

TEXT MINING

Bioinformatics and Computational Biology

Summer School – University Complutense of Madrid

LECTURE OVERVIEW

- The Biomedical literature
- Introduction to Natural Language Processing
- Information Retrieval in Biology
- Functional annotations: Gene Ontology
- Information Extraction in Biology
- Evaluation of Text mining tools
- Conclusions and outlook
- Useful links, reviews and articles

FROM EXPERIMENTS TO ARTICLES

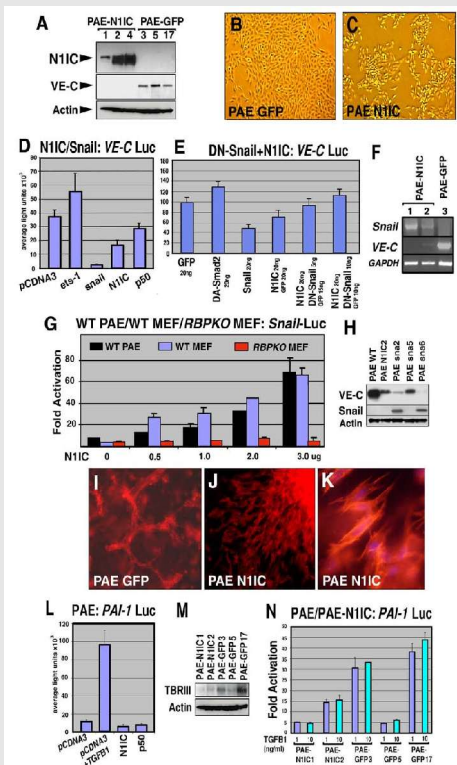
1- Experiments

Planning and carrying out experiments (lab work)



2- Results

Processing and interpretation of obtained results



3- Scientific articles

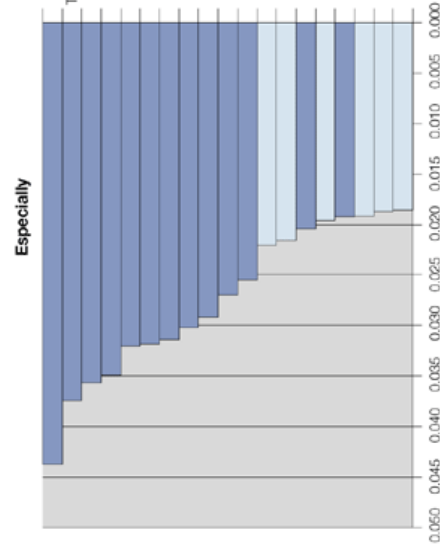
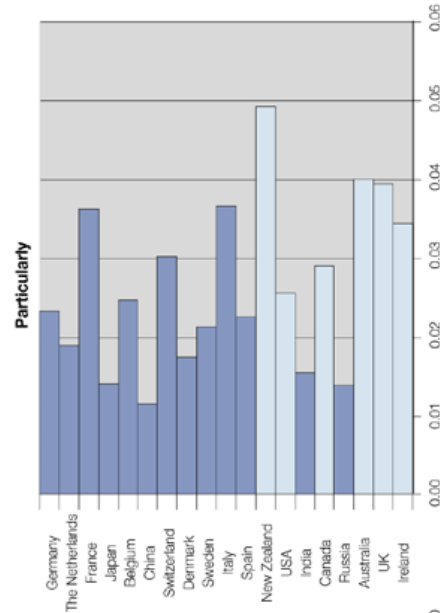
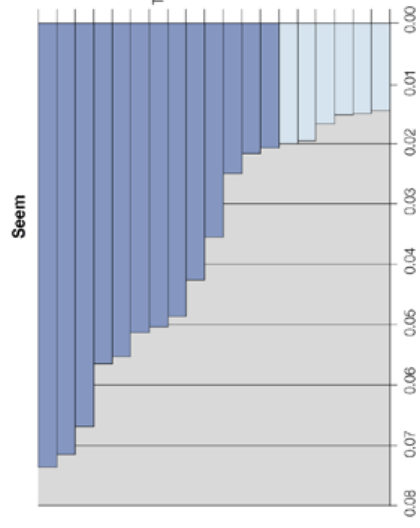
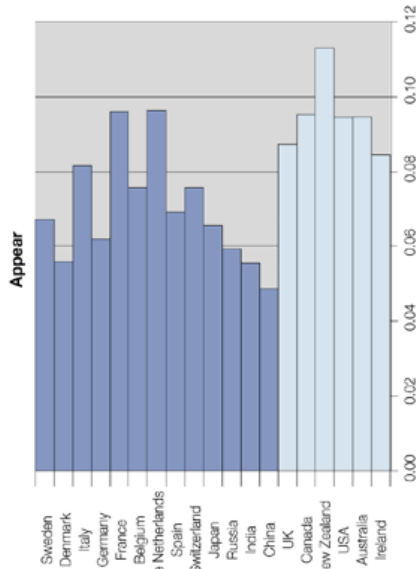
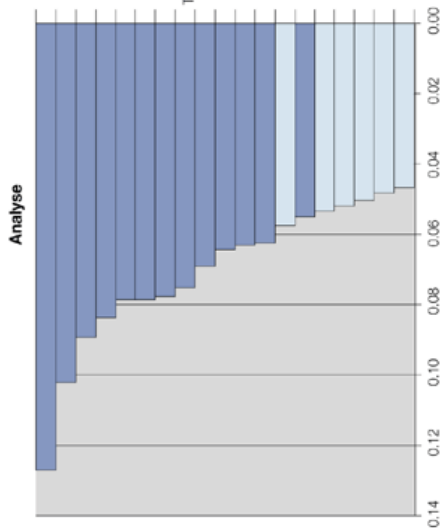
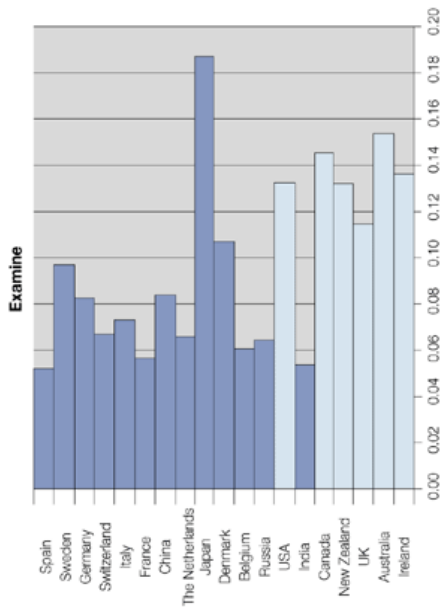
'Relevant' results are published in scientific journals



BIOMEDICAL LITERATURE CHARACTERISTICS

- Heavy use of domain specific terminology (12% biochemistry related technical terms).
- Polysemic words (word sense disambiguation), e.g. Drosophila genes like 'archipelago', 'capicua' or 'ebony'.
- Most words with low frequency (data sparseness).
- New names and terms created.
- Typographical variants
- Different writing styles (native languages)

SCIENTIFIC ENGLISH ?



- Most in English.
- Different native languages.
- Different word usage (preferences)

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

DIFFERENT COUNTRIES – DIFFERENT WORD USAGE

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 the agglomerates.	Hydroxypropylmethylcellulose (HPMC) was used to impart strength and sphericity to	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



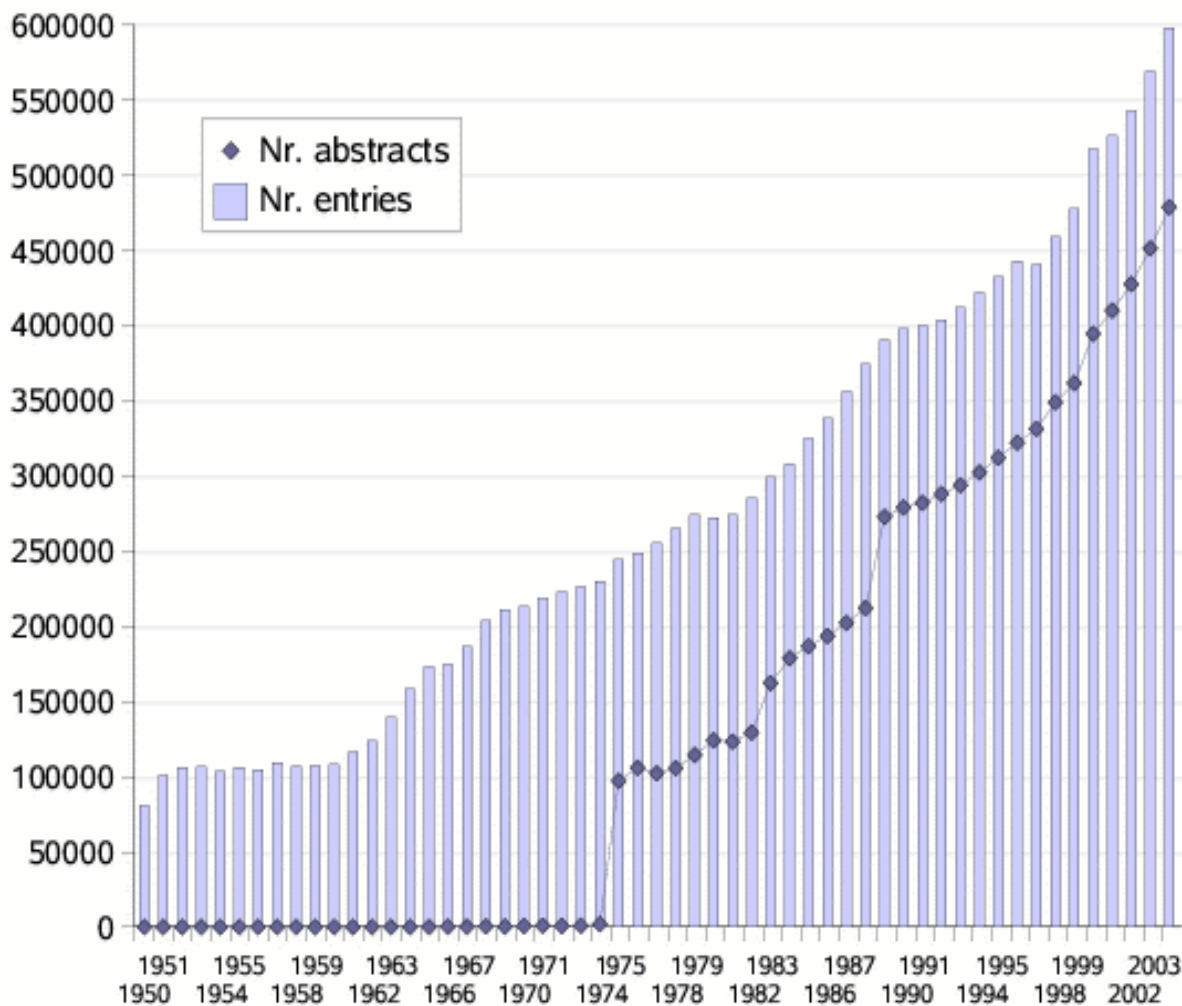
PubMed DATABASE



- Developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine NLM.
- Devoted mainly life science literature.
- Access through NCBI Entrez retrieval system:
<http://www.ncbi.nlm.nih.gov/entrez/>
- Entrez: text-based search and retrieval system.
- Publishers submit their citations electronically to PubMed.
- Over 14 million citations from the 50th until today.
- More than 48,000 journals
- Some articles are indexed with MeSH terms publication types and GenBank Accession nr.

PubMed GROWTH

PubMed growth



~ 450,000 new abstracts/a

> 4,800 biomedical journals

PubMed web-interface



Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi Search

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

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

- ◆ Enter one or more search terms, or click [Preview/Index](#) for advanced searching.
- ◆ Enter [author names](#) as smith jc. Initials are optional.
- ◆ Enter [journal titles](#) in full or as MEDLINE abbreviations. Use the [Journals Database](#) to find journal titles.

PubMed, a service of the National Library of Medicine, includes over 15 million citations for biomedical articles back to the 1950's. These citations are from MEDLINE and additional life science journals. PubMed includes links to many sites providing full text articles and other related resources.

Bookshelf Additions

Molecular Biology of the Cell, 4th Ed. and *The Genetic Landscape of Diabetes* are now available for interactive searching on the [Bookshelf](#).

PubMed Enhancements!

[Full author](#) searching is now available and the Single Citation Matcher has been enhanced to include first author searching and an autocomplete feature for journal titles.

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

PubMed retrieval



Entrez PubMed - Mozilla

File Edit View Go Bookmarks Tools Window Help

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

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

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 Summary Show 20 Sort by Send to

All: 101 Review: 1

Items 1 - 20 of 101 Page 1 of 6 Next

1: [Kim SS, Park RY, Jeon HJ, Kwon YS, Chun W.](#) [Related Articles, Links](#)
 Neuroprotective effects of 3,5-dicaffeoylquinic acid on hydrogen peroxide-induced cell death in SH-SY5Y cells.
 Phytother Res. 2005 Jun 2; 19(3): 243-245 [Epub ahead of print]
 PMID: 15934031 [PubMed - as supplied by publisher]

2: [Ruffels J, Griffin M, Dickenson JM.](#) [Related Articles, Links](#)
 Activation of ERK1/2, JNK and PKB by hydrogen peroxide in human SH-SY5Y neuroblastoma cells: role of ERK1/2 in H2O2-induced cell death.
 Eur J Pharmacol. 2004 Jan 12; 483(2-3): 163-73.
 PMID: 14729104 [PubMed - indexed for MEDLINE]

3: [De Sarno P, Shestopal SA, King TD, Zmijewska A, Song L, Ioppe RS.](#) [Related Articles, Links](#)
 Muscarinic receptor activation protects cells from apoptotic effects of DNA damage, oxidative stress, and mitochondrial inhibition.
 J Biol Chem. 2003 Mar 28; 278(13): 11086-93. Epub 2003 Jan 21.
 PMID: 12538580 [PubMed - indexed for MEDLINE]

Left sidebar menu:

- 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

PubMed retrieval

The screenshot shows a web browser window displaying a PubMed search result. The browser's address bar contains the URL: `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15784250&query_hl=7`. The page header includes the NCBI logo, the PubMed logo, and the National Library of Medicine logo. A navigation bar at the top lists various databases: All Databases, PubMed, Nucleotide, Protein, Genome, Structure, OMIM, PMC, Journals, and Books. The search bar shows "PubMed" and "for" with "Go" and "Clear" buttons. Below the search bar, there are tabs for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The main content area displays a search result for "1: J Mol Biol. 2006 Apr 15;347(5):895-902." with a "Link to full text" button. The title of the article is "Structure of the connector of bacteriophage T7 at 8A resolution: structural homologies of a basic component of a DNA translocating machinery." The authors listed are "Agirrezabala X, Martin-Benito J, Valle M, Gonzalez JM, Valencia A, Valpuesta JM, Carrascosa JL." The abstract text begins with "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." The PMID is 15784250. The page also includes a "Find similar entries" link and a "Related Articles, Links" link. The footer contains links to the Help Desk, NCBI, NLM, and NIH, and a Privacy Statement.

Annotations on the screenshot:

- Find similar entries**: Points to the "Find similar entries" link.
- Journal and publication date**: Points to the citation "1: J Mol Biol. 2006 Apr 15;347(5):895-902."
- Title**: Points to the article title "Structure of the connector of bacteriophage T7 at 8A resolution: structural homologies of a basic component of a DNA translocating machinery."
- Authors**: Points to the author list "Agirrezabala X, Martin-Benito J, Valle M, Gonzalez JM, Valencia A, Valpuesta JM, Carrascosa JL."
- Abstract**: Points to the beginning of the abstract text.
- PubMed identifier (unique document ID)**: Points to the PMID "15784250 [PubMed - indexed for MEDLINE]".

P

Exercise 1.1 : PubMed

1.1. Carry out a PubMed search for 'HIV' using the 'Limits' option. How many articles did you retrieve? Now try to follow the research interest in HIV over time through the associated publications deposited in PubMed by constructing a 'Publication time period' vs 'number of retrieved publications' table. Start from 1980 and use time intervals of 5 years (e.g. 1980-1984, 1985-1990,...). Describe your results?

Comment: The aim of search exercise is to explore an easy way to monitor research interests related to a certain topic of research. For instance pharmaceutical companies are often interested in monitoring research interests of other companies to obtain competitive intelligence.

http://www.pdg.cnb.uam.es/martink/LINKS/tm_sc_ucm2005.htm

P

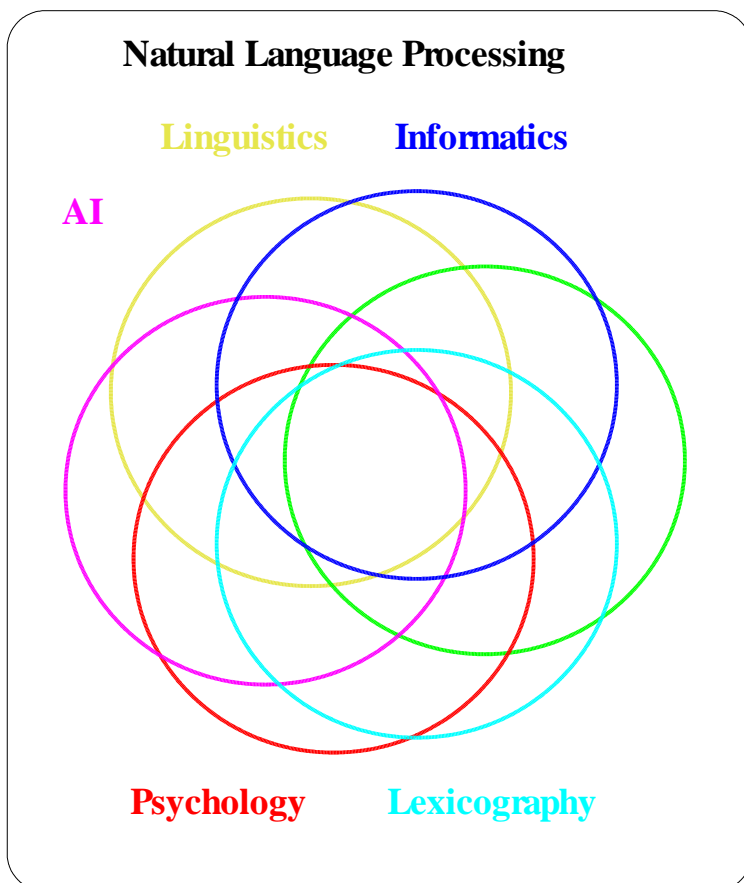
Exercises 1.2 - 1.3. : PubMed

1.2. Retrieve articles from PubMed for the Escherichia coli gene [TRME_ECOLI](#). How many articles did you retrieve? Which problems did you encounter? Comment the obtained results. Notice that you worked with this gene before in the

1.3. Perform the same search for the Escherichia coli gene [MRZ_ECOLI](#). Notice that you worked with that protein before, in the "[Redes de Interaccion de Proteinas](#)" session; and for the yeast gene [RPE_YEAST](#) (used in the [Analisis de Secuencias](#) session). What are the difficulties you encountered? How many documents did you retrieve?

http://www.pdg.cnb.uam.es/martink/LINKS/tm_sc_ucm2005.htm

Natural language processing (NLP)



- **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.

MAIN NLP TOPICS

- 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.

GOOGLE SCHOLAR

- Search engine for scholarly literature.
- URL: <http://scholar.google.com/>
- Include:
 - peer-reviewed papers
 - theses
 - books
 - preprints
 - abstracts
 - technical reports,...
- Return a ranked list according to relevance to user query.
- Ranking uses: full text, authors, publication type/journal, nr of citations in scholarly literature.

P

EXERCISE 2: GOOGLE SCHOLAR

Google developed **Google Scholar**, in order to provide a search engine specifically for academic and research users. Try out the search queries proposed in exercises 1.3 and 1.4. using the advanced Scholar Search. Compare the results with the results of PubMed. What are the advantages and disadvantages when using Google Scholar?

http://www.pdg.cnb.uam.es/martink/LINKS/tm_sc_ucm2005.htm

SELECTIVE DISSEMINATION OF INFORMATION SERVICES (SDI)

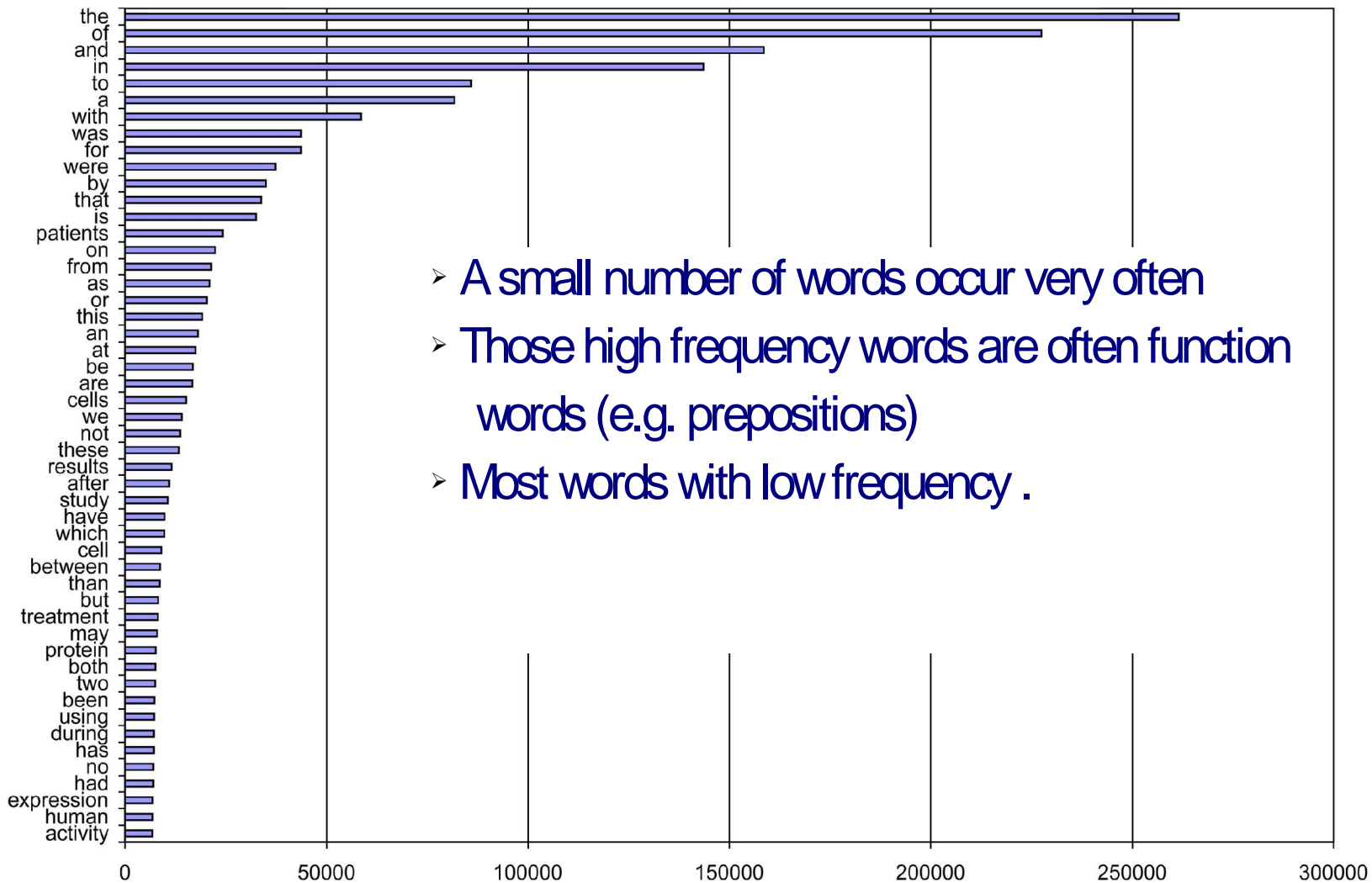
- Service provided by a library or data repository institution which periodically alerts users of new publications.
- New publications can be associated to certain subjects or information demands
- Often based on automated iterative/periodical IR queries.
- Advantages: new publications are automatically announced (e.g. using e-mail alerts)
- Disadvantages: implicit to IR based on Boolean queries, often irrelevant articles.
- Free SDI services based on PubMed / Biomedical literature:
 - Cubby (NCBI)
 - PubCrawler
 - BioMail

EXERCISE 3: SDI

1.2. Set up your own selective dissemination of information service (SDI) query using the [My NCBI Cubby](#) service for a query topic of your own interest or of one of the genes discussed before.

http://www.pdg.cnb.uam.es/martink/LINKS/tm_sc_ucm2005.htm

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

STOP WORD FILTERING

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 (VSM)

➤ Measure **similarity** between query and documents.

➤ (1) Document indexing , (2) Term weighting,
(3) similarity coefficient

➤ Query: a list of terms or even whole documents.

➤ Query as **vectors of terms**.

➤ **Term weighting** (w) according to their frequency:

➤ within the document (i)

➤ within the document collection (d)

➤ Widespread term weighting: $tf \times idf$.

➤ Calculate similarity between those vectors.

➤ Cosine similarity often used.

➤ Return a ranked list.

➤ Example: related article search in PubMed

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

w: term weight

tf: term frequency

$$idf_j = \log \left(\frac{N}{df_j} \right)$$

idf: inverted document frequency

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

sim(Q,D): similarity between query and document

PubMed related article search

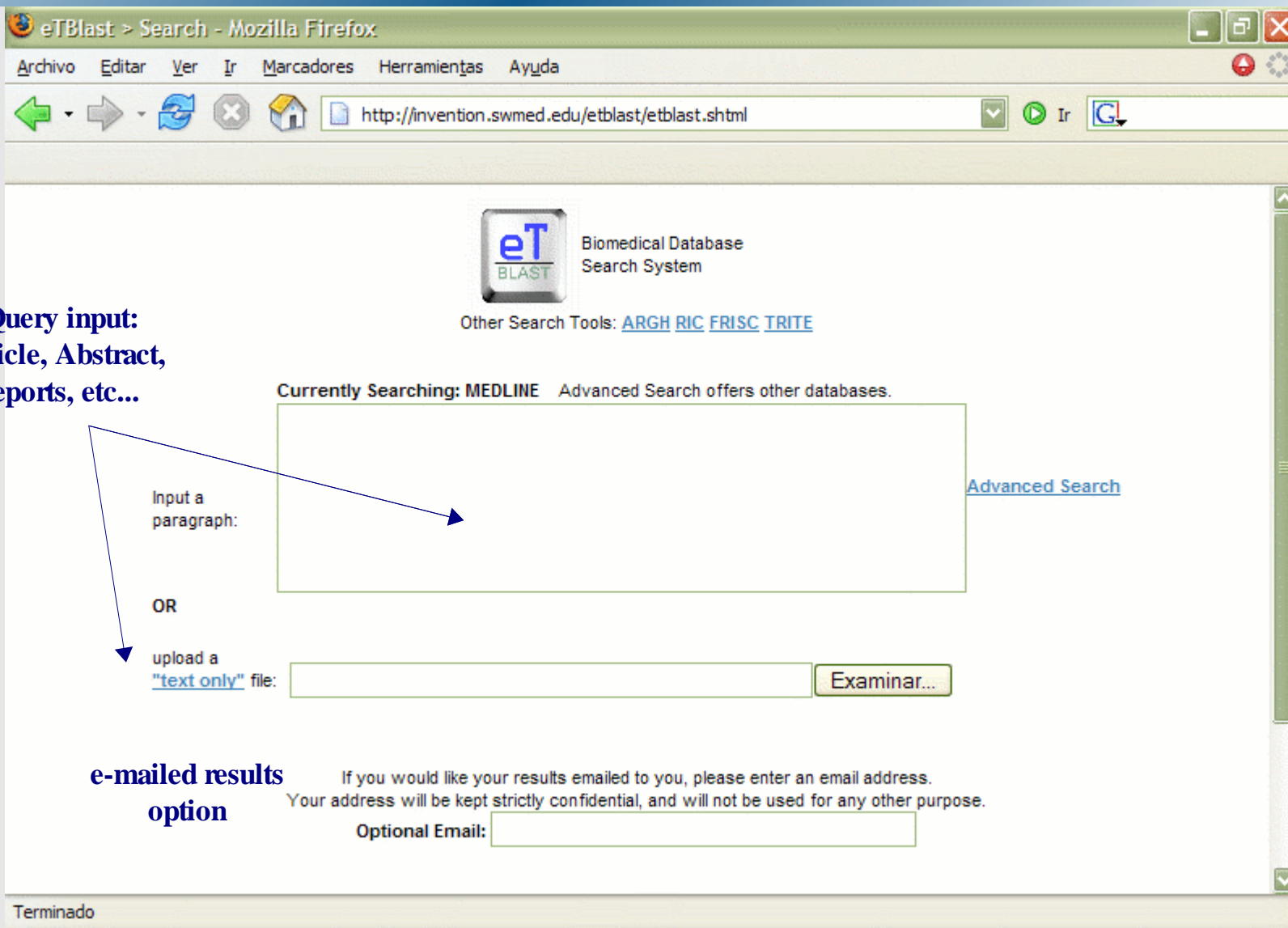
The screenshot shows a web browser window displaying a PubMed search result. The browser's address bar shows the URL: `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15784250&query_hl=7`. The page header includes the NCBI logo, the PubMed logo, and the National Library of Medicine (NLM) logo. A navigation bar at the top lists various databases: All Databases, PubMed, Nucleotide, Protein, Genome, Structure, OMIM, PMC, Journals, and Books. The search bar contains the text "PubMed for" and "Go" and "Clear" buttons. Below the search bar, there are options for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The main content area shows a search result for "J Mol Biol. 2005 Apr 15;347(5):895-902." with a "Link to full text" button. The title of the article is "Structure of the connector of bacteriophage T7 at 8A resolution: structural homologies of a basic component of a DNA translocating machinery." The authors listed are "Agirrezabala X, Martin-Benito J, Valle M, Gonzalez JM, Valencia A, Valpuesta JM, Carrascosa JL." The abstract text follows, starting with "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." At the bottom of the page, there is a "PubMed identifier" section with the PMID: 15784250 [PubMed - indexed for MEDLINE].

Annotations:

- Find similar entries:** A red box highlights the "Find similar entries" button in the top right corner of the search result area.
- Journal and publication date:** An arrow points to the citation "J Mol Biol. 2005 Apr 15;347(5):895-902.".
- Title:** An arrow points to the article title "Structure of the connector of bacteriophage T7 at 8A resolution: structural homologies of a basic component of a DNA translocating machinery.".
- Authors:** An arrow points to the author list "Agirrezabala X, Martin-Benito J, Valle M, Gonzalez JM, Valencia A, Valpuesta JM, Carrascosa JL.".
- Abstract:** An arrow points to the beginning of the abstract text "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.".
- PubMed identifier:** An arrow points to the PMID: 15784250 [PubMed - indexed for MEDLINE] at the bottom of the page.

eTBlast system

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



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

Currently Searching: MEDLINE Advanced Search offers other databases.

Input a paragraph:

OR

upload a "text only" file:

[Advanced Search](#)

e-mailed results option

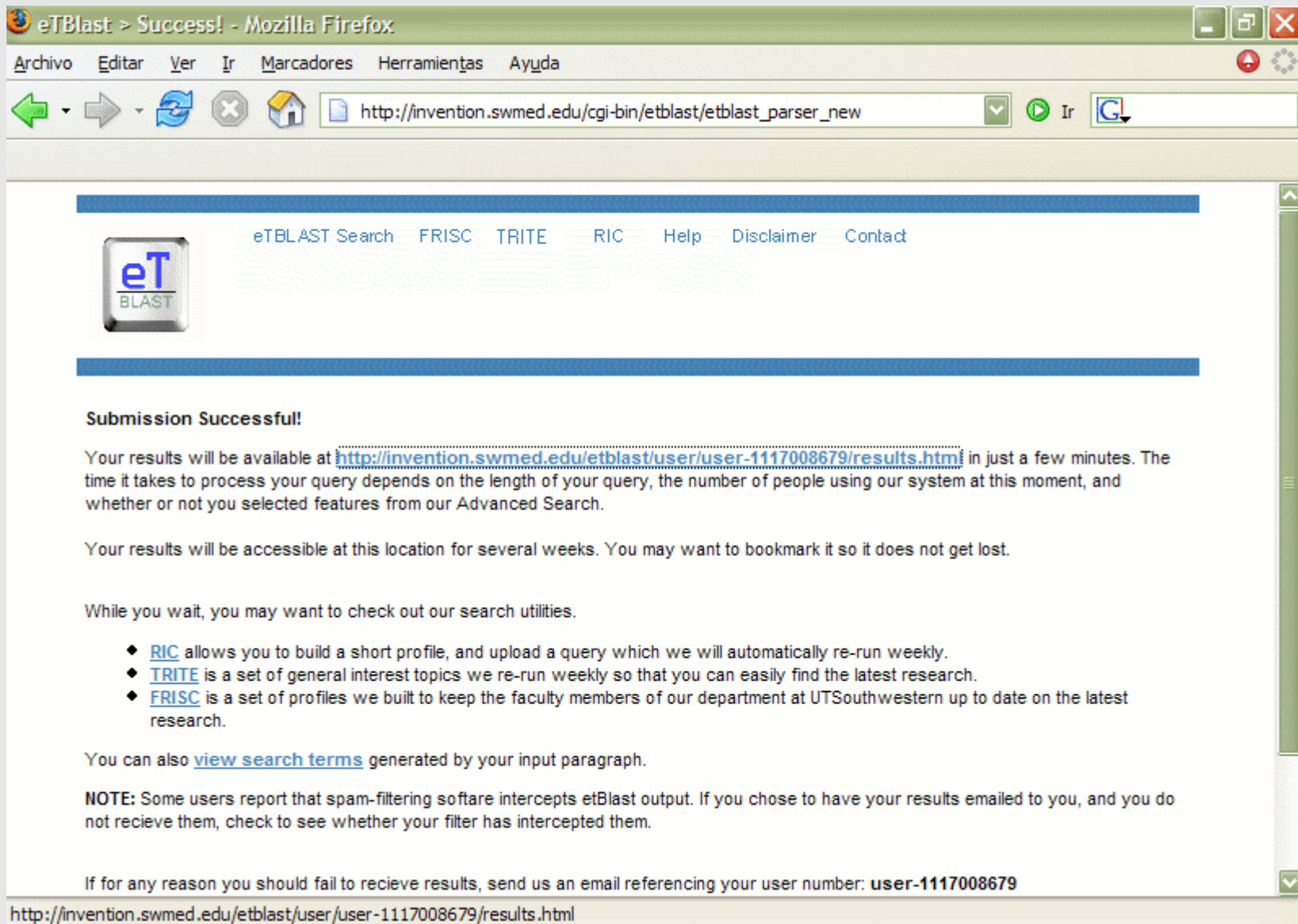
If you would like your results emailed to you, please enter an email address.
Your address will be kept strictly confidential, and will not be used for any other purpose.

Optional Email:

Terminado

eTBlast submission

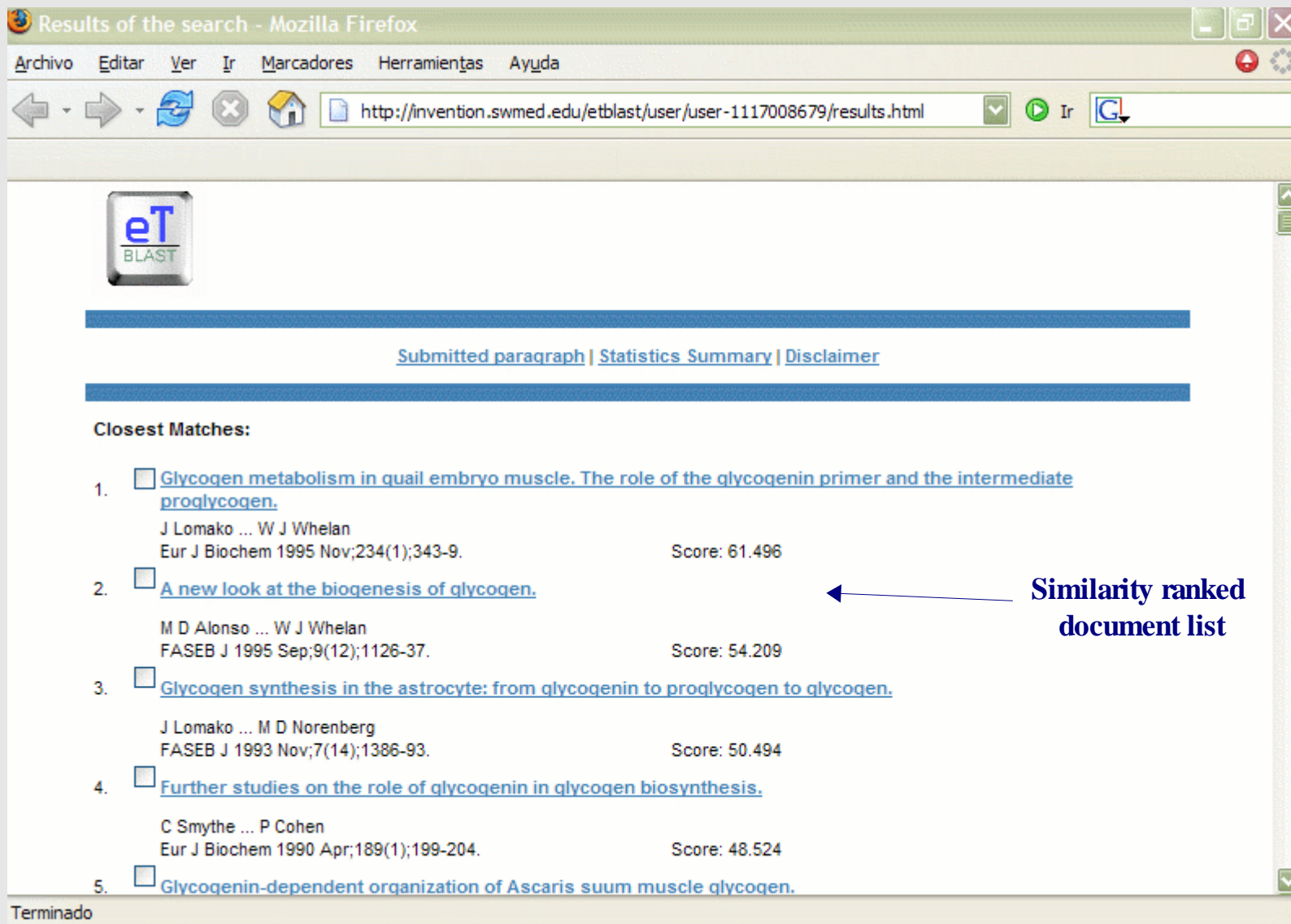
<http://invention.swmed.edu/etblast/index.shtml>



The screenshot shows a Mozilla Firefox browser window with the title "eTBlast > Success! - Mozilla Firefox". The address bar contains the URL http://invention.swmed.edu/cgi-bin/etblast/etblast_parser_new. The page content includes a navigation menu with links for eTBlast Search, FRISC, TRITE, RIC, Help, Disclaimer, and Contact. A logo for eTBlast is displayed on the left. The main content area features a "Submission Successful!" heading, followed by instructions on where to find the results and a list of search utilities (RIC, TRITE, FRISC). A note about spam-filtering software is also present, along with a contact email for failed submissions. The browser's status bar at the bottom shows the URL <http://invention.swmed.edu/etblast/user/user-1117008679/results.html>.

eTBLAST results


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



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](#)

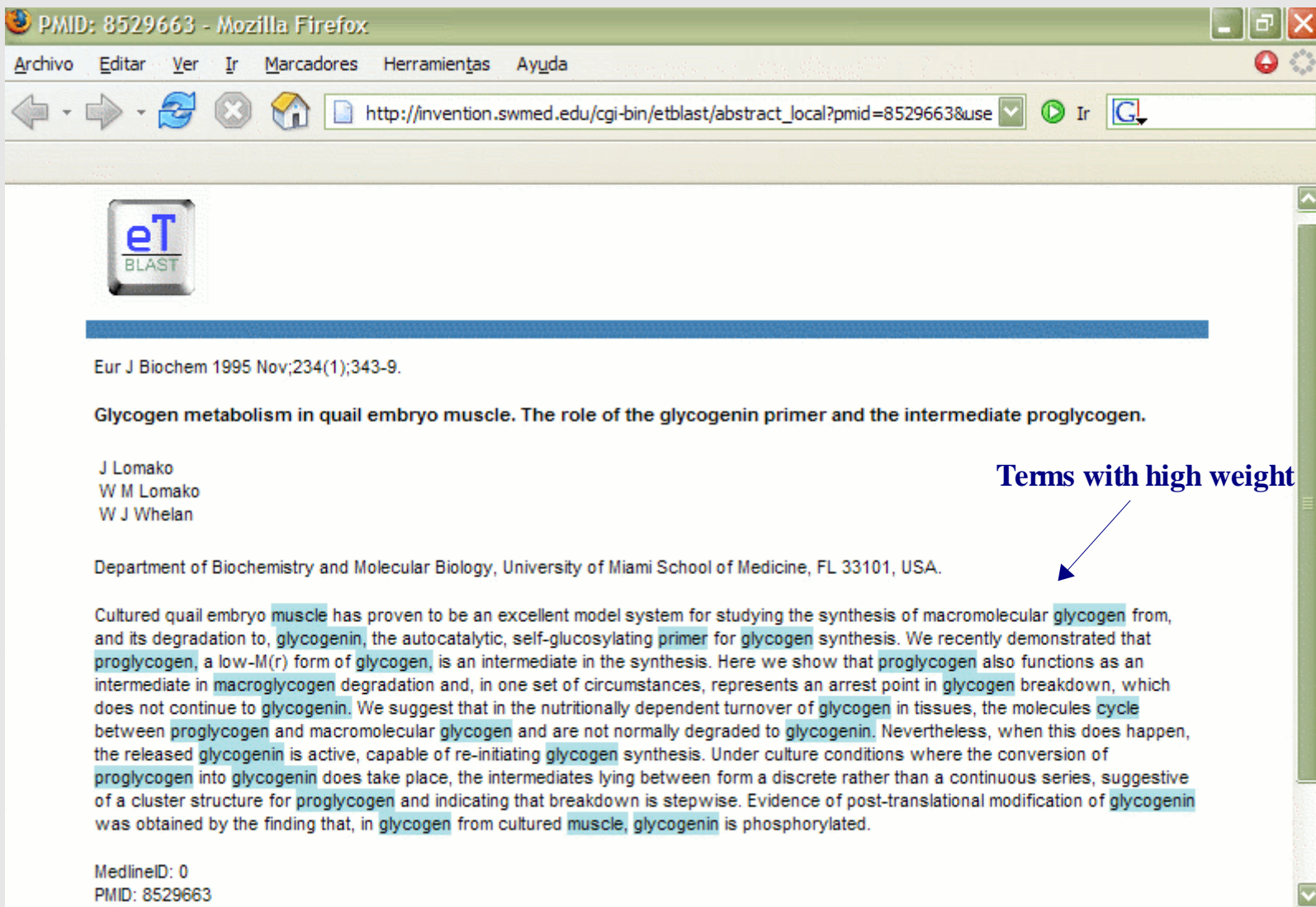
Closest Matches:

- [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
- [A new look at the biogenesis of glycogen.](#) ← Similarity ranked document list
M D Alonso ... W J Whelan
FASEB J 1995 Sep;9(12);1126-37. Score: 54.209
- [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
- [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
- [Glycogenin-dependent organization of Ascaris suum muscle glycogen.](#)

Terminado

eTBlast results


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



PMID: 8529663 - Mozilla Firefox

Archivo Editar Ver Ir Marcadores Herramientas Ayuda

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



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-glycosylating **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

Terms with high weight

P

EXERCISE 4. eTBlast

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

While writing a scientific article, report or a grant application, people often want to retrieve a set of documents which are related/relevant to this given work. What could/should you do in such situations? A PubMed search using alternative Boolean queries? Typically people use Boolean queries against PubMed to obtain their set of references.

You can use [eTBlast](#) instead and upload or past your free text to obtain similar articles. You can even iterate the search by selecting a subset of relevant documents retrieved in the first eTBlast round.

In case you have your own input document or are interested in certain PubMed article you can use it as your query text (or else try some of the following files: [etblast_sample1.txt](#), [etblast_sample1_trmE.txt](#)).

Notice that eTBlast is relatively slow. Use the advance search mode, you can try out different metrics for calculating the document similarity. You can try out uploading your own stop word file: [stop_word_list.txt](#) to filter those for when calculating the document similarity.

Explain the output (ranked list). Compare the list of similar documents for a given abstract in PubMed (related article search) with the results of eTBlast. What are the advantages of using eTBlast and what are the disadvantages. Are the highlighted word (with high weight) according to your opinion relevant and discriminative?

http://www.pdg.cnb.uam.es/martink/LINKS/tm_sc_ucm2005.htm

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

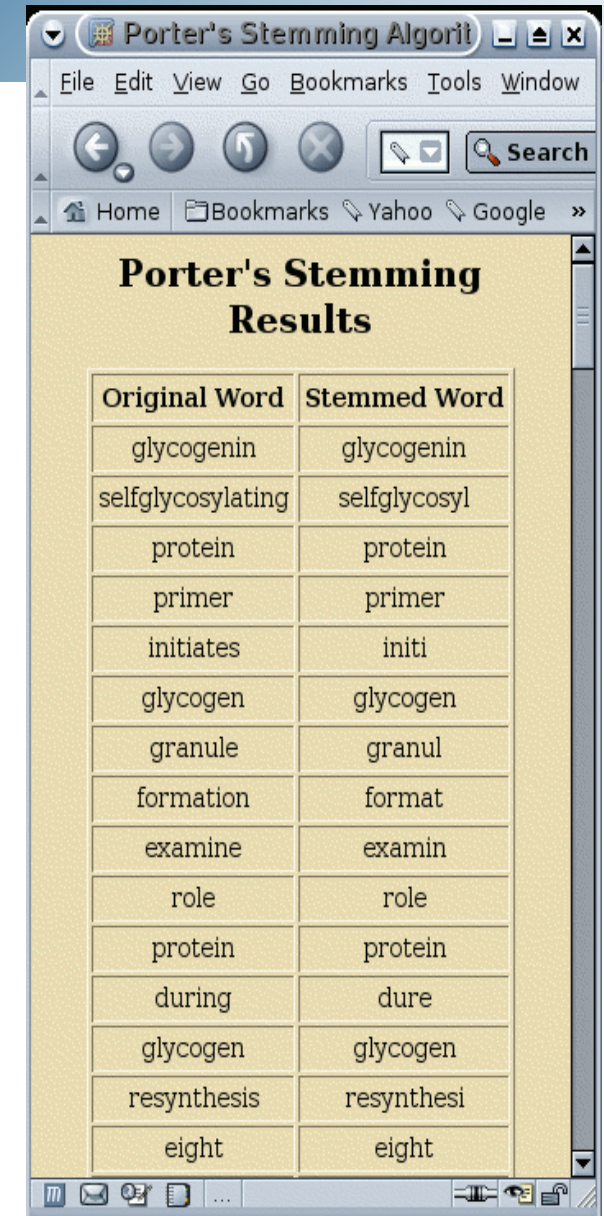
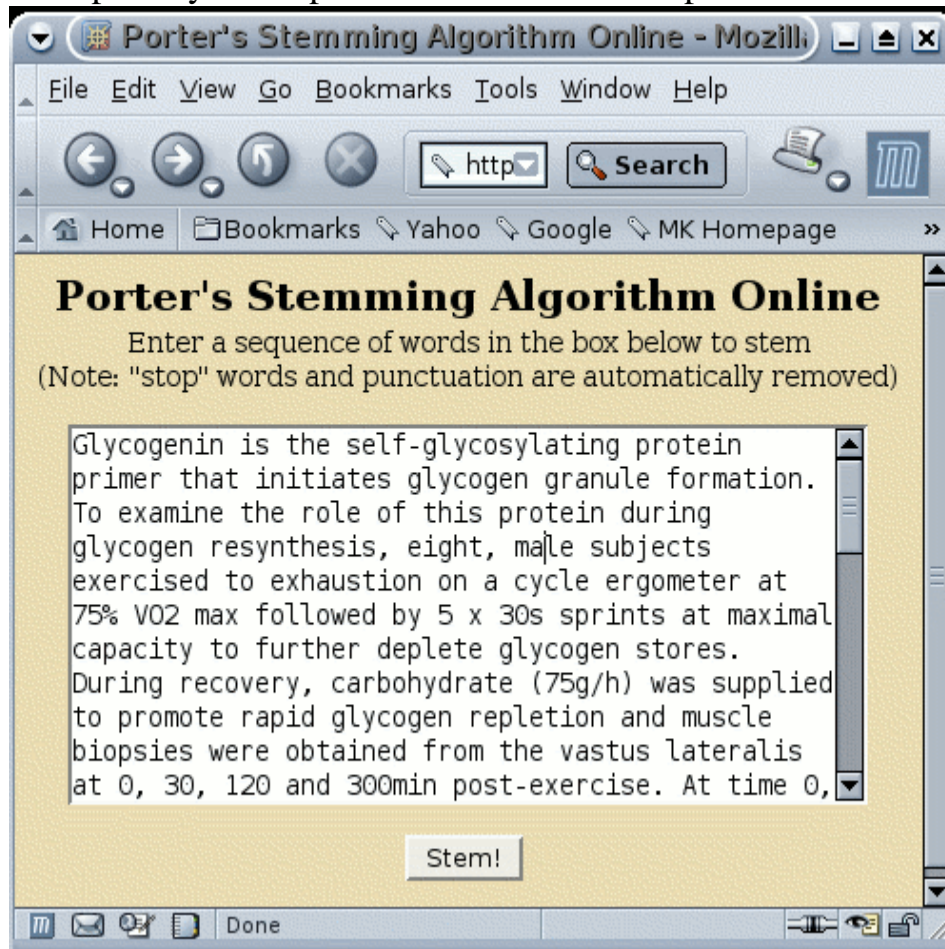
Information extraction and text mining

- 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

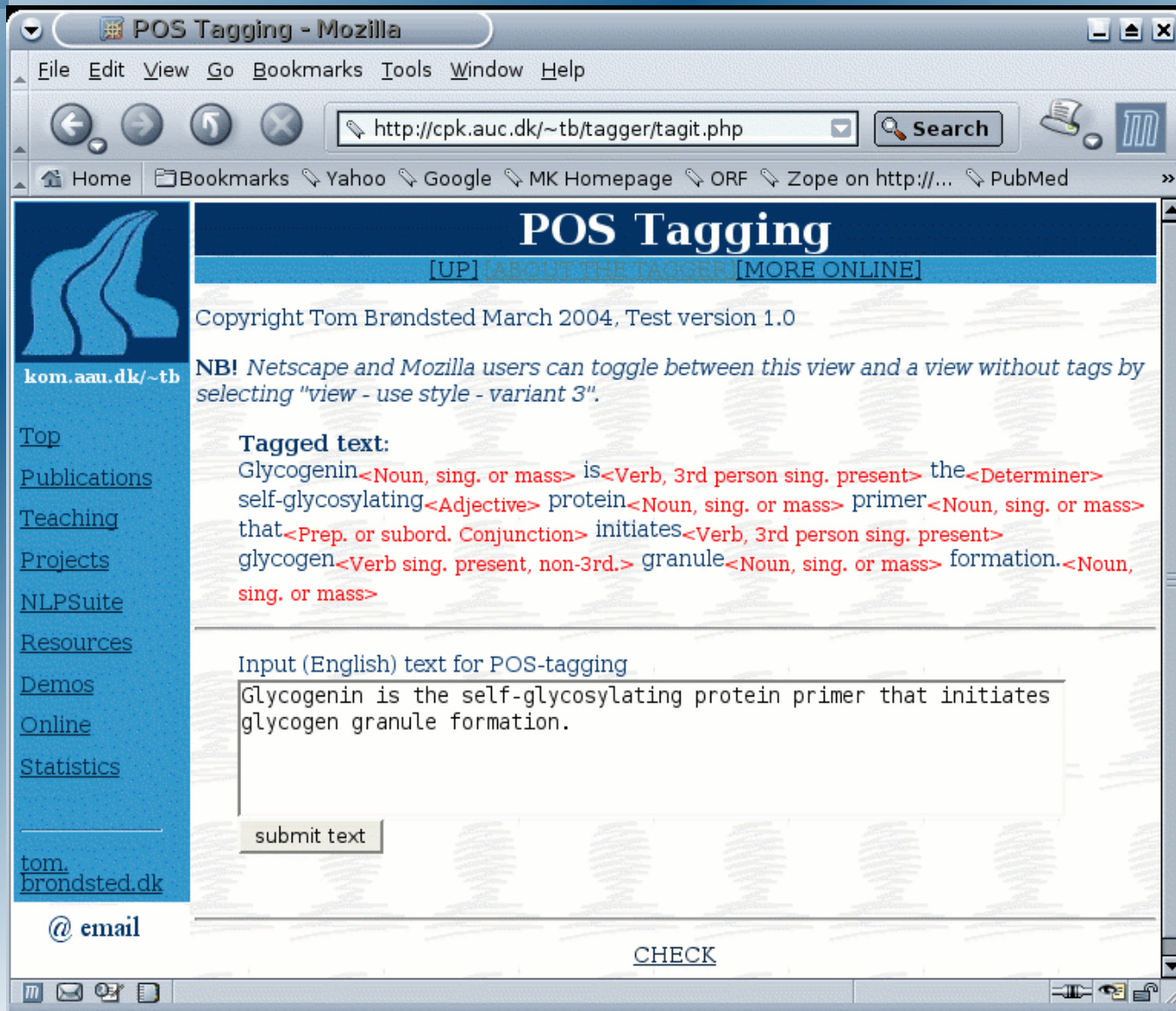
- Process of removing affixes of words transforming them to their corresponding morphological base form or root.

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



POS tagging

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



The screenshot shows a Mozilla browser window titled "POS Tagging - Mozilla". The address bar contains the URL "http://cpk.auc.dk/~tb/tagger/tagit.php". The browser's menu bar includes "File", "Edit", "View", "Go", "Bookmarks", "Tools", "Window", and "Help". The browser's toolbar shows navigation buttons (back, forward, home, stop) and a search box. The browser's status bar shows "Home", "Bookmarks", "Yahoo", "Google", "MK Homepage", "ORF", "Zope on http://...", and "PubMed".

The main content area of the browser displays the "POS Tagging" web application. The page has a blue header with the title "POS Tagging" and navigation links: "[UP]", "ABOUT THE TAGGER", and "[MORE ONLINE]". Below the header, the text reads: "Copyright Tom Brøndsted March 2004, Test version 1.0". A note states: "NB! Netscape and Mozilla users can toggle between this view and a view without tags by selecting 'view - use style - variant 3'".

The application shows a "Tagged text:" section with the following 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 subordin. Conjunction> initiates<Verb, 3rd person sing. present> glycogen<Verb sing. present, non-3rd.> granule<Noun, sing. or mass> formation.<Noun, sing. or mass>".

Below the tagged text is an "Input (English) text for POS-tagging" section with a text area containing the sentence: "Glycogenin is the self-glycosylating protein primer that initiates glycogen granule formation." and a "submit text" button.

The browser's status bar shows "CHECK".

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: 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.

Annotation of gene products: Gene Ontology

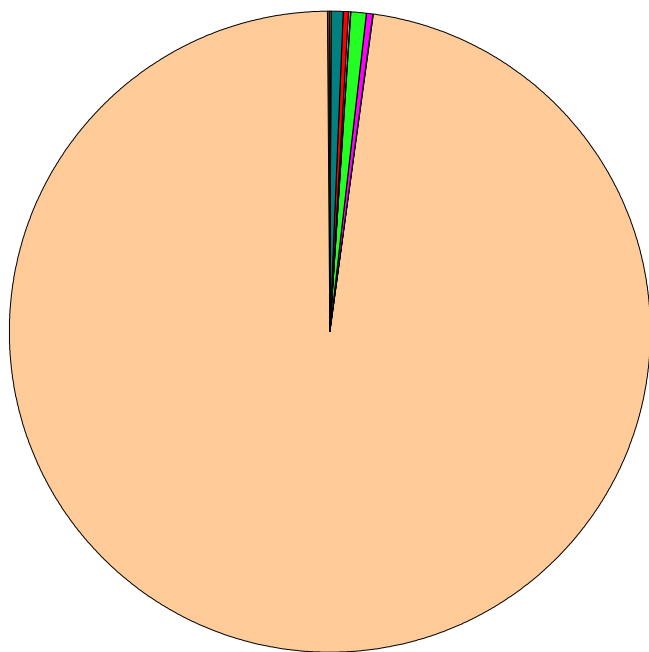
<http://www.geneontology.org/>

- 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.



Gene Ontology Annotation

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



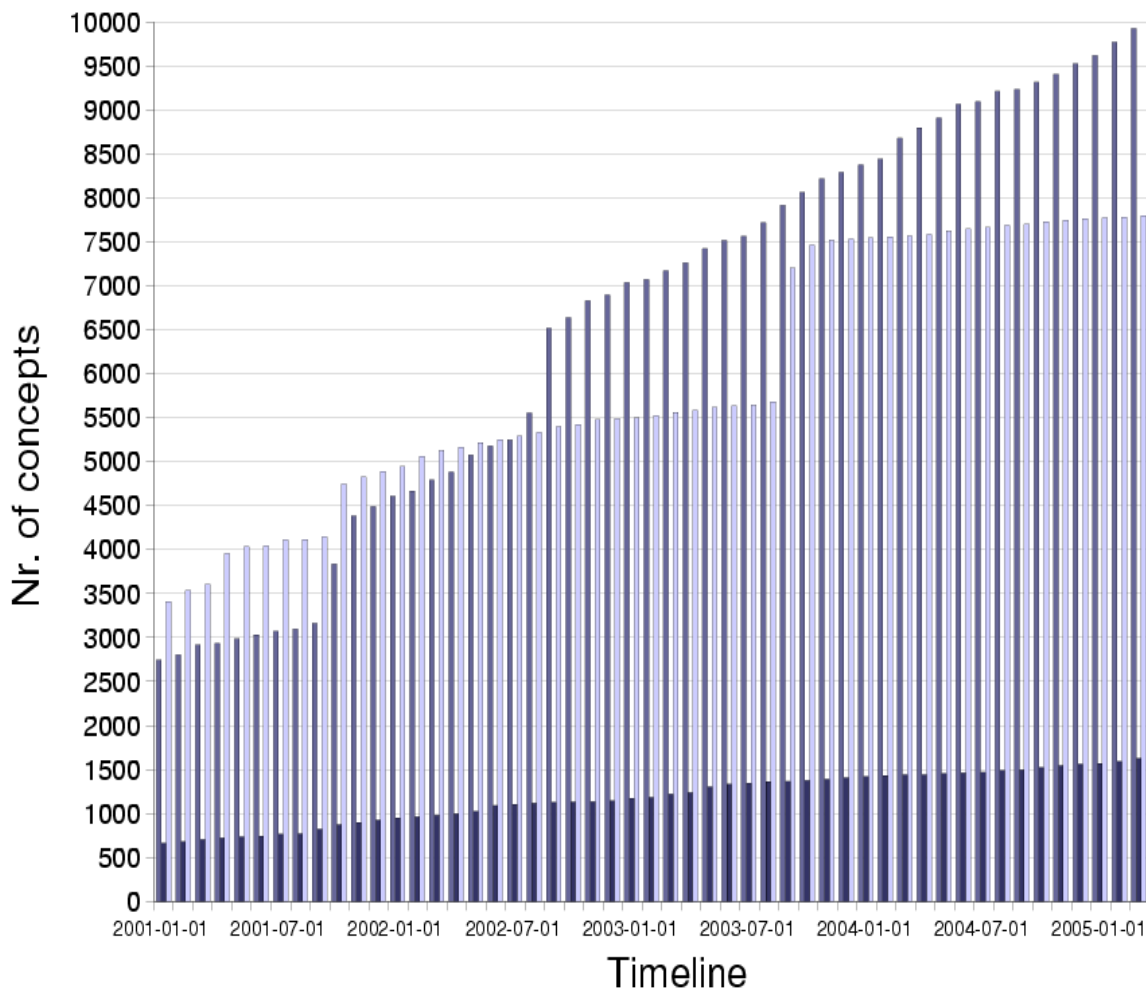
Ev.C.	Annot	Perc.	
IEA	6421817	0.97529	Electronic/ sequence- based annotation
ISS	19576	0.00297	
NR	2191	0.00033	
ND	4433	0.00067	
IPI	7130	0.00108	Experimental evidence
IGI	3014	0.00046	
IMP	19072	0.00290	
IDA	38862	0.00590	
IEP	1495	0.00023	Curator knowledge
IC	831	0.00013	
TAS	49630	0.00754	
NAS	16456	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 Growth

GO growth



- BP
- CC
- MF

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

Exercise 5: Gene Ontology

Gene Ontology (GO) aims to provide standardized concepts or terms to describe relevant biological aspects. Try to use [GO retrieve](#) the ontology sub-structure for a set of terms: apoptosis, caspase, glycogenin, transcription factor (or in case you are interested in some particular function/process/compartiment use your own query instead). What did you retrieve. Browse through the results and visualize the corresponding ontology graphs. What kind of relationships between terms did you find? What are the advantages of using such an ontology?

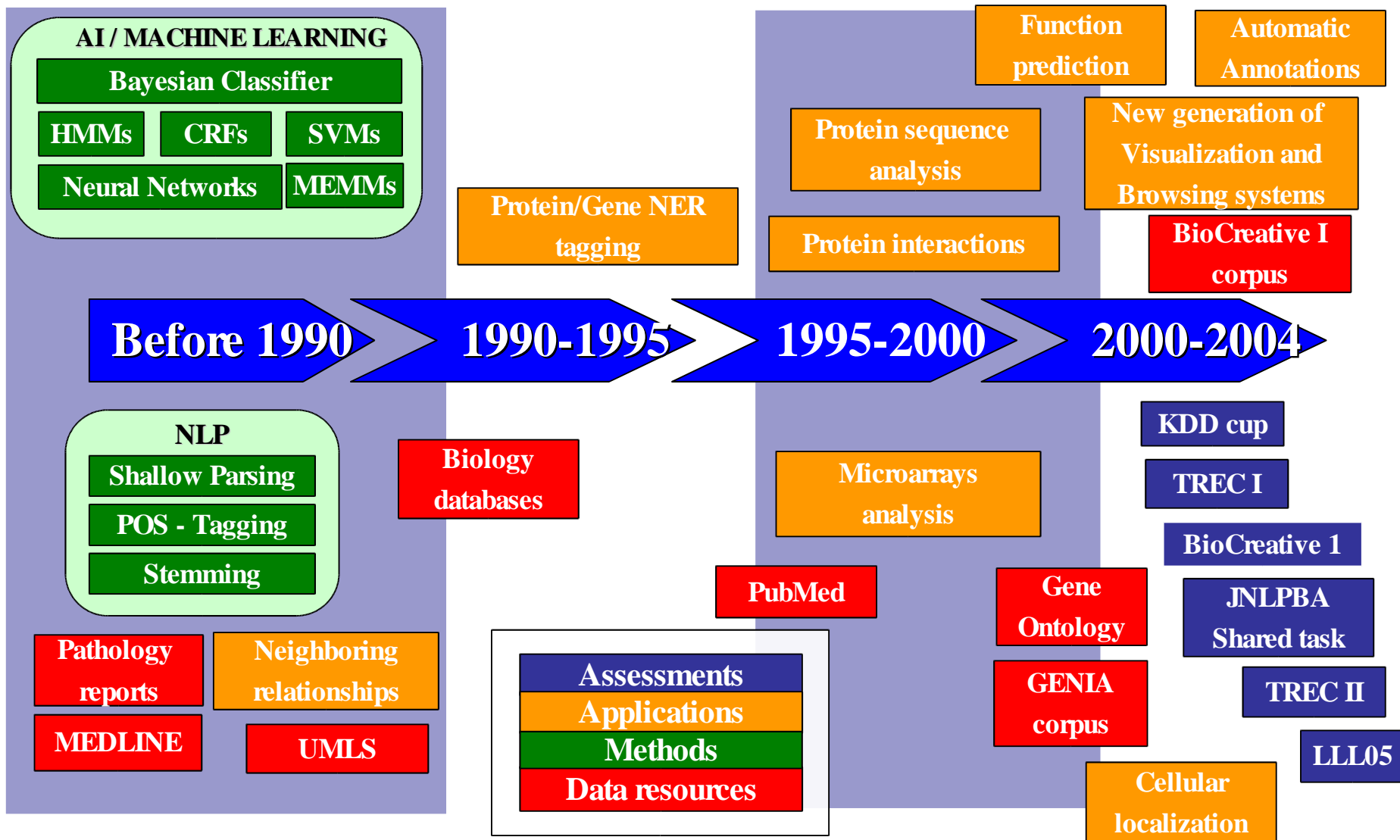
Try to explore annotation for a set of proteins, namely:

- 1) [CASP9_HUMAN](#) (P55211) (formerly known as ICE9_HUMAN),
- 2) [Y1333_MYCTU](#) (P64811) formerly known as YD33_MYCTU
- 3) [RPE_YEAST](#) (P46969)

by [Searching](#) the Gene Ontology Annotation database [GOA](#). Those proteins have been used in the practical part of the [Patrones, perfiles y dominios](#) session. What are one of the weak points when using GO annotations for bioinformatics annotations? (Hint: think about domains).

http://www.pdg.cnb.uam.es/martink/LINKS/tm_sc_ucm2005.htm

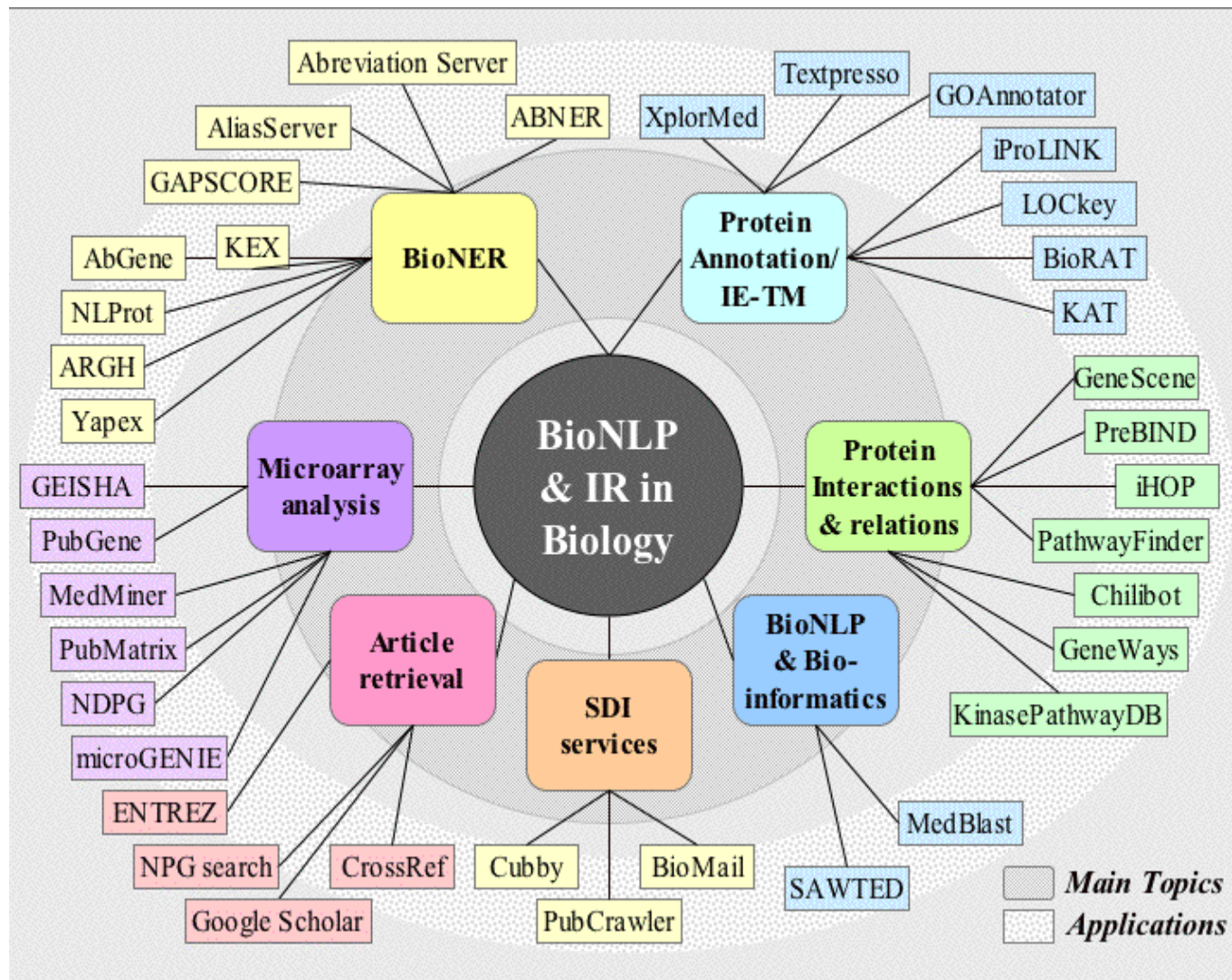
NLP in Molecular Biology - timeline



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.

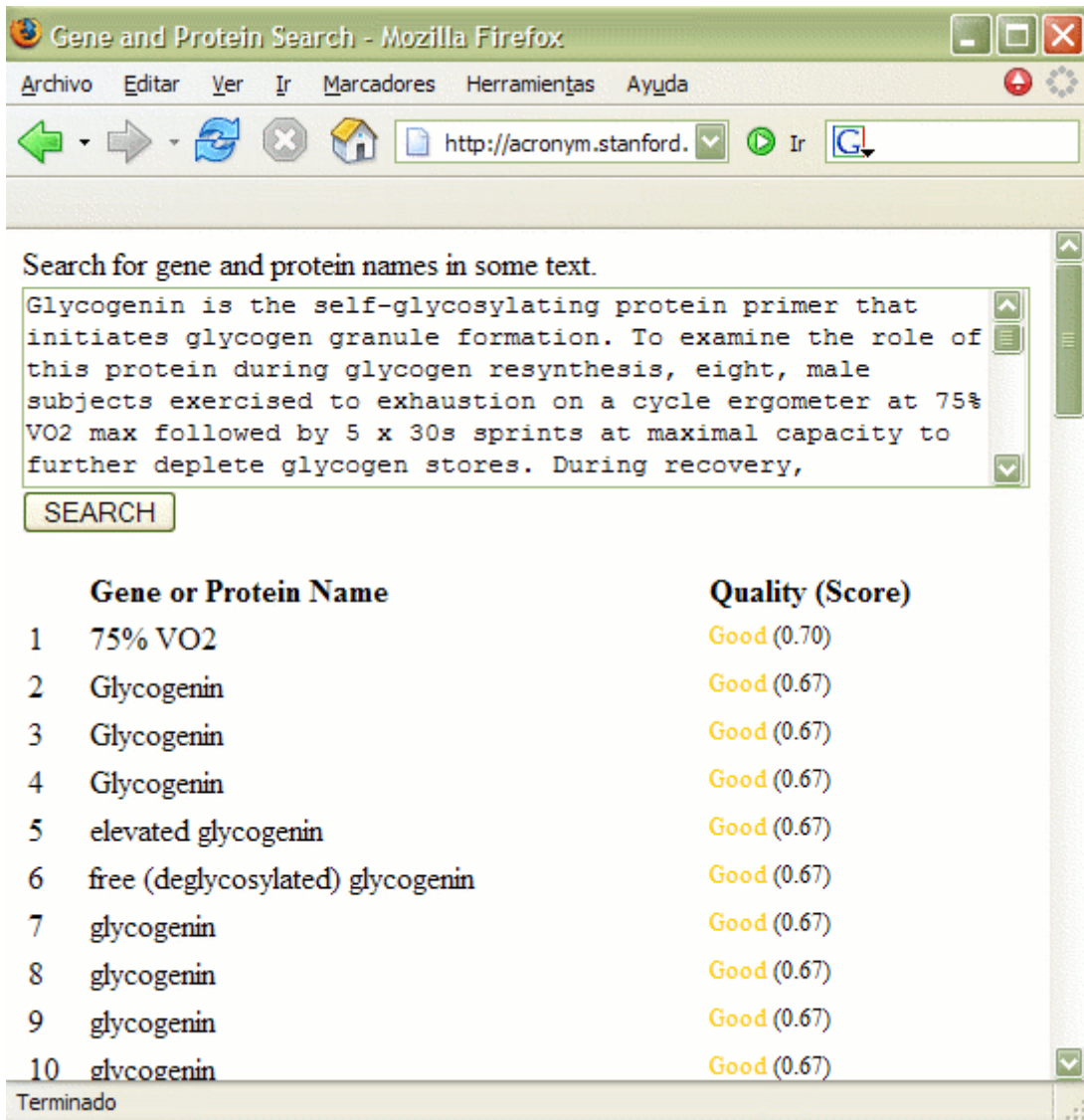
Text Mining applications in Biology



Tagging biological names

- 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



Gene and Protein Search - Mozilla Firefox

Archivo Editar Ver Ir Marcadores Herramientas Ayuda

http://acronym.stanford.

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,

SEARCH

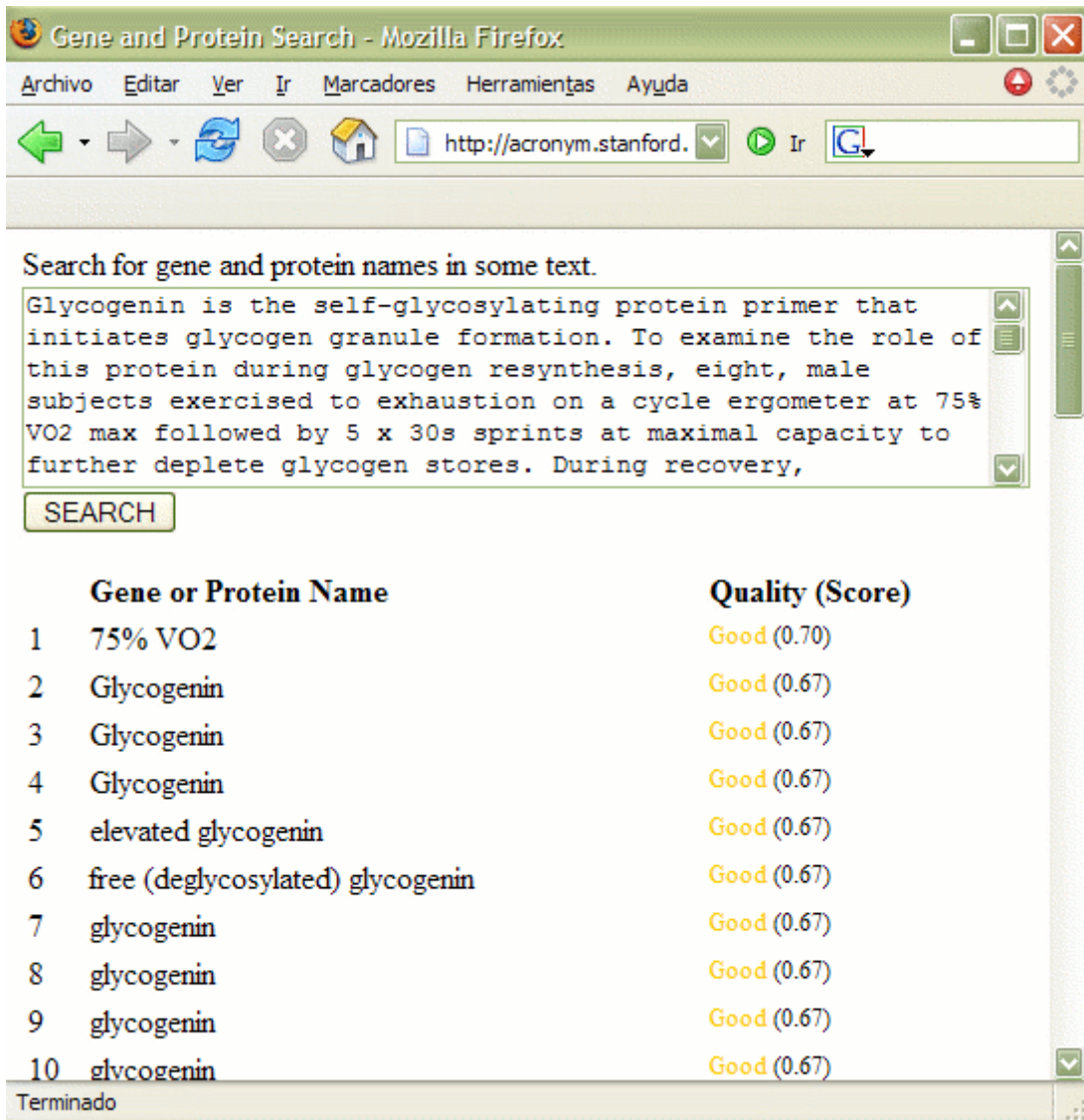
	Gene or Protein Name	Quality (Score)
1	75% VO2	Good (0.70)
2	Glycogenin	Good (0.67)
3	Glycogenin	Good (0.67)
4	Glycogenin	Good (0.67)
5	elevated glycogenin	Good (0.67)
6	free (deglycosylated) glycogenin	Good (0.67)
7	glycogenin	Good (0.67)
8	glycogenin	Good (0.67)
9	glycogenin	Good (0.67)
10	glvcogenin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

Terminado

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

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

GAPSCORE



Gene and Protein Search - Mozilla Firefox

Archivo Editar Ver Ir Marcadores Herramientas Ayuda

http://acronym.stanford.edu

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,

SEARCH

	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	glvcogenin	Good (0.67)

Terminado

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

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

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

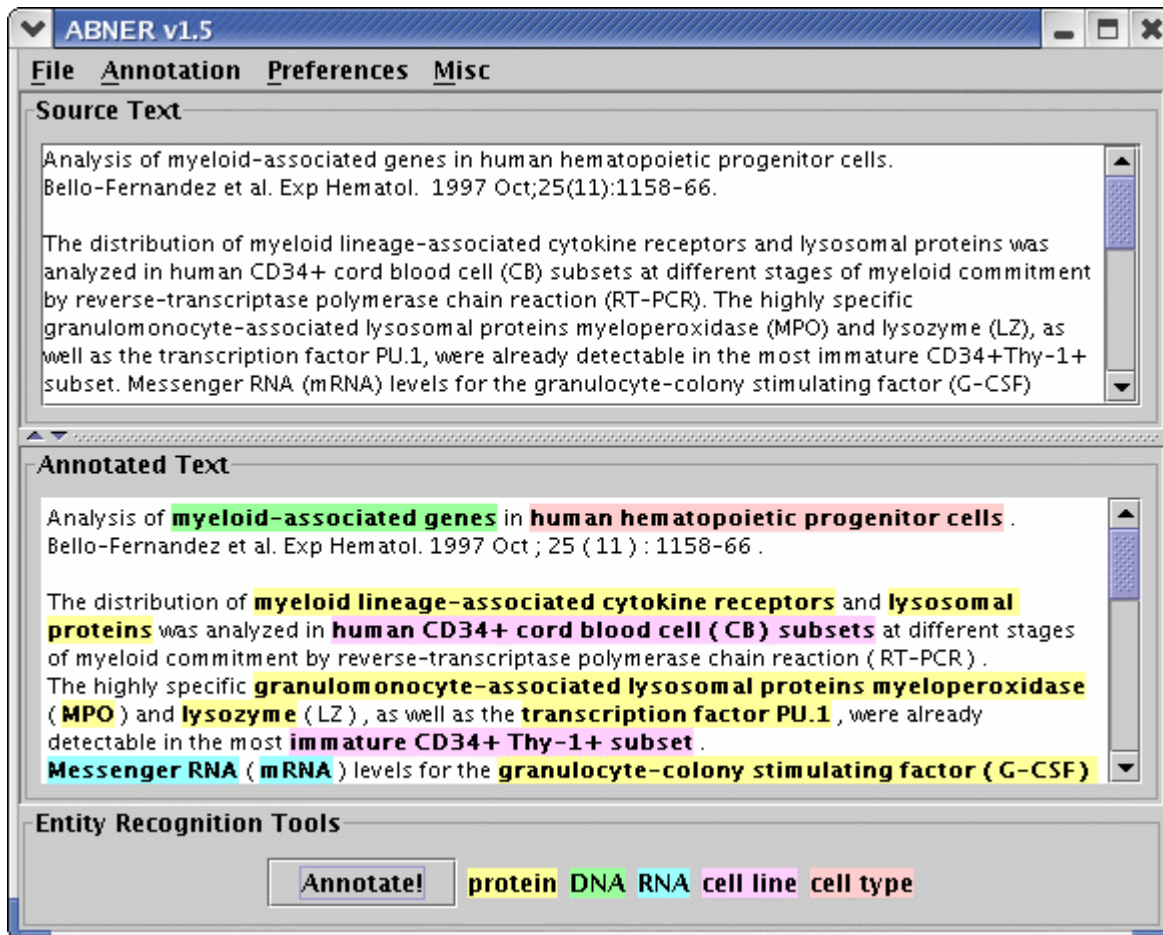
NAME	ORGANISM	TXT-POS	SCORE	METHOD	DB-ID(S)
Glycogenin	homo sapiens	1	1.040	SVM	GYG2 HUMAN (86%)
glycogenin	homo sapiens	96	0.856	SVM	GYG2 HUMAN (91%)
glycogenin	homo sapiens	103	1.040	SVM	GYG2 HUMAN (91%)
Glycogenin	homo sapiens	109	0.871	SVM	GYG2 HUMAN (86%)
glycogenin	homo sapiens	138	0.980	SVM	GYG2 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



- 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.

iHOP

iHOP - Information Hyperlinked over Proteins - Mozilla Firefox

Archivo Editar Ver Ir Marcadores Herramientas Ayuda

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

iHOP
information hyperlinked
Over proteins

Search Gene

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

Concept & Implementation

PHYSIOLOGY

INTERACTION

PATHOLOGY

PHENOTYPE

CD4

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

Search for a gene *synonym* or *accession number*...

Terminado

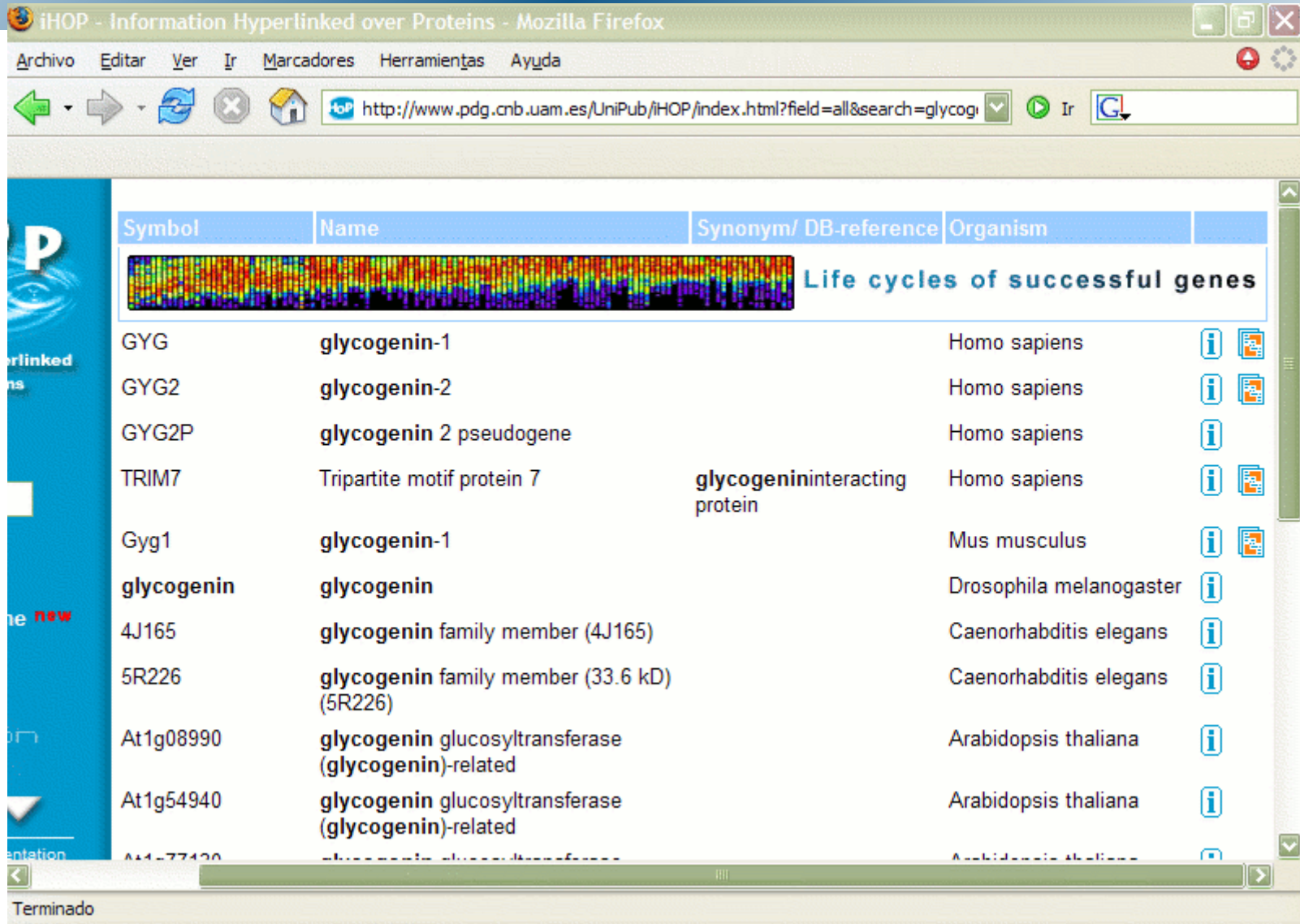
Hoffmann R, Valencia A. A gene network for navigating the literature *Nat Genet.* 2004 Jul;36(7):664.

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 Genet.* 2004 Jul;36(7):664.

iHOP



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	glycogenininteracting 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

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


iHOP

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

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



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

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 [glycoqen](#), [protein phosphatase inhibitor-1](#), and [glycogenin](#): lessons from a search for genetic variability of the insulin-stimulated [glycoqen](#) 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 [glycoqen](#) 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 37.2 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](#), [glycoqen](#) and [GLUT-4](#) protein content as well as GLUT-4, [glycoqen](#) 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.

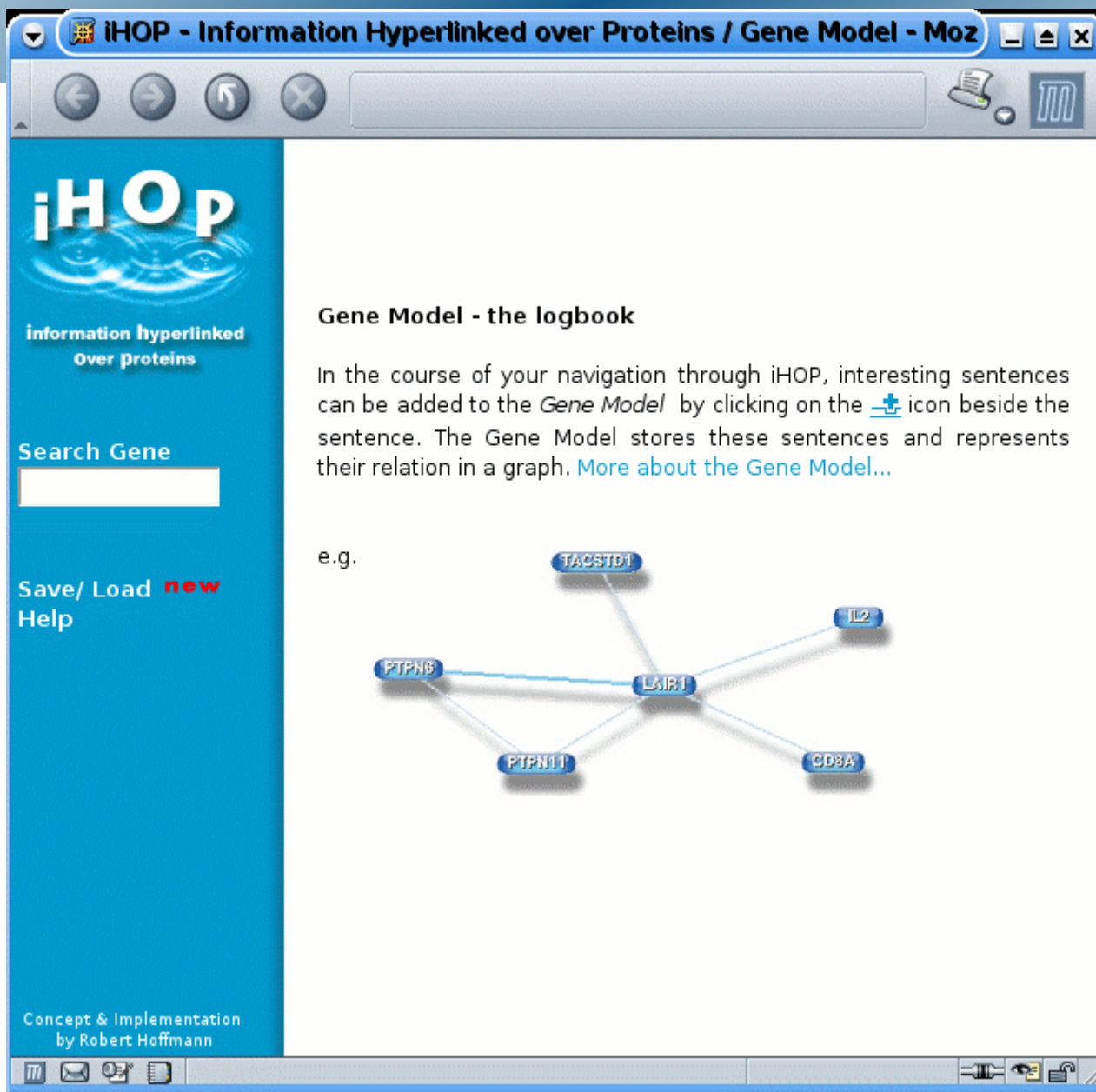
The discovery of a second human gene, [GYG2](#), encoding a liver-specific isoform of [glycogenin](#), the self-glucosylating initiator of [glycoqen](#) biosynthesis, raised the possibility for differential controls of this protein in [liver](#) and muscle.

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...

TEXT MINING (2005)

iHOP



The screenshot shows a Mozilla browser window titled "iHOP - Information Hyperlinked over Proteins / Gene Model - Moz". The interface features a blue sidebar on the left with the iHOP logo, the text "Information hyperlinked over proteins", a "Search Gene" input field, and "Save/ Load **new** Help" buttons. The main content area is titled "Gene Model - the logbook" and contains a paragraph explaining that interesting sentences can be added to the Gene Model by clicking on a plus icon. Below this text is a network graph with a central node labeled "LAI1" connected to five other nodes: "TACSTD1", "IL2", "CD3A", "PTPN11", and "PTPN6". The text "e.g." precedes the graph. At the bottom of the browser window, the text "Concept & Implementation by Robert Hoffmann" is visible.

iHOP:
Visualization
of protein
interactions
using network
graphs

EXERCISE 6: BIO-NER

Retrieve a given abstract from [PubMed](#) searching for genes of your own research interest or alternatively for some of the following genes gene names: Caspase-9 (CASP-9, APAF-3), RPE1 (EPI1, POS18), Orc-1, Bcl-2, glycogenin, p53. Then try to tag gene and protein names from some those abstracts using different gene/protein NER tools and compare the results. If you need GenBank ids (e.g. gi:20986531) or SwissProt accession numbers (Q07817 / BCLX_HUMAN) use: [NCBI](#) or [UniProt](#) Use some of the online applications [NLProt](#), [GAPSCORE](#), [Yapex](#) or [BioNE recognizer](#) (you can also download [ABNER](#)).

How do they perform? What are the common error? Which differences do you encounter? What are the main difficulties ? Which taggers do you think are useful in practice?

Explore for some of the gene symbols previously used (e.g. CASP-9, RPE1, Orc-1, glycogenin, p53) or for genes of your own research interest [iHOP](#). This tool was developed at our group (PDG) at the CNB. Create a gene model for your query gene, check the results carefully, and surf through the virtual gene network of iHOP. What kind of results are obtained by iHOP? What are the advantages/disadvantages when using iHOP instead of other bio-NER tools or the PubMed retrieval search?

http://www.pdg.cnb.uam.es/martink/LINKS/tm_sc_ucm2005.htm

P

EXERCISE 7: From sequence to abstracts

You have been using protein sequences for a range of analysis purposes in previous lectures of this course. Traditionally in case you want to obtain information related to a query sequence, after doing a sequence search (e.g. Blast against NCBI), retrieving the query genes, extracting their gene names or symbols and searching with those names PubMed you obtained the associated literature. This is a lot of work, with a lot of corresponding working steps. Those steps are integrated in the MedBlast tool.

Lets try to obtain the corresponding literature for some of the protein sequences used in other lectures (or your own query sequence of interest) for this exercise..

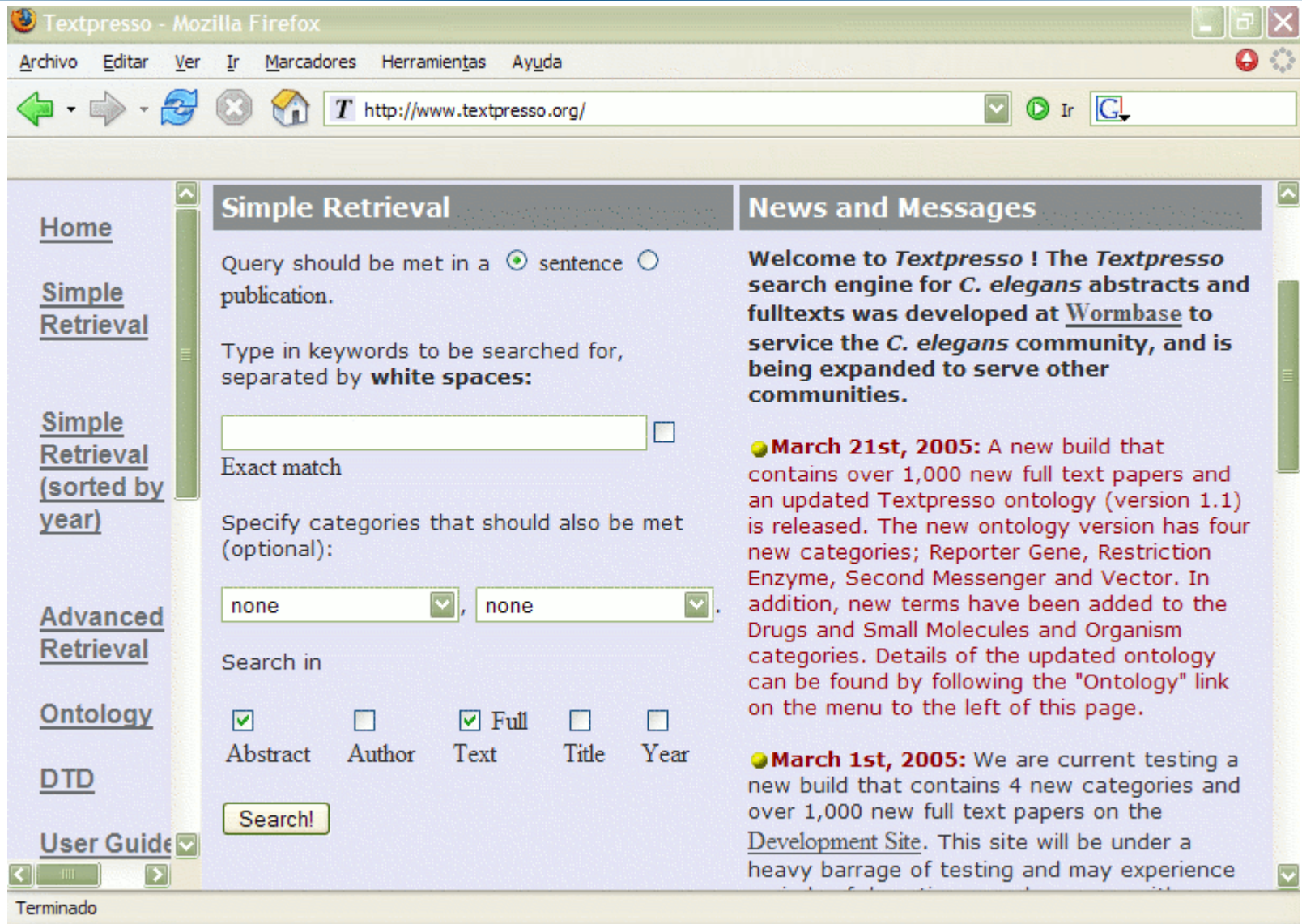
Use [MedBlast](#), a NLP based retrieval system to return relevant articles for your sequence. Notice that this system is low and sensitive to server overload! Describe the obtained results. What are the main difficulties when linking a query sequence to scientific articles?

http://www.pdg.cnb.uam.es/martink/LINKS/tm_sc_ucm2005.htm

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.

Textpresso



Textpresso - Mozilla Firefox

Archivo Editor Ver Ir Marcadores Herramientas Ayuda

http://www.textpresso.org/

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):

,

Search in

Abstract Author Text Title Year

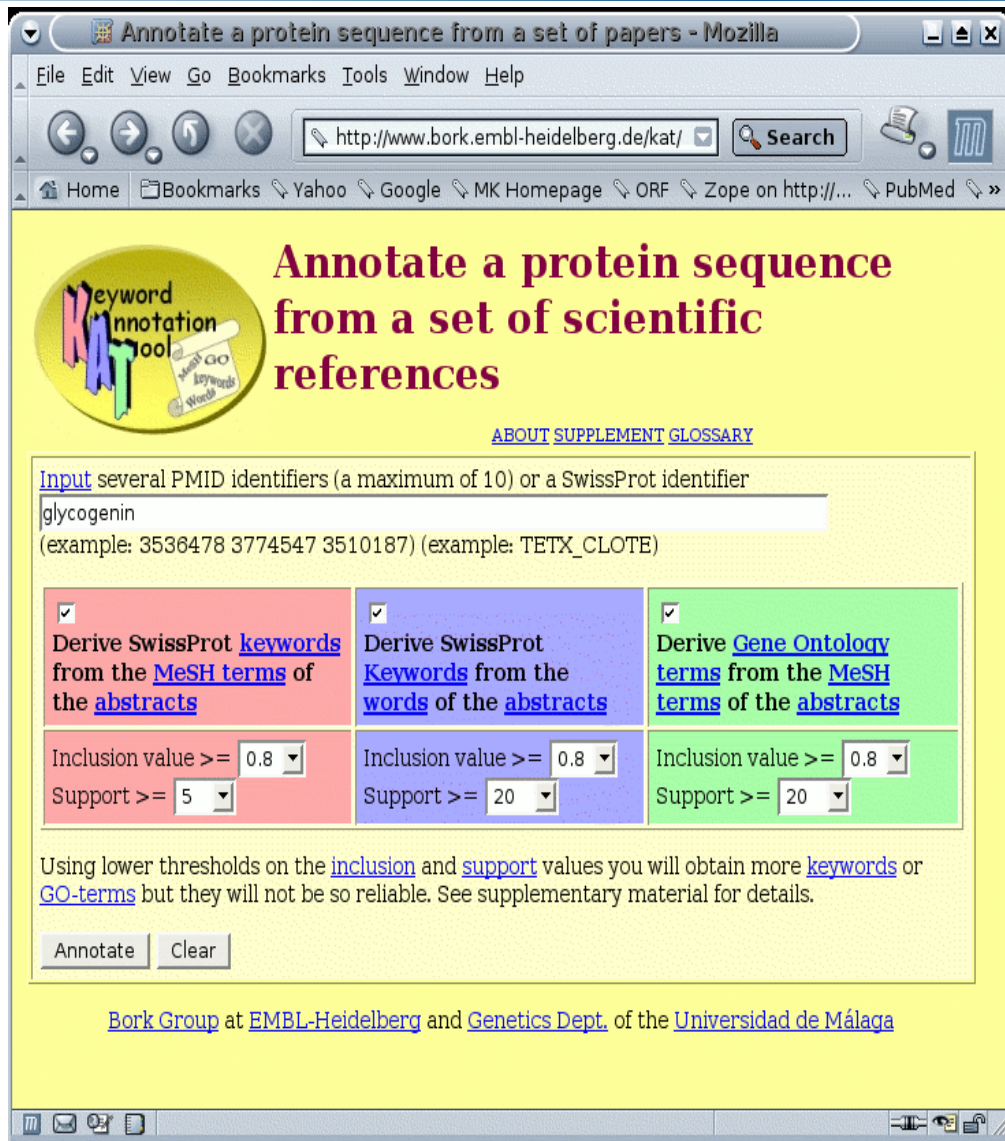
News and Messages

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)



Input several PMID identifiers (a maximum of 10) or a SwissProt identifier
glycogenin
(example: 3536478 3774547 3510187) (example: TETX_CLOTE)

<input checked="" type="checkbox"/> Derive SwissProt keywords from the MeSH terms of the abstracts	<input checked="" type="checkbox"/> Derive SwissProt Keywords from the words of the abstracts	<input checked="" type="checkbox"/> 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.

Buttons: 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.

Suppl. EXERCISE 8: PROTEIN FUNCTION

- The functional annotations contained in databases such as Gene
- Ontology annotation (GOA) was directly or indirectly extracted from the literature.
- Several applications have been developed to associate proteins with functional terms.
- Try to use text mining applications and GOA annotations to find functional information for your query proteins:
 - GOAnnotator
 - iHOP
 - GOA

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, NMR/FRET, chemical cross-linking, ..
- Bioinformatics methods for protein characterization:
 - Genome-based
 - Sequence-based

PROTEIN INTERACTION DATABASES

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

TEXT MINING & 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.

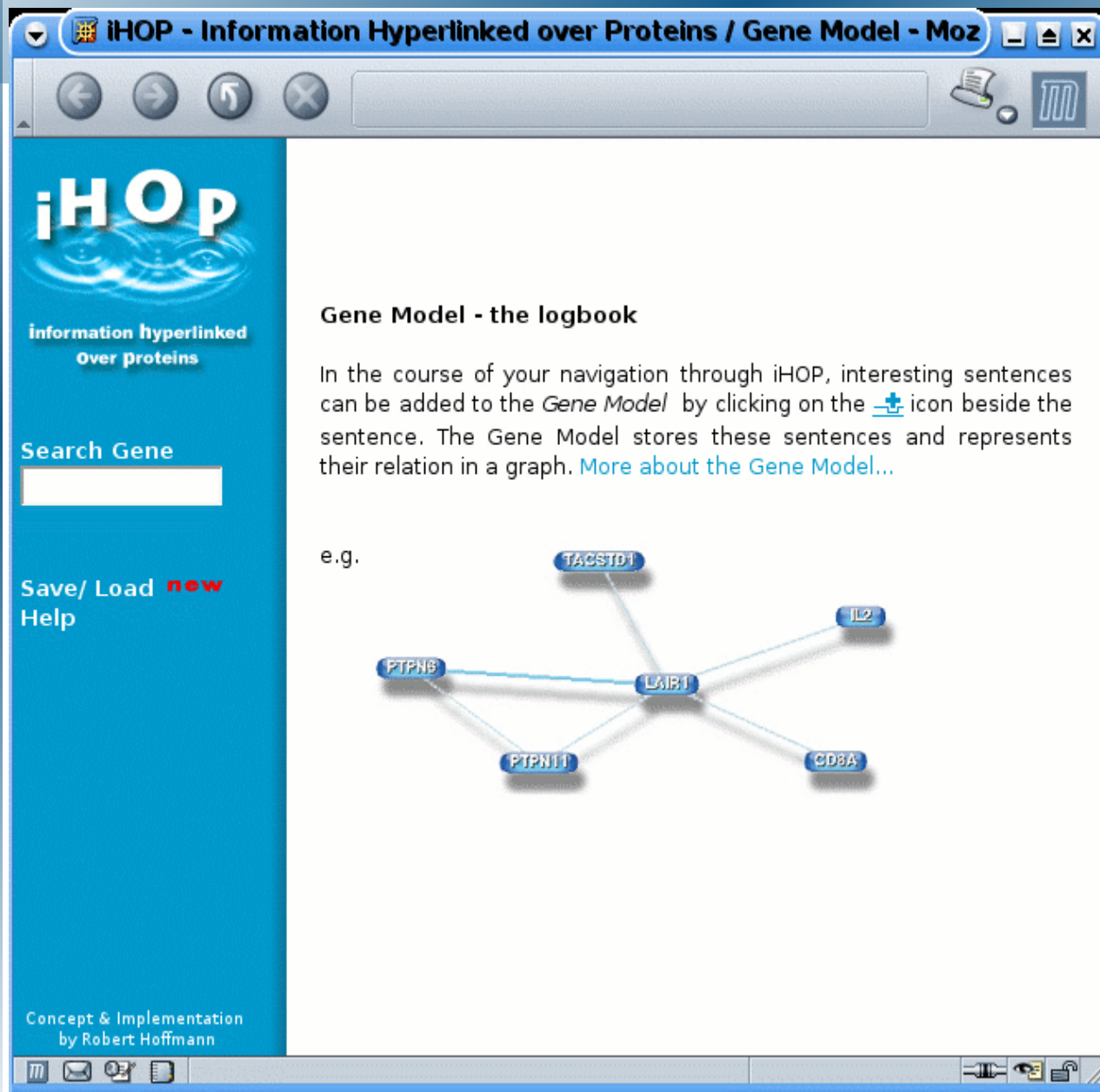
PubGene

- 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



iHOP - Information Hyperlinked over Proteins / Gene Model - Moz

iHOP
information hyperlinked over proteins

Search Gene


Save/ Load **now**
Help

Concept & Implementation
by Robert Hoffmann

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.



```

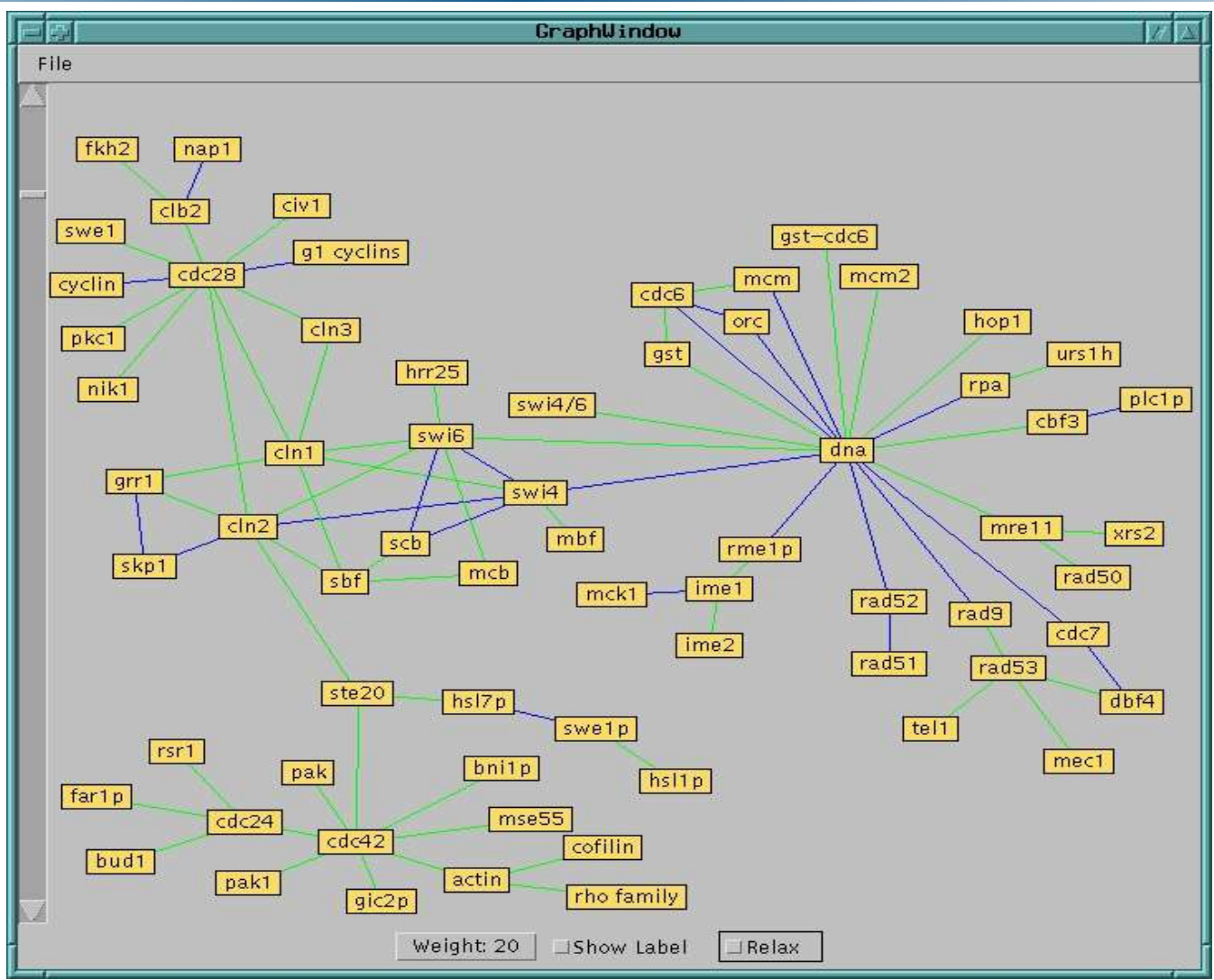
graph TD
    LAR1 --- TACSTD1
    LAR1 --- IL2
    LAR1 --- CD3A
    LAR1 --- PTPN11
    LAR1 --- PTPN6
    
```

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



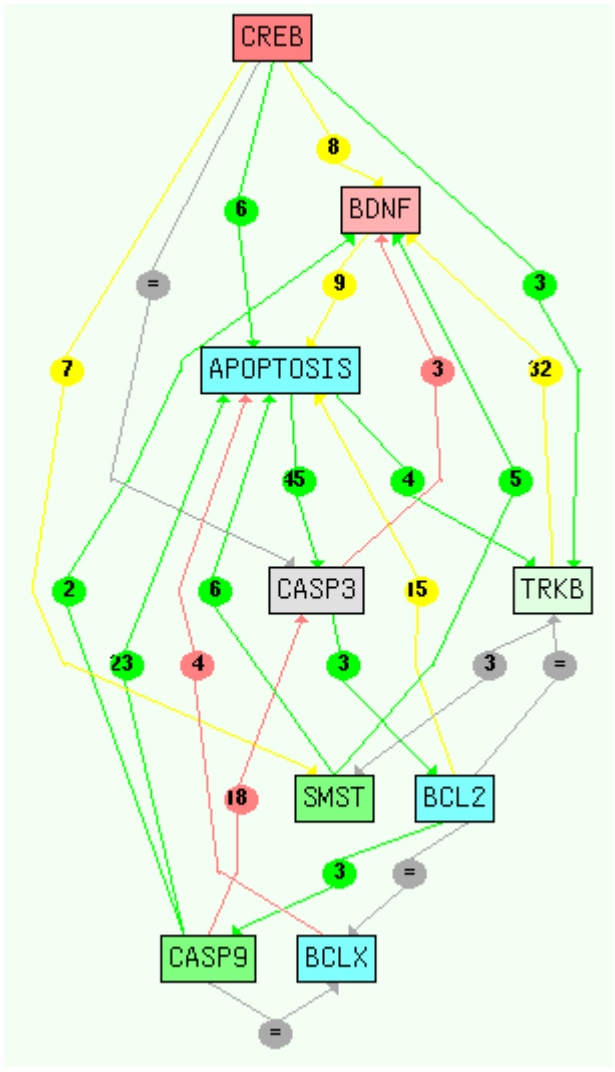
CHILIBOT

- 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/>

CHILBOT

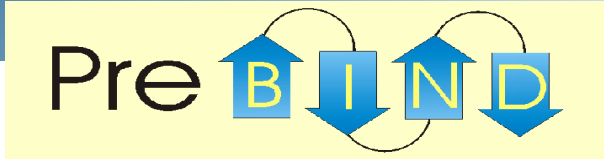


- Need registration.
- Hypothesis generation.

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/>

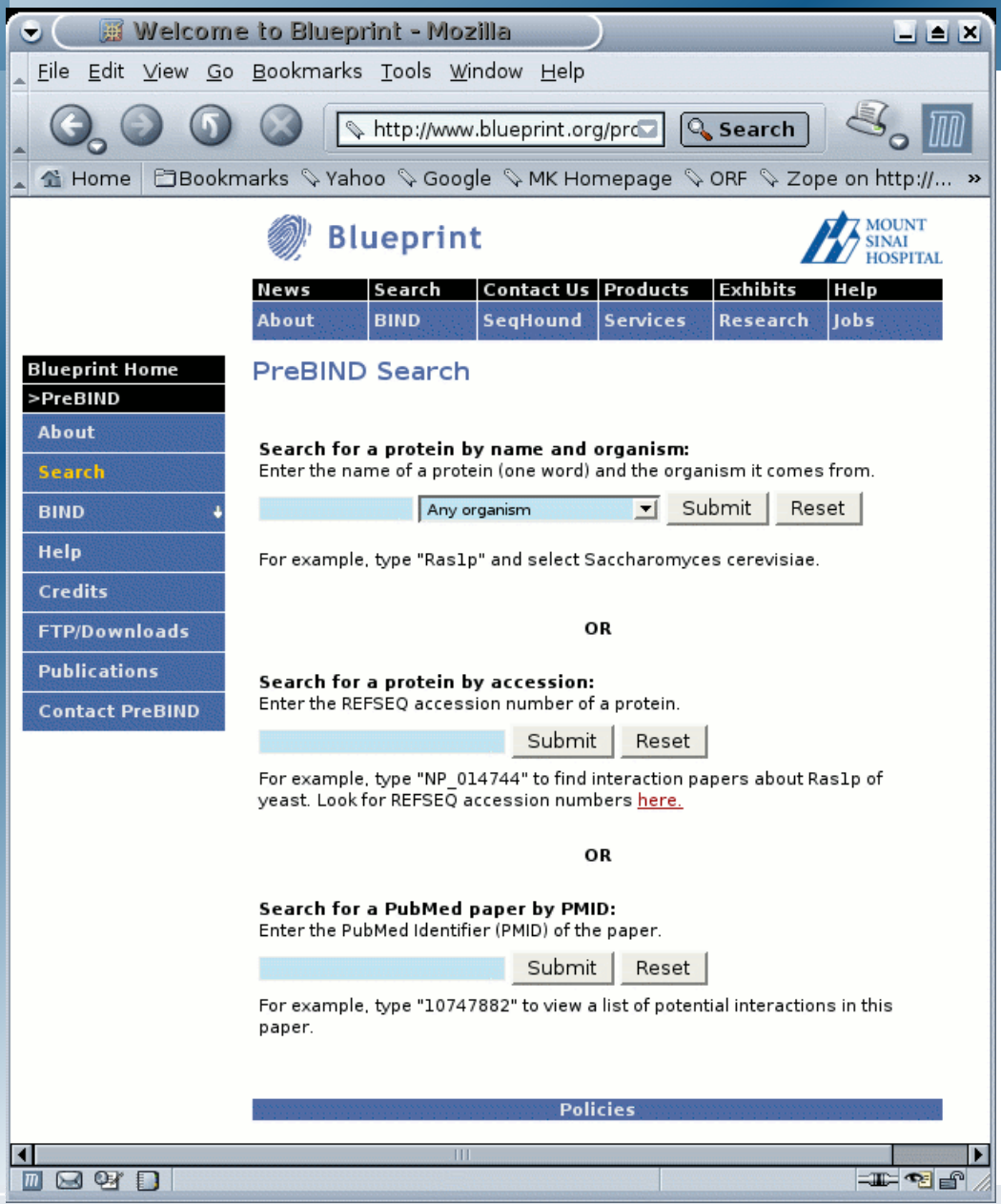
PreBIND



- 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>



PreBIND Search

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

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

OR

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

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.

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

[Policies](#)

EXERCISE 6.4.: PROTEIN INTERACTIONS

- Proteins instantiate their function through interactions with other bio-molecules.
- Use different text mining tools which try to extract protein interactions for a given query protein/s (caspase, glycogenin, p53 etc...) from texts: iHOP, PreBIND, Chilibot.
- Compare your results with entries in interaction databases:
- BIND, DIP , GRID , HPID, HPRD, IntAct, MINT and STRING.
- What kind of output is produced by each tool?
- Which differences do you encounter?
- What are the difficulties encountered by those tools?
-

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.

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.

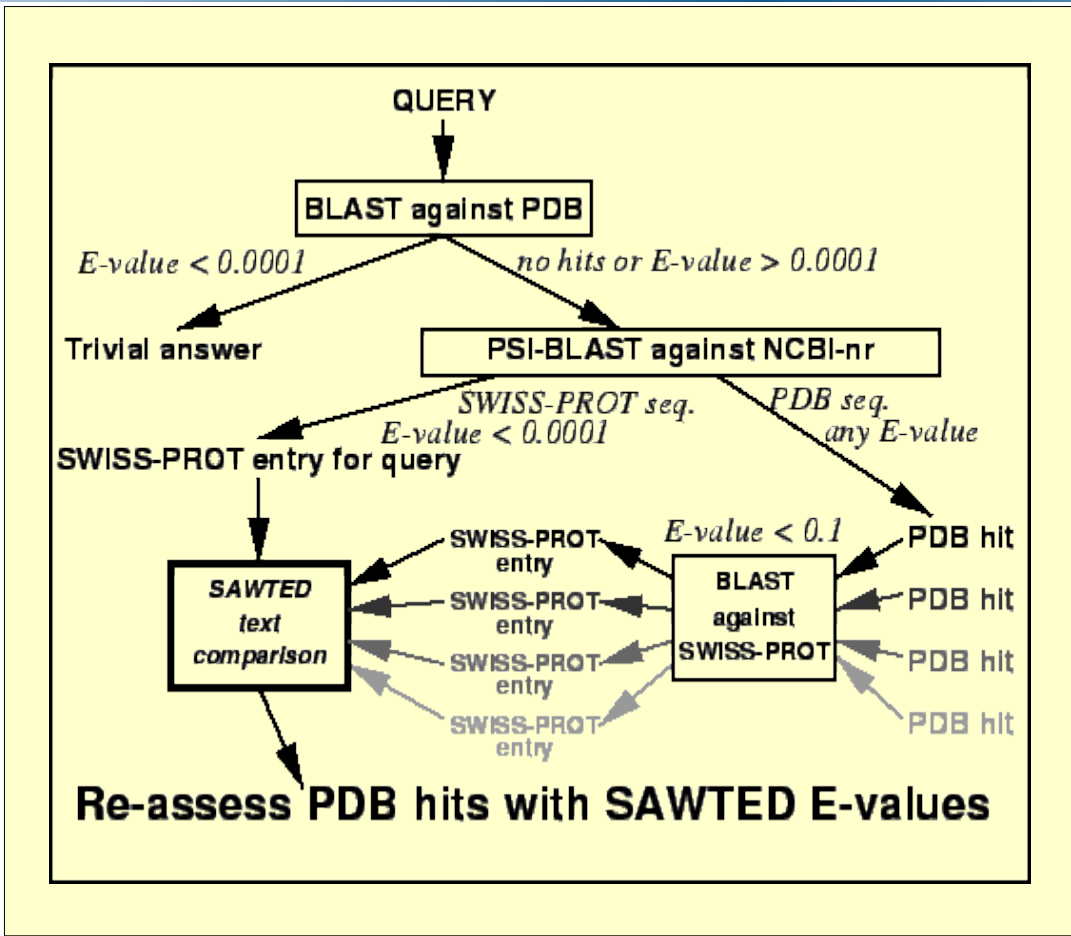
PROTEIN LOCALIZATION

- 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

NLP AND SEQUENCE ANALYSIS: SAWTED



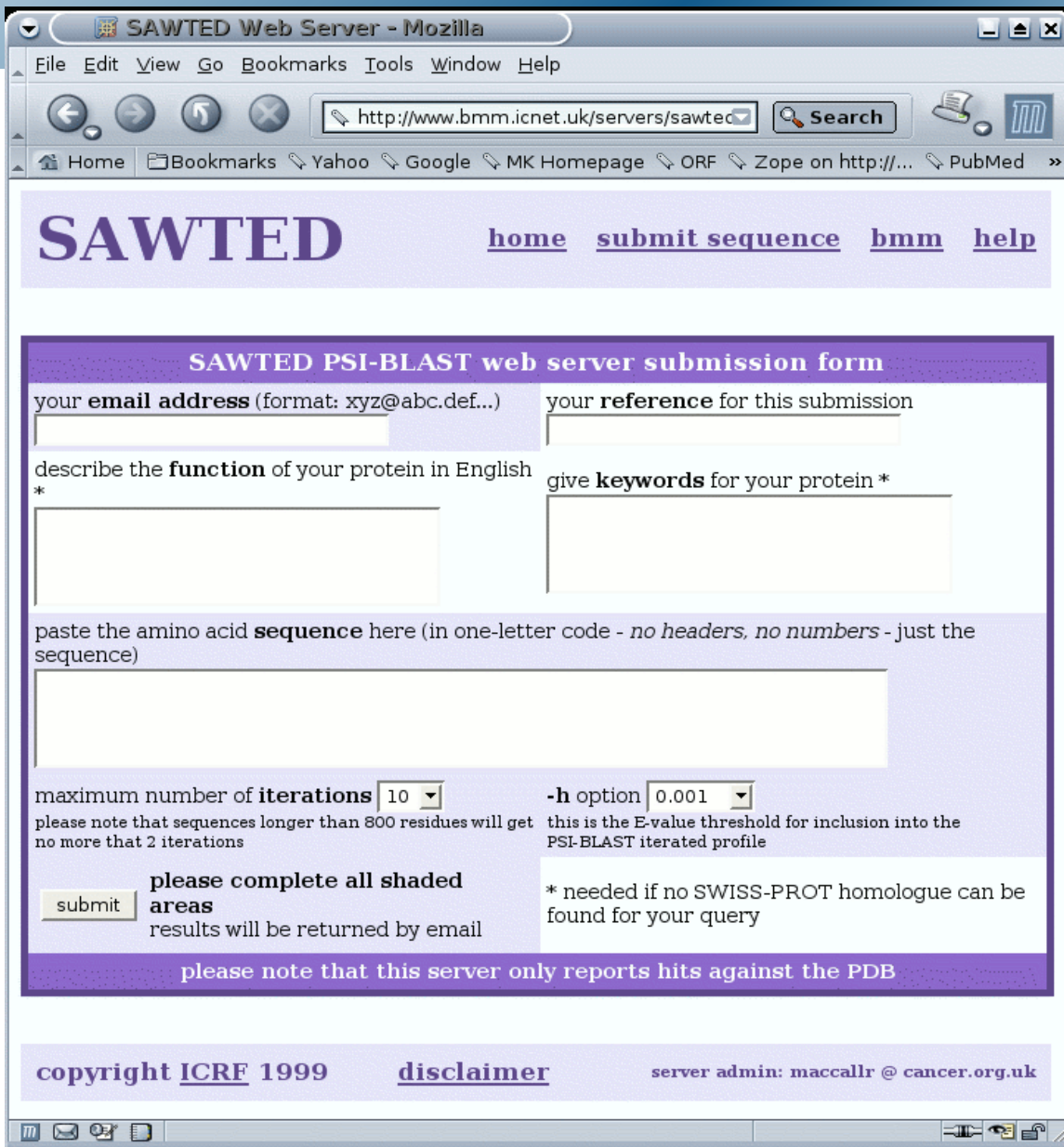
Sequence similarity
the base for identifying
structure templates
for query sequence

Structure Assignment
With Text Description

Document comparison
algorithms

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

NLP AND SEQUENCE ANALYSIS: 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

SAWTED [home](#) [submit sequence](#) [bmm](#) [help](#)

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** **-h option**

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

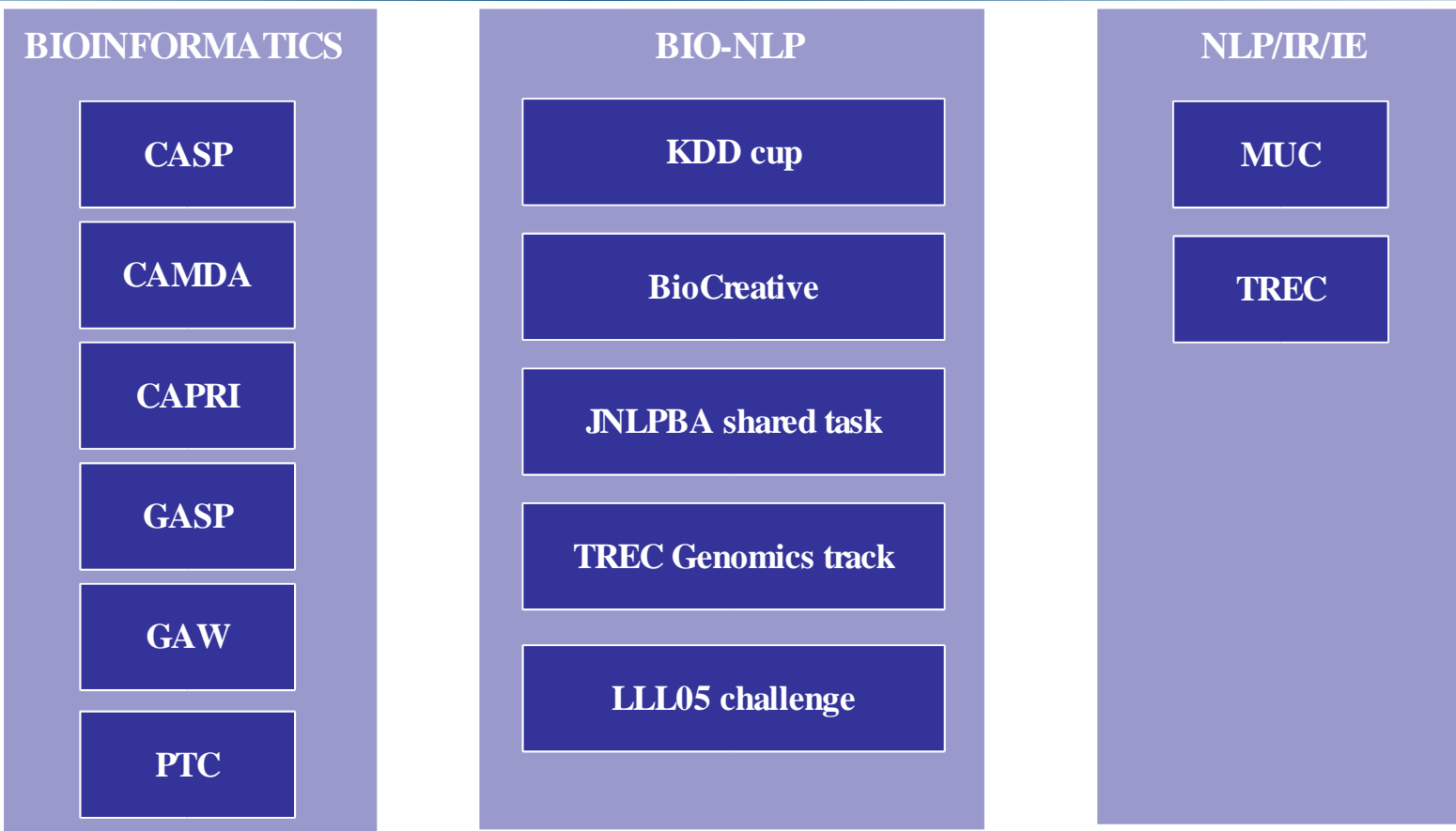
please note that this server only reports hits against the PDB

copyright [ICRF](#) 1999 [disclaimer](#) server admin: maccallr @ cancer.org.uk

Use information contained in text descriptions of SwissProt annotations

identification of remote homologues

COMMUNITY WIDE 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 Biomedicine
TREC: Text Retrieval conference
MUC: Message Understanding conference
LLL05: Genic interaction extraction challenge

CONCLUSIONS AND OUTLOOK

- BIO-NLP VERY RECENT DISCIPLINE (MAINLY 2003-TODAY).
- GROWING INTEREST
- NEW TECHNIQUES AND DATASETS
- NEED OF USER FEEDBACK AND INTERACTIVE LEARNING

SELECTED REVIEW REFERENCES

- M. Krallinger and A. Valencia. text mining and information retrieval services for Molecular Biology. *Genome Biology*, 6 (7), 224 (2005)
- 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 Erhardt 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

I would like to thank Alfonso Valencia for his supervisions and suggestions, the Protein Design Group at the National Biotechnology Centre (CNB) for interesting discussions.