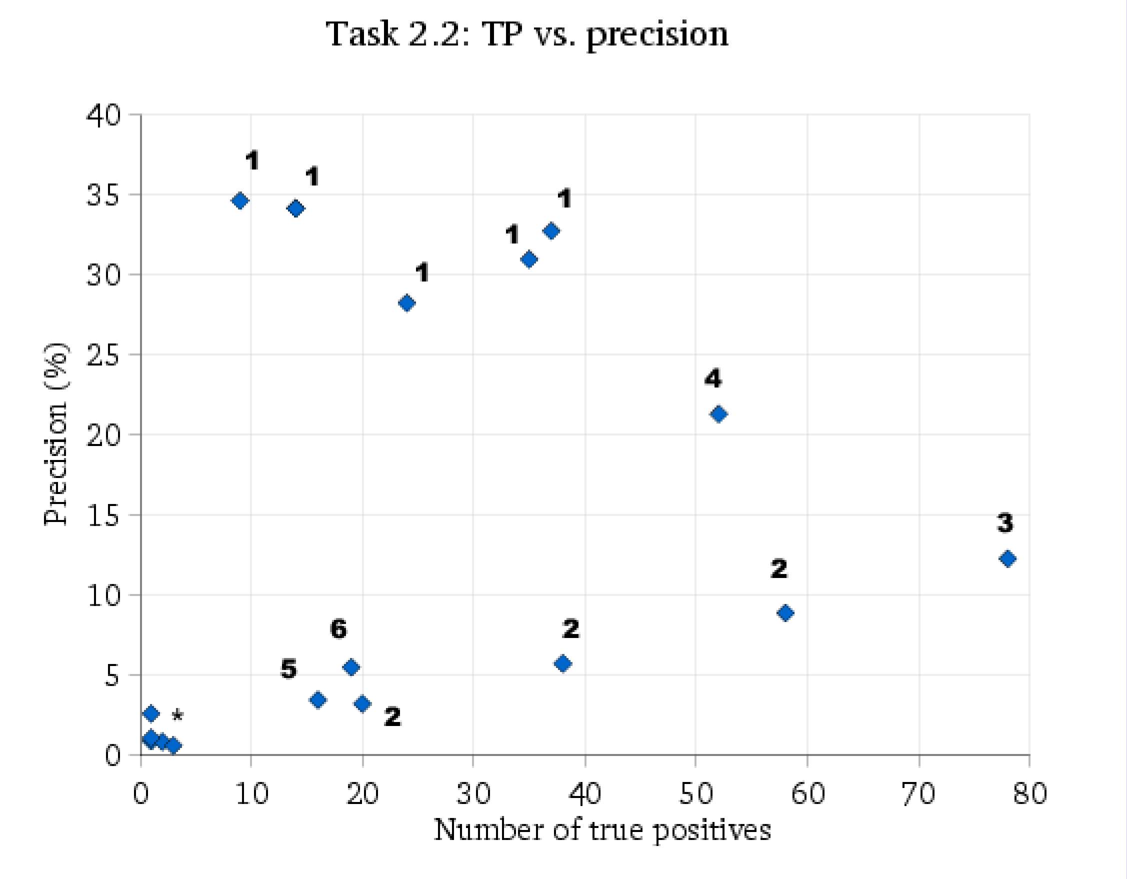
Teams:
1: Chiang et al.
2: Couto et al.
3: Ehrler et al.
4: Krallinger et al.
5: Krymolowski et al.
6: Ray et al.
7: Rice et al.
8: Verspoor et al.

BioCreative Task 2.1 examples

Query_id	Q96PH1_0000910_11483596
PMID:	11483596
UserId:	user20_1
UserName:	Couto et al.
ProteinName:	NADPH oxidase 5 gamma
AccessionNr:	Q96PH1
GO_term:	cytokinesis
GO_id:	0000910
EvalProtein:	high
EvalGO:	high
EvalAnnot:	highhigh
Len_GO:	1
GO_cat:	P

EvidenceText: Thus, <GLOSREF RID="G8">NOX5</GLOSREF> might have a function in the early stages of spermatogenesis such as cell division, induction of apoptosis, or DNA compaction.

BioCreative Task 2.2 results



TP: prediction evaluated as protein and GO terms correct

Precision: TP / Total nr. of evaluated submissions

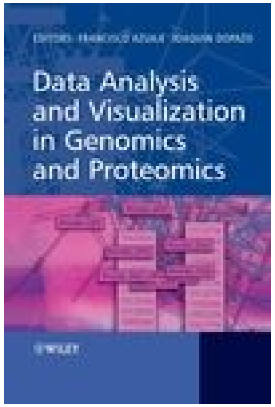
1: Chiang et al.
2: Couto et al.
3: Ehrler et al.
4: Ray et al.
5: Rice et al.
6: Verspoor et al.

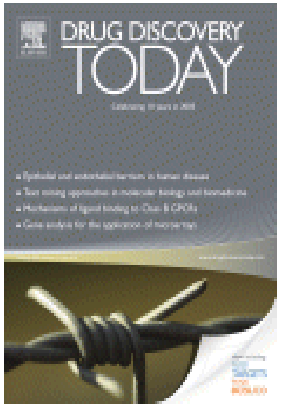
Selected review references

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.





Selected links

http://www.pdg.cnb.uam.es/martink/LINKS/bionlp_tools_links.htm

http://www.pdg.cnb.uam.es/martink/links.htm

Acknowledgements

- To the audience for paying attention
- To Alfonso Valencia for his supervision, support and discussions.
- To Belen Bañeres for organisational aid.
- The the Protein Design group at CNB and especially Maria Padron for discussions and suggestions.