Text Mining Biomedical Repositories

Ashish Vijay Tendulkar

School of Technology and Computer Science,
Tata Institute of Fundamental Research, Mumbai.
http://www.tifr.res.in/∼ashishvt/

Dec 20, 2011

Acknowledgement

The tutorial slides were jointly prepared with Martin Krallinger of Spanish National Cancer Research Center, Madrid.

Table of Contents

1. General Background
2. BioText Mining
3. Applications of BioText Mining
4. Evaluation of Text Mining Systems
5. Practical Case Studies

Biology 101

Biological Entities

- Genes, Genome
- mRNA, transcriptome
- Protein, Proteome
- Cell
- Tissues
- Organisms

Biomedical Literature

Generation and Form

- Biologists conduct experiments and generate heterogeneous data types such as sequence, structure, expression etc.
- A large fraction in form of Natural Language (NL) report of experimental findings in form of research papers, reports, patents, newswire articles
- Structured database records (in relatively less proportion than NL)(e.g. Genes/Protein sequence, structure etc.)

Biomedical Literature

Generation and Form

- Biologists conduct experiments and generate heterogeneous data types such as sequence, structure, expression etc.
- A large fraction in form of Natural Language (NL) report of experimental findings in form of research papers, reports, patents, newswire articles
- Structured database records (in relatively less proportion than NL)(e.g. Genes/Protein sequence, structure etc.)

Importance

- Curation of structured databases (UniProt)
- Deriving functional annotations beyond what is present in DBs.
- Contextual information about experimental results and conditions(Cell lines, tissues, etc.)

Literature and Scientific Discovery Process

Biology

- Define the biological question
- Select the actual target being studied
- Extract information relevant for experimental set up
- Locate relevant resources
- Essential to understand and interpret the resulting data
- Draw conclusions about new discoveries
- Communicated to the scientific community using publications in peer-reviewed journals

Clinics

- Resource for clinical decision support in evidence-based clinical practice
- Useful information for diagnostic aids

Literature and Scientific Discovery Process

Pharma

- Drug discovery and target selection
- Identifying adverse drug effect
- Competitive intelligence and knowledge management

Funding

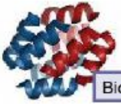
- Global view of the current research state and monitor trends to ensure optimal resource allocation

Publishing Groups

- Find domain experts for specific topics for the peer-review process and detecting potential cases of plagiarism

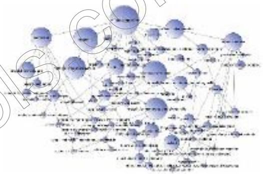
Relevance of Literature in Bioinformatics

Scientific Literature

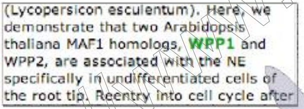


Bio-entities

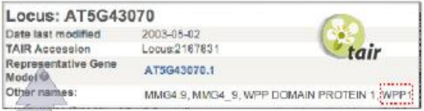
Database curator



Controlled vocabularies



(Lycopersicon esculentum). Here, we demonstrate that two Arabidopsis thaliana MAF1 homologs, WPP1 and WPP2, are associated with the NE specifically in undifferentiated cells of the root tip. Reentry into cell cycle after



Locus: AT5G43070

Date last modified	2003-05-02
TAIR Accession	Locus:2167831
Representative Gene Model	AT5G43070.1
Other names:	MMG4 9, MMG4_9, WPP DOMAIN PROTEIN 1, WPP1

tair

tair FlyBase MGI GDB SGD swissprot

MaizeGDB Maize Genetics and Genomics Database RGD EcoCyc MINT

Challenges in Exploring Biomedical Literature

Rapid growth of literature data poses following challenges:

- Efficient methods for extraction of information
- Effective ways of querying the information

Curation of Biological Databases from Literature

Classical Method: Manual Curation

- Trained human experts reads scientific literature and extracts information of interest
- Manual time consuming and labor intensive process
- Accurate through human inference and background knowledge
- Example DBs: Uniprot, GOA, SGD, MGI etc.

Text Mining assisted Curation

- Retrieval of relevant literature from literature repositories
- Textual evidence and entity detection
- Revision and editing of manual records
- E.g. TextPresso, Rodriguez-Penagos et al (gene regulation), Grover el at (PPI), Chang et al (Pathways), Ongenaert et al (methylation), Shtatland (peptides), Miotto (allergen cross-reactivity).

Overview of Current Literature Repositories

- e-Books: NCBI Bookshelf
- Citation of Biomedical Research Articles + Abstract: PubMed
- Full text research articles:
  - PubMed Central (PMC)
  - Highwire Press
  - BioMed Central

PubMed

Overview

- Developed by NCBI
- Citation entries of scientific articles of all biomedical sciences
- Each entry is characterized by a unique identifier, the PubMed identifier: PMID
- Often links to the full text articles are displayed

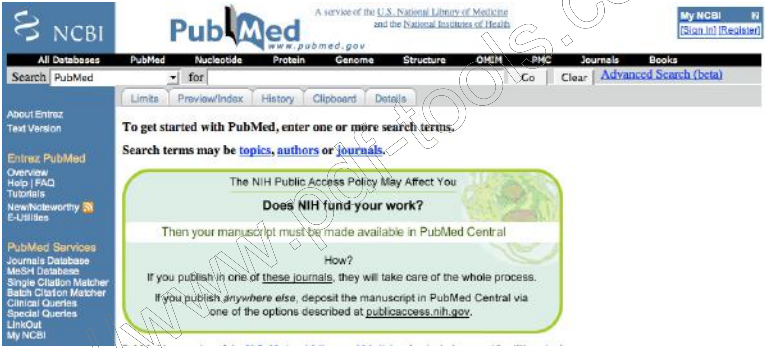
Statistics

No. of Citations	16 million
No. of Indexed Journals	approx. 5000
No. of English Articles	12 million
No. of Articles with Abstracts	7,000,000

Importance of PubMed in Biomedical Text Mining

- Approximately 1 million entries refer to gene descriptions
- Author, journal and title information of the publication
- Some records with gene symbols and molecular sequence databank numbers
- Indexed with Medical Subject Headings (MeSH)
- Accessed online through a text-based search query system called Entrez
- Offers additional programming utilities, the Entrez Programming Utilities (eUtils)
- NLM also leases the content of the PubMed/ Medline database on a yearly basis

Entrez



To get started with PubMed, enter one or more search terms.

Search terms may be topics, authors or journals.

The NIH Public Access Policy May Affect You

Does NIH fund your work?

Then your manuscript must be made available in PubMed Central

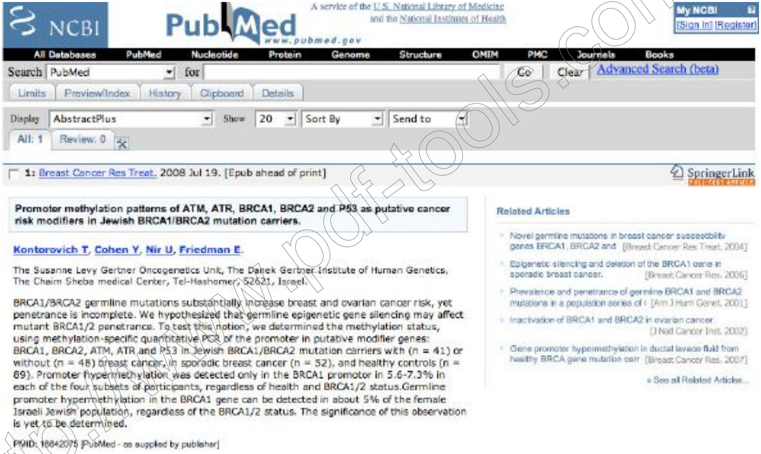
How?

If you publish in one of these journals, they will take care of the whole process.

If you publish *anywhere else*, deposit the manuscript in PubMed Central via one of the options described at publicaccess.nih.gov.

Figure: Source http://www.ncbi.nlm.nih.gov/pubmed/

PubMed Search Results



□ 1: Breast Cancer Res Treat. 2008 Jul 19. [Epub ahead of print]

Promoter methylation patterns of ATM, ATR, BRCA1, BRCA2 and P53 as putative cancer risk modifiers in Jewish BRCA1/BRCA2 mutation carriers.

Kontorovich T, Cohen Y, Nir U, Friedman E.

The Susanne Levy Gertner Oncogenetics Unit, The Danek Gertner Institute of Human Genetics, The Chaim Sheba medical Center, Tel-Hashomer, 52621, Israel.

BRCA1/BRCA2 germline mutations substantially increase breast and ovarian cancer risk, yet penetrance is incomplete. We hypothesized that germline epigenetic gene silencing may affect mutant BRCA1/2 penetrance. To test this notion, we determined the methylation status, using methylation-specific quantitative PCR of the promoter in putative modifier genes: BRCA1, BRCA2, ATM, ATR and P53 in Jewish BRCA1/BRCA2 mutation carriers with (n = 41) or without (n = 48) breast cancer, in sporadic breast cancer (n = 52), and healthy controls (n = 89). Promoter hypermethylation was detected only in the BRCA1 promotor in 5.6-7.3% in each of the four subsets of participants, regardless of health and BRCA1/2 status.Germline promoter hypermethylation in the BRCA1 gene can be detected in about 5% of the female Israeli Jewish population, regardless of the BRCA1/2 status. The significance of this observation is yet to be determined.

PMID: 18642075 [PubMed - as supplied by publisher]

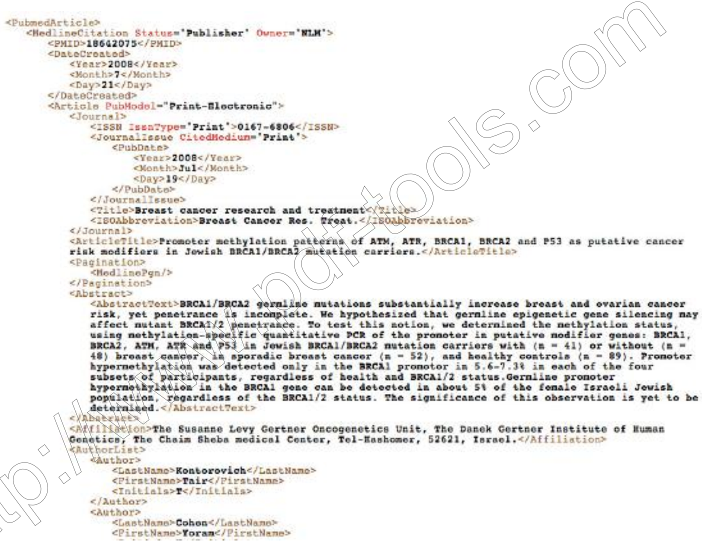
Related Articles

- Novel germline mutations in breast cancer susceptibility genes BRCA1, BRCA2 and [Breast Cancer Res Treat. 2004]
- Epigenetic silencing and deletion of the BRCA1 gene in sporadic breast cancer. [Breast Cancer Res. 2006]
- Prevalence and penetrance of germline BRCA1 and BRCA2 mutations in a population series of [Am J Hum Genet. 2001]
- Inactivation of BRCA1 and BRCA2 in ovarian cancer. [J Natl Cancer Inst. 2002]
- Gene promoter hypermethylation in ductal lavage fluid from healthy BRCA gene mutation carr [Breast Cancer Res. 2007]

» See all Related Articles...

Figure: Source http://www.ncbi.nlm.nih.gov/pubmed/

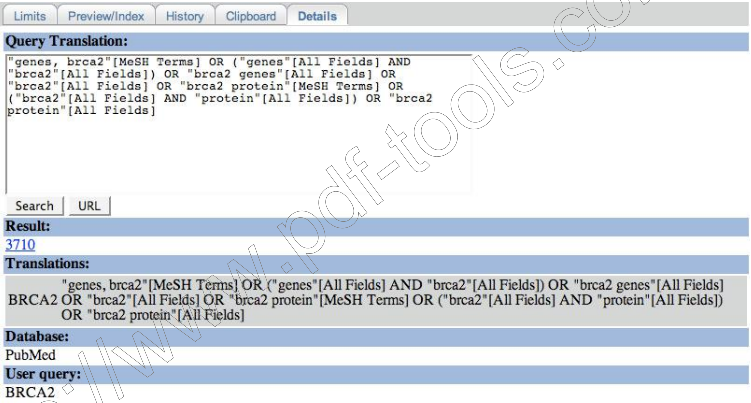
PubMed XML Record



```
<PubmedArticle>
   <MedlineCitation Status='Publisher' Owner='NLM'>
      <PMID>18642075</PMID>
      <DateCreated>
         <Year>2008</Year>
         <Month>7</Month>
         <Day>21</Day>
      </DateCreated>
      <Article PubModel="Print-Electronic">
         <Journal>
            <ISSN IssnType='Print'>0167-6806</ISSN>
            <JournalIssue CitedMedium='Print'>
               <PubDate>
                  <Year>2008</Year>
                  <Month>Jul</Month>
                  <Day>19</Day>
               </PubDate>
            </JournalIssue>
            <Title>Breast cancer research and treatment</Title>
            <ISOAbbreviation>Breast Cancer Res. Treat.</ISOAbbreviation>
         </Journal>
         <ArticleTitle>Promoter methylation patterns of ATM, ATR, BRCA1, BRCA2 and P53 as putative cancer
         risk modifiers in Jewish BRCA1/BRCA2 mutation carriers.</ArticleTitle>
         <Pagination>
            <MedlinePgn/>
         </Pagination>
         <Abstract>
            <AbstractText>BRCA1/BRCA2 germline mutations substantially increase breast and ovarian cancer
            risk, yet penetrance is incomplete. We hypothesized that germline epigenetic gene silencing may
            affect mutant BRCA1/2 penetrance. To test this notion, we determined the methylation status,
            using methylation-specific quantitative PCR of the promoter in putative modifier genes: BRCA1,
            BRCA2, ATM, ATR and P53 in Jewish BRCA1/BRCA2 mutation carriers with (n = 41) or without (n =
            48) breast cancer, in sporadic breast cancer (n = 52), and healthy controls (n = 89). Promoter
            hypermethylation was detected only in the BRCA1 promotor in 5.6-7.3% in each of the four
            subsets of participants, regardless of health and BRCA1/2 status.Germline promoter
            hypermethylation in the BRCA1 gene can be detected in about 5% of the female Israeli Jewish
            population, regardless of the BRCA1/2 status. The significance of this observation is yet to be
            determined.</AbstractText>
         </Abstract>
         <Affiliation>The Susanne Levy Gertner Oncogenetics Unit, The Danek Gertner Institute of Human
         Genetics, The Chaim Sheba medical Center, Tel-Hashomer, 52621, Israel.</Affiliation>
         <AuthorList>
            <Author>
               <LastName>Kontorovich</LastName>
               <FirstName>Tair</FirstName>
               <Initials>T</Initials>
            </Author>
            <Author>
               <LastName>Cohen</LastName>
               <FirstName>Yoram</FirstName>
```

Figure: Source http://www.ncbi.nlm.nih.gov/pubmed/

PubMed Query Translation



Query Translation:

```
"genes, brca2"[MeSH Terms] OR ("genes"[All Fields] AND
"brca2"[All Fields]) OR "brca2 genes"[All Fields] OR
"brca2"[All Fields] OR "brca2 protein"[MeSH Terms] OR
("brca2"[All Fields] AND "protein"[All Fields]) OR "brca2
protein"[All Fields]
```

Result:

3710

Translations:

BRCA2	"genes, brca2"[MeSH Terms] OR ("genes"[All Fields] AND "brca2"[All Fields]) OR "brca2 genes"[All Fields] OR "brca2"[All Fields] OR "brca2 protein"[MeSH Terms] OR ("brca2"[All Fields] AND "protein"[All Fields]) OR "brca2 protein"[All Fields]

Database:

PubMed

User query:

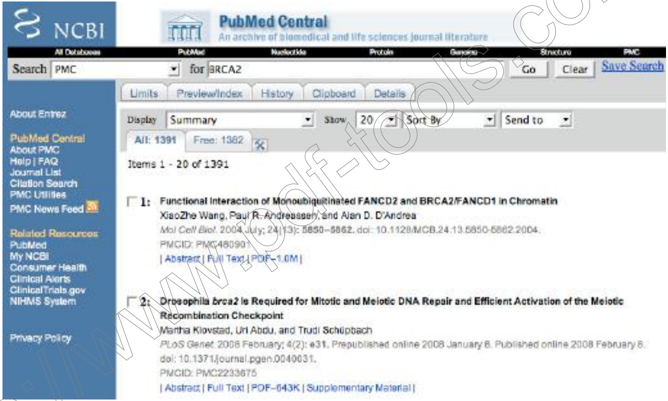
BRCA2

Figure: Source http://www.ncbi.nlm.nih.gov/pubmed/

PubMedCentral

- Digital archive of full text life science journals
- Articles have a unique PMCID
- Allows Boolean query search
- Offers free full text articles
- Journal Publishing XML DTD, but also other widely used DTD in life science

Example PubMedCentral Query



Search PMC for BRCA2

Display Summary Show 20 Sort By

All: 1391 Free: 1382

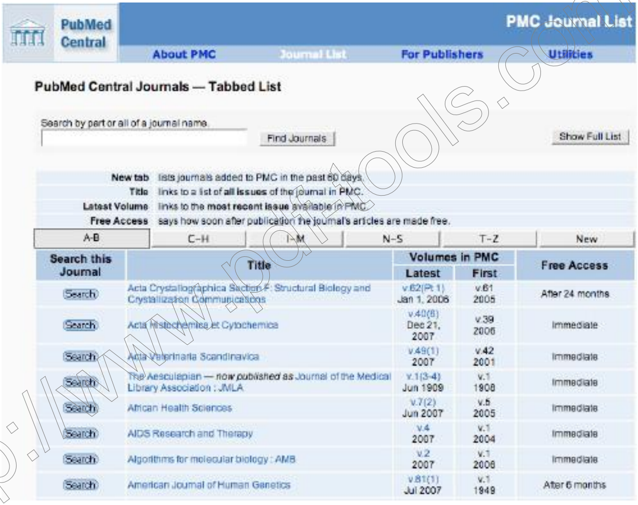
Items 1 - 20 of 1391

□ 1: **Functional Interaction of Monoubiquitinated FANCD2 and BRCA2/FANCD1 in Chromatin**
XiaoZhe Wang, Paul R. Andreassen, and Alan D. D'Andrea
Mol Cell Biol. 2004 July; 24(13): 5850–5862. doi: 10.1128/MCB.24.13.5850-5862.2004.
PMCID: PMC480901
| Abstract | Full Text | PDF–1.0M |

□ 2: **Drosophila *brca2* Is Required for Mitotic and Meiotic DNA Repair and Efficient Activation of the Meiotic Recombination Checkpoint**
Martha Klovstad, Uri Abdu, and Trudi Schüpbach
PLoS Genet. 2008 February; 4(2): e31. Prepublished online 2008 January 8. Published online 2008 February 8. doi: 10.1371/journal.pgen.0040031.
PMCID: PMC2233675
| Abstract | Full Text | PDF–643K | Supplementary Material |

Figure: Source:http://www.ncbi.nlm.nih.gov/pmc/

PubMed Journals



PubMed Central Journals — Tabbed List

Search by part or all of a journal name

New tab	lists journals added to PMC in the past 60 days.
Title	links to a list of **all issues** of the journal in PMC.
Latest Volume	links to the **most recent issue** available in PMC.
Free Access	says how soon after publication the journal's articles are made free.

Search this Journal	Title	Volumes in PMC		Free Access
		Latest	First	
Search	Acta Crystallographica Section F: Structural Biology and Crystallization Communications	v.62(Pt 1) Jan 1, 2006	v.61 2005	After 24 months
Search	Acta Histochemica et Cytochemica	v.40(6) Dec 21, 2007	v.39 2006	Immediate
Search	Acta Veterinaria Scandinavica	v.49(1) 2007	v.42 2001	Immediate
Search	The Aesculepian — *now published as* Journal of the Medical Library Association : JMLA	v.1(3-4) Jun 1909	v.1 1908	Immediate
Search	African Health Sciences	v.7(2) Jun 2007	v.5 2005	Immediate
Search	AIDS Research and Therapy	v.4 2007	v.1 2004	Immediate
Search	Algorithms for molecular biology : AMB	v.2 2007	v.1 2006	Immediate
Search	American Journal of Human Genetics	v.81(1) Jul 2007	v.1 1949	After 6 months

Figure: Source:http://www.ncbi.nlm.nih.gov/pmc/

NCBI Bookshelf

- Collection of biomedical text books
- Allows boolean query searches
- Offers free full text articles
- Direct searching the books or from PubMed abstract

Retrieving Electronic Literature Data from Web

- Get a local copy of some centralized literature repository (PubMed, PubMed Central, journals, etc): Leasing PubMed
- Use literature retrieval modules:
  - BioPython/BioPerl: Gazelle Z39.50 interface to PubMed
  - pubmed.pm by J. Smyser
  - Pubmed crawler written in Perl (http://pubcrawler.ie/)
  - eUtils: Entrez programming utilities
- Adopt web crawler, spiders or focussed crawler such as DataparkSearch, GNU Wget, Heritrix, Mutch.

Preprocessing Scientific Articles

1. **Document Standardization**: variety of formats (ASCII, HTML, XML, PDF, scanned PDF, SGML), convert them into a common format and encoding.
2. XML /Extensible Markup language, standard way to insert tags onto a text to identify its parts
3. OCR (Optical Character Recognition), used to digitize older literature (PMC Back Issue Digitization initiative)
4. Recover article Structure and content using pdftotext, PDFLib,PDF Converter

Preprocessing Scientific Articles

1. **Tokenization** break a stream of characters into words (tokens), e.g. white space, special chars. Each token is an instance of a type
2. **Stemming and lemmatization** standardize word tokens (e.g. Morphological analysis and Inflectional stemming, convert words to their corresponding root form)
3. Lexical analysis of the text with the objective of treating digits, hyphens, punctuation marks, and the case of letters
4. Elimination of stop-words
5. Selection of index terms

Journal Specific Characteristics of Literature

- Journal/Article format
- Paper structure (Section types)
- Article types (Review, Clinical Study, etc.)
- Target audience of journal/article

Processing of Full Text Articles

- Extract title, authors, abstract, text body, references
- Extract tables and tables legends
- Extract figures and figure legends

Basic Features of Biomedical Literature Data

- Heavy use of domain specific terminology (12% biochemistry related technical terms). E.g. chemoattractant, fibroblasts, angiogenesis
- Polysemic words (Word Sense Disambiguation). For example, APC means either:
  - Argon Plasma Coagulation
  - Activated Protein C

  Teashirt means:
  - a type of cloth
  - tsh gene
- Heavy use of acronyms, e.g. Activated protein C (APC), or vascular endothelial growth factor (VEGF)
- Data sparseness: Many words occur with low frequency

BioTerms Characteristics

- Novelty: New names and terms are frequently created. E.g. This disorder maps to chromosome 7q11-21, and this locus was named CLAM.[PMID:12771259]
- Typographical variants. E.g. TNF-Alpha and TNF Alpha
- Different writing styles
- Heavy use of referring expressions (anaphora, cataphora and ellipsis) and inference, example: Glycogenin is a glycosyltransferase. It functions as the autocatalytic initiator for the synthesis of glycogen in eukaryotic organisms.

Biomedical Corpora and Text Collections

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

Features of Natural Language Processing

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

Grammatical Features

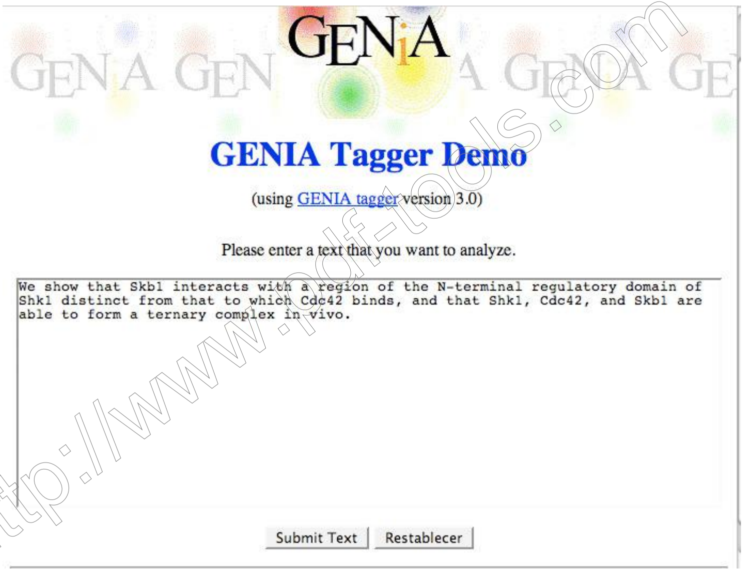
- Grammar: rules governing a particular language.
- Rules for correct formulation of a specific language
- Grammatical features in NLP, e.g. part of speech (POS)
- POS of a word depends on sentence context. E.g. noun, verb, adjective, adverb or preposition. E.g. [PMID 12700631]

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

POS Taggers

- Programs to label words with POS
- POS taggers are usually based on machine learning
- Trained with a set of manually POS-tagged sentences
- POS useful for gene name identification and protein interactions detection from text
- MedPost POS tagger for biomedical domain. MedPost: 97% accuracy in PubMed abstracts (86.8% general POS tagger)

GENIA POS Tagger



GENIA Tagger Demo

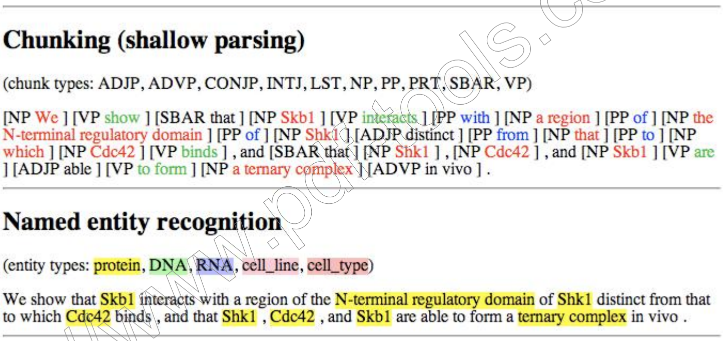
(using GENIA tagger version 3.0)

Please enter a text that you want to analyze.

```
We show that Skbl interacts with a region of the N-terminal regulatory domain of
Shkl distinct from that to which Cdc42 binds, and that Shkl, Cdc42, and Skbl are
able to form a ternary complex in vivo.
```

Figure: Source: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

GENIA POS Tagger Output



Chunking (shallow parsing)

(chunk types: ADJP, ADVP, CONJP, INTJ, LST, NP, PP, PRT, SBAR, VP)

[NP We] [VP show] [SBAR that] [NP Skb1] [VP interacts] [PP with] [NP a region] [PP of] [NP the N-terminal regulatory domain] [PP of] [NP Shk1] [ADJP distinct] [PP from] [NP that] [PP to] [NP which] [NP Cdc42] [VP binds] , and [SBAR that] [NP Shk1] , [NP Cdc42] , and [NP Skb1] [VP are] [ADJP able] [VP to form] [NP a ternary complex] [ADVP in vivo] .

Named entity recognition

(entity types: protein, DNA, RNA, cell_line, cell_type)

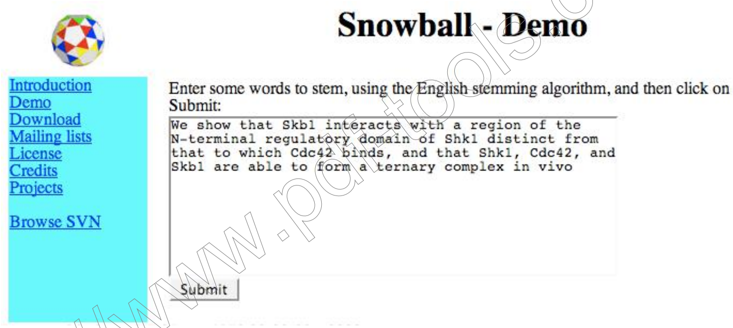
We show that Skb1 interacts with a region of the N-terminal regulatory domain of Shk1 distinct from that to which Cdc42 binds , and that Shk1 , Cdc42 , and Skb1 are able to form a ternary complex in vivo .

Figure: Source: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

Morphological Features

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

Online Stemmer: SnowBall



Snowball - Demo

Enter some words to stem, using the English stemming algorithm, and then click on Submit:

We show that Skbl interacts with a region of the N-terminal regulatory domain of Shk1 distinct from that to which Cdc42 binds, and that Shk1, Cdc42, and Skbl are able to form a ternary complex in vivo

Figure: Source: http://snowball.tartarus.org/demo.php

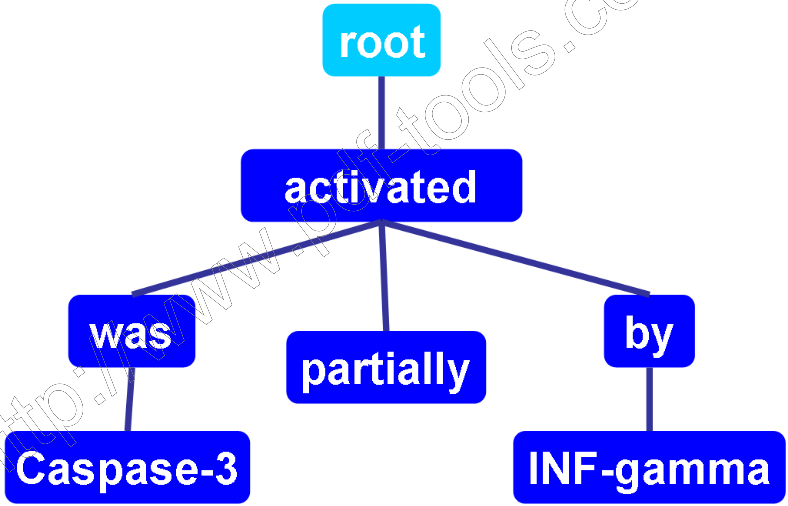
Syntactic features

- Relationships between words in a sentence: syntactic structure
- Shallow parsers analyze such relations at a coarse level, identification of phrases (groups of words which function as a syntactic unit).
- Output of Connexor shallow parser

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

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

Syntactic Features



Syntactic Features

- Identification of subject-object relationships
- NP-VP-NP E.g.

  ```
  Overexpression of <gene>IMEl</gene> induced an
  <GO> early meiotic event (recombination) </GO> in
  rich medium, but later meiotic events did not
  occur (i.e., they detected [no spore formation])
  ```

  Subject: IMEI gene and object is GO term early meiotic event

Semantic features

- Associations of words with their corresponding meaning in a given context.
- Semantics (meanings) of a word → understand meaning sentence.
- Dictionaries and thesauri provide such associations.
- Gene Ontology (GO) provides concepts for biological aspects of genes
- Gene names and symbols contained in SwissProt (symbol dict.)

Token	GO Symbols
Caspase-3	GENE PRODUCT
was	
partially	
activated	INTERACTION VERB
by	
IFN-gamma	GENE PRODUCT

Contextual Features

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

Part 2: BioText Mining

Key Technologies in BioText Mining

- Information Retrieval (IR)
- Information Extraction (IE)
- Text Classification
- Text Clustering

Information Retrieval

IR

- It is a process of recovery of those documents from a collection of documents which satisfy a given information demand.
- Information demand is posed in form of a query

Efficient Indexing is required to reduce vocabulary of terms and query formulation.

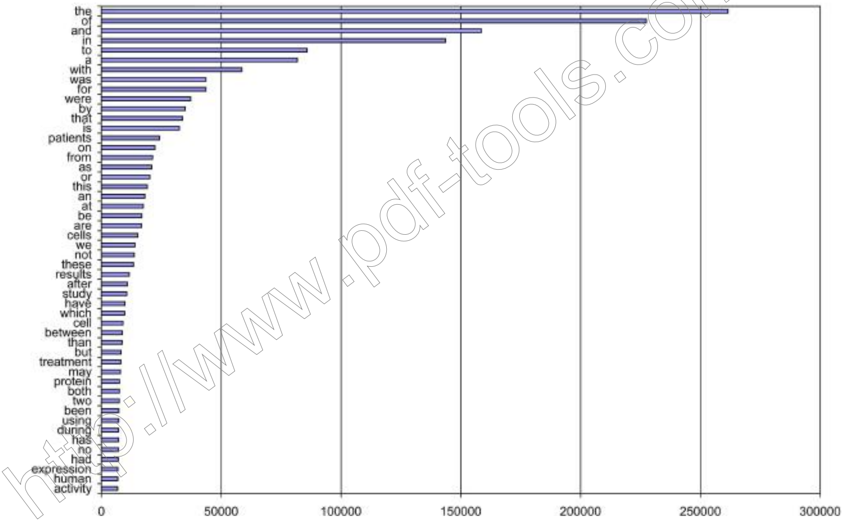
Important Steps in Indexing Document Collection

- Tokenization
- Case folding
- Stemming
- Stop word removal

Query Types:

- Boolean queries

Zipf's Law



the
of
and
in
to
a
with
was
for
were
by
that
is
patients
on
from
as
or
this
an
at
be
are
cells
we
not
these
results
after
study
have
which
cell
between
than
but
treatment
may
protein
both
two
been
using
during
has
no
had
expression
human
activity
0
50000
100000
150000
200000
250000
300000

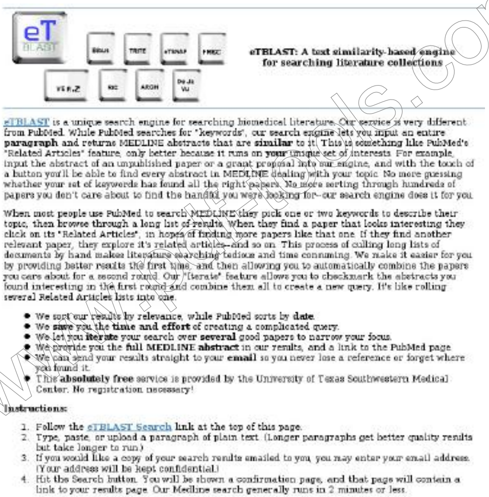
Boolean Queries

- Based on combination of terms using Boolean operators
- Basic Boolean operators: **AND**, **OR**, **NOT**
- Queries matched against the terms in the inverted index file
- E.g. Entrez - Boolean search in PubMed
- Fast and easy to implement

Vector Space Model

- Measure similarity between query and documents
- Query may be a list of terms or even whole documents
- Represent document and query using a vector of terms.
- Each term t is weighted according to its frequency in the document d and in the whole document collection D.
- Calculate cosine similarity between query and document vector.
- Return ranked list of documents
- E.g. Related article search in PubMed

eTBLAST



eTBLAST: A text similarity-based engine for searching literature collections

eTBLAST is a unique search engine for searching biomedical literature. Our service is very different from PubMed. While PubMed searches for "keywords", our search engine lets you input an entire **paragraph** and returns MEDLINE abstracts that are **similar** to it. This is something like PubMed's "Related Articles" feature, only better because it runs on **your** unique set of interests. For example, input the abstract of an unpublished paper or a grant proposal into our engine, and with the touch of a button you'll be able to find every abstract in MEDLINE dealing with your topic. No more guessing whether your set of keywords has found all the right papers. No more sorting through hundreds of papers you don't care about to find the handful you were looking for--our search engine does it for you.

When most people use PubMed to search MEDLINE they pick one or two keywords to describe their topic, then browse through a long list of results. When they find a paper that looks interesting they click on its "Related Articles", in hopes of finding more papers like that one. If they find another relevant paper, they explore it's related articles--and so on. This process of culling long lists of documents by hand makes literature searching tedious and time consuming. We make it easier for you by providing better results the first time, and then allowing you to automatically combine the papers you care about for a second round. Our "Iterate" feature allows you to checkmark the abstracts you found interesting in the first round and combine them all to create a new query. It's like rolling several Related Articles lists into one.

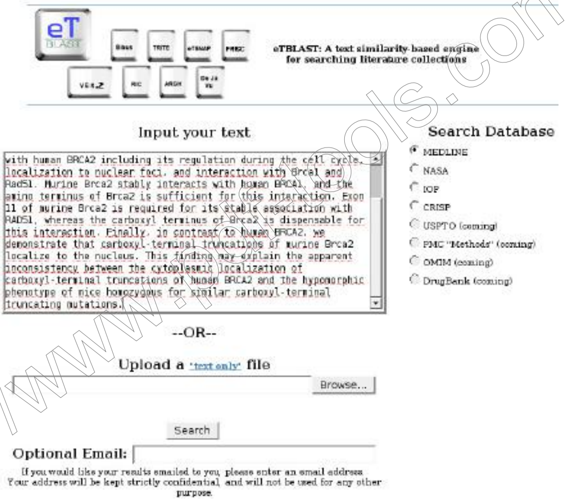
- We sort our results by relevance, while PubMed sorts by **date**.
- We **save** you the **time and effort** of creating a complicated query.
- We let you **iterate** your search over **several** good papers to narrow your focus.
- We provide you the **full MEDLINE abstract** in our results, and a link to the PubMed page
- We can send your results straight to your **email** so you never lose a reference or forget where you found it.
- This **absolutely free** service is provided by the University of Texas Southwestern Medical Center. No registration necessary!

Instructions:

1. Follow the eTBLAST Search link at the top of this page.
2. Type, paste, or upload a paragraph of plain text. (Longer paragraphs get better quality results but take longer to run.)
3. If you would like a copy of your search results emailed to you, you may enter your email address. (Your address will be kept confidential.)
4. Hit the Search button. You will be shown a confirmation page, and that page will contain a link to your results page. Our Medline search generally runs in 2 minutes or less.

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

eTBLAST



eTBLAST: A text similarity-based engine for searching literature collections

Input your text

with human BRCA2 including its regulation during the cell cycle, localization to nuclear foci, and interaction with Brca1 and Rad51. Murine Brca2 stably interacts with human BRCA1, and the amino terminus of Brca2 is sufficient for this interaction. Exon 11 of murine Brca2 is required for its stable association with RAD51, whereas the carboxyl terminus of Brca2 is dispensable for this interaction. Finally, in contrast to human BRCA2, we demonstrate that carboxyl-terminal truncations of murine Brca2 localize to the nucleus. This finding may explain the apparent inconsistency between the cytoplasmic localization of carboxyl-terminal truncations of human BRCA2 and the hypomorphic phenotype of mice homozygous for similar carboxyl-terminal truncating mutations.

--OR--

Upload a 'text only' file

Browse...

Search

Optional Email:

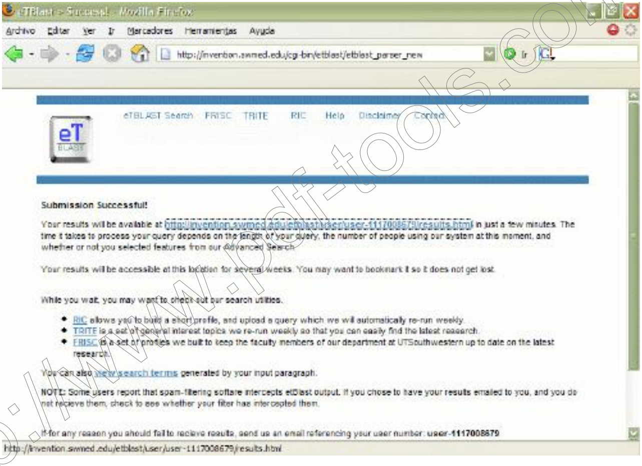
If you would like your results emailed to you, please enter an email address. Your address will be kept strictly confidential, and will not be used for any other purpose.

Search Database

● MEDLINE
○ NASA
○ IOP
○ CRISP
○ USPTO (coming)
○ PMC "Methods" (coming)
○ OMIM (coming)
○ DrugBank (coming)

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

eTBLAST


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

Submission Successful!

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

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

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

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

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

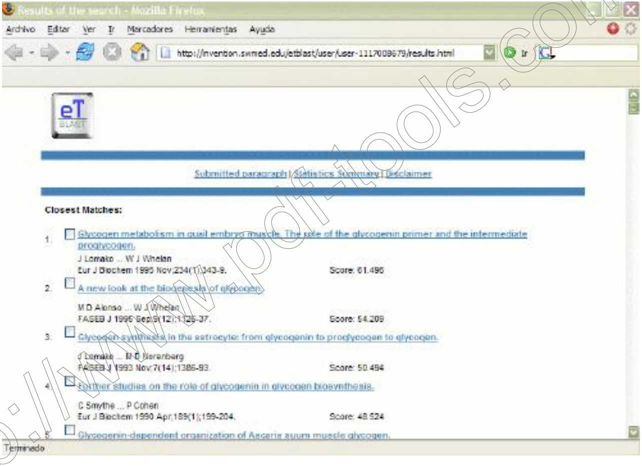
NOTE: Some users report that spam-filtering softare intercepts etBlast output. If you chose to have your results emailed to you, and you do not recieve them, check to see whether your filter has intercepted them.

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

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

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

eTBLAST Results



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

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

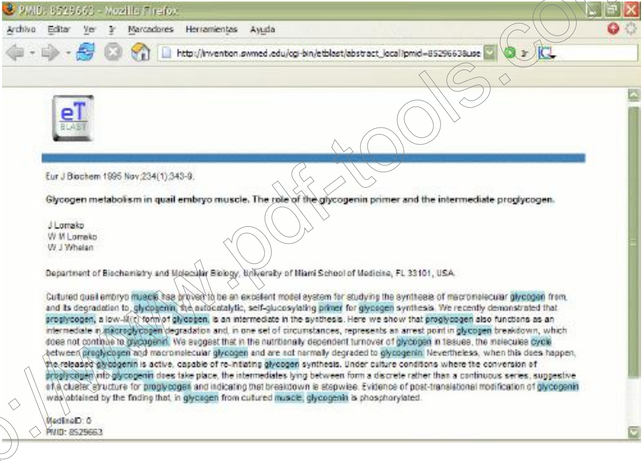
3. □ Glycogen synthesis in the astrocyte: from glycogenin to proglycogen to glycogen.
J Lomako ... M D Neranberg
FASEB J 1993 Nov;7(14):1386-93. Score: 50.494

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.

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

eTBLAST


PMID: 8529663 - Mozilla Firefox
Archivo
Editar
Ver
Ir
Marcadores
Herramientas
Ayuda
http://invention.swmed.edu/cgi-bin/etblast/abstract_local?pmid=8529663&use
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

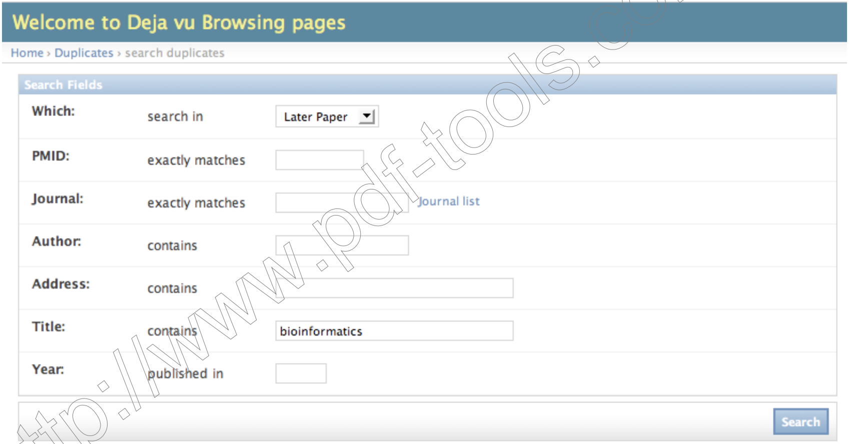
Department of Biochemistry and Molecular Biology, University of Miami School of Medicine, FL 33101, USA.

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

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

Text Similarity and Deja Vu



Welcome to Deja vu Browsing pages

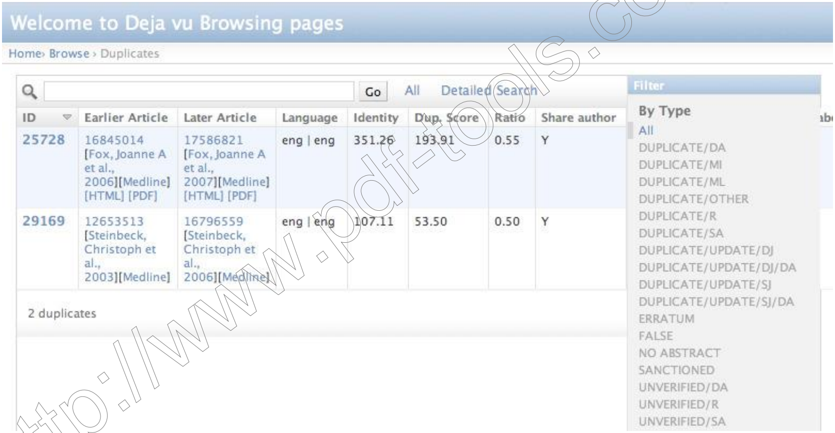
Home › Duplicates › search duplicates

Search Fields

Which:	search in	Later Paper
PMID:	exactly matches	
Journal:	exactly matches	
Author:	contains	
Address:	contains	
Title:	contains	bioinformatics
Year:	published in	

Figure: Source: http://spore.swmed.edu/dejavu/

Text Similarity and Deja Vu



ID	Earlier Article	Later Article	Language	Identity	Dup. Score	Ratio	Share author
25728	16845014 [Fox, Joanne A et al., 2006][Medline] [HTML] [PDF]	17586821 [Fox, Joanne A et al., 2007][Medline] [HTML] [PDF]	eng | eng	351.26	193.91	0.55	Y
29169	12653513 [Steinbeck, Christoph et al., 2003][Medline]	16796559 [Steinbeck, Christoph et al., 2006][Medline]	eng | eng	107.11	53.50	0.50	Y

2 duplicates

Figure: Source: http://spore.swmed.edu/dejavu/

IR Evaluation

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

Text Clustering

- Find which documents have many words in common, and place the documents with the most words in common into the same groups.
- Similarity of documents instead of similarity of sequences, expression profiles or structures
- Cluster documents into topics, for instance: clinical, biochemical and microbiology articles
- A clustering program tries to find the groups in the data.

Text Clustering

- Clustering programs often choose first the documents that seem representative of the middle of each of the clusters (candidate centers of the clusters).
- Then it compares all the documents to these initial representatives.
- Each documents is assigned to the cluster it is most similar to.
- Similarity is based on how many words the documents have in common, and how strongly they are weighted.
- The topical terms of the clusters are chosen from words that represent the center of the cluster.
- The best clustering is one in which the average difference of the documents to their cluster centers smallest.
- Agglomerative clustering: first comparing every pair of documents, and finding the pair of documents which are most similar to each other.

Text Classification

- Common problem in information science.
- Assignment of an electronic document to one or more categories, based on its contents (words).
- Supervised document classification where training examples of document classification are provided and the correct classification model is learnt based on one of the following techniques:
  - naive Bayes classifier
  - tf-idf
  - latent semantic indexing
  - support vector machines
  - artificial neural network
  - kNN
  - decision trees, such as ID3
  - Concept Mining
- Classification techniques have been applied to spam filtering
- Can use the bow toolkit, SVMlight, LibSVM etc

Information Extraction (IE)

IE refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources [1].

Applications of IE

- Enterprise applications such as news tracking, customer care, data cleaning, classified Ads
- Personal Information Management such as Emails, Documents, Presentations
- Scientific applications such as BioIE
- Web oriented applications such as citation databases, opinion databases, community websites, comparison shopping, ad-placement

[1] Sarawagi, S. (2008) Information Extraction, FnT Databases, 1(3).

IE Taxonomy

1. Structure extracted (entities, relationships, lists, tables, attributes etc.)
2. Unstructure source (short strings or documents, templatized or open ended.)
3. Input resources available for extraction (structured databases, labeled unstructured data, linguistic tags, dictionaries etc.)
4. Extraction method (rule based or statistical, manually coded or trained from examples)
5. Output of extraction (annotated unstructured text or a database)

Key Challenges in IE

Accuracy

- Diversity of clues such as orthographic features of words, POS, similarity with existing entries in database, presence of trigger words and so on
- Difficulty of detecting missed extractions: High recall models are desirable, but how do we ensure high recall without extensive labeled data?
- Increased complexity of the structures extracted: Extraction of longer entities where the boundary is not defined clearly. e.g. extracting name of restaurant from Blog.

Key Challenges in IE

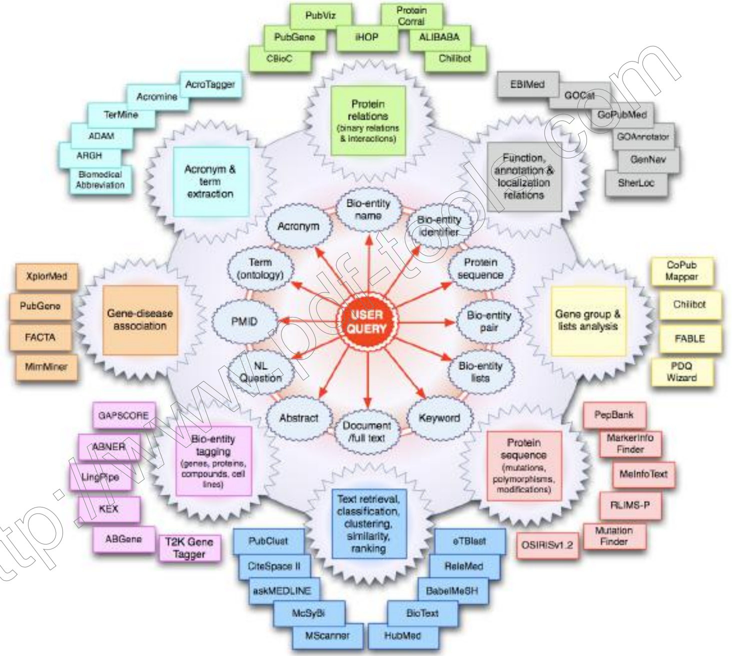
Running Time

- Efficiently filtering right subset of documents that are likely to contain structured information of interest
- Efficiently locating portion of document containing relevant information of interest
- Efficient extraction of information

Other Systems Issues

- Dynamically Changing Sources
- Data Integration
- Extraction Errors

Part 3: Applications of Text Mining



USER QUERY

Bio-entity name

Bio-entity identifier

Protein sequence

Bio-entity pair

Bio-entity lists

Keyword

Document /full text

Abstract

NL Question

PMID

Term (ontology)

Acronym

Protein relations (binary relations & interactions)

PubViz

Protein Corral

PubGene

iHOP

ALIBABA

CBioC

Chilibot

Function, annotation & localization relations

EBIMed

GOCat

GoPubMed

GOAnnotator

GenNav

SherLoc

Gene group & lists analysis

CoPub Mapper

Chilibot

FABLE

PDQ Wizard

Protein sequence (mutations, polymorphisms, modifications)

PepBank

MarkerInfo Finder

MeInfoText

RLIMS-P

Mutation Finder

OSIRISv1.2

Text retrieval, classification, clustering, similarity, ranking

PubClust

CiteSpace II

askMEDLINE

McSyBi

MScanner

eTBlast

ReleMed

BabelMeSH

BioText

HubMed

Bio-entity tagging (genes, proteins, compounds, cell lines)

GAPSCORE

ABNER

LingPipe

KEX

ABGene

T2K Gene Tagger

Gene-disease association

XplorMed

PubGene

FACTA

MimMiner

Acronym & term extraction

AcroTagger

Acromine

TerMine

ADAM

ARGH

Biomedical Abbreviation

Applications of BioText Mining

- Named entity recognition of biological entities (BioNER)
- Gene normalization
- Protein-Protein interaction
- Functional Analysis of genes and gene sets
- Extraction of gene-disease association
- Extraction of mutations and epigenetic characteristics
- Extraction of protein location information
- Building terminology resource for a specific domain
- Knowledge discovery and pathways

Bio-Named Entity Recognition (BioNER)

Objective

Identify biological entities in articles and to link them to entries in biological databases.

Challenges

- more complex (synonyms, disambiguation, typographical variants, official symbols not used)
- Performance organism dependent

Methods

- POS tagging,
- Rule-based,
- Flexible matching,
- Statistical and Machine Learning (naive Bayes, ME, SVM, CRF, HMM).

BioNER Example

Input: Abstract/Full Text Article

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).

Output: Input + Tagged Entities

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).

Protein Cell Line

BioNER Challenges

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

Tricky Issues in Gene Tagging

- The nightcap mutation caused severe defects in these cells [PMID:12399306].
- In the present investigation, we have discovered that Piccolo, a CAZ (cytoskeletal matrix associated with the active zone) protein in neurons that is structurally related to Rim2, [PMID:12401793]
- The Drosophila takeout gene is regulated by the somatic sex-determination pathway and affects male courtship behavior. [PMID:12435630]
- This function is independent of Chico, the Drosophila insulin receptor substrate (IRS) homolog [PMID:12702880].
- A new longevity gene, Indy (for I'm not dead yet), which doubles the average. [PMID:12391301]
- The Drosophila peanut gene is required for cytokinesis and encodes a protein similar to yeast putative bud neck filament proteins [PMID 8181057].
- Ambiguity of PKC: Protein kinase C and Pollution kerato-conjunctivitis

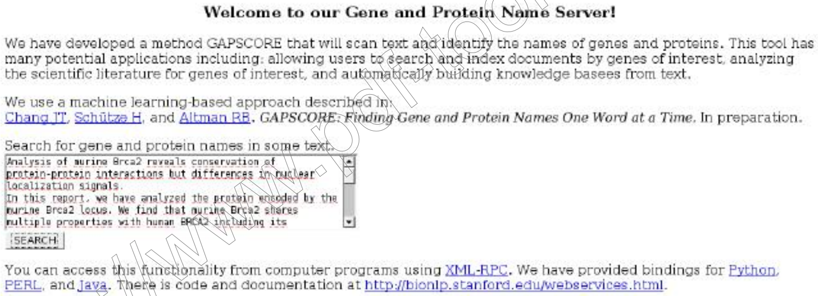
GapScore

- Scores words based on a statistical model of gene names
- Quantifies: Appearance, Morphology, Context.
- Performance on Yapex corpus:

Match Type	**Precision**	**Recall**	**F1 Score**
Partial Match	81.5%	83.3%	82.5%
Exact Match	56.7%	58.5%	57.6%

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

GapScore



Welcome to our Gene and Protein Name Server!

We have developed a method GAPSCORE that will scan text and identify the names of genes and proteins. This tool has many potential applications including: allowing users to search and index documents by genes of interest, analyzing the scientific literature for genes of interest, and automatically building knowledge basees from text.

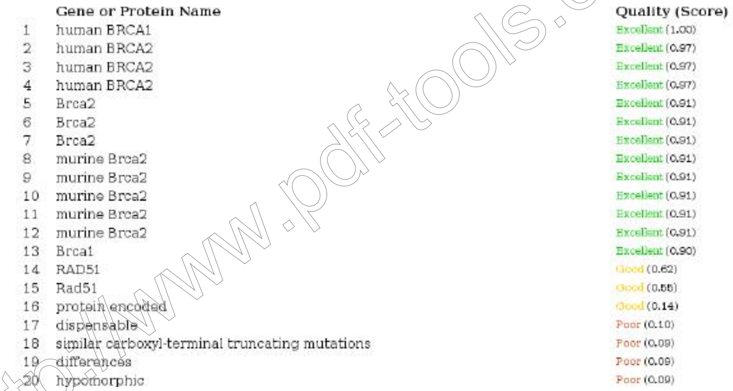
We use a machine learning-based approach described in:
Chang JT, Schütze H, and Altman RB. *GAPSCORE: Finding Gene and Protein Names One Word at a Time.* In preparation.

Search for gene and protein names in some text:

```
Analysis of murine Brca2 reveals conservation of
protein-protein interactions but differences in nuclear
localization signals.
In this report, we have analyzed the protein encoded by the
murine Brca2 locus. We find that murine Brca2 shares
multiple properties with human BRCA2 including its
```

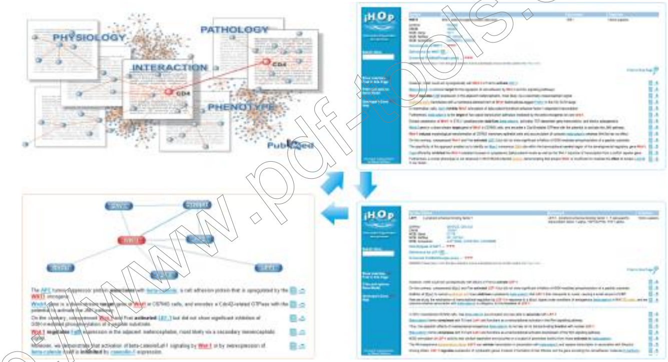
You can access this functionality from computer programs using XML-RPC. We have provided bindings for Python, PERL, and Java. There is code and documentation at http://bionlp.stanford.edu/webservices.html.

GapScore Output



	Gene or Protein Name	Quality (Score)
1	human BRCA1	Excellent (1.00)
2	human BRCA2	Excellent (0.97)
3	human BRCA2	Excellent (0.97)
4	human BRCA2	Excellent (0.97)
5	Brca2	Excellent (0.91)
6	Brca2	Excellent (0.91)
7	Brca2	Excellent (0.91)
8	murine Brca2	Excellent (0.91)
9	murine Brca2	Excellent (0.91)
10	murine Brca2	Excellent (0.91)
11	murine Brca2	Excellent (0.91)
12	murine Brca2	Excellent (0.91)
13	Brca1	Excellent (0.90)
14	RAD51	Good (0.62)
15	Rad51	Good (0.55)
16	protein encoded	Good (0.14)
17	dispensable	Poor (0.10)
18	similar carboxyl-terminal truncating mutations	Poor (0.09)
19	differences	Poor (0.09)
20	hypomorphic	Poor (0.09)

iHOP: Gene Name Tagging



The APC tumour suppressor protein associates with beta-catenin, a cell adhesion protein that is upregulated by the Wnt1 oncogene.

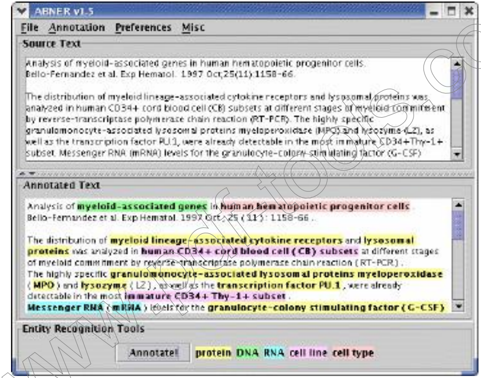
Wrch1 gene is a downstream target gene of Wnt1 in C57MG cells, and encodes a Cdc42-related GTPase with the potential to activate the JNK pathway.

On the contrary, coexpressed Wnt-1 and Fzd activated LEF-1 but did not show significant inhibition of GSK-3-mediated phosphorylation of a peptide substrate.

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

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

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)**

- ABNER is a software tool for molecular biology text analysis.
- It uses linear chain conditional random fields (CRFs) with a variety of orthographic and contextual features.
- URL: http://pages.cs.wisc.edu/ bsettles/abner/

Gene Normalization

Objective

- Linking genes or gene products mentioned in the literature to biological databases.
- Key step in enabling accurate search of biological literature and linking database information to passages in research articles.

Challenges

- Genes are often described rather than referred to by gene symbol
- One gene name may refer to different genes (often from different organism)
- Incomplete dictionaries of gene names

Example

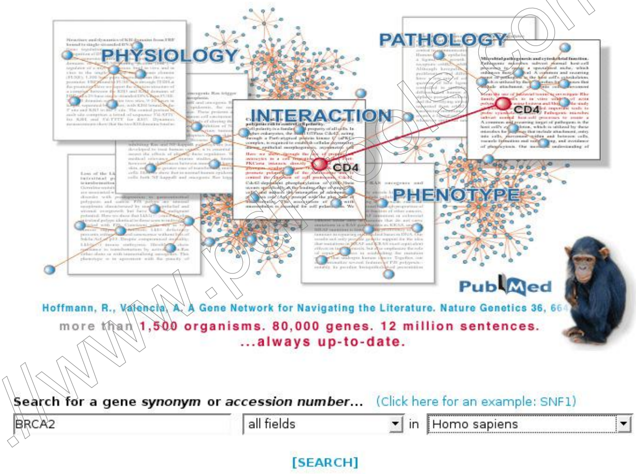
Input: Abstract/Full Text Articles with Entites Tagged

The double-stranded (ds) RNA-activated protein kinase from human cells is a 68 kd protein (p68 kinase) induced by interferon. On activation by dsRNA in the presence of ATP, the kinase becomes autophosphorylated and can catalyze the phosphorylation of the alpha subunit of eIF2, which leads to an inhibition of the initiation of protein synthesis.

Gene Normalized to Database Records

Entity	**Normalized Gene Name**	UniProt ID
RNA-activated protein kinase	E2AK2_HUMAN	P19525
p68 kinase	E2AK2_HUMAN	P19525
alpha subunit of eIF2	IF2A_HUMAN	P05198

iHOP System



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

more than 1,500 organisms. 80,000 genes. 12 million sentences.
...always up-to-date.

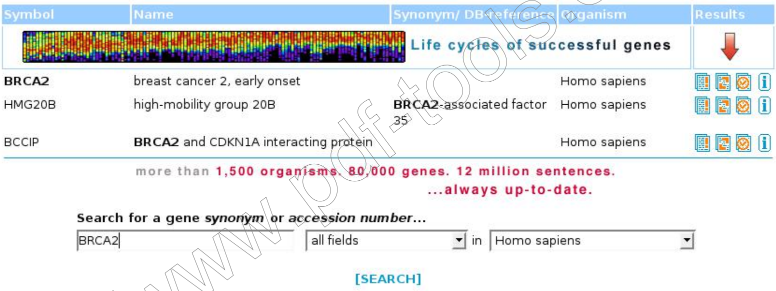
Search for a gene *synonym* or *accession number*... (Click here for an example: SNF1)

BRCA2 | all fields ▾ in Homo sapiens ▾

[SEARCH]

Figure: Source: http://www.ihop-net.org/

iHOP: Query to DB Records



Symbol	Name	Synonym/ DB-reference	Organism	Results
		Life cycles of successful genes		↓
BRCA2	breast cancer 2, early onset		Homo sapiens	▤ ▤ ◎ ℹ
HMG20B	high-mobility group 20B	**BRCA2**-associated factor 35	Homo sapiens	▤ ▤ ◎ ℹ
BCCIP	**BRCA2** and CDKN1A interacting protein		Homo sapiens	▤ ▤ ◎ ℹ

more than **1,500 organisms. 80,000 genes. 12 million sentences.**

...always up-to-date.

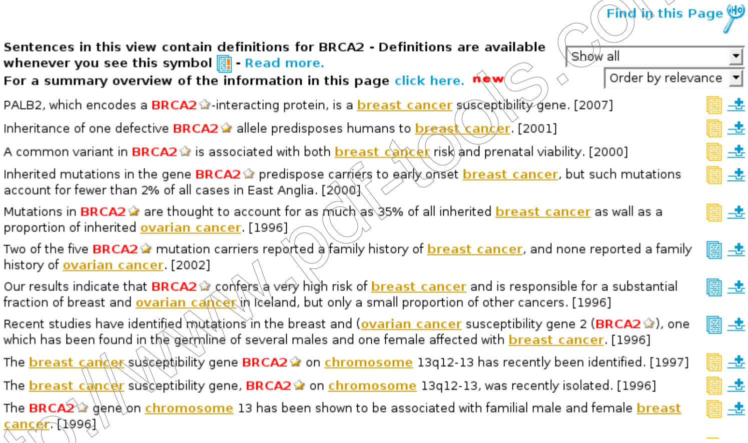
Search for a gene *synonym* or *accession number*...

BRCA2 | all fields ▾ in Homo sapiens ▾

[SEARCH]

Figure: Source: http://www.ihop-net.org/

iHOP: Information about Gene



Sentences in this view contain definitions for BRCA2 - Definitions are available whenever you see this symbol - Read more.
For a summary overview of the information in this page click here. new

Show all

Order by relevance

PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. [2007]

Inheritance of one defective BRCA2 allele predisposes humans to breast cancer. [2001]

A common variant in BRCA2 is associated with both breast cancer risk and prenatal viability. [2000]

Inherited mutations in the gene BRCA2 predispose carriers to early onset breast cancer, but such mutations account for fewer than 2% of all cases in East Anglia. [2000]

Mutations in BRCA2 are thought to account for as much as 35% of all inherited breast cancer as wall as a proportion of inherited ovarian cancer. [1996]

Two of the five BRCA2 mutation carriers reported a family history of breast cancer, and none reported a family history of ovarian cancer. [2002]

Our results indicate that BRCA2 confers a very high risk of breast cancer and is responsible for a substantial fraction of breast and ovarian cancer in Iceland, but only a small proportion of other cancers. [1996]

Recent studies have identified mutations in the breast and (ovarian cancer susceptibility gene 2 (BRCA2), one which has been found in the germline of several males and one female affected with breast cancer. [1996]

The breast cancer susceptibility gene BRCA2 on chromosome 13q12-13 has recently been identified. [1997]

The breast cancer susceptibility gene, BRCA2 on chromosome 13q12-13, was recently isolated. [1996]

The BRCA2 gene on chromosome 13 has been shown to be associated with familial male and female breast cancer. [1996]

Figure: Source: http://www.ihop-net.org/

Interaction Extraction

Objective

Extract interaction information between biological entities from literature. For example, protein-protein interaction.

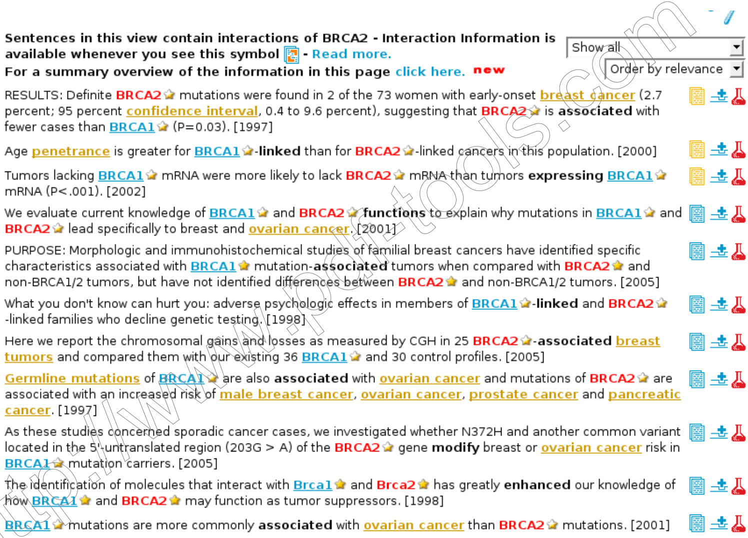
Key Techniques

- Co-occurrence of bioentities within close vicinity
- Machine learning based methods (Relationship extraction)
- Linguistic methods (Dependency parsers, link parsers)
- Rule based

Manually curated Interaction Databases

- MINT
- BioGRID
- IntAct

iHOP: Interaction Information



Sentences in this view contain interactions of BRCA2 - Interaction Information is available whenever you see this symbol - Read more.
For a summary overview of the information in this page click here. new

RESULTS: Definite BRCA2 mutations were found in 2 of the 73 women with early-onset breast cancer (2.7 percent; 95 percent confidence interval, 0.4 to 9.6 percent), suggesting that BRCA2 is **associated** with fewer cases than BRCA1 (P=0.03). [1997]

Age penetrance is greater for BRCA1**-linked** than for BRCA2-linked cancers in this population. [2000]

Tumors lacking BRCA1 mRNA were more likely to lack BRCA2 mRNA than tumors **expressing** BRCA1 mRNA (P<.001). [2002]

We evaluate current knowledge of BRCA1 and BRCA2 **functions** to explain why mutations in BRCA1 and BRCA2 lead specifically to breast and ovarian cancer. [2001]

PURPOSE: Morphologic and immunohistochemical studies of familial breast cancers have identified specific characteristics associated with BRCA1 mutation-**associated** tumors when compared with BRCA2 and non-BRCA1/2 tumors, but have not identified differences between BRCA2 and non-BRCA1/2 tumors. [2005]

What you don't know can hurt you: adverse psychologic effects in members of BRCA1**-linked** and BRCA2 -linked families who decline genetic testing. [1998]

Here we report the chromosomal gains and losses as measured by CGH in 25 BRCA2**-associated** breast tumors and compared them with our existing 36 BRCA1 and 30 control profiles. [2005]

Germline mutations of BRCA1 are also **associated** with ovarian cancer and mutations of BRCA2 are associated with an increased risk of male breast cancer, ovarian cancer, prostate cancer and pancreatic cancer. [1997]

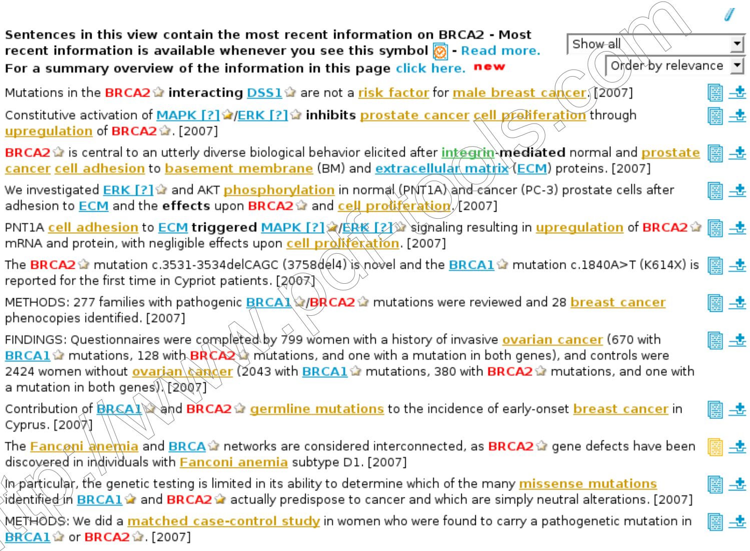
As these studies concerned sporadic cancer cases, we investigated whether N372H and another common variant located in the 5'-untranslated region (203G > A) of the BRCA2 gene **modify** breast or ovarian cancer risk in BRCA1 mutation carriers. [2005]

The identification of molecules that interact with Brca1 and Brca2 has greatly **enhanced** our knowledge of how BRCA1 and BRCA2 may function as tumor suppressors. [1998]

BRCA1 mutations are more commonly **associated** with ovarian cancer than BRCA2 mutations. [2001]

Figure: Source: http://www.ihop-net.org/

iHOP: Recent Information



Sentences in this view contain the most recent information on BRCA2 - Most recent information is available whenever you see this symbol - Read more. For a summary overview of the information in this page click here. new

Mutations in the BRCA2 **interacting** DSS1 are not a risk factor for male breast cancer. [2007]

Constitutive activation of MAPK [?]/ERK [?] **inhibits** prostate cancer cell proliferation through upregulation of BRCA2. [2007]

BRCA2 is central to an utterly diverse biological behavior elicited after integrin-**mediated** normal and prostate cancer cell adhesion to basement membrane (BM) and extracellular matrix (ECM) proteins. [2007]

We investigated ERK [?] and AKT phosphorylation in normal (PNT1A) and cancer (PC-3) prostate cells after adhesion to ECM and the **effects** upon BRCA2 and cell proliferation. [2007]

PNT1A cell adhesion to ECM **triggered** MAPK [?]/ERK [?] signaling resulting in upregulation of BRCA2 mRNA and protein, with negligible effects upon cell proliferation. [2007]

The BRCA2 mutation c.3531-3534delCAGC (3758del4) is novel and the BRCA1 mutation c.1840A>T (K614X) is reported for the first time in Cypriot patients. [2007]

METHODS: 277 families with pathogenic BRCA1/BRCA2 mutations were reviewed and 28 breast cancer phenocopies identified. [2007]

FINDINGS: Questionnaires were completed by 799 women with a history of invasive ovarian cancer (670 with BRCA1 mutations, 128 with BRCA2 mutations, and one with a mutation in both genes), and controls were 2424 women without ovarian cancer (2043 with BRCA1 mutations, 380 with BRCA2 mutations, and one with a mutation in both genes). [2007]

Contribution of BRCA1 and BRCA2 germline mutations to the incidence of early-onset breast cancer in Cyprus. [2007]

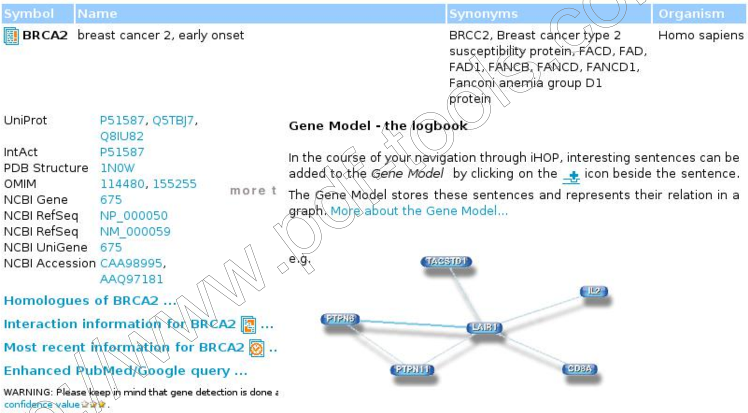
The Fanconi anemia and BRCA networks are considered interconnected, as BRCA2 gene defects have been discovered in individuals with Fanconi anemia subtype D1. [2007]

In particular, the genetic testing is limited in its ability to determine which of the many missense mutations identified in BRCA1 and BRCA2 actually predispose to cancer and which are simply neutral alterations. [2007]

METHODS: We did a matched case-control study in women who were found to carry a pathogenetic mutation in BRCA1 or BRCA2. [2007]

Figure: Source: http://www.ihop-net.org/

iHOP System: Gene Model/Graph



Symbol	Name	Synonyms	Organism
BRCA2	breast cancer 2, early onset	BRCC2, Breast cancer type 2 susceptibility protein, FACD, FAD, FAD1, FANCB, FANCD, FANCD1, Fanconi anemia group D1 protein	Homo sapiens

UniProt	P51587, Q5TBJ7, Q8IU82
IntAct	P51587
PDB Structure	1N0W
OMIM	114480, 155255
NCBI Gene	675
NCBI RefSeq	NP_000050
NCBI RefSeq	NM_000059
NCBI UniGene	675
NCBI Accession	CAA98995, AAQ97181

more t

Homologues of BRCA2 ...

Interaction information for BRCA2 ...

Most recent information for BRCA2 ..

Enhanced PubMed/Google query ...

WARNING: Please keep in mind that gene detection is done a

confidence value ☆★★.

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...

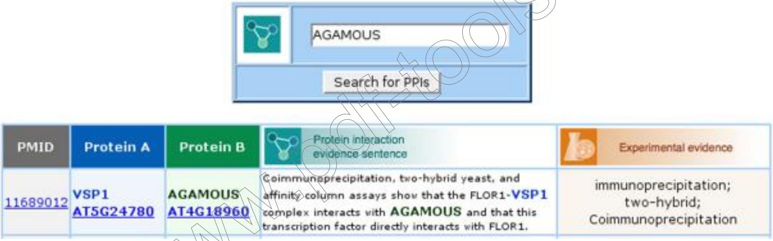
e.g.

Figure: Source: http://www.ihop-net.org/

PLAN2L: Plant Annotation to Literature

- Web tool for integrated text mining and literature-derived bio-entity relation extraction
- Provides following searches
  - Searching *Arabidopsis* bibliome
  - Searching for PPI
  - Searching for gene regulation association
  - Searching for location sentences
  - Searching for cell cycle association
  - Association retrieval

PLAN2L: PPI Extraction



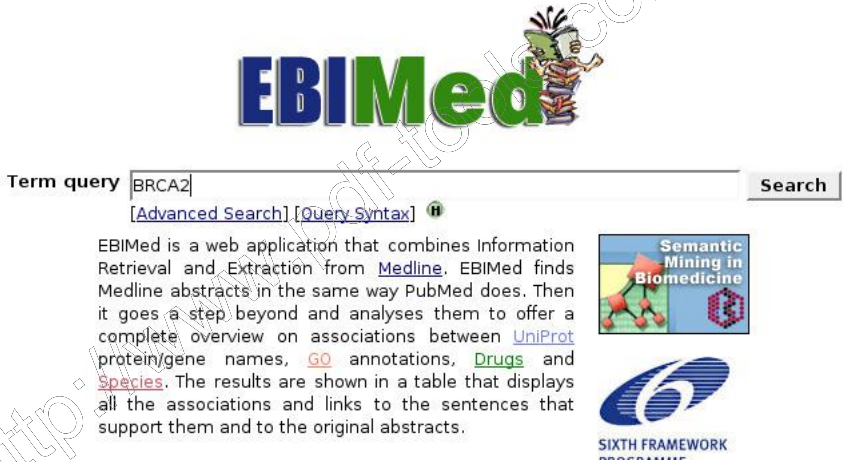
AGAMOUS

PMID	Protein A	Protein B	Protein interaction evidence sentence	Experimental evidence

11689012	VSP1
	AT5G24780
		AGAMOUS
		AT4G18960
			Coimmunoprecipitation, two-hybrid yeast, and affinity column assays show that the FLOR1-VSP1 complex interacts with AGAMOUS and that this transcription factor directly interacts with FLOR1.	immunoprecipitation;
				two-hybrid;
				Coimmunoprecipitation

Figure: Source:http://zope.bioinfo.cnio.es/plan2l/plan2l.html

EBIMed

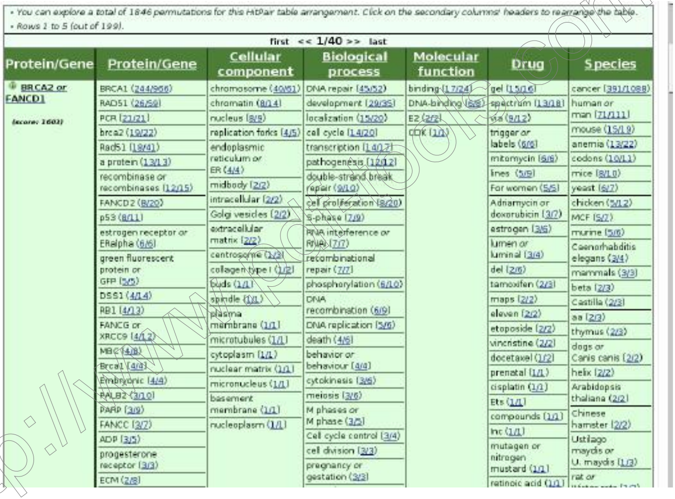


EBIMed
Term query
BRCA2
Search
[Advanced Search] [Query Syntax]
Semantic Mining in Biomedicine
SIXTH FRAMEWORK

EBIMed is a web application that combines Information Retrieval and Extraction from Medline. EBIMed finds Medline abstracts in the same way PubMed does. Then it goes a step beyond and analyses them to offer a complete overview on associations between UniProt protein/gene names, GO annotations, Drugs and Species. The results are shown in a table that displays all the associations and links to the sentences that support them and to the original abstracts.

Figure: Source:http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp

EBIMed



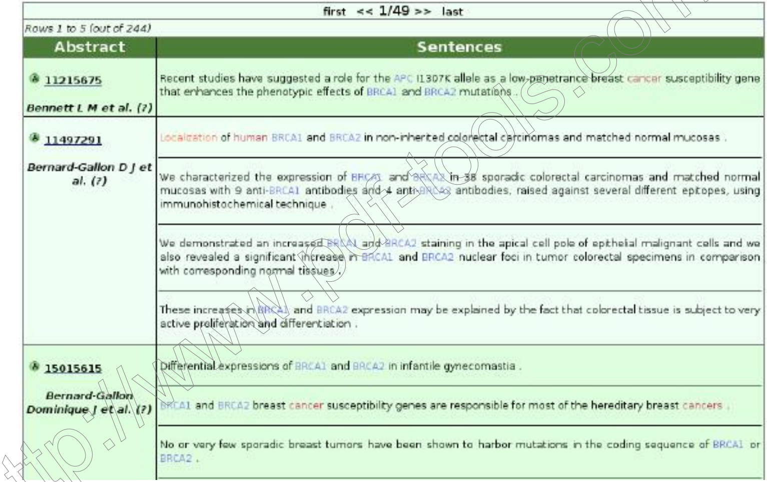
- *You can explore a total of 1846 permutations for this HitPair table arrangement. Click on the secondary columns' headers to rearrange the table.*
- *Rows 1 to 5 (out of 199).*

first << 1/40 >> last

Protein/Gene	Protein/Gene	Cellular component	Biological process	Molecular function	Drug	Species
BRCA2 or FANCD1 *(score: 1602)*	BRCA1 (244/966)	chromosome (40/61)	DNA repair (45/52)	binding (17/24)	gel (15/16)	cancer (391/1088)
	RAD51 (26/59)	chromatin (8/14)	development (29/35)	DNA-binding (6/8)	spectrum (13/18)	human or man (71/111)
	PCR (21/21)	nucleus (8/9)	localization (15/20)	E2 (2/2)	via (9/12)	mouse (15/19)
	brca2 (19/22)	replication forks (4/5)	cell cycle (14/20)	CDK (1/1)	trigger or labels (6/6)	anemia (13/22)
	Rad51 (18/41)	endoplasmic reticulum or ER (4/4)	transcription (14/17)		mitomycin (6/6)	codons (10/11)
	a protein (13/13)	midbody (2/2)	pathogenesis (12/12)		lines (5/9)	mice (8/10)
	recombinase or recombinases (12/15)	intracellular (2/2)	double-strand break repair (9/10)		For women (5/5)	yeast (6/7)
	FANCD2 (8/20)	Golgi vesicles (2/2)	cell proliferation (8/20)		Adriamycin or doxorubicin (3/7)	chicken (5/12)
	p53 (8/11)	extracellular matrix (2/2)	S-phase (7/9)		estrogen (3/5)	MCF (5/7)
	estrogen receptor or ERalpha (6/6)	centrosome (1/3)	RNA interference or RNAi (7/7)		lumen or luminal (3/4)	murine (5/6)
	green fluorescent protein or GFP (5/5)	collagen type I (1/2)	recombinational repair (7/7)		del (2/6)	Caenorhabditis elegans (3/4)
	DSS1 (4/14)	buds (1/1)	phosphorylation (6/10)		tamoxifen (2/3)	mammals (3/3)
	RB1 (4/13)	spindle (1/1)	DNA recombination (6/9)		maps (2/2)	beta (2/3)
	FANCG or XRCC9 (4/12)	plasma membrane (1/1)	DNA replication (5/6)		eleven (2/2)	Castilla (2/3)
	MBC (4/8)	microtubules (1/1)	death (4/6)		etoposide (2/2)	aa (2/3)
	Brca1 (4/4)	cytoplasm (1/1)	behavior or behaviour (4/4)		vincristine (2/2)	thymus (2/3)
	Embryonic (4/4)	nuclear matrix (1/1)	cytokinesis (3/6)		docetaxel (1/2)	dogs or Canis canis (2/2)
	PALB2 (3/10)	micronucleus (1/1)	meiosis (3/6)		prenatal (1/1)	helix (2/2)
	PARP (3/9)	basement membrane (1/1)	M phases or M phase (3/5)		cisplatin (1/1)	Arabidopsis thaliana (2/2)
	FANCC (3/7)	nucleoplasm (1/1)	Cell cycle control (3/4)		Ets (1/1)	Chinese hamster (2/2)
	ADP (3/5)		cell division (3/3)		compounds (1/1)	Ustilago maydis or U. maydis (1/3)
	progesterone receptor (3/3)		pregnancy or gestation (3/3)		Inc (1/1)	rat or
	ECM (2/8)				mutagen or nitrogen mustard (1/1)	
					retinoic acid (1/1)	

Figure: Source:http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp

EBIMed



first << 1/49 >> last

Rows 1 to 5 (out of 244)

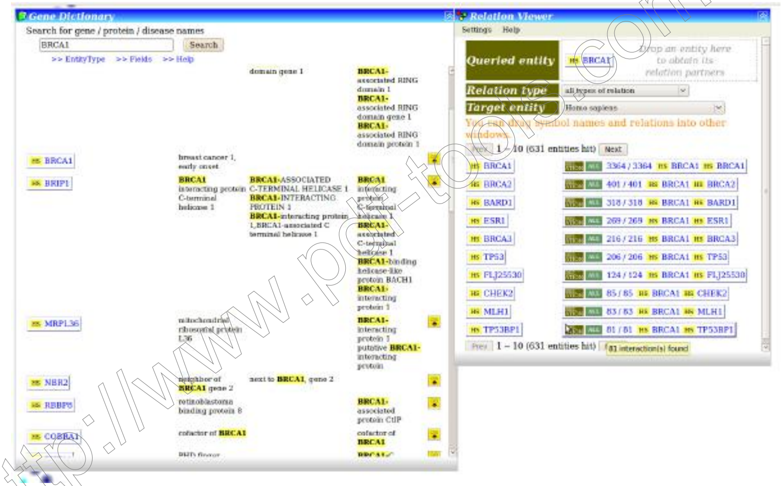
Abstract	Sentences
11215675 *Bennett L M et al. (?)*	Recent studies have suggested a role for the APC I1307K allele as a low-penetrance breast cancer susceptibility gene that enhances the phenotypic effects of BRCA1 and BRCA2 mutations .
11497291 *Bernard-Gallon D J et al. (?)*	Localization of human BRCA1 and BRCA2 in non-inherited colorectal carcinomas and matched normal mucosas .
	We characterized the expression of BRCA1 and BRCA2 in 38 sporadic colorectal carcinomas and matched normal mucosas with 9 anti-BRCA1 antibodies and 4 anti-BRCA2 antibodies, raised against several different epitopes, using immunohistochemical technique .
	We demonstrated an increased BRCA1 and BRCA2 staining in the apical cell pole of epithelial malignant cells and we also revealed a significant increase in BRCA1 and BRCA2 nuclear foci in tumor colorectal specimens in comparison with corresponding normal tissues .
	These increases in BRCA1 and BRCA2 expression may be explained by the fact that colorectal tissue is subject to very active proliferation and differentiation .
15015615 *Bernard-Gallon Dominique J et al. (?)*	Differential expressions of BRCA1 and BRCA2 in infantile gynecomastia .
	BRCA1 and BRCA2 breast cancer susceptibility genes are responsible for most of the hereditary breast cancers .
	No or very few sporadic breast tumors have been shown to harbor mutations in the coding sequence of BRCA1 or BRCA2 .

Figure: Source:http://www.ebi.ac.uk/Rebholz-srv/ebimed/index.jsp

InfoPubMed

- Info-PubMed provides information from Medline on protein-protein interactions.
- Given the name of a gene or protein, it shows a list of the names of other genes/proteins which co-occur in sentences from Medline, along with the frequency of co-occurrence.
- Uses information extraction techniques to identify interacting entities.
- URL: http://www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed/

InfoPubMed



Gene Dictionary

Search for gene / protein / disease names

BRCA1 | Search

>> EntityType >> Fields >> Help

		domain gene 1	BRCA1-associated RING domain 1 BRCA1-associated RING domain gene 1 BRCA1-associated RING domain protein 1	
HS BRCA1	breast cancer 1, early onset			
HS BRIP1	BRCA1 interacting protein C-terminal helicase 1	BRCA1-ASSOCIATED C-TERMINAL HELICASE 1 BRCA1-INTERACTING PROTEIN 1 BRCA1-interacting protein 1,BRCA1-associated C terminal helicase 1	BRCA1 interacting protein C-terminal helicase 1 BRCA1-associated C-terminal helicase 1 BRCA1-binding helicase-like protein BACH1 BRCA1-interacting protein 1	
HS MRPL36	mitochondrial ribosomal protein L36		BRCA1-interacting protein 1 putative BRCA1-interacting protein	
HS NBR2	neighbor of BRCA1 gene 2	next to BRCA1, gene 2		
HS RBBP8	retinoblastoma binding protein 8		BRCA1-associated protein CtIP	
HS COBRA1	cofactor of BRCA1		cofactor of BRCA1	
	PHD finger		BRCA1-	

Relation Viewer

Settings Help

Queried entity: HS BRCA1 | Drop an entity here to obtain its relation partners

Relation type: all types of relation

Target entity: Homo sapiens

You can drag symbol names and relations into other windows.

Prev 1 – 10 (631 entities hit) Next

HS BRCA1	ALL 3364 / 3364 HS BRCA1 HS BRCA1
HS BRCA2	ALL 401 / 401 HS BRCA1 HS BRCA2
HS BARD1	ALL 318 / 318 HS BRCA1 HS BARD1
HS ESR1	ALL 269 / 269 HS BRCA1 HS ESR1
HS BRCA3	ALL 216 / 216 HS BRCA1 HS BRCA3
HS TP53	ALL 206 / 206 HS BRCA1 HS TP53
HS FLJ25530	ALL 124 / 124 HS BRCA1 HS FLJ25530
HS CHEK2	ALL 85 / 85 HS BRCA1 HS CHEK2
HS MLH1	ALL 83 / 83 HS BRCA1 HS MLH1
HS TP53BP1	ALL 81 / 81 HS BRCA1 HS TP53BP1

Prev 1 – 10 (631 entities hit)

81 interaction(s) found

Figure: Source:http://www-tsujii.is.s.u-tokyo.ac.jp/info-pubmed/

PPI Finder: Extraction of Human PPI

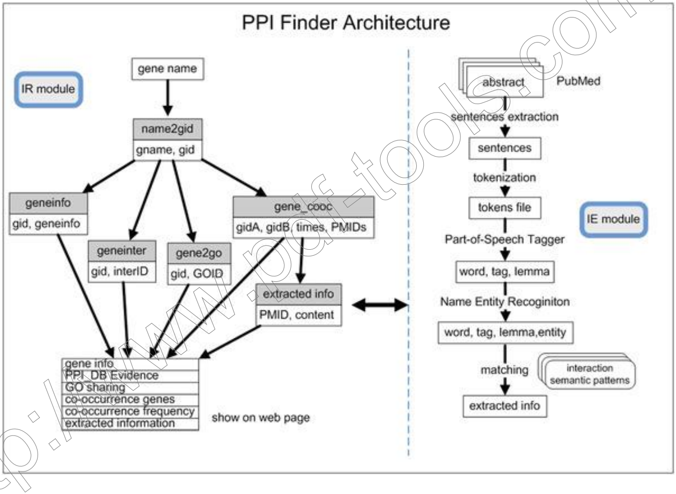


Figure: Source:He et. al. (2009) PLoS ONE 4(2):e4554

STITCH

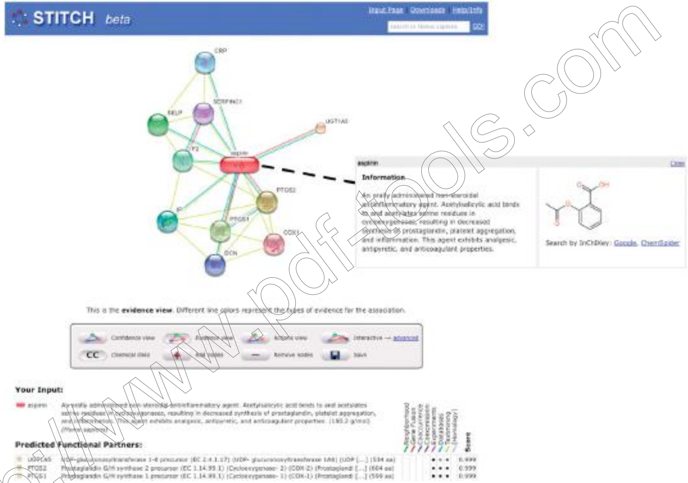
Salient Features

- A resource to explore known and predicted interactions of chemicals and proteins.
- Chemicals are linked to other chemicals and proteins by evidence derived from experiments, databases and the literature.
- Contains interactions for over 74,000 small molecules and over 2.5 million proteins in 630 organisms.

Search Types

- Protein/chemical name
- Chemical structure
- Protein sequence
- Multiple names and sequences

STITCH URL: http://stitch.embl.de/



aspirin

Information

An orally administered non-steroidal antiinflammatory agent. Acetylsalicylic acid binds to and acetylates serine residues in cyclooxygenases, resulting in decreased synthesis of prostaglandin, platelet aggregation, and inflammation. This agent exhibits analgesic, antipyretic, and anticoagulant properties.

Search by InChIKey: Google, ChemSpider

This is the **evidence view**. Different line colors represent the types of evidence for the association.

Your Input:

aspirin An orally administered non-steroidal antiinflammatory agent. Acetylsalicylic acid binds to and acetylates serine residues in cyclooxygenases, resulting in decreased synthesis of prostaglandin, platelet aggregation, and inflammation. This agent exhibits analgesic, antipyretic, and anticoagulant properties. (180.2 g/mol) (Homo sapiens)

Predicted Functional Partners:

		Neighborhood	Gene Fusion	Cooccurrence	Coexpression	Experiments	Databases	Textmining	[Homology]	Score
UGT1A6	UDP-glucuronosyltransferase 1-6 precursor (EC 2.4.1.17) (UDP- glucuronosyltransferase 1A6) (UDP [...] (534 aa)					•	•	•		0.999
PTGS2	Prostaglandin G/H synthase 2 precursor (EC 1.14.99.1) (Cyclooxygenase- 2) (COX-2) (Prostaglandi [...] (604 aa)					•	•	•		0.999
PTGS1	Prostaglandin G/H synthase 1 precursor (EC 1.14.99.1) (Cyclooxygenase- 1) (COX-1) (Prostaglandi [...] (599 aa)					•	•	•		0.999

Figure: Source: Kuhn et. al. (2009) Nucleic Acid Res.:Database Issue

Functional Analysis of Genes and Gene Sets

Objective

Extract information about gene/protein function and gene sets.

Key Techniques

- Rule based
- Dictionary based

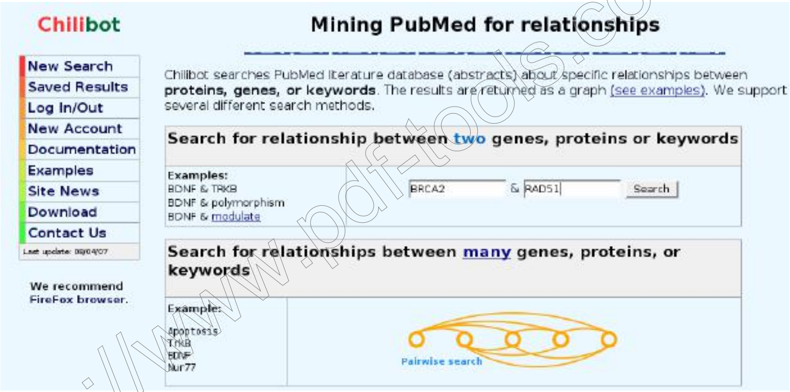
SciMiner

Salient Features

- Web based literature mining and functional analysis tool
- Identifies gene and protein names via context specific analysis of MEDLINE abstract
- Accepts query in form of a list of PubMed IDs or Entrez style free text search
- Scans biomedical literature for gene/protein of user's interest
- Finds significant enrichment in
  - Target gene list
  - GO terms
  - Mesh Terms
  - Pathways
  - PPI Network

http://jdrf.neurology.med.umich.edu/SciMiner/

Chilibot



Mining PubMed for relationships

Chilibot searches PubMed literature database (abstracts) about specific relationships between **proteins, genes, or keywords**. The results are returned as a graph (see examples). We support several different search methods.

Search for relationship between two genes, proteins or keywords

Examples:
BDNF & TRKB
BDNF & polymorphism
BDNF & modulate

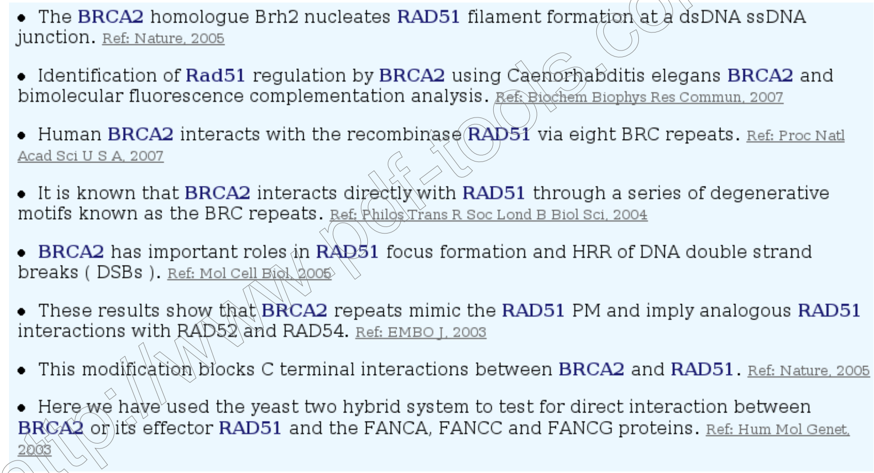
BRCA2 & RAD51

Search for relationships between many genes, proteins, or keywords

Example:
Apoptosis
TrkB
BDNF
Nur77

Figure: Source:http://www.chilibot.net/

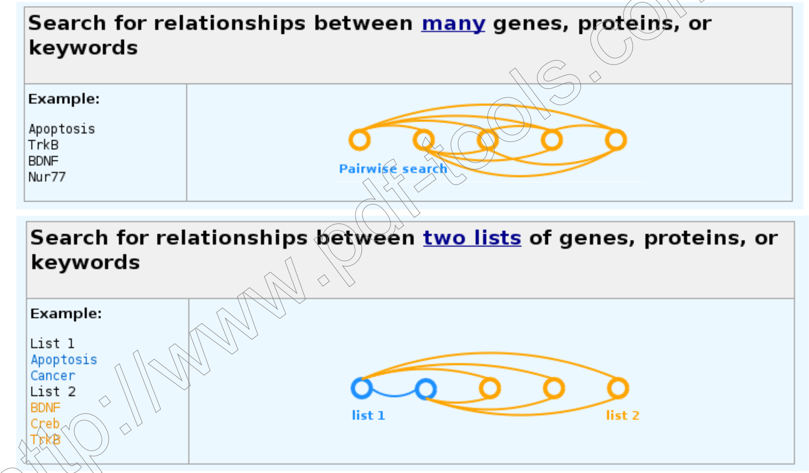
Chilibot



- The **BRCA2** homologue Brh2 nucleates **RAD51** filament formation at a dsDNA ssDNA junction. Ref: Nature, 2005
- Identification of **Rad51** regulation by **BRCA2** using Caenorhabditis elegans **BRCA2** and bimolecular fluorescence complementation analysis. Ref: Biochem Biophys Res Commun, 2007
- Human **BRCA2** interacts with the recombinase **RAD51** via eight BRC repeats. Ref: Proc Natl Acad Sci U S A, 2007
- It is known that **BRCA2** interacts directly with **RAD51** through a series of degenerative motifs known as the BRC repeats. Ref: Philos Trans R Soc Lond B Biol Sci, 2004
- **BRCA2** has important roles in **RAD51** focus formation and HRR of DNA double strand breaks (DSBs). Ref: Mol Cell Biol, 2005
- These results show that **BRCA2** repeats mimic the **RAD51** PM and imply analogous **RAD51** interactions with RAD52 and RAD54. Ref: EMBO J, 2003
- This modification blocks C terminal interactions between **BRCA2** and **RAD51**. Ref: Nature, 2005
- Here we have used the yeast two hybrid system to test for direct interaction between **BRCA2** or its effector **RAD51** and the FANCA, FANCC and FANCG proteins. Ref: Hum Mol Genet, 2003

Figure: Source:http://www.chilibot.net/

Chilibot



Search for relationships between many genes, proteins, or keywords

Example:

Apoptosis
TrkB
BDNF
Nur77

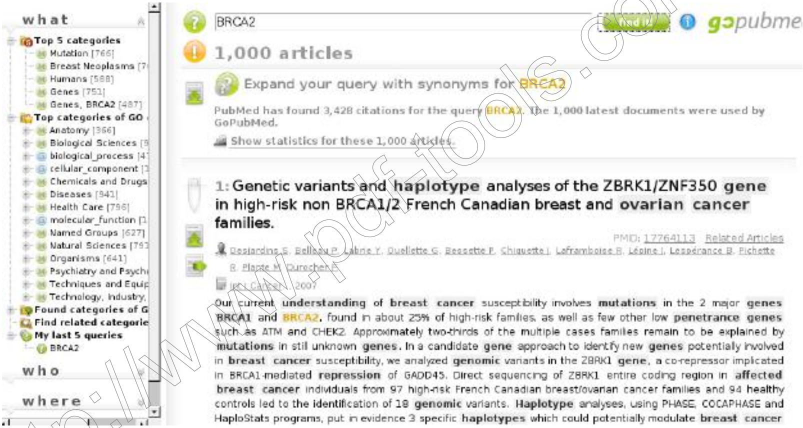
Search for relationships between two lists of genes, proteins, or keywords

Example:

List 1
Apoptosis
Cancer
List 2
BDNF
Creb
TrkB

Figure: Source:http://www.chilibot.net/

GoPubmed



1,000 articles

Expand your query with synonyms for BRCA2

PubMed has found 3,428 citations for the query BRCA2. The 1,000 latest documents were used by GoPubMed.

Show statistics for these 1,000 articles.

1: Genetic variants and haplotype analyses of the ZBRK1/ZNF350 gene in high-risk non BRCA1/2 French Canadian breast and ovarian cancer families.

PMID: 17764113 Related Articles

Desjardins S, Belleau P, Labrie Y, Ouellette G, Bessette P, Chiquette J, Laframboise R, Lépine J, Lespérance B, Pichette R, Plante M, Durocher F.

Int J Cancer. 2007

Our current understanding of breast cancer susceptibility involves mutations in the 2 major genes BRCA1 and BRCA2, found in about 25% of high-risk families, as well as few other low penetrance genes such as ATM and CHEK2. Approximately two-thirds of the multiple cases families remain to be explained by mutations in still unknown genes. In a candidate gene approach to identify new genes potentially involved in breast cancer susceptibility, we analyzed genomic variants in the ZBRK1 gene, a co-repressor implicated in BRCA1-mediated repression of GADD45. Direct sequencing of ZBRK1 entire coding region in affected breast cancer individuals from 97 high-risk French Canadian breast/ovarian cancer families and 94 healthy controls led to the identification of 18 genomic variants. Haplotype analyses, using PHASE, COCAPHASE and HaploStats programs, put in evidence 3 specific haplotypes which could potentially modulate breast cancer

Figure: Source: http://www.gopubmed.org/

Extraction of Gene-Disease Association

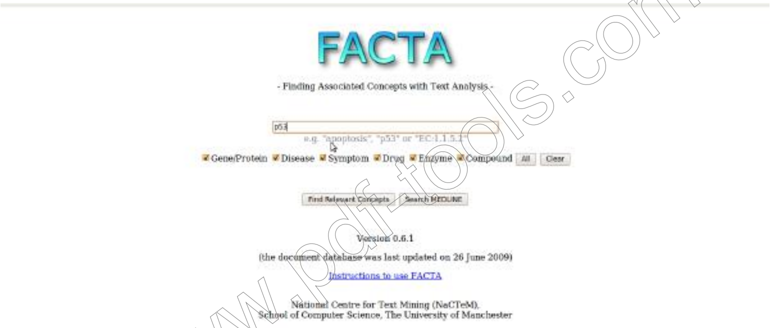
Objective

Given a biomedical research article/abstract, extract gene disease associations.

Tools

- Facta: Finding Associated Concepts with Text Analytics

Facta Query



FACTA

- Finding Associated Concepts with Text Analysis -

p53

e.g. "apoptosis", "p53" or "EC:1.1.5.1"

☑ Gene/Protein ☑ Disease ☑ Symptom ☑ Drug ☑ Enzyme ☑ Compound

Version 0.6.1

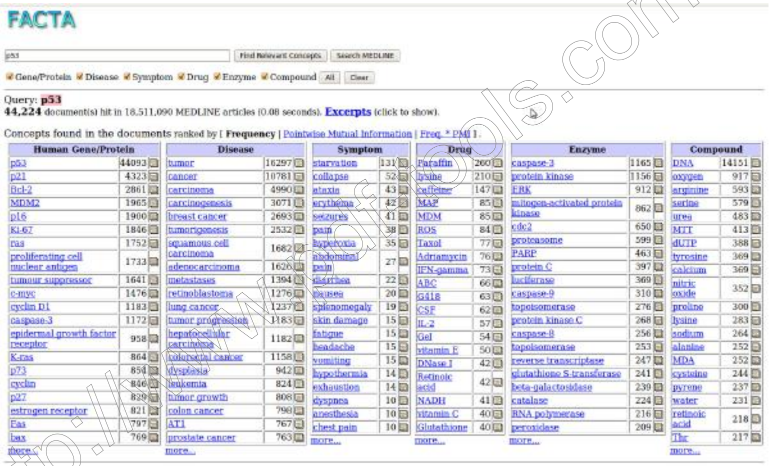
(the document database was last updated on 26 June 2009)

Instructions to use FACTA

National Centre for Text Mining (NaCTeM),
School of Computer Science, The University of Manchester

Figure: Source:http://text0.mib.man.ac.uk/software/facta/main.html

Facta Relevant Concepts



FACTA

p53

Find Relevant Concepts
Search MEDLINE
☑ Gene/Protein ☑ Disease ☑ Symptom ☑ Drug ☑ Enzyme ☑ Compound
All
Clear

Query: **p53**

44,224 document(s) hit in 18,511,090 MEDLINE articles (0.08 seconds). **Excerpts** (click to show).

Concepts found in the documents ranked by [**Frequency** | Pointwise Mutual Information | Freq. * PMI].

Human Gene/Protein	
p53	44093
p21	4323
Bcl-2	2861
MDM2	1965
p16	1900
Ki-67	1846
ras	1752
proliferating cell nuclear antigen	1733
tumour suppressor	1641
c-myc	1476
cyclin D1	1183
caspase-3	1172
epidermal growth factor receptor	958
K-ras	864
p73	854
cyclin	846
p27	829
estrogen receptor	821
Fas	797
bax	769

more...

Disease	
tumor	16297
cancer	10781
carcinoma	4990
carcinogenesis	3071
breast cancer	2693
tumorigenesis	2532
squamous cell carcinoma	1682
adenocarcinoma	1626
metastases	1394
retinoblastoma	1276
lung cancer	1237
tumor progression	1183
hepatocellular carcinoma	1182
colorectal cancer	1158
dysplasia	942
leukemia	824
tumor growth	808
colon cancer	798
AT1	767
prostate cancer	763

more...

Symptom	
starvation	131
collapse	52
ataxia	43
erythema	42
seizures	41
pain	38
hyperoxia	35
abdominal pain	27
diarrhea	22
nausea	20
splenomegaly	19
skin damage	15
fatigue	15
headache	15
vomiting	15
hypothermia	14
exhaustion	14
dyspnea	10
anesthesia	10
chest pain	10

more...

Drug	
Paraffin	260
lysine	210
caffeine	147
MAP	85
MDM	85
ROS	84
Taxol	77
Adriamycin	76
IFN-gamma	73
ABC	66
G418	63
CSF	62
IL-2	57
Gel	54
vitamin E	50
DNase I	42
Retinoic acid	42
NADH	41
vitamin C	40
Glutathione	40

more...

Enzyme	
caspase-3	1165
protein kinase	1156
ERK	912
mitogen-activated protein kinase	862
cdc2	650
proteasome	599
PARP	463
protein C	397
luciferase	369
caspase-9	310
topoisomerase	276
protein kinase C	268
caspase-8	256
topoisomerase	253
reverse transcriptase	247
glutathione S-transferase	241
beta-galactosidase	239
catalase	224
RNA polymerase	216
peroxidase	209

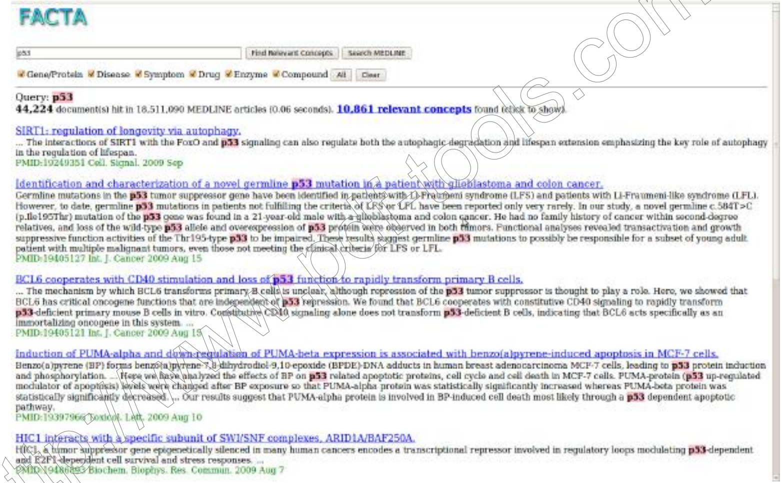
more...

Compound	
DNA	14151
oxygen	917
arginine	593
serine	579
urea	483
MTT	413
dUTP	388
tyrosine	369
calcium	369
nitric oxide	352
proline	300
lysine	283
sodium	264
alanine	252
MDA	252
cysteine	244
pyrene	237
water	231
retinoic acid	218
The	217

more...

Figure: Source:http://text0.mib.man.ac.uk/software/facta/main.html

Facta Medline Search


FACTA
p53
Find Relevant Concepts
Search MEDLINE
☑ Gene/Protein ☑ Disease ☑ Symptom ☑ Drug ☑ Enzyme ☑ Compound
All
Clear

Query: **p53**
44,224 document(s) hit in 18,511,090 MEDLINE articles (0.06 seconds). **10,861 relevant concepts** found (click to show).

SIRT1: regulation of longevity via autophagy.
... The interactions of SIRT1 with the FoxO and **p53** signaling can also regulate both the autophagic degradation and lifespan extension emphasizing the key role of autophagy in the regulation of lifespan.
PMID:19249351 Cell. Signal. 2009 Sep

Identification and characterization of a novel germline **p53** mutation in a patient with glioblastoma and colon cancer.
Germline mutations in the **p53** tumor suppressor gene have been identified in patients with Li-Fraumeni syndrome (LFS) and patients with Li-Fraumeni-like syndrome (LFL). However, to date, germline **p53** mutations in patients not fulfilling the criteria of LFS or LFL have been reported only very rarely. In our study, a novel germline c.584T>C (p.Ile195Thr) mutation of the **p53** gene was found in a 21-year-old male with a glioblastoma and colon cancer. He had no family history of cancer within second-degree relatives, and loss of the wild-type **p53** allele and overexpression of **p53** protein were observed in both tumors. Functional analyses revealed transactivation and growth suppressive function activities of the Thr195-type **p53** to be impaired. These results suggest germline **p53** mutations to possibly be responsible for a subset of young adult patient with multiple malignant tumors, even those not meeting the clinical criteria for LFS or LFL.
PMID:19405127 Int. J. Cancer 2009 Aug 15

BCL6 cooperates with CD40 stimulation and loss of **p53** function to rapidly transform primary B cells.
... The mechanism by which BCL6 transforms primary B cells is unclear, although repression of the **p53** tumor suppressor is thought to play a role. Here, we showed that BCL6 has critical oncogene functions that are independent of **p53** repression. We found that BCL6 cooperates with constitutive CD40 signaling to rapidly transform **p53**-deficient primary mouse B cells in vitro. Constitutive CD40 signaling alone does not transform **p53**-deficient B cells, indicating that BCL6 acts specifically as an immortalizing oncogene in this system. ...
PMID:19405121 Int. J. Cancer 2009 Aug 15

Induction of PUMA-alpha and down-regulation of PUMA-beta expression is associated with benzo(a)pyrene-induced apoptosis in MCF-7 cells.
Benzo(a)pyrene (BP) forms benzo(a)pyrene-7,8-dihydrodiol-9,10-epoxide (BPDE)-DNA adducts in human breast adenocarcinoma MCF-7 cells, leading to **p53** protein induction and phosphorylation. ... Here we have analyzed the effects of BP on **p53** related apoptotic proteins, cell cycle and cell death in MCF-7 cells. PUMA-protein (**p53** up-regulated modulator of apoptosis) levels were changed after BP exposure so that PUMA-alpha protein was statistically significantly increased whereas PUMA-beta protein was statistically significantly decreased. ... Our results suggest that PUMA-alpha protein is involved in BP-induced cell death most likely through a **p53** dependent apoptotic pathway.
PMID:19397966 Toxicol. Lett. 2009 Aug 10

HIC1 interacts with a specific subunit of SWI/SNF complexes, ARID1A/BAF250A.
HIC1, a tumor suppressor gene epigenetically silenced in many human cancers encodes a transcriptional repressor involved in regulatory loops modulating **p53**-dependent and E2F1-dependent cell survival and stress responses. ...
PMID:19486893 Biochem. Biophys. Res. Commun. 2009 Aug 7

Figure: Source:http://text0.mib.man.ac.uk/software/facta/main.html

Extraction of Mutations and Epigenetic Characteristics

Objective

Given a biomedical research article/abstract, extract mutation and epigenetic characteristics.

Challenges

- Deluge of experimental data from high throughput screens such as microarray and RNAi
- Traditionally hundreds of genes are clustered via enrichment of GO terms.
- Analysis results in GO annotations and not always available for all genes in the model organism
- Deluge of literature data calls for high through put techniques to extract mutations and epigenetic characteristics

GoGene

Salient Features

- Extract co-occurrences of genes and ontology terms from literature
- Combines disease, compounds, techniques, and mutations information
- Claims to provide most recent facts about genes and rank them according to novelty and importance

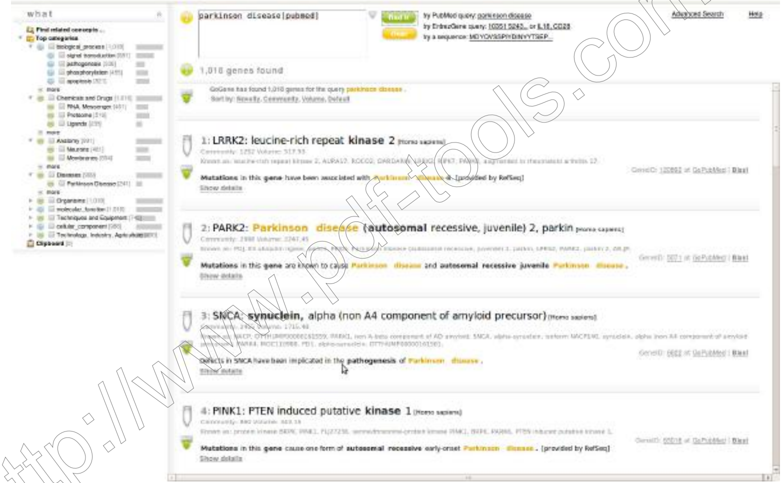
No. of Associations	4,000,000
No. of Model Organisms	10
No. of PubMed Articles	18,000,000

Query Types

PubMed, Entrez, Sequence

http://gopubmed2.biotec.tu-dresden.de/gogene/gogene/

GoGene PubMed Query



parkinson disease[pubmed]

try PubMed query: parkinson disease
try EntrezGene query: 10351 5243... or IL18, CD28
try a sequence: MDYOVSSPIYDINYYTSEP...

1,018 genes found

GoGene has found 1,018 genes for the query parkinson disease .
Sort by: Novelty, Community, Volume, Default

1: LRRK2: leucine-rich repeat kinase 2 (Homo sapiens)
Community: 1252 Volume: 317.93
Known as: leucine-rich repeat kinase 2, AURA17, ROCO2, DARDARIN, LRRK2, RIPK7, PARK8, augmented in rheumatoid arthritis 17
Mutations in this gene have been associated with Parkinson disease 8. [provided by RefSeq]
Show details
GeneID: 120892 at GoPubMed | Blast

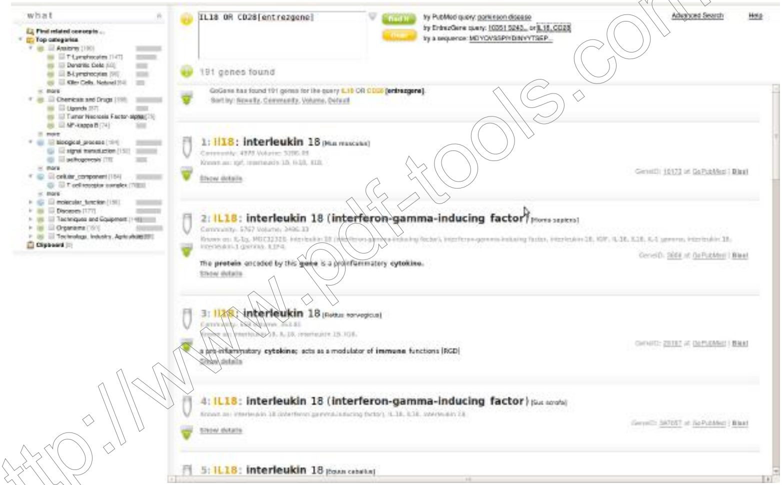
2: PARK2: Parkinson disease (autosomal recessive, juvenile) 2, parkin (Homo sapiens)
Community: 2988 Volume: 2747.45
Known as: PDJ, E3 ubiquitin ligase, parkin, PRKN, Parkinson disease (autosomal recessive, juvenile) 2, parkin, LPRS2, PARK2, parkin 2, AR-JP
Mutations in this gene are known to cause Parkinson disease and autosomal recessive juvenile Parkinson disease .
Show details
GeneID: 5071 at GoPubMed | Blast

3: SNCA: synuclein, alpha (non A4 component of amyloid precursor) (Homo sapiens)
Community: 2455 Volume: 1715.48
Known as: NACP, OTTHUMP00000161559, PARK1, non A-beta component of AD amyloid, SNCA, alpha-synuclein, isoform NACP140, synuclein, alpha (non A4 component of amyloid precursor), PARK4, MGC110988, PD1, alpha-synuclein, OTTHUMP00000161560
Defects in SNCA have been implicated in the pathogenesis of Parkinson disease .
Show details
GeneID: 6622 at GoPubMed | Blast

4: PINK1: PTEN induced putative kinase 1 (Homo sapiens)
Community: 880 Volume: 463.18
Known as: protein kinase BRPK, PINK1, FLJ27236, serine/threonine-protein kinase PINK1, BRPK, PARK6, PTEN induced putative kinase 1
Mutations in this gene cause one form of autosomal recessive early-onset Parkinson disease . [provided by RefSeq]
Show details
GeneID: 65018 at GoPubMed | Blast

Figure: Source:http://gopubmed2.biotec.tu-dresden.de/gogene/gogene/

GoGene Entrez Query



what

Find related concepts ...

Top categories

- Anatomy [190]
  - T-Lymphocytes [147]
  - Dendritic Cells [40]
  - B-Lymphocytes [90]
  - Killer Cells, Natural [64]
  - more
- Chemicals and Drugs [199]
  - Ligands [87]
  - Tumor Necrosis Factor-alpha [73]
  - NF-kappa B [74]
  - more
- biological_process [184]
  - signal transduction [150]
  - pathogenesis [78]
  - more
- cellular_component [184]
  - T cell receptor complex [70]
  - more
- molecular_function [186]
- Diseases [177]
- Techniques and Equipment [148]
- Organisms [191]
- Technology, Industry, Agriculture [81]

Clipboard [0]

`IL18 OR CD28[entrezgene]`

Find it

Clear

try PubMed query: parkinson disease

try EntrezGene query: 10351 5243... or IL18, CD28

try a sequence: MDYOVSSPIYDINYYTSEP...

Advanced Search

Help

191 genes found

GoGene has found 191 genes for the query IL18 OR CD28 [entrezgene].

Sort by: Novelty, Community, Volume, Default

1: Il18: interleukin 18 [Mus musculus]

Community: 4978 Volume: 3295.88

Known as: Igif, interleukin 18, Il-18, Il18.

Show details

GeneID: 16173 at GoPubMed | Blast

2: IL18: interleukin 18 (interferon-gamma-inducing factor) [Homo sapiens]

Community: 5767 Volume: 3496.33

Known as: IL-1g, MGC12320, interleukin 18 (interferon-gamma-inducing factor), interferon-gamma-inducing factor, interleukin-18, IGIF, IL 18, IL18, IL-1 gamma, interleukin 18, interleukin-1 gamma, IL1F4.

GeneID: 3606 at GoPubMed | Blast

The **protein** encoded by this **gene** is a proinflammatory **cytokine**.

Show details

3: Il18: interleukin 18 [Rattus norvegicus]

Community: 659 Volume: 363.81

Known as: interleukin 18, IL 18, interleukin 18, IGIF.

GeneID: 29197 at GoPubMed | Blast

a pro-inflammatory **cytokine**; acts as a modulator of **immune** functions [RGD]

Show details

4: IL18: interleukin 18 (interferon-gamma-inducing factor) [Sus scrofa]

Known as: interleukin 18 (interferon gamma-inducing factor), IL 18, IL18, interleukin 18

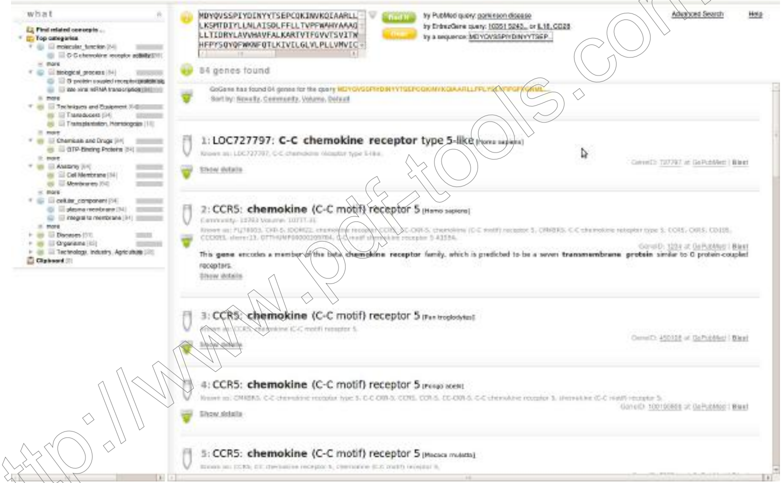
GeneID: 397057 at GoPubMed | Blast

Show details

5: IL18: interleukin 18 [Equus caballus]

Figure: Source:http://gopubmed2.biotec.tu-dresden.de/gogene/gogene/

GoGene Sequence Query



what

Find related concepts ...

Top categories

- molecular_function [34]
  - C-C chemokine receptor activity [34]
  - more
- biological_process [34]
  - G-protein coupled receptor protein sig
  - anti-viral mRNA transcription [34]
  - more
- Techniques and Equipment [34]
  - Transducers [34]
  - Transplantation, Homologous [18]
  - more
- Chemicals and Drugs [34]
  - GTP-Binding Proteins [34]
  - more
- Anatomy [34]
  - Cell Membrane [34]
  - Membranes [34]
  - more
- cellular_component [34]
  - plasma membrane [34]
  - integral to membrane [34]
  - more
- Diseases [17]
- Organisms [33]
- Technology, Industry, Agriculture [20]

Clipboard [2]

```
MDYQVSSPIYDINYYTSEPCQKINVKQIAARLL
LKSMTDIYLLNLAISDLFFLLTVPFWAHYAAAQ
LLTIDRYLAVVHAVFALKARTVTFGVVTSVITW
HFPYSQYQFWKNFQTLKIVILGLVLPLLVMVIC
```

try PubMed query: parkinson disease
try EntrezGene query: 10351 5243... or IL18, CD28
try a sequence: MDYQVSSPIYDINYYTSEP...

34 genes found

GoGene has found 34 genes for the query MDYQVSSPIYDINYYTSEPCQKINVKQIAARLLPPLYSLVFIFGFVGNML...

Sort by: Novelty, Community, Volume, Default

1: LOC727797: **C-C chemokine receptor** type 5-like [Homo sapiens]

Known as: LOC727797, C-C chemokine receptor type 5-like.

GeneID: 727797 at GoPubMed | Blast

Show details

2: CCR5: **chemokine** (C-C motif) **receptor** 5 [Homo sapiens]

Community: 13793 Volume: 10777.31

Known as: FLJ78003, CKR-5, IDDM22, chemokine receptor CCR5, CC-CKR-5, chemokine (C-C motif) receptor 5, CMKBR5, C-C chemokine receptor type 5, CCR5, CKR5, CD195, CCCKR5, chemr13, OTTHUMP00000209784, C-C motif chemokine receptor 5 A3556.

GeneID: 1234 at GoPubMed | Blast

This **gene** encodes a member of the beta **chemokine receptor** family, which is predicted to be a seven **transmembrane protein** similar to G protein-coupled receptors.

Show details

3: CCR5: **chemokine** (C-C motif) receptor 5 [Pan troglodytes]

Known as: CCR5, chemokine (C-C motif) receptor 5.

GeneID: 450128 at GoPubMed | Blast

Show details

4: CCR5: **chemokine** (C-C motif) receptor 5 [Pongo abelii]

Known as: CMKBR5, C-C chemokine receptor type 5, C-C CKR-5, CCR5, CCR-5, CC-CKR-5, C-C chemokine receptor 5, chemokine (C-C motif) receptor 5.

GeneID: 100190868 at GoPubMed | Blast

Show details

5: CCR5: **chemokine** (C-C motif) receptor 5 [Macaca mulatta]

Known as: CCR5, CC chemokine receptor 5, chemokine (C-C motif) receptor 5.

Figure: Source:http://gopubmed2.biotec.tu-dresden.de/gogene/gogene/

MeInfoText: Salient Features

- Extracts associated gene mythylation and cancer information from biomedical text and integrates it with biological pathways and protein protein interaction.
- DNA methylation, occurring predominantly in CpG islands, is an important epigenetic modification of the genome that is involved in mediating various cellular processes (Robertson, 2005).
- Abnormal methylation of DNA may result in increased transcription of oncogenes or silencing of tumor suppressor genes and is common in a variety of human cancer cells (Esteller, 2005).

MeInfoText: Search Types

- Associations among gene, methylation and cancer
- Gene methylation associations
- Profile of gene methylation across human cancer types
- Gene methylation of a specific cancer type
- http://mit.lifescience.ntu.edu.tw/

Extraction of Protein Location from Literature

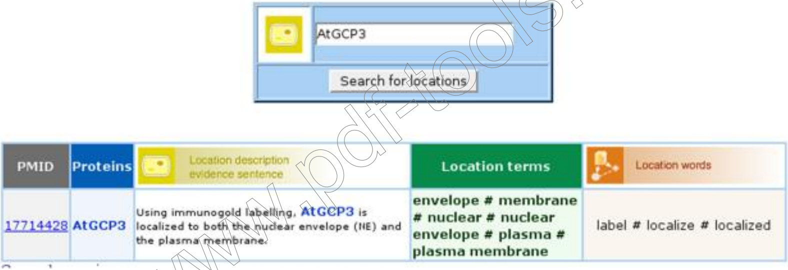
Objective

Given a biomedical research article/abstract, extract protein localization information

Significance

The role of proteins in biochemical reactions depends on its location.

Location extraction from PLAN2L



AtGCP3

PMID	Proteins	Location description evidence sentence	Location terms	Location words
17714428	AtGCP3	Using immunogold labelling, AtGCP3 is localized to both the nuclear envelope (NE) and the plasma membrane.	envelope # membrane # nuclear # nuclear envelope # plasma # plasma membrane	label # localize # localized

Figure: Source:http://zope.bioinfo.cnio.es/plan2l/plan2l.html

Automatic Construction of Lexical Resources

- Automatic construct lexical resource such as dictionaries, thesauri from biomedical literature
- Rapidly expanding biomedical literature makes it hard to manually maintain dictionaries and thesauri
- High throughput and automated methods are required for creation of lexical resources
- Highly domain specific and virtually open terminology used in biomedical domain makes it challenging to extract appropriate lexical resources

Termine

- Uses C-value term extraction and Acromine acronym recognition
- Statistical analysis assigns termhood to a candidate term using the following characteristics:
  - occurrence frequency of the candidate term
  - frequency of the candidate term as part of other longer candidate terms
  - number of these longer candidate terms
  - length of the candidate term
- URL: http://www.nactem.ac.uk/software/termine/

Acronym Finder

- Abbreviation is a short form of a word or a phrase.
- Identification of correct abbreviation and its long form pair is crucial for IR and IE applications
- Fast rate of growth of biomedical literature makes it hard for thesauruses to keep track of all abbreviations
- Abbreviations are of two types:
  - Acronyms: Word formed by initial letters or letters of each successive parts or major parts of long form. E.g. CKB stands for Brain Creatin Kinase
  - Non-acronyms do not follow lexical patterns with long forms. E.g. 11p stands for the short arm of chromosome 11

ADAM: Another Database of Abbreviations in MEDLINE

- URL: http://128.248.65.210/arrowsmith_uic/adam.html
- Input: Short form or long form
- Output: Corresponding long form or short form

MedlineRanker

- Flexible ranking system for MEDLINE abstract
- Given an abstract related to a specific topic, MedlineRanker returns the most discriminative words in comparison with a random selection
- These words are used to score other abstracts.
- URL: http://cbdm.mdc-berlin.de/tools/medlineranker

Part 4: Issues and Challenges in Evaluation of Text Mining Systems

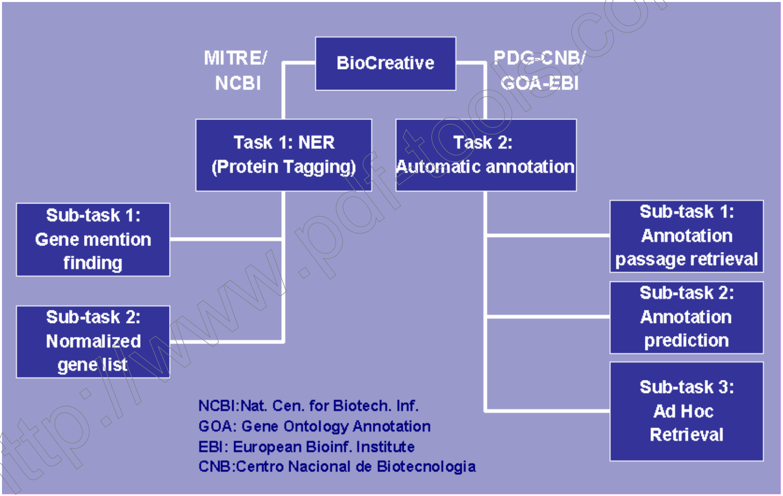
Why Community Assessment?

- Compare different methods and strategies
- Reproduce performance of systems on common data
- Provide useful data collections: Gold Standard data
- Explore meaningful evaluation strategies and tools
- Determine the state of the art
- Monitor improvements in the field
- Point out needs of the user community
- Promote collaborative efforts

BioCreative Challenge

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

BioCreative Challenge 1



BioCreative

MITRE/ NCBI

PDG-CNB/ GOA-EBI

Task 1: NER (Protein Tagging)

Sub-task 1: Gene mention finding

Sub-task 2: Normalized gene list

Task 2: Automatic annotation

Sub-task 1: Annotation passage retrieval

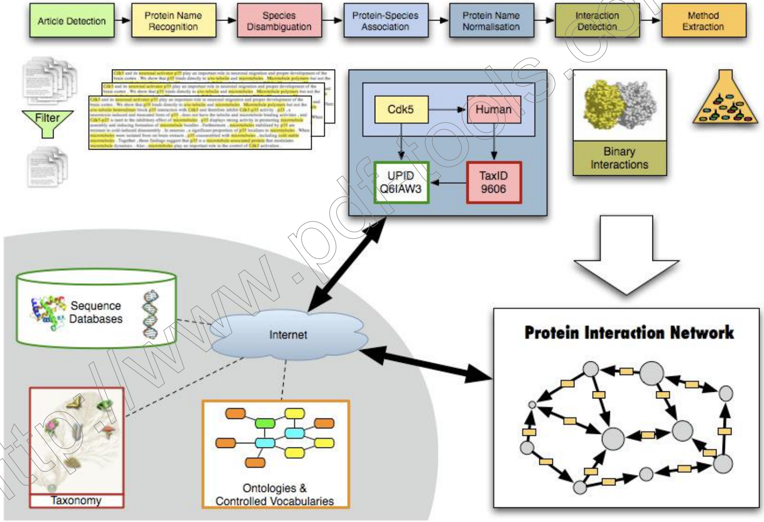
Sub-task 2: Annotation prediction

Sub-task 3: Ad Hoc Retrieval

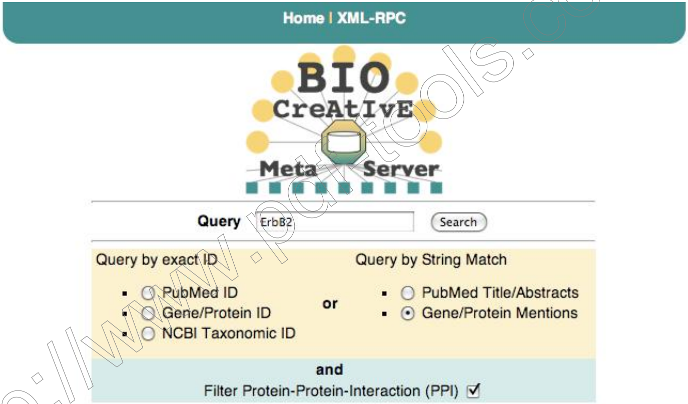
NCBI:Nat. Cen. for Biotech. Inf.
GOA: Gene Ontology Annotation
EBI: European Bioinf. Institute
CNB:Centro Nacional de Biotecnologia

BioCreative Challenge 2

BioCreative Challenge 2.5



BioCreative Meta-Server (BCMS)



Query ErbB2

Query by exact ID

- ○ PubMed ID
- ○ Gene/Protein ID
- ○ NCBI Taxonomic ID

or

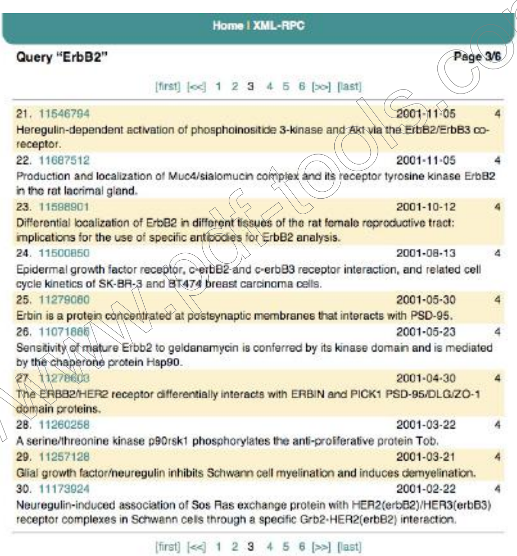
Query by String Match

- ○ PubMed Title/Abstracts
- ● Gene/Protein Mentions

and

Filter Protein-Protein-Interaction (PPI) ☑

BioCreative Meta-Server (BCMS)



Query "ErbB2" Page 3/6

[first] [<<] 1 2 3 4 5 6 [>>] [last]

21. 11546794 2001-11-05 4
Heregulin-dependent activation of phosphoinositide 3-kinase and Akt via the ErbB2/ErbB3 co-receptor.

22. 11687512 2001-11-05 4
Production and localization of Muc4/sialomucin complex and its receptor tyrosine kinase ErbB2 in the rat lacrimal gland.

23. 11598901 2001-10-12 4
Differential localization of ErbB2 in different tissues of the rat female reproductive tract: implications for the use of specific antibodies for ErbB2 analysis.

24. 11500850 2001-08-13 4
Epidermal growth factor receptor, c-erbB2 and c-erbB3 receptor interaction, and related cell cycle kinetics of SK-BR-3 and BT474 breast carcinoma cells.

25. 11279080 2001-05-30 4
Erbin is a protein concentrated at postsynaptic membranes that interacts with PSD-95.

26. 11071886 2001-05-23 4
Sensitivity of mature Erbb2 to geldanamycin is conferred by its kinase domain and is mediated by the chaperone protein Hsp90.

27. 11278603 2001-04-30 4
The ERBB2/HER2 receptor differentially interacts with ERBIN and PICK1 PSD-95/DLG/ZO-1 domain proteins.

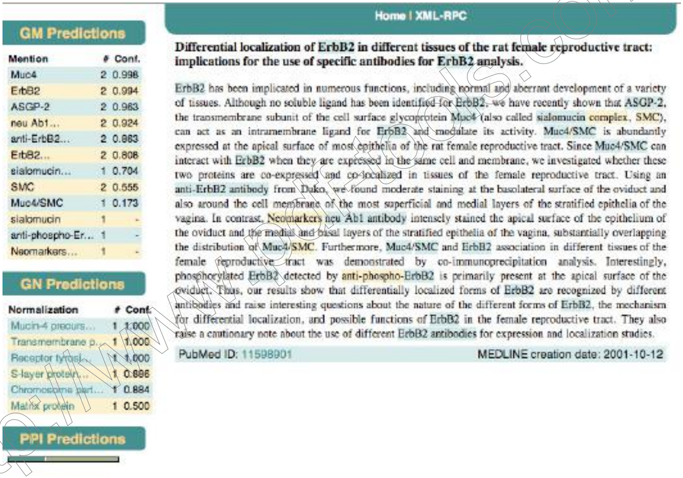
28. 11260258 2001-03-22 4
A serine/threonine kinase p90rsk1 phosphorylates the anti-proliferative protein Tob.

29. 11257128 2001-03-21 4
Glial growth factor/neuregulin inhibits Schwann cell myelination and induces demyelination.

30. 11173924 2001-02-22 4
Neuregulin-induced association of Sos Ras exchange protein with HER2(erbB2)/HER3(erbB3) receptor complexes in Schwann cells through a specific Grb2-HER2(erbB2) interaction.

[first] [<<] 1 2 3 4 5 6 [>>] [last]

BioCreative Meta-Server (BCMS)



GM Predictions

Mention	#	Conf.
Muc4	2	0.998
ErbB2	2	0.994
ASGP-2	2	0.963
neu Ab1...	2	0.924
anti-ErbB2...	2	0.863
ErbB2...	2	0.808
sialomucin...	1	0.704
SMC	2	0.555
Muc4/SMC	1	0.173
sialomucin	1	-
anti-phospho-Er...	1	-
Neomarkers...	1	-

GN Predictions

Normalization	#	Conf.
Mucin-4 precurs...	1	1.000
Transmembrane p...	1	1.000
Receptor tyrosi...	1	1.000
S-layer protein...	1	0.896
Chromosome part...	1	0.884
Matrix protein	1	0.500

PPI Predictions

Differential localization of ErbB2 in different tissues of the rat female reproductive tract: implications for the use of specific antibodies for ErbB2 analysis.

ErbB2 has been implicated in numerous functions, including normal and aberrant development of a variety of tissues. Although no soluble ligand has been identified for ErbB2, we have recently shown that ASGP-2, the transmembrane subunit of the cell surface glycoprotein Muc4 (also called sialomucin complex, SMC), can act as an intramembrane ligand for ErbB2 and modulate its activity. Muc4/SMC is abundantly expressed at the apical surface of most epithelia of the rat female reproductive tract. Since Muc4/SMC can interact with ErbB2 when they are expressed in the same cell and membrane, we investigated whether these two proteins are co-expressed and co-localized in tissues of the female reproductive tract. Using an anti-ErbB2 antibody from Dako, we found moderate staining at the basolateral surface of the oviduct and also around the cell membrane of the most superficial and medial layers of the stratified epithelia of the vagina. In contrast, Neomarkers neu Ab1 antibody intensely stained the apical surface of the epithelium of the oviduct and the medial and basal layers of the stratified epithelia of the vagina, substantially overlapping the distribution of Muc4/SMC. Furthermore, Muc4/SMC and ErbB2 association in different tissues of the female reproductive tract was demonstrated by co-immunoprecipitation analysis. Interestingly, phosphorylated ErbB2 detected by anti-phospho-ErbB2 is primarily present at the apical surface of the oviduct. Thus, our results show that differentially localized forms of ErbB2 are recognized by different antibodies and raise interesting questions about the nature of the different forms of ErbB2, the mechanism for differential localization, and possible functions of ErbB2 in the female reproductive tract. They also raise a cautionary note about the use of different ErbB2 antibodies for expression and localization studies.

PubMed ID: 11598901 MEDLINE creation date: 2001-10-12

BioCreative Conclusion

- Repeatability of experimental results
- Comparability of the experimental results
- Take into account potential user community: Biologists and Interaction databases
- Estimate how hard the task actually is and the quality of the training data: Inter-annotator agreement, e.g. kappa score (will be done as well, GB article)
- Evaluate also sub-aspects: bio-entity, functional term, relationships, sub-categories, organism source, sampling

Part 5: Practical Case Studies

PLAN2L: Flowchart

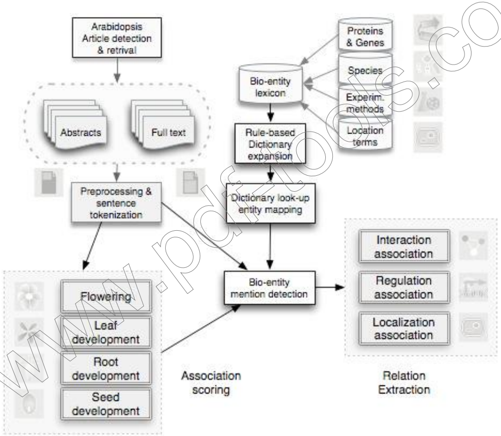


Figure: Source:http://zope.bioinfo.cnio.es/plan2l/plan2l.html

PLAN2L: Protein Normalization

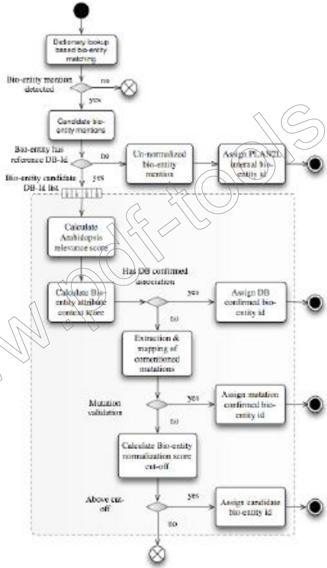
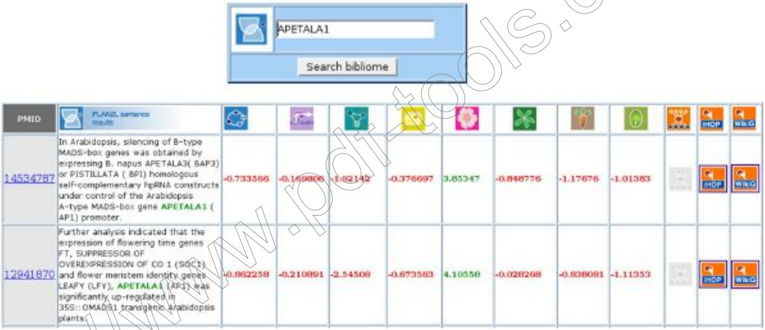

Dictionary lookup based bio-entity matching
Bio-entity mention detected
no
yes
Candidate bio-entity mentions
Bio-entity has reference DB-Id
no
Un-normalized bio-entity mention
Assign PLAN2L internal bio-entity id
Bio-entity candidate DB-Id list
yes
Calculate Arabidopsis relevance score
Has DB confirmed association
Calculate Bio-entity attribute context score
yes
Assign DB confirmed bio-entity id
no
Extraction & mapping of comentioned mutations
Mutation validation
yes
Assign mutation confirmed bio-entity id
no
Calculate Bio-entity normalization score cut-off
Above cut-off
yes
Assign candidate bio-entity id
no

Figure: Source:http://zope.bioinfo.cnio.es/plan2l/plan2l.html

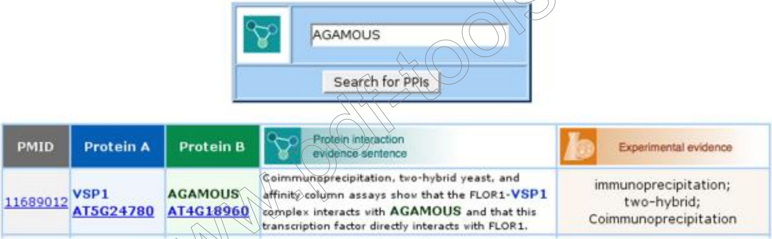
PLAN2L: Search Bibliome



PMID	PLAN2L sentence results										iHOP	WikiG
14534787	In Arabidopsis, silencing of B-type MADS-box genes was obtained by expressing B. napus APETALA3(BAP3) or PISTILLATA (BPI) homologous self-complementary hpRNA constructs under control of the Arabidopsis A-type MADS-box gene **APETALA1** (AP1) promoter.	-0.733566	-0.169806	-1.92142	-0.376697	3.85347	-0.848776	-1.17676	-1.01383		iHOP	WikiG
12941870	Further analysis indicated that the expression of flowering time genes FT, SUPPRESSOR OF OVEREXPRESSION OF CO 1 (SOC1) and flower meristem identity genes LEAFY (LFY), **APETALA1** (AP1) was significantly up-regulated in 35S::OMADS1 transgenic Arabidopsis plants.	-0.882258	-0.210891	-2.54508	-0.673563	4.10558	-0.028268	-0.838081	-1.11353		iHOP	WikiG

Figure: Source:http://zope.bioinfo.cnio.es/plan2l/plan2l.html

PLAN2L: Search PPI

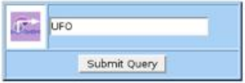


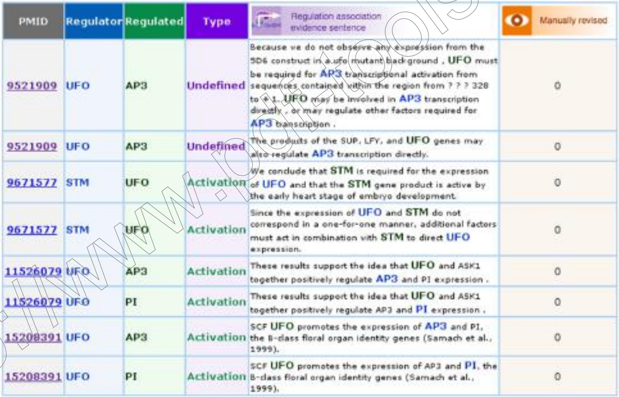
PMID	Protein A	Protein B	Protein interaction evidence sentence	Experimental evidence

11689012	VSP1
	AT5G24780
		AGAMOUS
		AT4G18960
			Coimmunoprecipitation, two-hybrid yeast, and affinity column assays show that the FLOR1-VSP1 complex interacts with AGAMOUS and that this transcription factor directly interacts with FLOR1.	immunoprecipitation;
				two-hybrid;
				Coimmunoprecipitation

Figure: Source:http://zope.bioinfo.cnio.es/plan2l/plan2l.html

PLAN2L: Search Gene Regulation Association

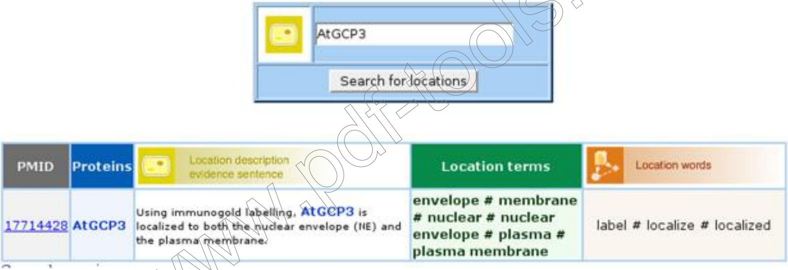




PMID	Regulator	Regulated	Type	Regulation association evidence sentence	Manually revised
9521909	UFO	AP3	Undefined	Because ve do not observe any expression from the 5D6 construct in a ufo mutant background , UFO must be required for AP3 transcriptional activation from sequences contained within the region from ? ? ? 328 to + 1. UFO may be involved in AP3 transcription directly , or may regulate other factors required for AP3 transcription .	0
9521909	UFO	AP3	Undefined	The products of the SUP, LFY, and UFO genes may also regulate AP3 transcription directly.	0
9671577	STM	UFO	Activation	We conclude that STM is required for the expression of UFO and that the STM gene product is active by the early heart stage of embryo development.	0
9671577	STM	UFO	Activation	Since the expression of UFO and STM do not correspond in a one-for-one manner, additional factors must act in combination vith STM to direct UFO expression.	0
11526079	UFO	AP3	Activation	These results support the idea that UFO and ASK1 together positively regulate AP3 and PI expression .	0
11526079	UFO	PI	Activation	These results support the idea that UFO and ASK1 together positively regulate AP3 and PI expression .	0
15208391	UFO	AP3	Activation	SCF UFO promotes the expression of AP3 and PI, the B-class floral organ identity genes (Samach et al., 1999).	0
15208391	UFO	PI	Activation	SCF UFO promotes the expression of AP3 and PI, the B-class floral organ identity genes (Samach et al., 1999).	0

Figure: Source:http://zope.bioinfo.cnio.es/plan2l/plan2l.html

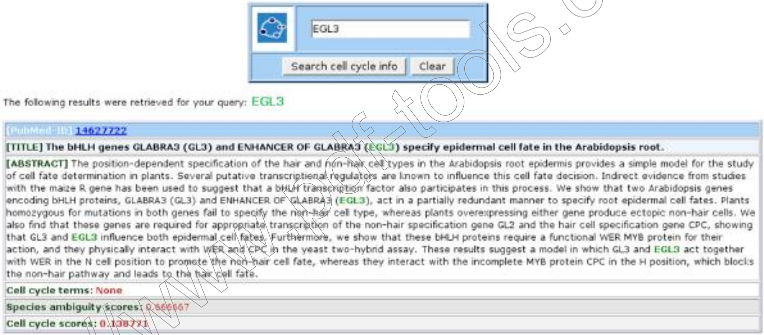
PLAN2L: Search Location Sentences



PMID	Proteins	Location description evidence sentence	Location terms	Location words
17714428	AtGCP3	Using immunogold labelling, AtGCP3 is localized to both the nuclear envelope (NE) and the plasma membrane.	envelope # membrane # nuclear # nuclear envelope # plasma # plasma membrane	label # localize # localized

Figure: Source:http://zope.bioinfo.cnio.es/plan2l/plan2l.html

PLAN2L: Search Cell Sycle Associations



The following results were retrieved for your query: EGL3

[PubMed-ID] 14627722

[TITLE] The bHLH genes GLABRA3 (GL3) and ENHANCER OF GLABRA3 (EGL3) specify epidermal cell fate in the Arabidopsis root.

[ABSTRACT] The position-dependent specification of the hair and non-hair cell types in the Arabidopsis root epidermis provides a simple model for the study of cell fate determination in plants. Several putative transcriptional regulators are known to influence this cell fate decision. Indirect evidence from studies with the maize R gene has been used to suggest that a bHLH transcription factor also participates in this process. We show that two Arabidopsis genes encoding bHLH proteins, GLABRA3 (GL3) and ENHANCER OF GLABRA3 (EGL3), act in a partially redundant manner to specify root epidermal cell fates. Plants homozygous for mutations in both genes fail to specify the non-hair cell type, whereas plants overexpressing either gene produce ectopic non-hair cells. We also find that these genes are required for appropriate transcription of the non-hair specification gene GL2 and the hair cell specification gene CPC, showing that GL3 and EGL3 influence both epidermal cell fates. Furthermore, we show that these bHLH proteins require a functional WER MYB protein for their action, and they physically interact with WER and CPC in the yeast two-hybrid assay. These results suggest a model in which GL3 and EGL3 act together with WER in the N cell position to promote the non-hair cell fate, whereas they interact with the incomplete MYB protein CPC in the H position, which blocks the non-hair pathway and leads to the hair cell fate.

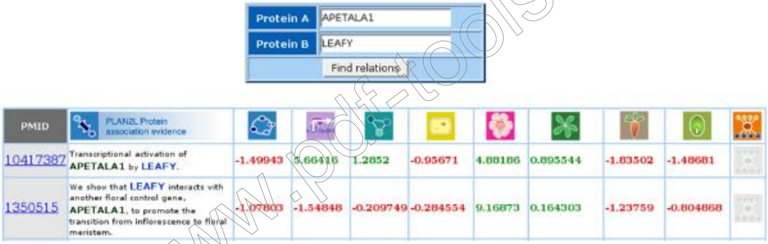
Cell cycle terms: None

Species ambiguity scores: 0.666667

Cell cycle scores: 0.138771

Figure: Source:http://zope.bioinfo.cnio.es/plan2l/plan2l.html

PLAN2L: Association between Bio-entities



Protein A	APETALA1
Protein B	LEAFY

PMID	PLAN2L Protein association evidence									
10417387	Transcriptional activation of APETALA1 by LEAFY.	-1.49943	5.66416	1.2852	-0.95671	4.88186	0.895544	-1.83502	-1.48681	
1350515	We show that LEAFY interacts with another floral control gene, APETALA1, to promote the transition from inflorescence to floral meristem.	-1.07803	-1.54848	-0.209749	-0.284554	9.16873	0.164303	-1.23759	-0.804868	

Figure: Source:http://zope.bioinfo.cnio.es/plan2l/plan2l.html

References

http://zope.bioinfo.cnio.es/teaching/
http://zope.bioinfo.cnio.es/bionlp_tools/
http://www.tifr.res.in/∼ashishvt/biotextmining/