An Introduction to Bioinformatics Infrastructures:

Text Mining and Information Extraction Applications for Bioinformatics and Systems Biology

Plant Bioinformatics, Systems and Synthetic Biology Summer School
27-31 July 2009 - University of Nottingham, UK

Martin Krallinger,
Spanish National Cancer Research Centre - CNIO
mkrallinger@cnio.es

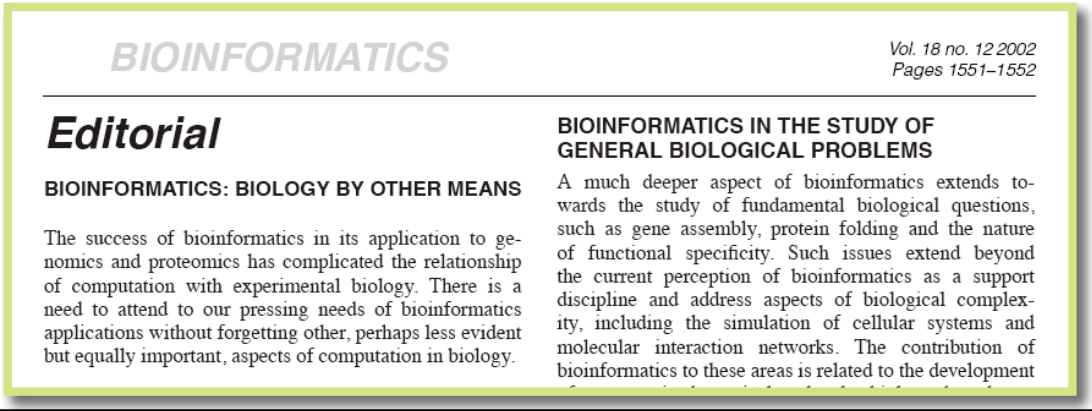
Talk Outline / Topcis (I)

- Bioinformatics infrastructures
- Integration of heterogeneous data types
- Bioinformatics resources
- Importance and use of scientific literature data
- Manual literature curation process for building systems biology resources
- Annotation types
- Building literature curation workflows
- Relevance of text mining strategies in the context of SB

Talk Outline / Topics (II)

- Short intro to text mining and NLP
- Short overview of existing BioNLP application types
- Implementing a text mining system: basic steps
- The PLAN2L literature mining tool

Bioinformatics & biological projects



BIOINFORMATICS

Vol. 18 no. 12 2002
Pages 1551–1552

Editorial

BIOINFORMATICS: BIOLOGY BY OTHER MEANS

The success of bioinformatics in its application to genomics and proteomics has complicated the relationship of computation with experimental biology. There is a need to attend to our pressing needs of bioinformatics applications without forgetting other, perhaps less evident but equally important, aspects of computation in biology.

BIOINFORMATICS IN THE STUDY OF GENERAL BIOLOGICAL PROBLEMS

A much deeper aspect of bioinformatics extends towards the study of fundamental biological questions, such as gene assembly, protein folding and the nature of functional specificity. Such issues extend beyond the current perception of bioinformatics as a support discipline and address aspects of biological complexity, including the simulation of cellular systems and molecular interaction networks. The contribution of bioinformatics to these areas is related to the development

ALL biological projects need or will need Bioinformatics (.. as soon as they enter into genomics):

- as **resource** (databases and software)
- as support for **design**, **organization** & **interpretation** of the data
- in the research team for the specific scientific project

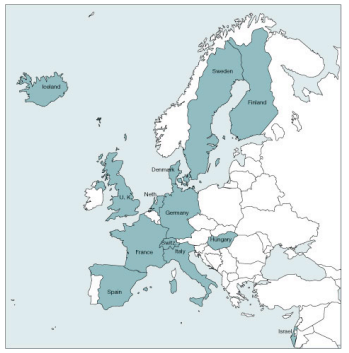
Bioinformaticians are scientists working in:

- developing **methods** (Bioinformatics as a research area)
- developing **resources** e.g. databases (Bioinformatics as technology)
- Embedded in biology/Biotech/Biomed *(the single bioinformatician syndrome)*

ELIXIR EUROPEAN LIFE SCIENCES INFRASTRUCTURE FOR BIOLOGICAL INFORMATION

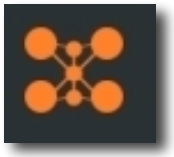
To construct and operate a <u>sustainable infrastructure for biological information</u> in Europe,

To <u>support</u> life science research and its translation to medicine and the environment, the bio-industries and society.



- **Partners:** 32 partners, 13 member states
- **Funding:** 4.5 M€ from EU FP7
- **Deliverable:** Consortium agreement to <u>define the scope</u> of the infrastructure and how it will be constructed

- **Optimal Data Management**
  - **Coordinated Data Resources with improved access**
  - **Integration and interoperability of diverse heterogeneous data**
  - **Good Value for Money**
- **Forge Links to data in other related domains**
- **A single European voice in international collaborations to influence global decisions and maintain open access to data**
- **Enhance European competitiveness in bioscience industries**
- **Address need for Increased Funding & its Coordination**



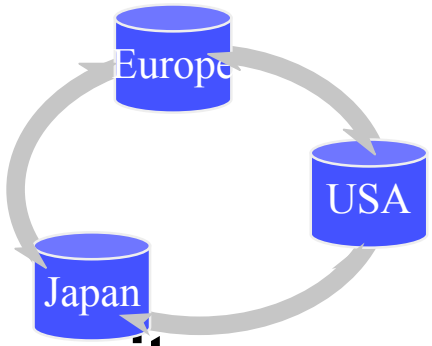
The Preparatory Phase project

Elixir is organised into 14 work packages which have committees of (mainly) European experts associated with them.

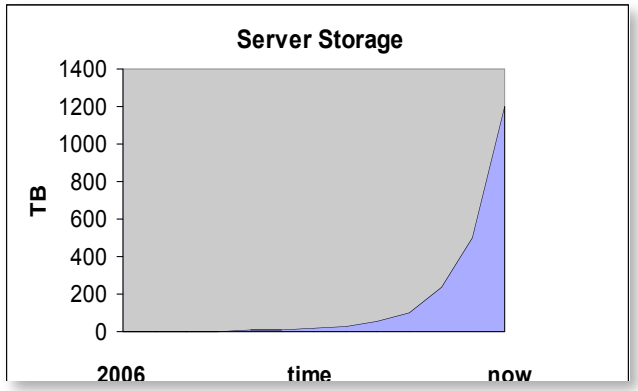
1. Project management
2. Data providers
3. **User communities**
4. Organisation and Legal
5. Funding
6. Physical infrastructure
7. Data interoperability
8. **Literature**
9. Healthcare
10. **Chemistry, Plants, Agriculture & Environment**
11. **Training**
12. Tools integration
13. Feasibility studies
14. Reporting and negotiation

Why do we need ELIXIR?

(Why do we need bioinformatics infrastructures)

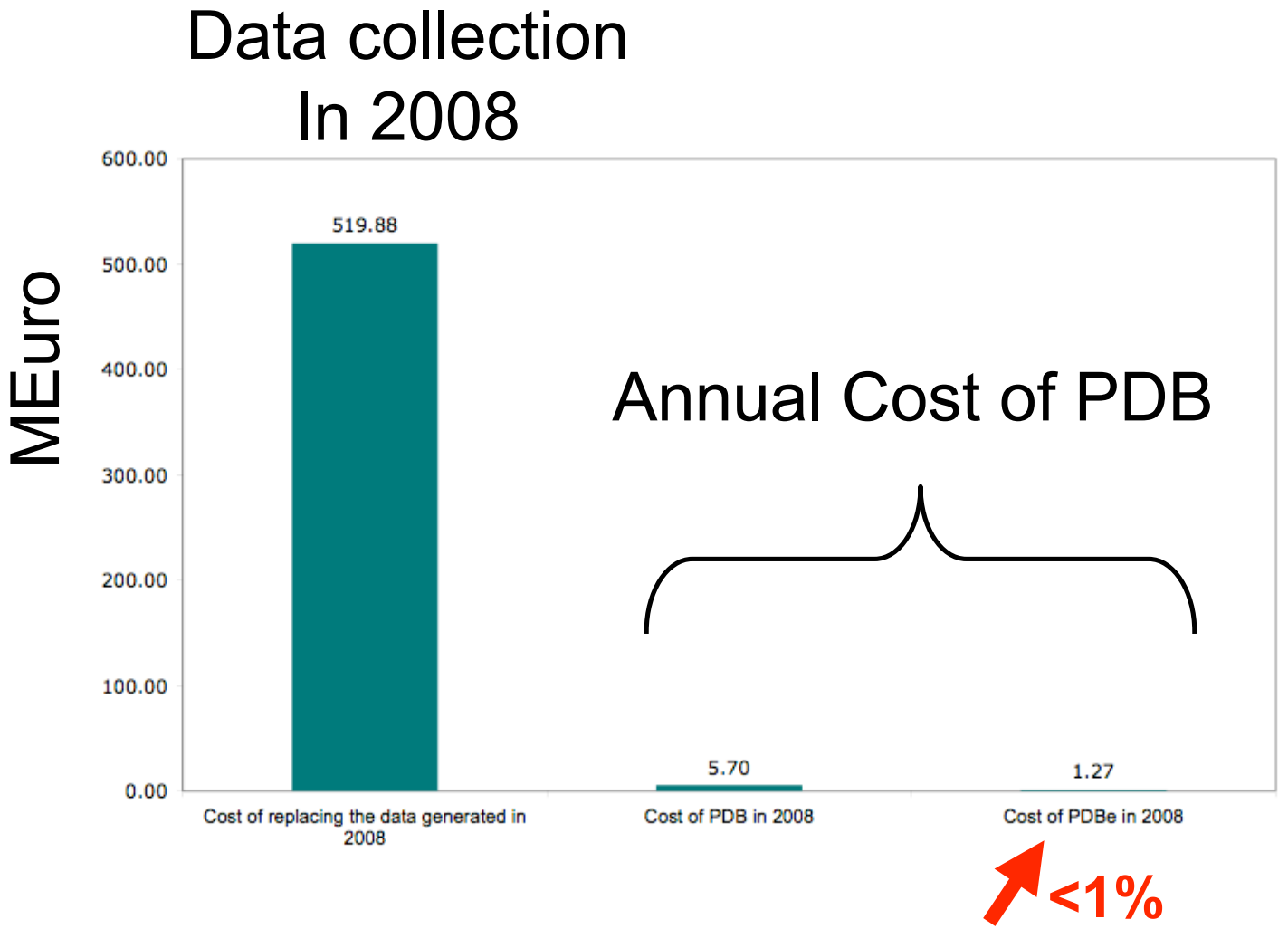


- **Data Growth**
- **Global context**
- **Very large user community:**
  - **3.3 m web hits/day**
  - **20,000 unique users per day**

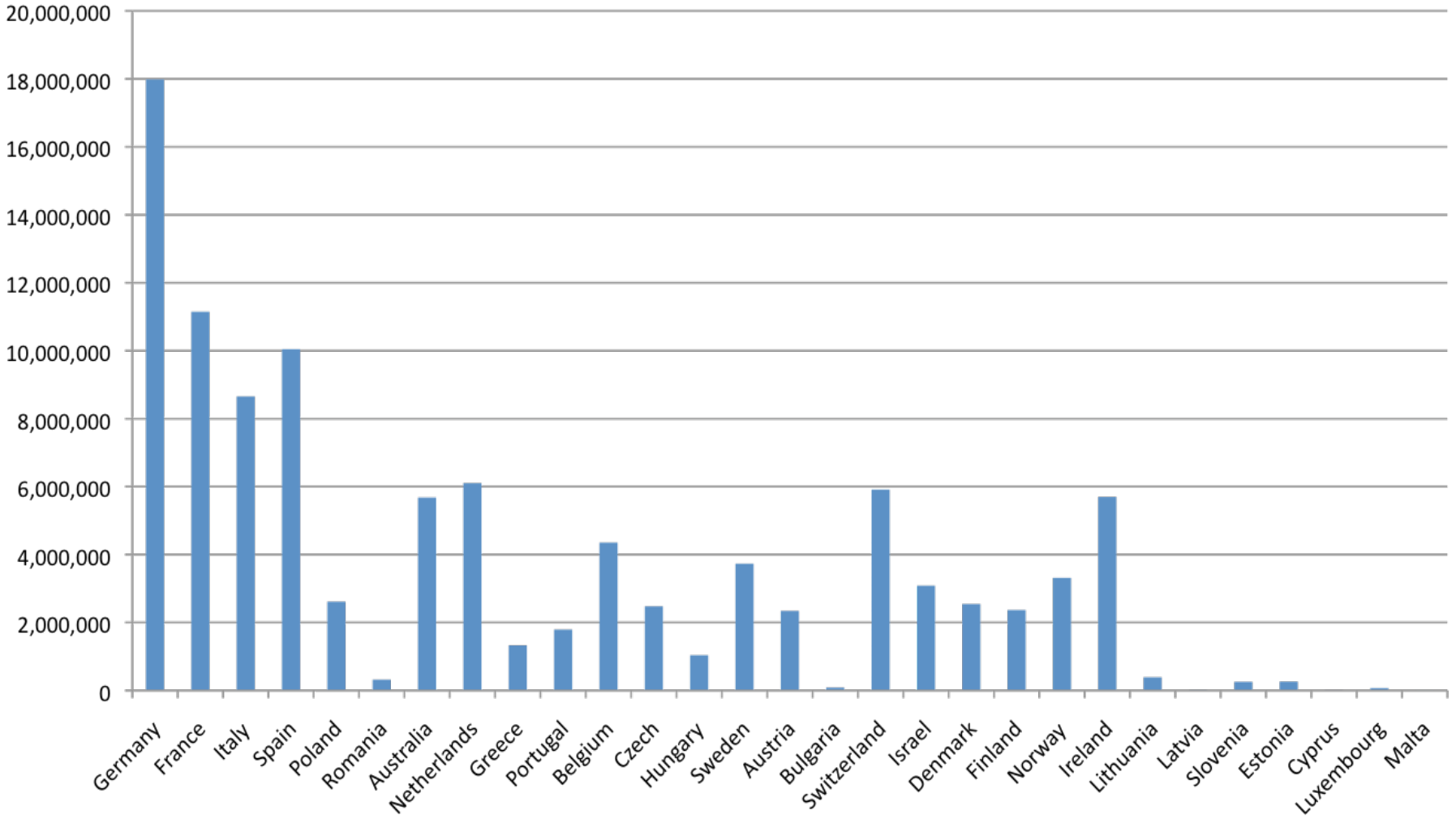


- **Need to preserve data and make accessible to all**
- **Impact on Medicine, Agriculture & Biotechnology**
- **Impact on society & bioindustries**
- **Need for increased funding for biodata resources**

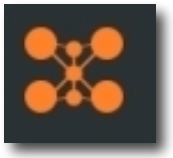
Good Value for Money e.g. PDB



EBI Hits in 2008

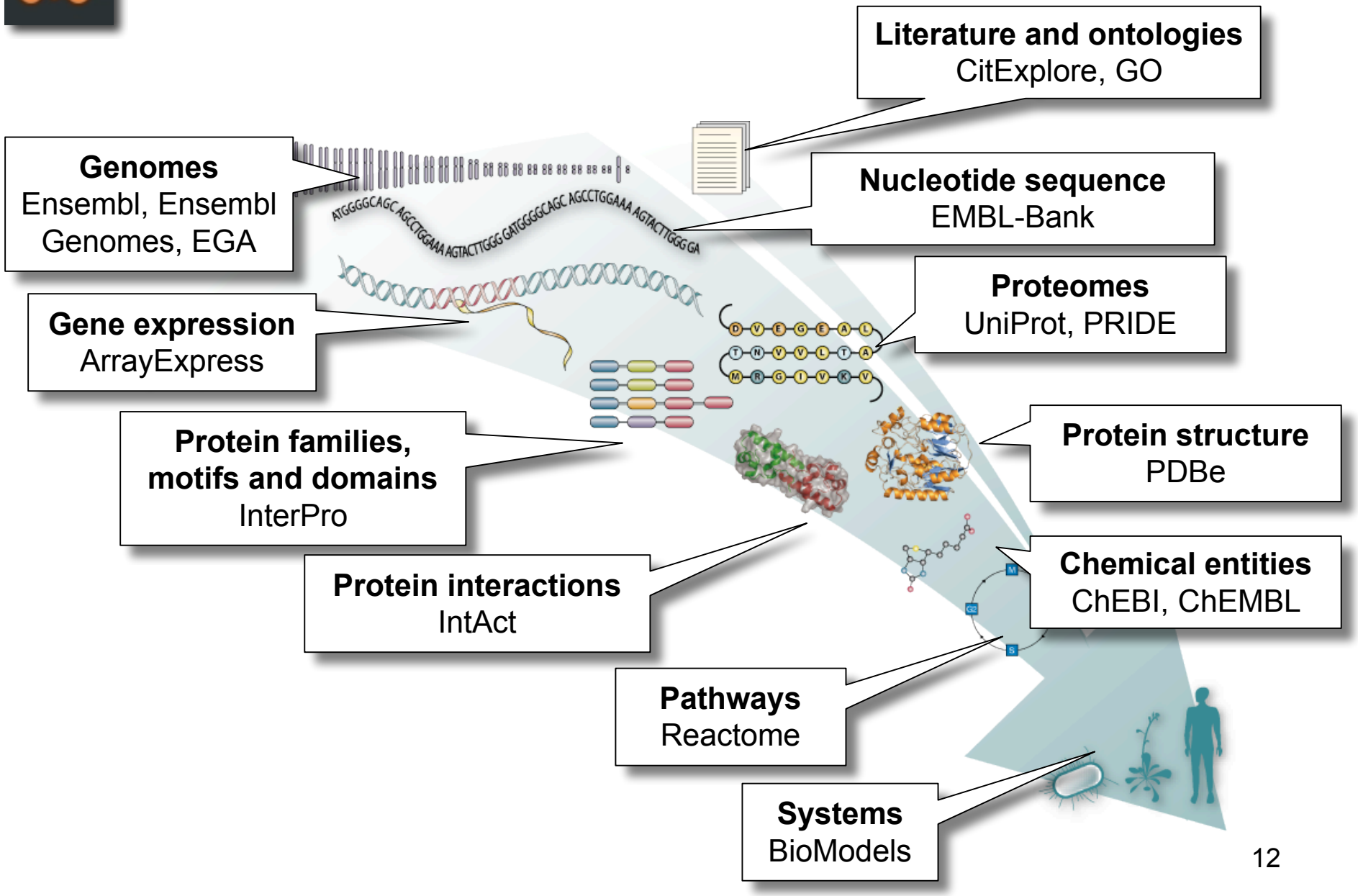


WP3: User Communities

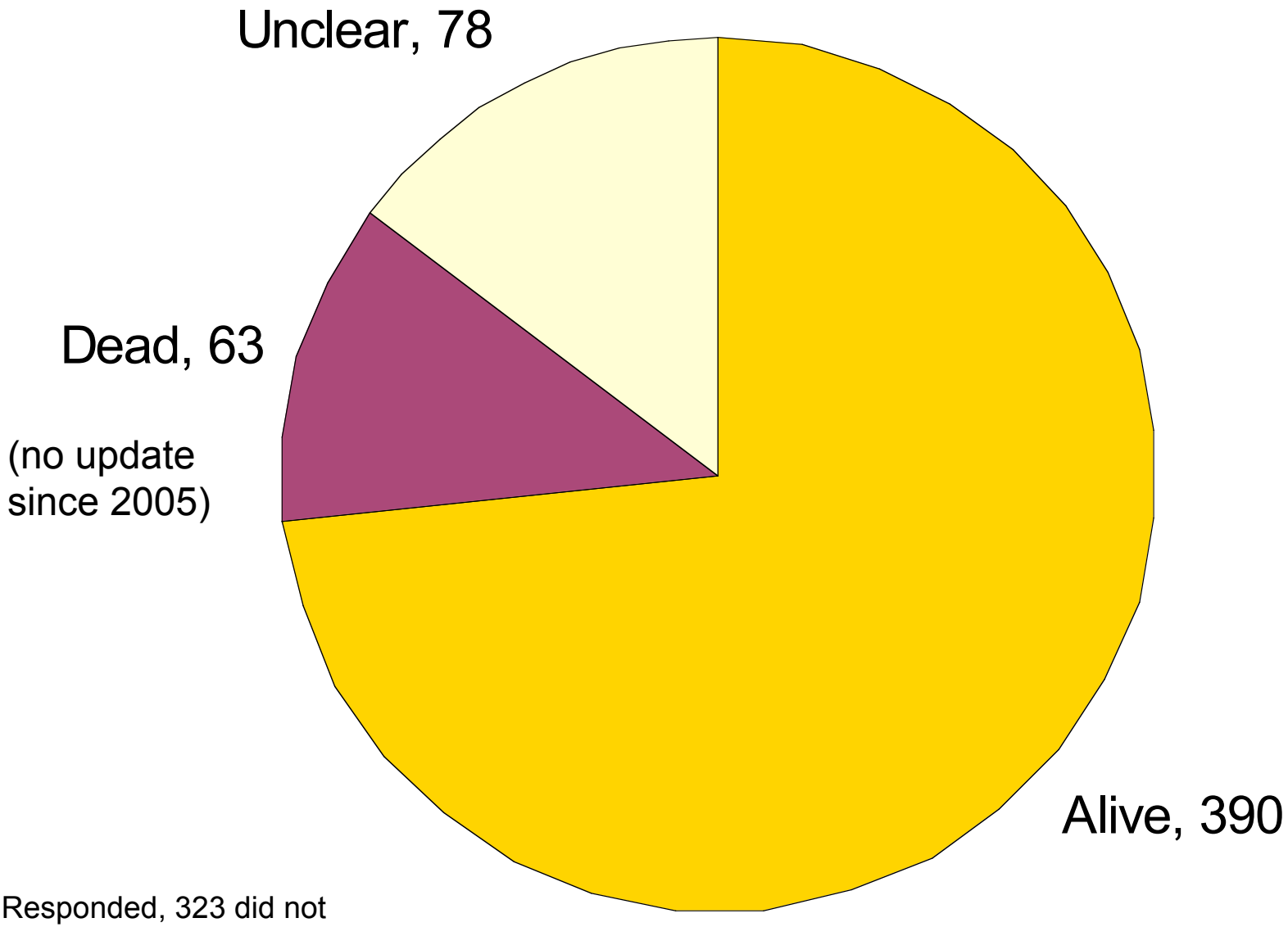


- User Survey: 1000 responses
  - Long term support essential
  - Top 3 challenges:
    - **Data integration**; **Format compatibility**; **Website usability**
  - Concerns
    - Data quality and measures; Quality of tools;Training
- Need to consider different needs in different countries
- Need for a plan for long-term maintenance of computational tools
  - Create mechanisms for long-term maintenance of bioinformatics tools
    - user-friendly & machine-friendly interfaces
- Need for standards for formats and integration
  - Increased integration of databases, tools and between infrastructure domains
- Need to provide mechanisms for prioritisation of need for resources

Databases: molecules to systems



531 Databases surveyed



208 Responded, 323 did not

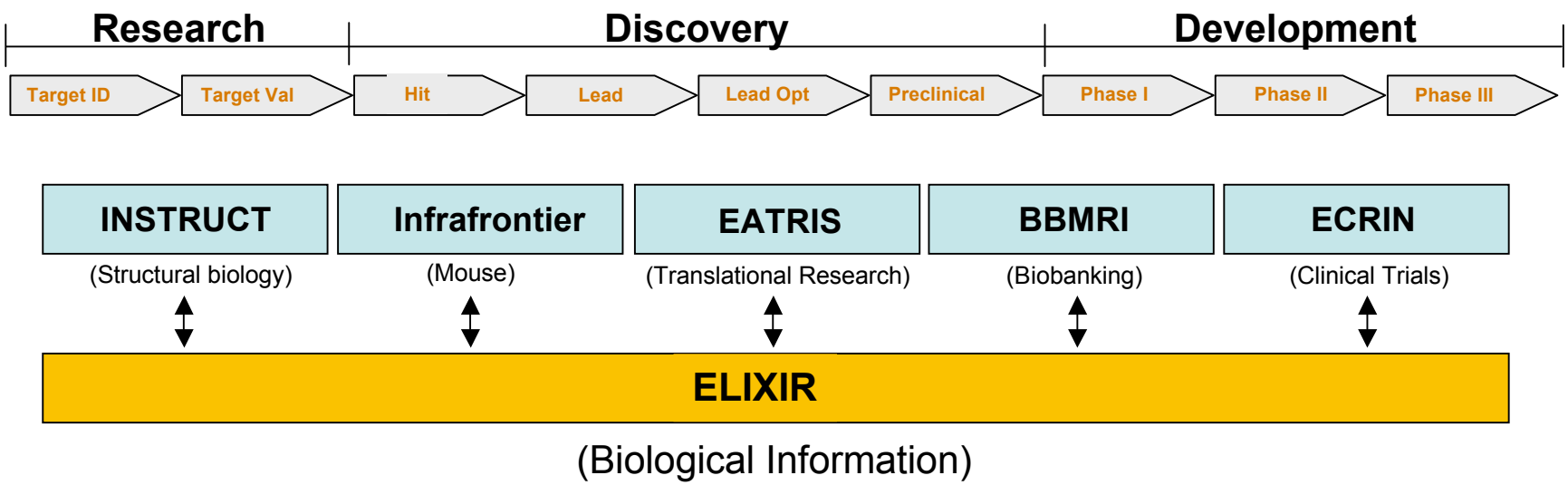
Total European effort

- 200 Databases
- 700 People
- 100 Institutions
- 60 million web hits per month
- Total investment to date €308 million
- Annual cost €35 million

RECOMMENDATION

Coordination and prioritisation, as well as stable funding, is needed for many of these resources

ESFRI
Biology Research Infrastructure proposals.



WP 10: Chemistry, Plants, Agriculture & Environment

- Support / extend current core resources for
  - Nucleotide/protein sequence, genomes, structures, interactions etc.
- Selected specialist resources migrated to Elixir infrastructure
  - Reduce complexity of informatics landscape, maintain functionality
  - Integration allows mining of combined data
- Adopt key data standards and work for common infrastructure
  - Link to other ESFRI, non ESFRI European projects
  - Link to non European initiatives (NSF/iPlant, DOE/Camera)
- Free access to Elixir data and core analysis tools
  - Web based queries, programmatic access, download

WP11: Training

Identified training issues in Europe:

- Little or no **coordination**
- Rapid **evolution** of bioinformatics resources
- Lack of a **centralised** body for guidance;
- Lack of **recognition** of the importance of bioinformatics user training, even within the bioinformatics community.

Elixir recommendations:

Link the **development** of data resources to the provision of **training** materials;

Create a training **support unit** that will:

a) provide a centralised training registry;

b) provide support for trainers throughout Europe

c) develop benchmarking and evaluation systems;

d) provide mechanisms for developing new training programmes

e) act as a single point of contact for national and pan-European training

Elixir WP8: Scientific Literature Interdisciplinary Interactions

Chair: Alfonso Valencia (CNIO)

Co-Chairs: Dietrich Rebholz-Schuhmann & Peter Stoehr (EMBL-EBI)

Initial committee

- Robert Kiley, Wellcome Trust
- Carole Goble, U. Manchester
- Larry Hunter, UCHSColorado
- Manuel Peitsch, SIB
- Matthew Cockerill, BMC
- Jun'ichi Tsujii, NaCTeM and U. Tokyo
- Timo Hannay, Nature PG

Addtional Contributions

Ian Dix, Astrazeneca
Ian Harrow, Pfizer
Udo Hahn, U. Jena
Sophia Ananiadou , NacTeM
Patrick Ruch, Geneva University
Christopher Bake, New Brunswick U.
Juliane Fluck, Fraunhofer
Anita Burgun, Rennes University
and Kostas Repanas (CNIO) WP Coordinator

European Life-science Infrastructure for Biological Information (Elixir) WP 8: Scientific Literature Interdisciplinary Interactions

D8.1 A report summarising the current

(1) status of literature repositories throughout Europe and recommendations for the future

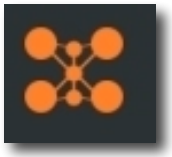
(2) infrastructure needs in Europe to establish an information-sharing platform to integrate databases and literature for **(*) experts and non-experts**, with

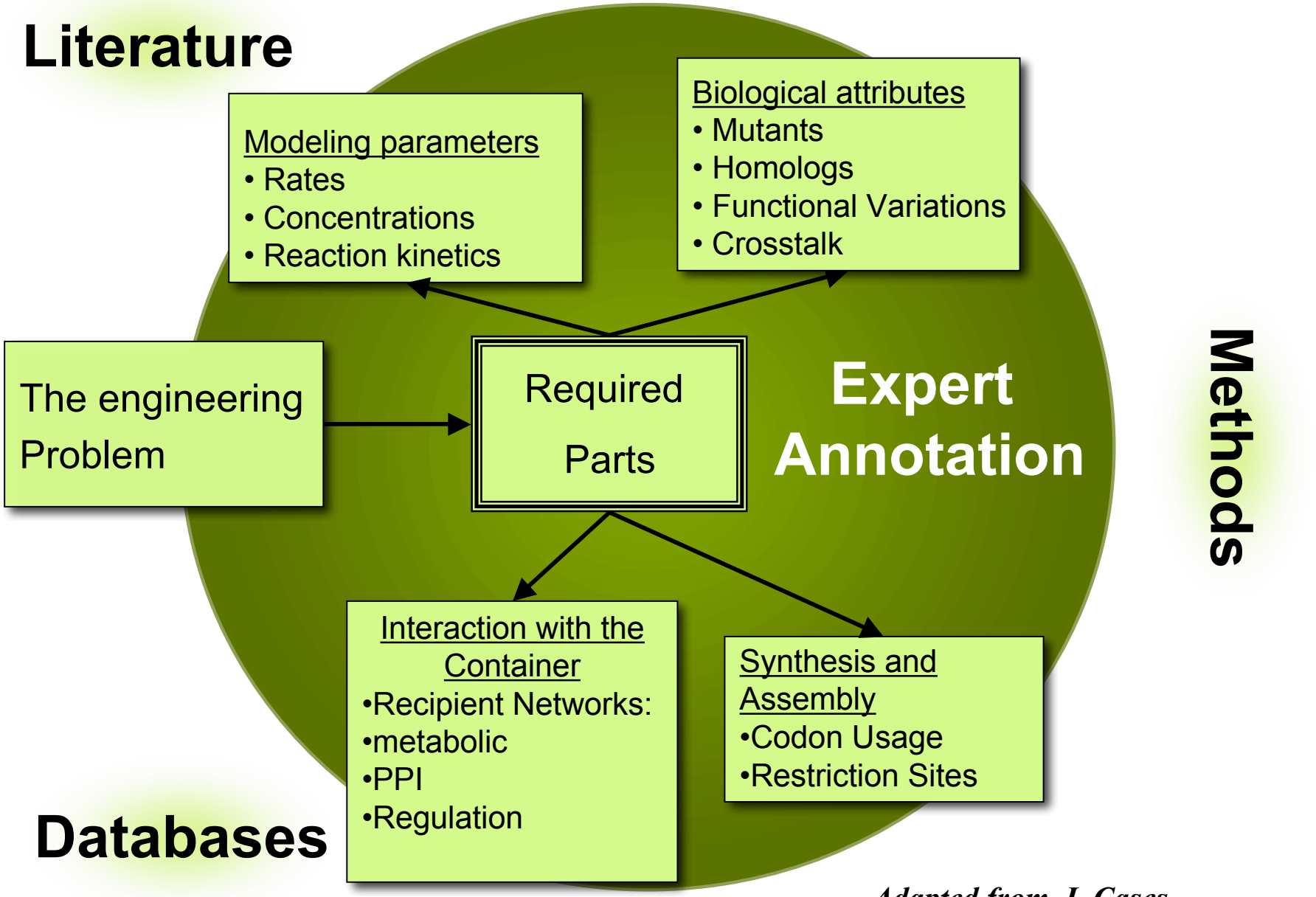
(3a) specific reference to the provision of literature from repositories commonly used in biological information extraction and

(3bi) tools for access to the literature, for

(3bii) data representation and for

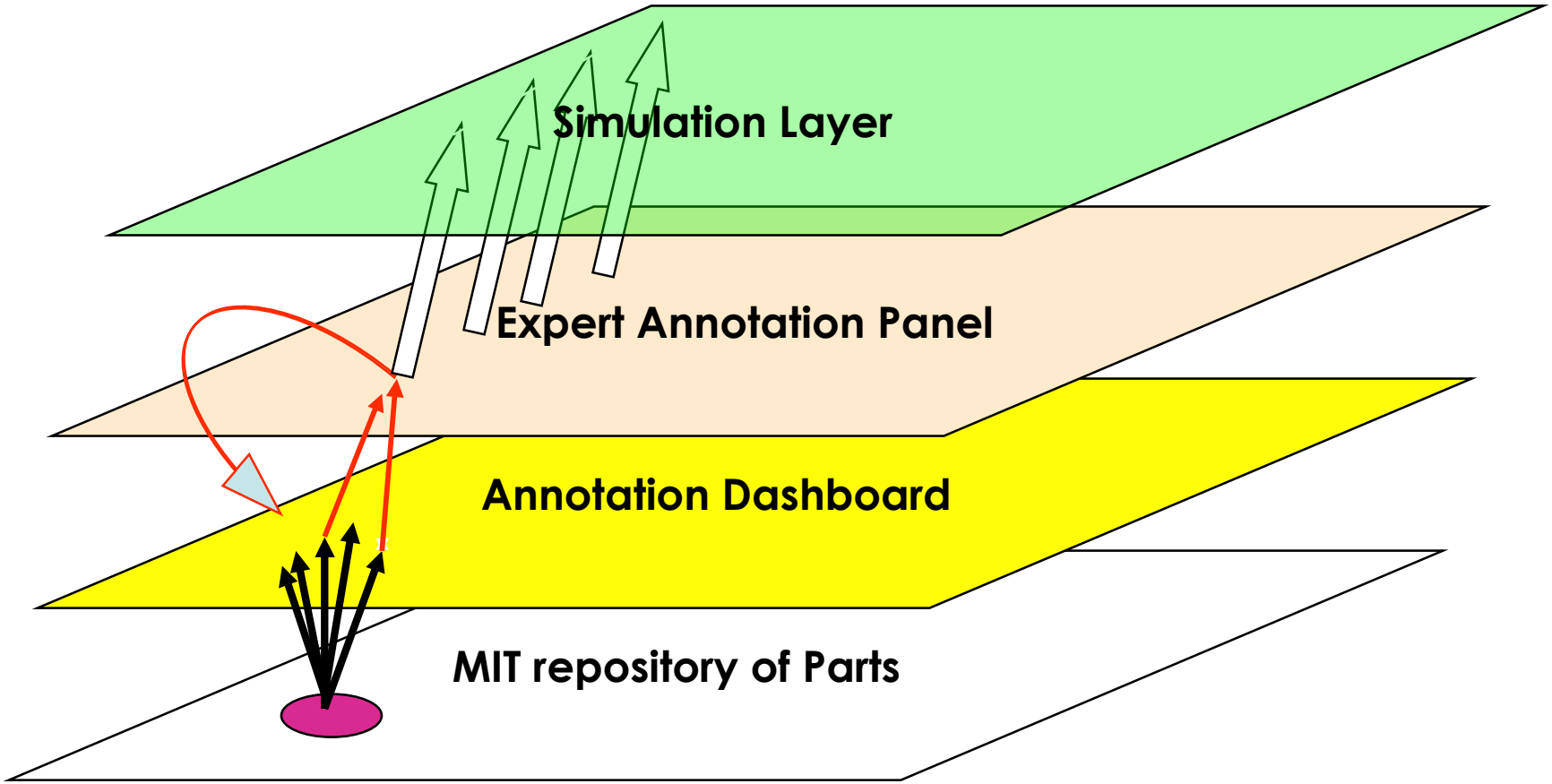
(3biii) interaction with end users.



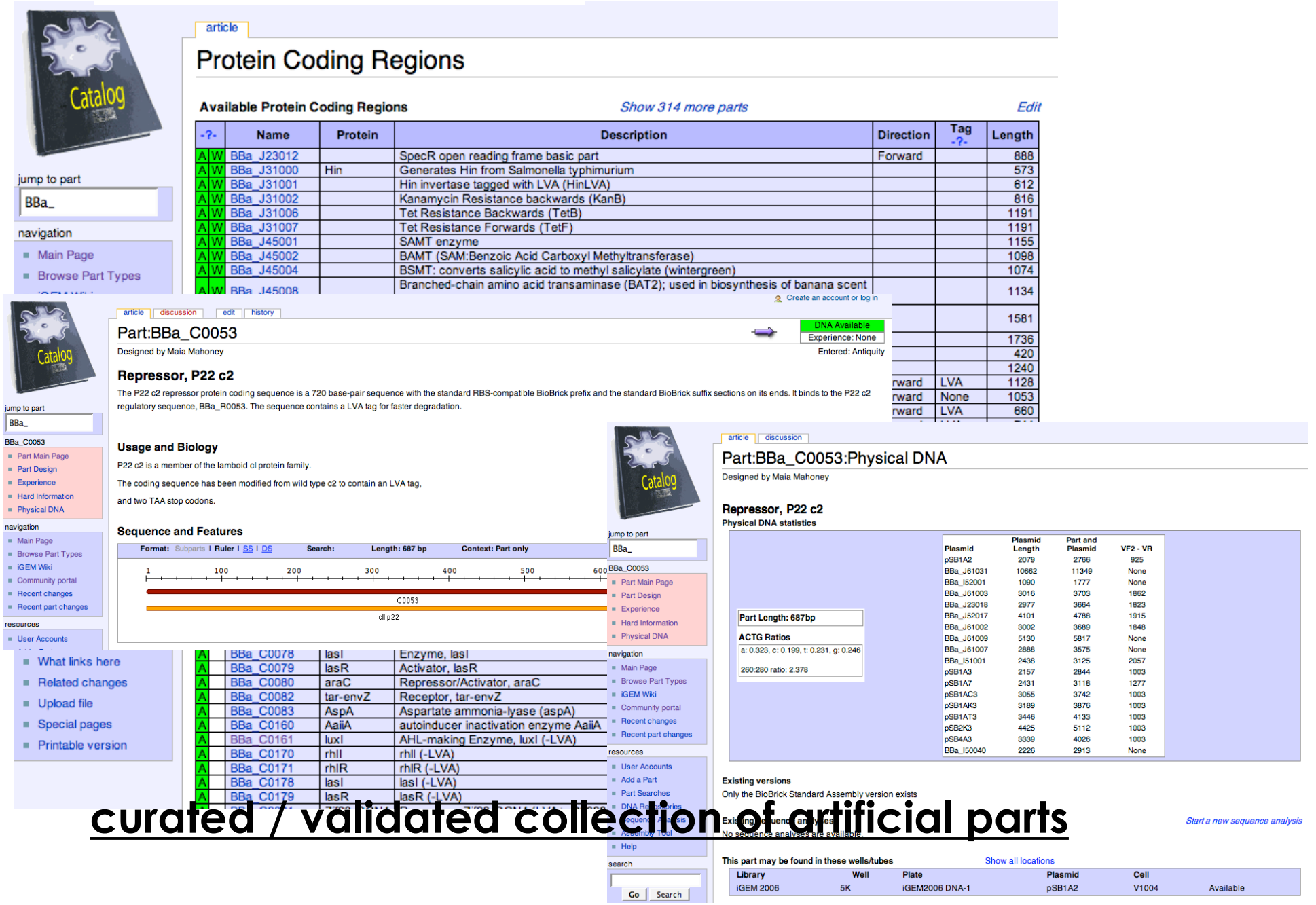


Adapted from I. Cases

Emergence IT layout



MIT Repository of Parts



curated / validated collection of artificial parts

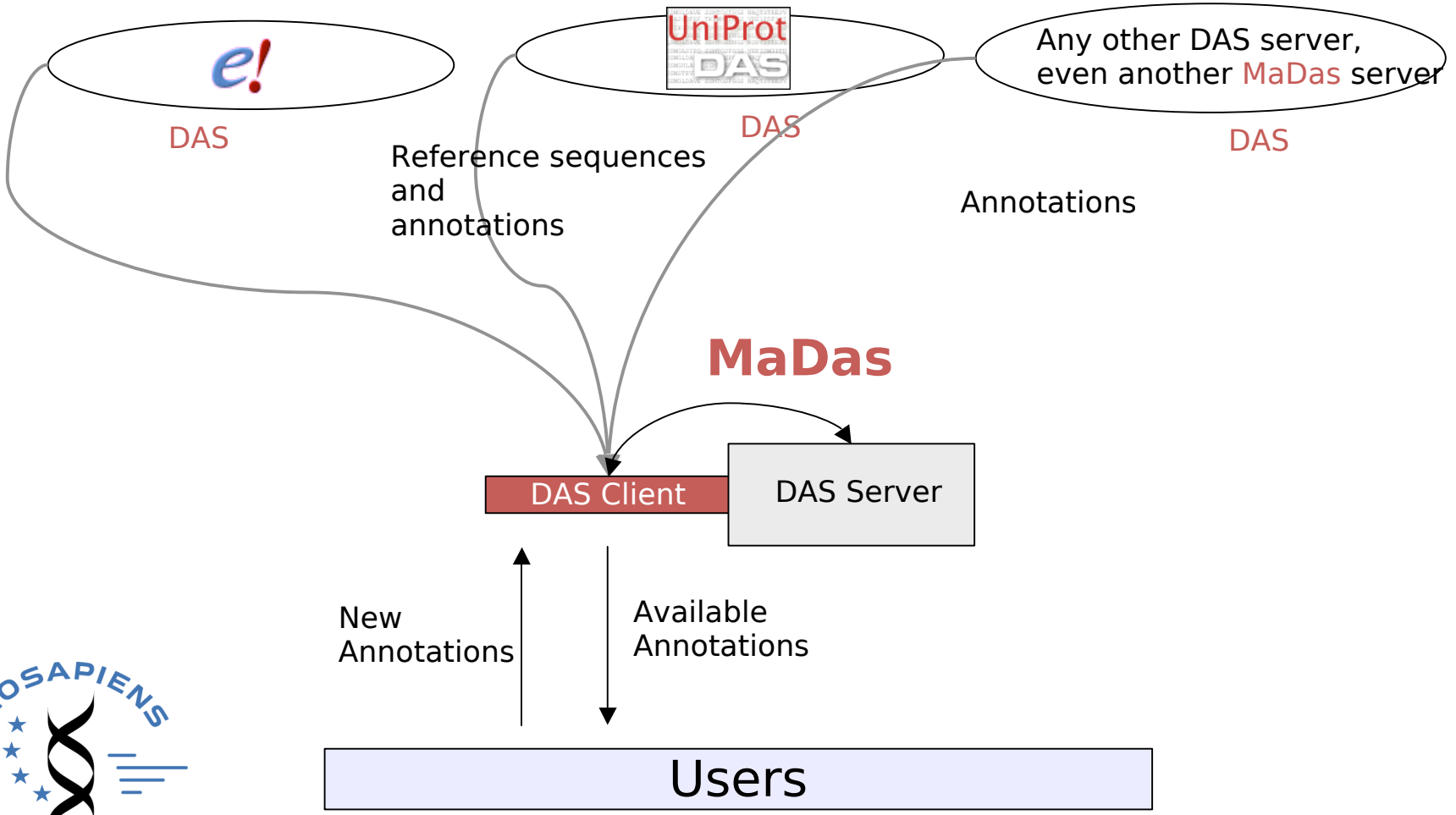
MaDAS principal features

- MaDAS allows users to add, edit, or remove self generated sequence annotations
- Allows to upload multiple annotations from different sources.
- Provides a security system based on projects. The annotations could be public or only available for the project members.
- Provides an interface to manage projects, users and collections of annotations.

Collaborative features

- **Project based** system. Users can create their own projects or participate in projects hosted in MaDAS.
- Projects can be public or private, in private projects the project leader decide who can view or edit the project annotations.
- The notification system inform about: new projects, new annotations, new users or new plugins.
- Searches by: category, project leader, institutions, etc

MaDas
Manual Sequence Annotation System



MaDAS modules

MaDAS is composed by:

- "The core" which provide different APIs in order to facilitated the development of plug-ins and the communication between them.
- Data Source plug-ins 
- DAS server plug-ins 
- Visualization plug-ins 

Data source plug-ins

Manage Reference plug-in: We use the DAS reference sequence concept (http://www.biodas.org/wiki/DAS/1/Overview#.5BReference.5D_Sequence) to describe a biological sequence that will be annotated.

Setup Ensembl genome, a collection of proteins , a new sequenced genome or just a DNA/protein fragment.

Load GFF plug-in: This plug-in allows users to upload GFF files to the system.

Manage DAS Tracks plug-in: Through this plug-in users can add annotations provided by any DAS server

Load chip plug-in: This plug-in allows experimentalist to map Affymetrix or Illumina microarray probes to a human reference sequence stored in MaDAS. Probe associated genes and proteins are also mapped.

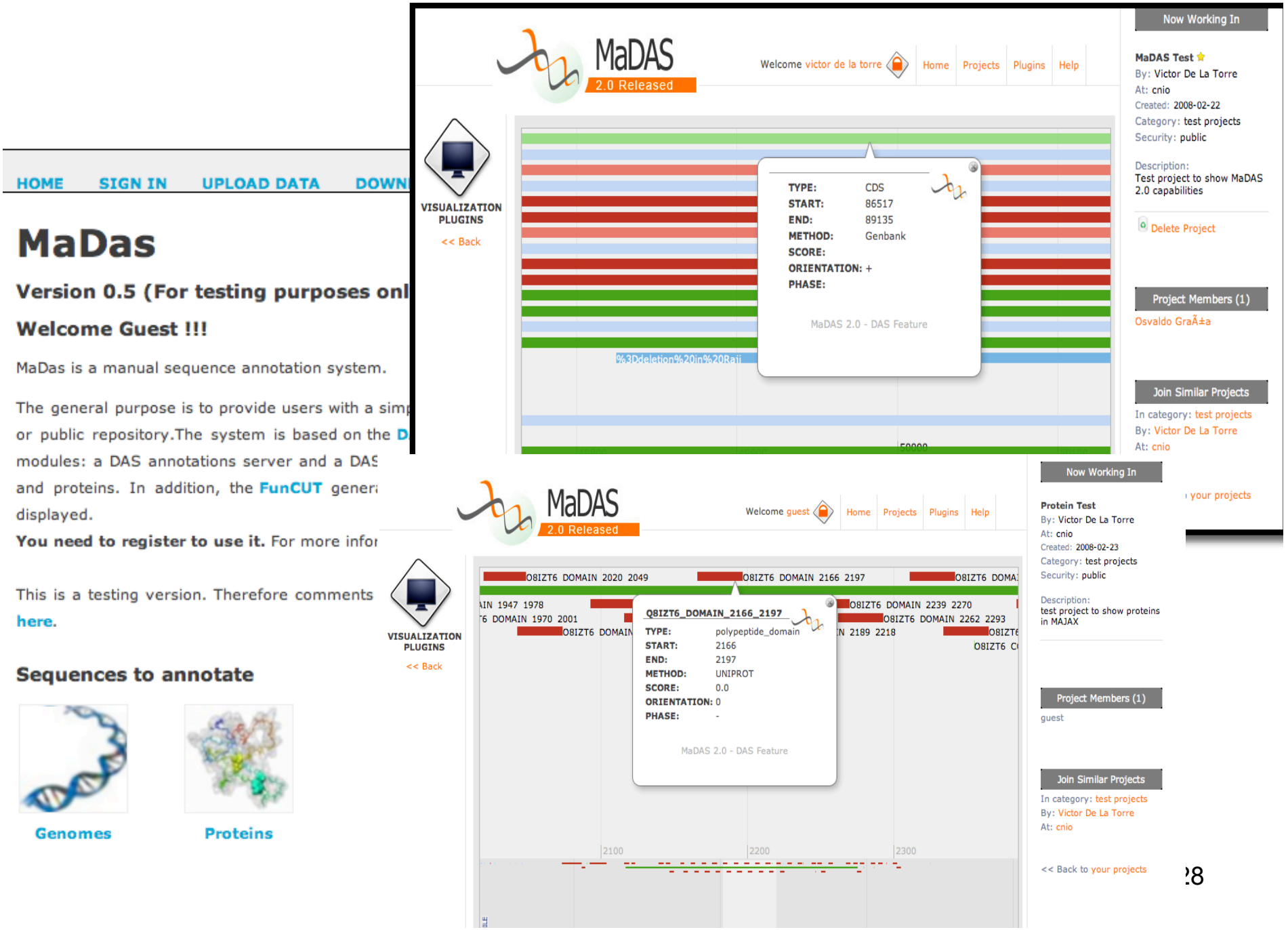
Load Gene expression plug-in: Allows users to upload data from a gene expression experiments.

Map Annotations plug-in: Using this plug-in is possible to add new annotations just mapping existing annotations to other online resource. For example if we have a gene track is possible to setup a disease track mapping these genes to OMIM diseases. This plug-in use several mapping services to map the annotations (Biomart, Uniprot Database mapping, PICR, ID converter)

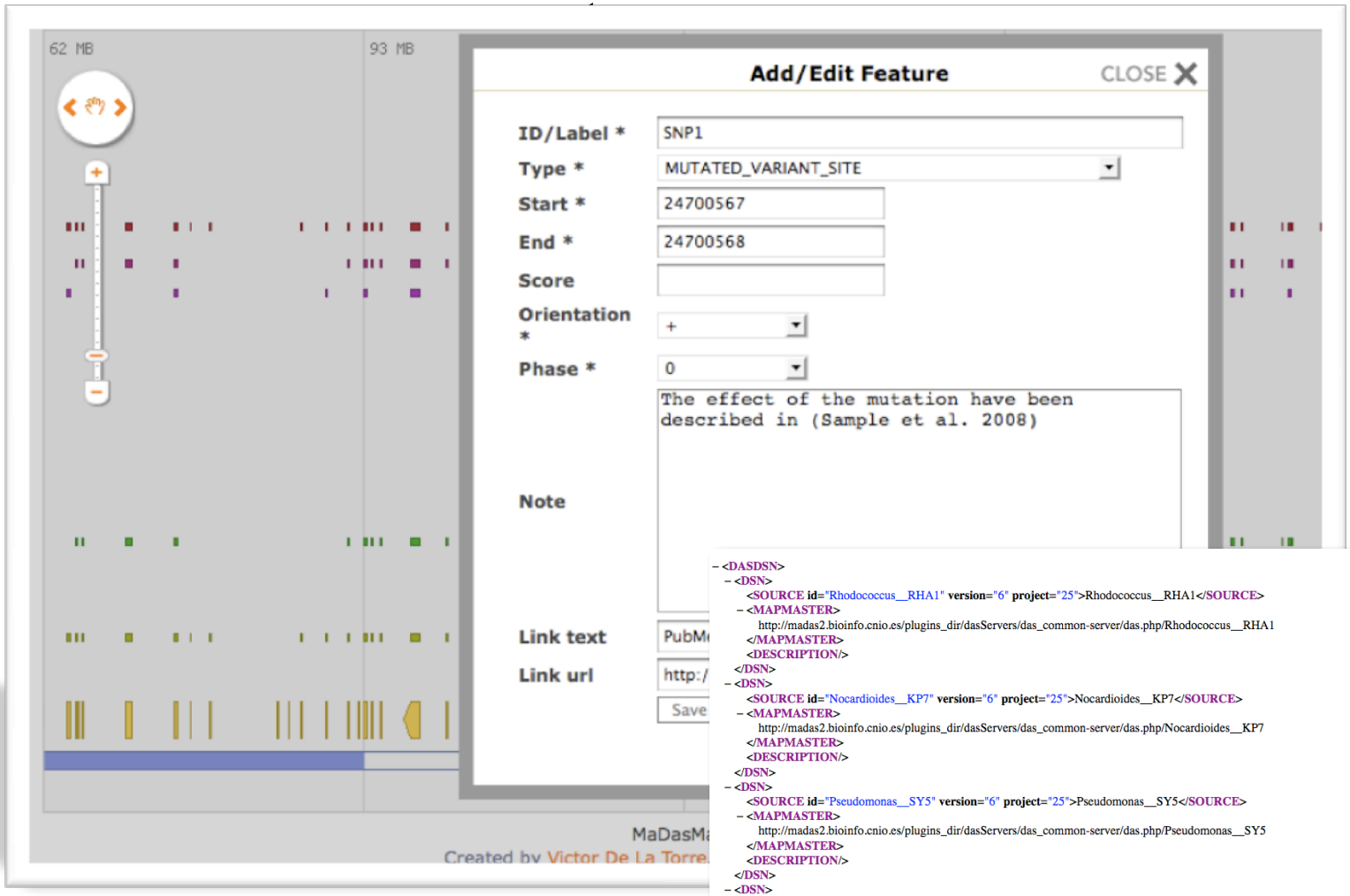
Treefam plug-in: This is an example of a very specific plug-in, which allows to information form Treefam).

Bionemo plug-in: import information stored in the Bionemo database (Bopdegradation and gene control reactions)

Manage annotations plug-in: to remove or inactivate an entire set of annotations.

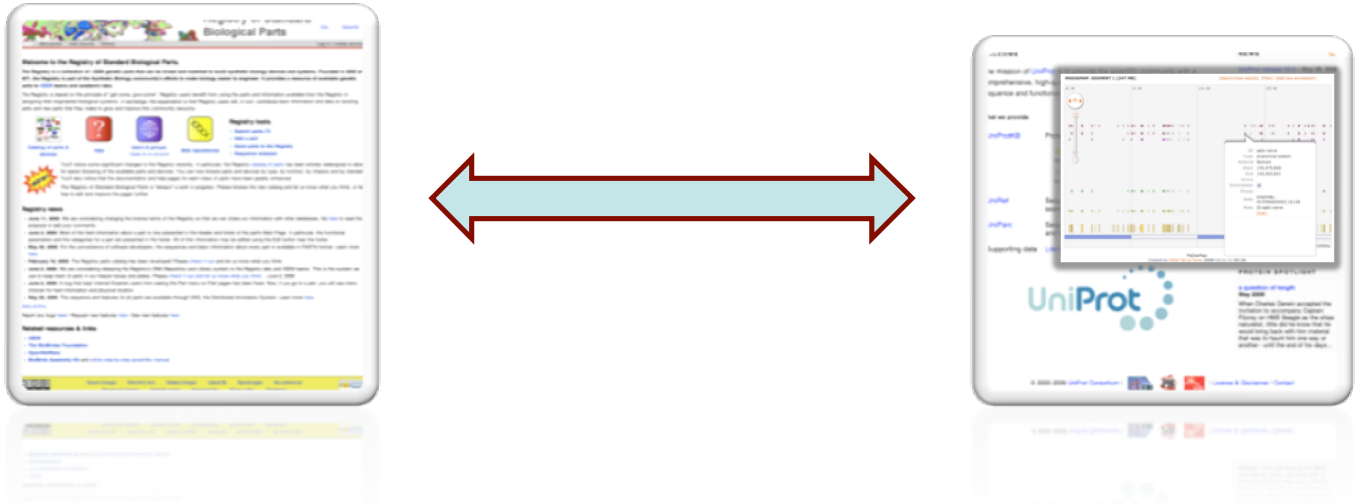

HOME
SIGN IN
UPLOAD DATA
MaDas
Version 0.5 (For testing purposes onl
Welcome Guest !!!
MaDas is a manual sequence annotation system.
The general purpose is to provide users with a simp
or public repository.The system is based on the D
modules: a DAS annotations server and a DAS
and proteins. In addition, the FunCUT gener
displayed.
You need to register to use it. For more infor
This is a testing version. Therefore comments
here.
Sequences to annotate
Genomes
Proteins
MaDAS
2.0 Released
Welcome victor de la torre
Home
Projects
Plugins
Help
VISUALIZATION PLUGINS
<< Back
TYPE: CDS
START: 86517
END: 89135
METHOD: Genbank
SCORE:
ORIENTATION: +
PHASE:
MaDAS 2.0 - DAS Feature
%3Ddeletion%20in%20Raji
50000
Now Working In
MaDAS Test
By: Victor De La Torre
At: cnio
Created: 2008-02-22
Category: test projects
Security: public
Description:
Test project to show MaDAS 2.0 capabilities
Delete Project
Project Members (1)
Osvaldo GraÃ±a
Join Similar Projects
In category: test projects
By: Victor De La Torre
At: cnio
your projects
MaDAS
2.0 Released
Welcome guest
Home
Projects
Plugins
Help
VISUALIZATION PLUGINS
<< Back
O8IZT6 DOMAIN 2020 2049
O8IZT6 DOMAIN 2166 2197
O8IZT6 DOMAIN 2239 2270
O8IZT6 DOMAIN 2262 2293
O8IZT6 DOMAIN 1970 2001
O8IZT6 DOMAIN 2189 2218
Q8IZT6_DOMAIN_2166_2197
TYPE: polypeptide_domain
START: 2166
END: 2197
METHOD: UNIPROT
SCORE: 0.0
ORIENTATION: 0
PHASE: -
MaDAS 2.0 - DAS Feature
2100
2200
2300
Now Working In
Protein Test
By: Victor De La Torre
At: cnio
Created: 2008-02-23
Category: test projects
Security: public
Description:
test project to show proteins in MAJAX
Project Members (1)
guest
Join Similar Projects
In category: test projects
By: Victor De La Torre
At: cnio
<< Back to your projects

Introducing expert annotations and consolidating them in databases/visualization



Added annotations are also available through DAS

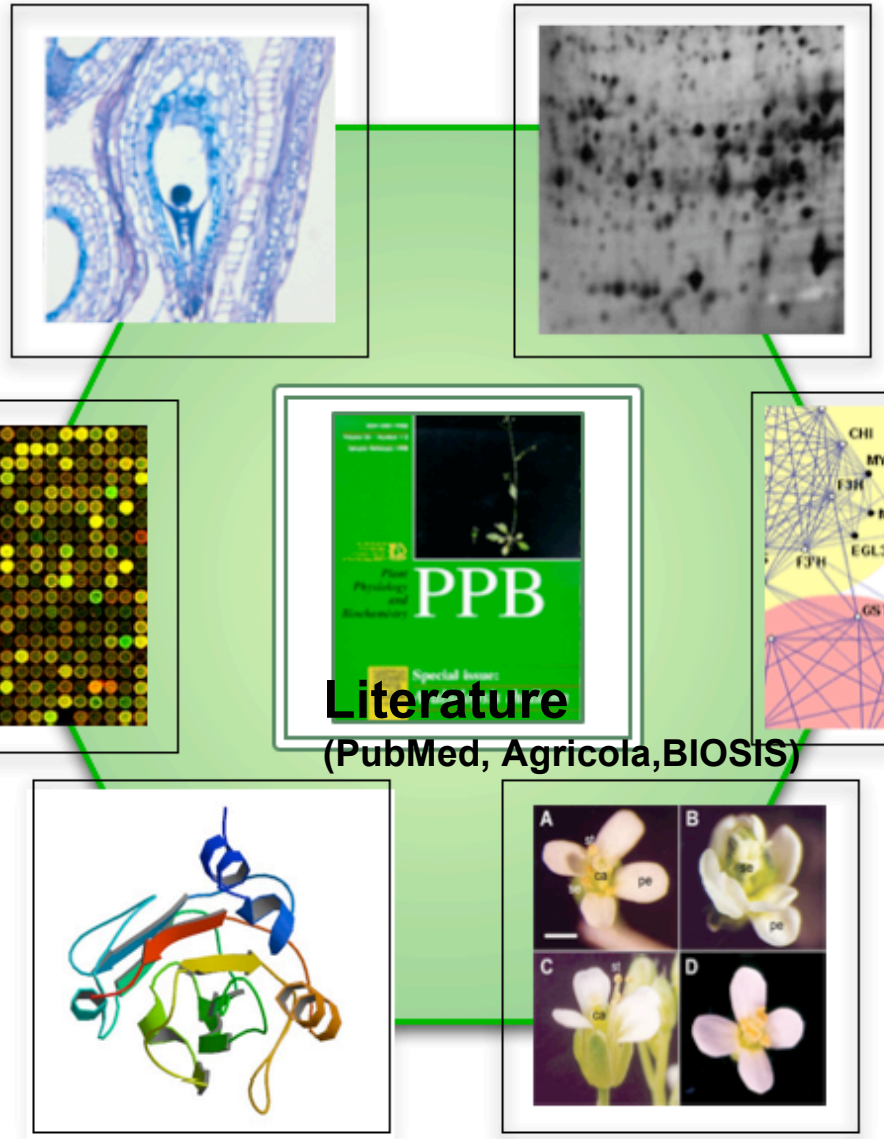
How to exchange annotations



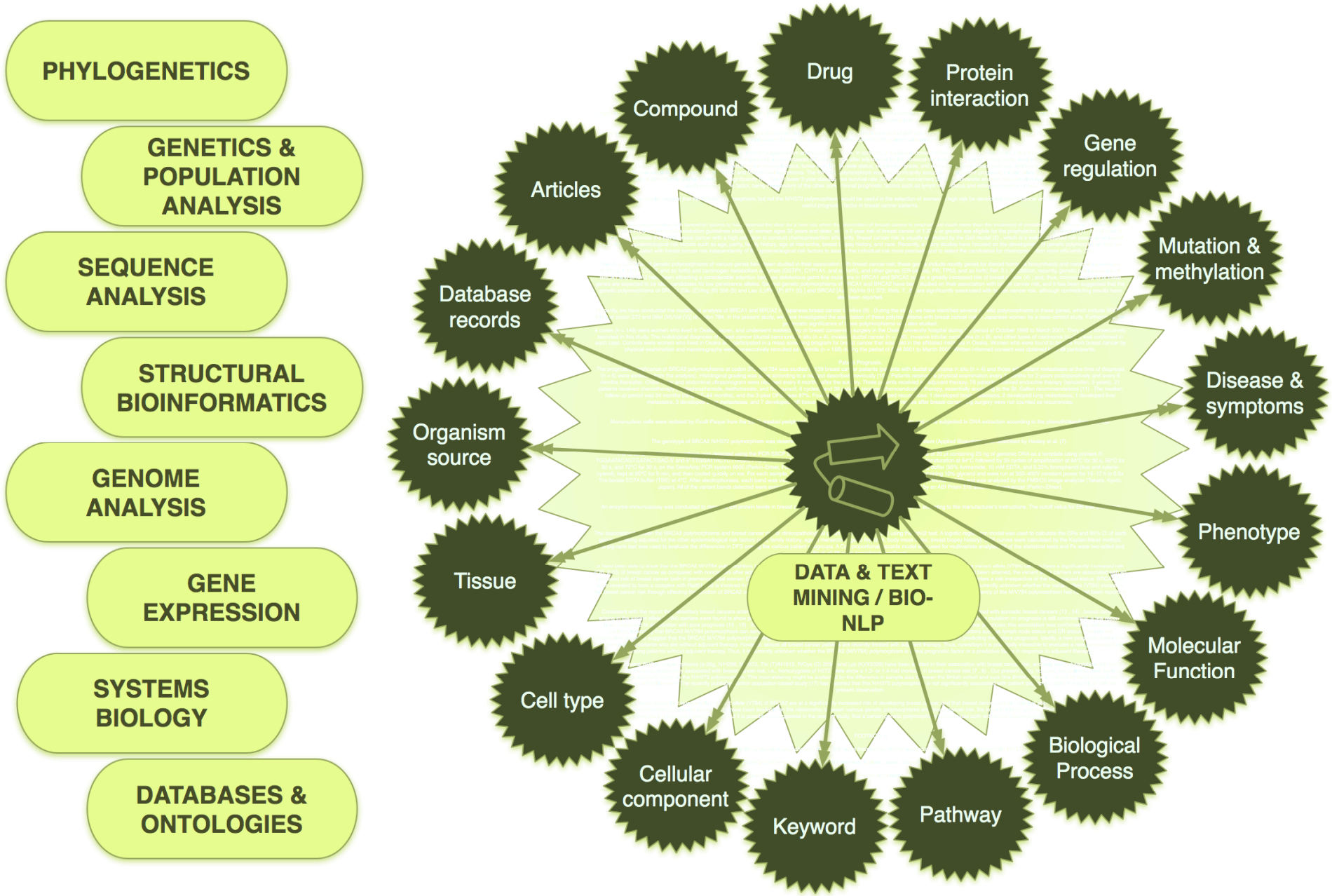
- ✓ Distributed annotation system (DAS) protocol. (MR)
- ✓ Web services. (MR)
- ✓ Database dump. (MR)
- ✓ Biological Web Elements and Registry Embed Code. (HR)

MR = Machine readable
HR = Human readable

Integration of heterogeneous data types



Text mining covers multiple topics



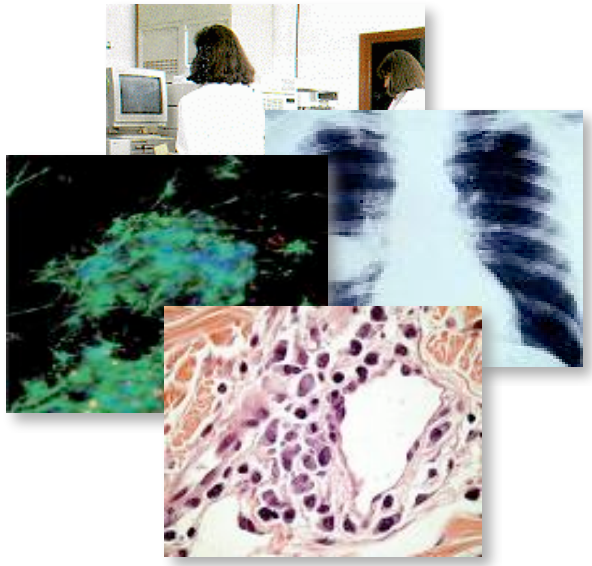
Importance of literature data for Biology

- Life sciences -> generates heterogeneous data types (sequence, structure,..)
- Natural language used for **communicating** scientific discoveries.
- Natural language texts amenable for direct human **interpretation**
- Natural language not only in scientific **articles**, but also patents, reports, newswire, database records, controlled vocabularies (GO terms),…
- **Functional** information & annotations directly or indirectly derived from the literature (curation and electronic annotation).
- **Databases** are generally only capable of covering a small fraction of the biological context information that can be encountered in the literature.
- **Contextual information** of experimental results (cell line, tissue, conditions).
- User demands of better information access (beyond keyword searches)
- Rapid growth of information, manual information extraction not efficient.

Literature and the 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

Pharma

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

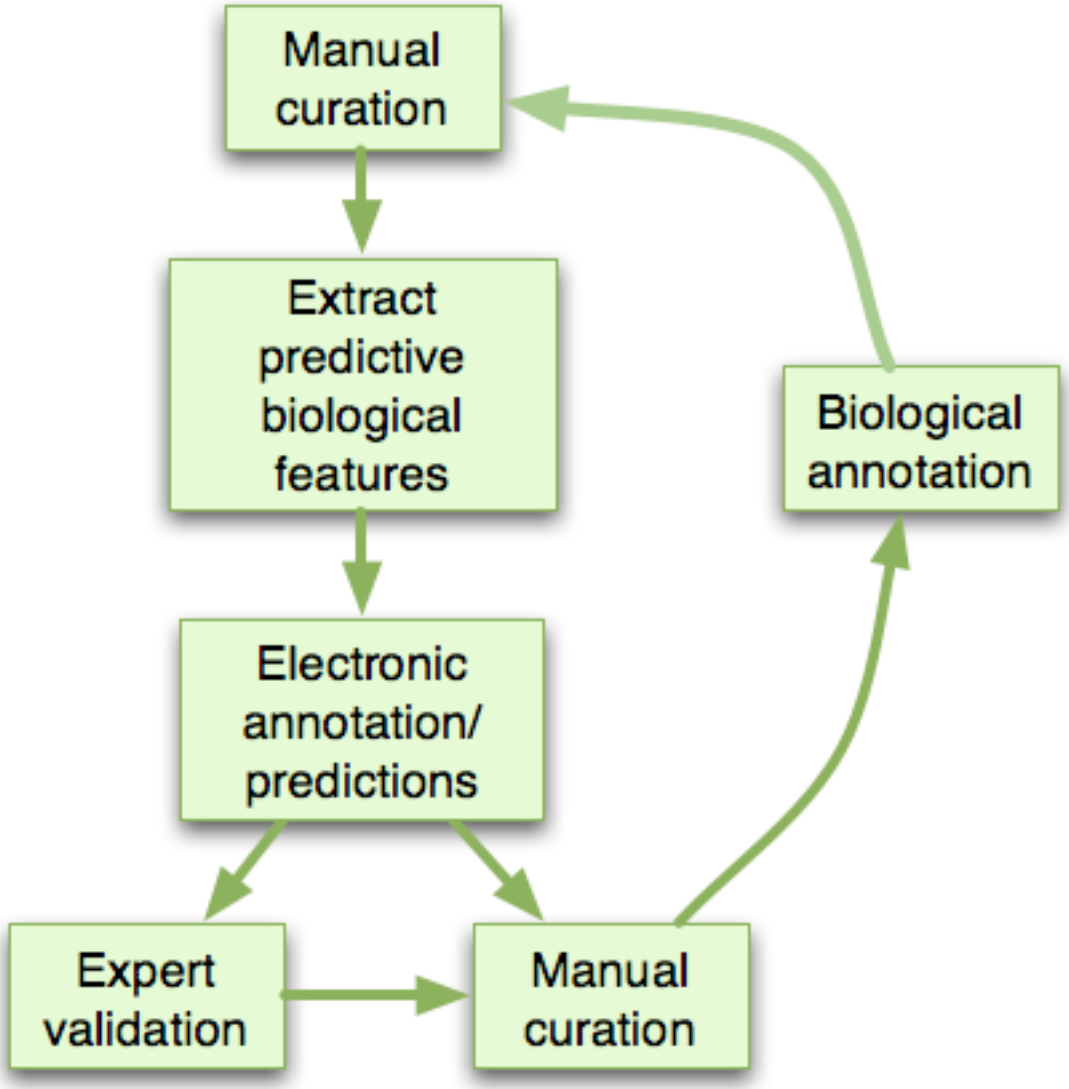
Funding

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

Publ.

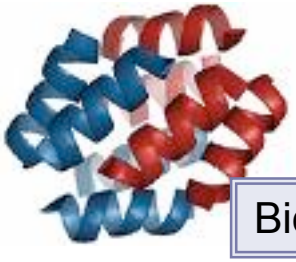
- Find domain experts for specific topics for the peer-review process & detecting potential cases of plagiarism

Literature Gold Standard datasets / DBs

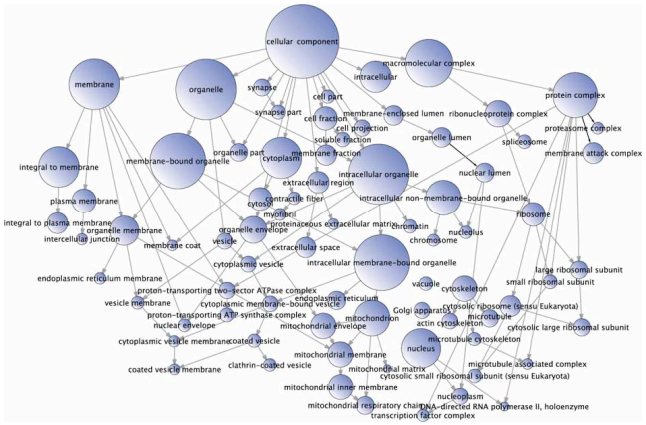


Biocuration: manual literature annotations & databases

Scientific Literature



Bio-entities



Controlled vocabularies

Database curator

(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

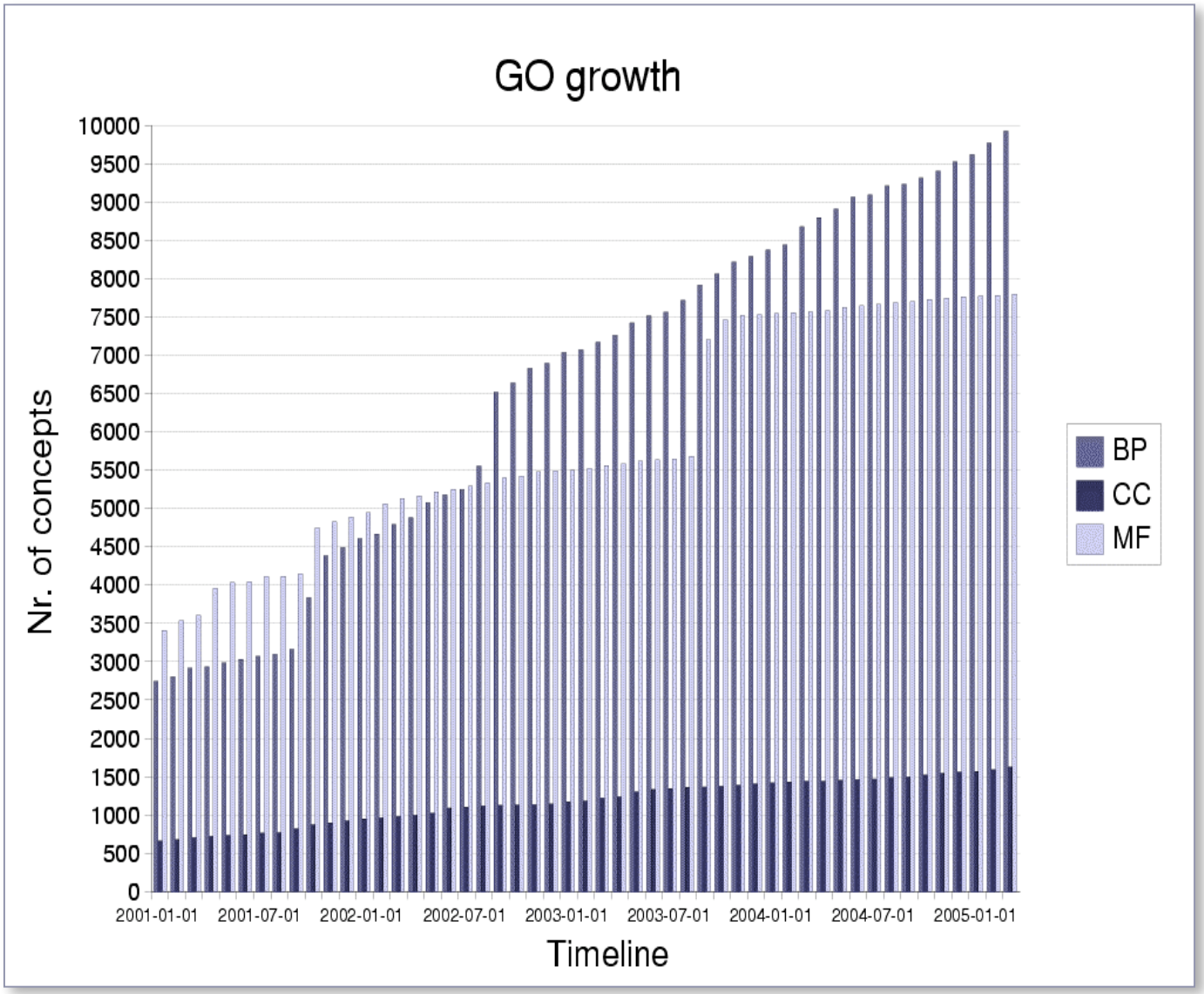
Locus: AT5G43070

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

Curation challenge I: growing number of CV terms



Curation challenge II: growing number of ontologies

> 130

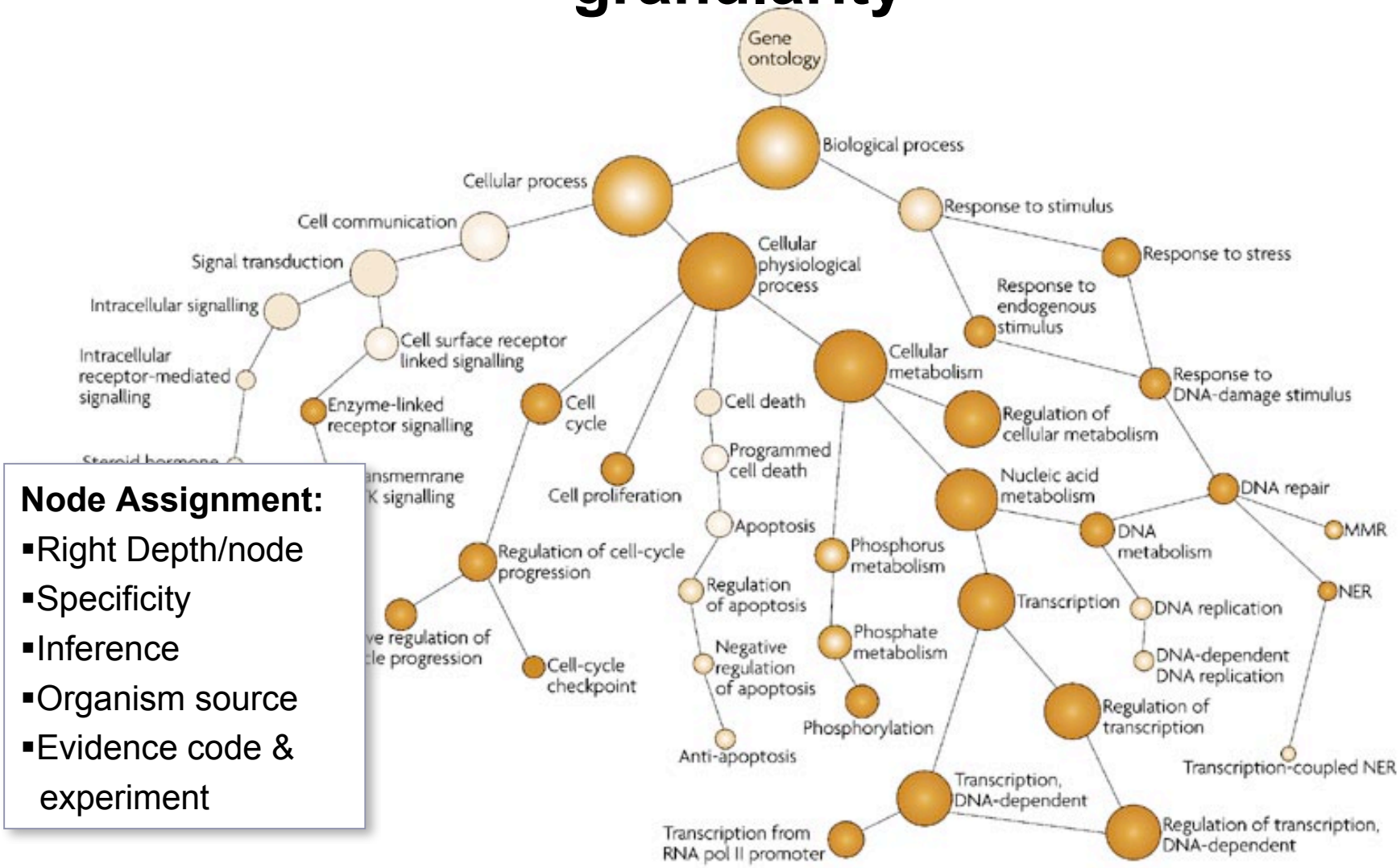
Domain	File
Biological imaging methods	image.obo
Biological process	gene ontology.obo
BRENDA tissue / enzyme source	
C. elegans development	worm development.obo
C. elegans gross anatomy	worm_anatomy.obo
C. elegans phenotype	worm phenotype.obo
Cell type	cell.obo
Cellular component	gene ontology.obo
Cereal plant development	cereals development.obo
Cereal plant trait	plant trait.obo
Chemical entities of biological interest	chebi.obo
Common Anatomy Reference Ontology	caro.obo
Dictyostelium discoideum anatomy	dictyostelium anatomy.obo
Drosophila development	fly development.obo
Drosophila gross anatomy	fly anatomy.obo
Environment Ontology	envo.obo
Event (INOH pathway ontology)	event.obo
Evidence codes	evidence code.obo
eVOC (Expressed Sequence Annotation for Humans)	evoc.obo.tar (v2.7)
Fly taxonomy	fly taxonomy.obo
FlyBase Controlled Vocabulary	flybase controlled vocabulary.obo
Foundational Model of Anatomy (subset)	fma_obo.obo
Fungal gross anatomy	fungal anatomy.obo
Habronattus courtship	protege source
Human developmental anatomy, abstract version	human dev anat abstract.obo
Human developmental anatomy, timed version	human dev anat staged.obo
Human disease	human_disease.obo
Loggerhead nesting	B8467OBO
Maize gross anatomy	zea mays anatomy.obo
Mammalian phenotype	mammalian phenotype.obo
Medaka fish anatomy and development	medaka ontology.obo
Microarray experimental conditions	MGEDOntology.owl
Molecular function	gene ontology.obo
Molecule role (INOH Protein name/family name ontology)	molecule role.obo
Mosquito gross anatomy	mosquito anatomy.obo
Mosquito insecticide resistance	mosquito_insecticide_resistance.obo
	adult mouse anatomy.obo

Mosquito insecticide resistance
Mouse adult gross anatomy
Mouse gross anatomy and develop
Mouse pathology
Multiple alignment
NCBI organismal classification
NCI Thesaurus
NMR-instrument specific component investigations
OBO relationship types
Ontology for biomedical investigatic
Pathway ontology
Phenotypic quality
Physico-chemical methods and prop
Physico-chemical process
Plant environmental conditions
Plant growth and developmental sta
Plant structure
Plasmodium life cycle
Protein domain
Protein modification
protein ontology
Protein-protein interaction
Proteomics data and process prove
Sample processing and separation
Sequence types and features
Suggested Ontology for Pharmacog
Systems Biology
Teleost anatomy and development
Tick gross anatomy
Unit
Xenopus anatomy and developmen

mosquito_insecticide_resistance.obo
adult mouse anatomy.obo
EMAP.obo
mouse pathology.obo
mao.obo
taxonomy.dat
EVS ftp site
nmr.owl
ro.obo
OBI.owl
pathway.obo
quality.obo
fix.obo
rex.obo
environment ontology.obo
po temporal.obo
po anatomy.obo
PLO.ontology
InterPro FTP directory
psi-mod.obo
pro.obo
psi-mi.obo
ProPreO.owl
sep.obo
so.obo
sopharm.owl
SBO_OWL.owl
teleost_anatomy.obo
tick_anatomy.obo

Formats (OBO, OWL, XML, RDF)
(http://www.obofoundry.org)

Curation challenge III: annotation granularity



Node Assignment:

- Right Depth/node
- Specificity
- Inference
- Organism source
- Evidence code & experiment

Computational prediction of cancer-gene function Pingzhao Hu, Gary Bader, Dennis A. Wigle and Andrew Emili Nature Reviews Cancer 7, 23-34 (January 2007)

Nature Reviews | Cancer

Creating reference datasets for Systems Biology applications using text mining

- Manually annotated data repositories: incomplete, fraction of knowledge in literature
- Text mining: to extract, organize and present information for topic of interest
- Enable topic-centric literature navigation
- Assist in construction of manually revised data repositories
- Prioritization of biological entities for experimental characterization
- Facilitate human interpretation of large scale experiments by providing direct literature pointers

- **Automatic retrieval of information relevant to human kinases.**
- **Linking kinase protein mentions to database records (i.e. sequences): protein mention normalization**
- **Extraction of Kinase mutations described in the literature**
- **Integration of information from full text articles, databases and genomic studies**

Krallinger,M et al. **Creating reference datasets for Systems Biology applications using text mining**. *Ann N Y Acad Sci*., (2009) 1158:14-28.


2009
3rd International
Biocuration Conference
April 16-19 Berlin, Germany
Meeting Schedule
Registration
Abstract Submission
Venue & Lodging
Sponsors
Contact
Workshop
Text Mining for the BioCuration Workflow
Organizers:
Lynette Hirschman, MITRE: lynette@mitre.org
Gully APC Burns, ISI/USC: GullyBurns@gmail.com
K. Bretonnel Cohen, University of Colorado: kevin.cohen@gmail.com
Martin Krallinger, CNIO: mkrallinger@cnio.es
Cathy Wu, Georgetown: wuc@georgetown.edu
biocurator.org
The goals of this workshop are to update the BioCurator community on the state of the art in text mining and to elicit the requirements from the BioCurator community for enhanced tools to support the curation workflow.
The workshop will be divided into two parts. The first part will be tutorial in nature and will cover what tools are available, how to integrate components into a curation workflow, and what kind of performance to expect based on available resources. We will also discuss models for curation, including structured digital abstracts.
The second part of the workshop will be interactive, with a focus on understanding the diversity of curation workflows and requirements. For this part, we will invite participants to submit short presentations (5-10 min) on their requirements and their experiences or needs integrating text mining into their curation workflow. We will also discuss how to create partnerships between the bio-text mining tool developers and the BioCurator community.

BIOCURATION WORKFLOW TASKS

TEXT CLASSIFICATION
INFORMATION RETRIEVAL

- FIND CURATION RELEVANT ARTICLE
- TRIAGE TASK
- CLASSIFY AND RANK DOCUMENTS
- FULL TEXT VS ABSTRACTS

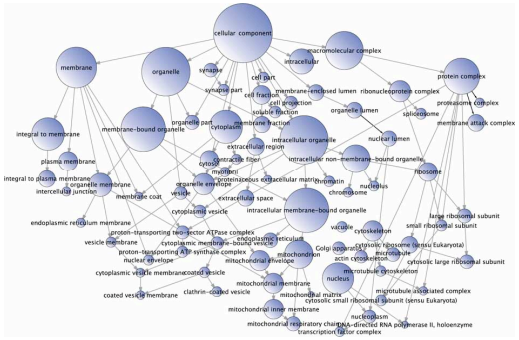
ARTICLE IDENTIFICATION TASK

RELATION, EVENT
EXTRACTION

ANNOTATION EVENT IDENTIFICATION TASK

- FIND RELATIONS: BETWEEN BIO-ENTITIES (PPI, GI), BETWEEN BIO-ENTITIES AND CONTROLLED VOCABULARY TERMS (E.G. GO)
- NEED SUPPORTING EVIDENCE PASSAGES
- COMPLEX PROCESS, OFTEN BASED ON DOMAIN KNOWLEDGE AND EXPERT INFERENCE

GENERAL ANNOTATION PROCESS



- FIND BIO-ENTITY MENTIONS (PROTEINS, GENES, PROTEIN FAMILIES, ENZYMES,...).
- IDENTIFY BIO-ENTITY SOURCE (E.G. ORGANISM)
- MAP TO UNIQUE REFERENCE DATABASE IDENTIFIER (UNIPROT, GENBANK, MODB,...)

BIO-ENTITY IDENTIFICATION TASK

EVIDENTIAL QUALIFIER IDENTIFICATION TASK

- IDENTIFY SUPPORTING EVIDENCE QUALIFIER, EXPERIMENTAL TECHNIQUE, RELEVANT CONTEXTUAL INFORMATION (E.G. GO EVIDENCE CODES, PSI-MI INTERACTION DETECTION METHODS,...)

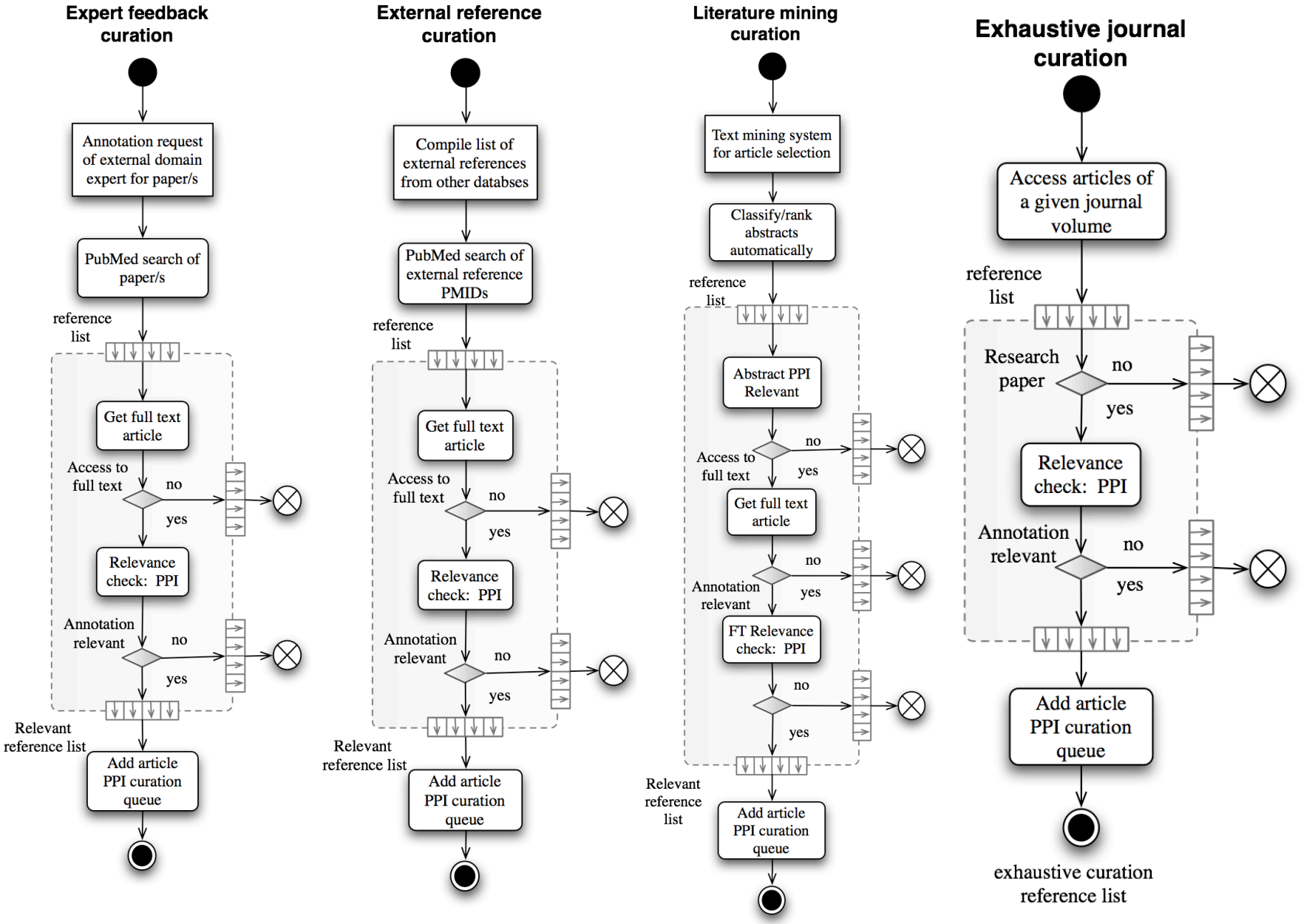
NAMED ENTITY RECOGNITION
& NORMALIZATION

TERM EXTRACTION,
CONTROLLED VOCABULARY
MAPPING

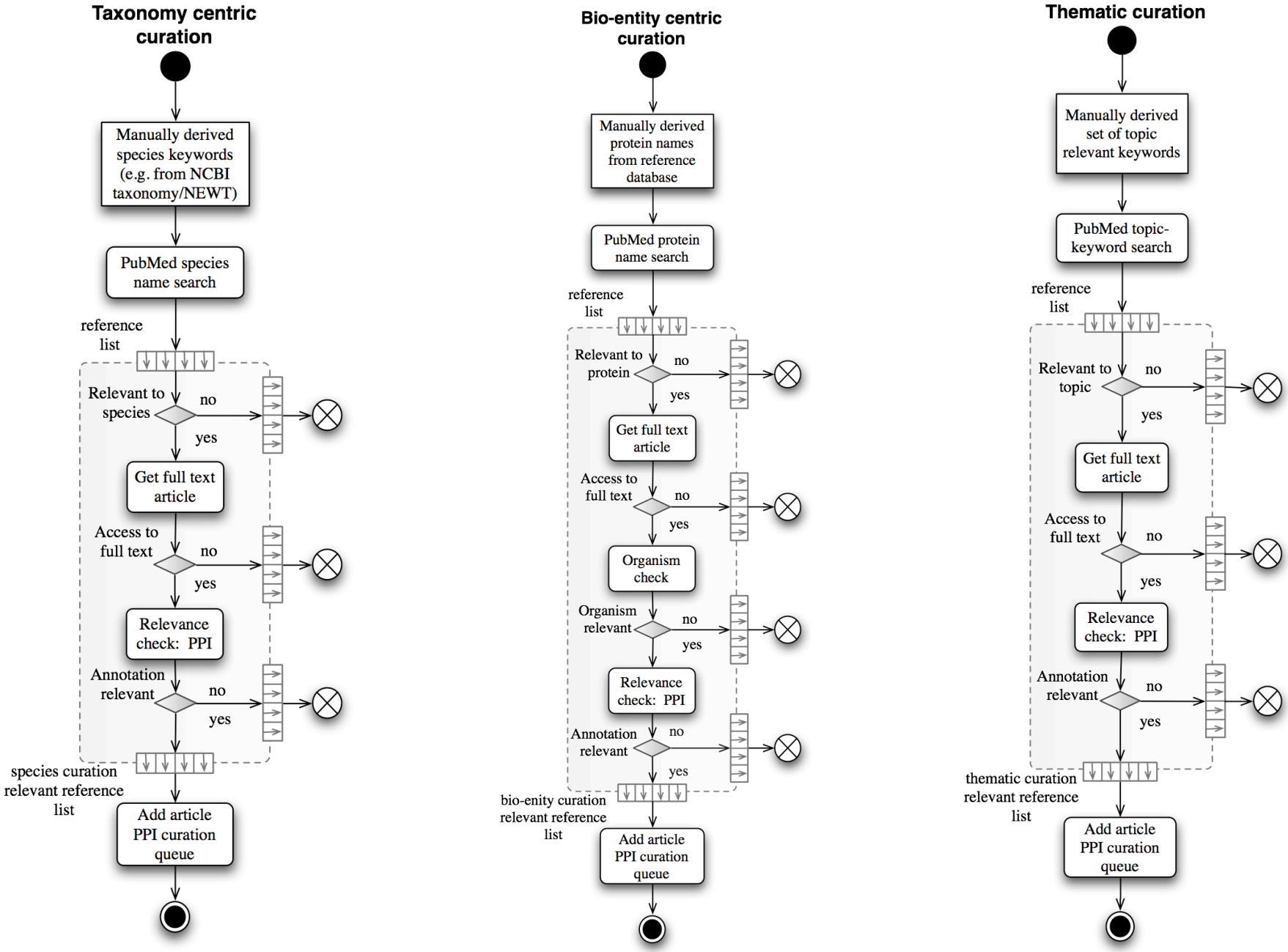
WORKFLOW TASKS AND TEXT MINING

- DEFINE & FORMALIZE INDIVIDUAL STEPS IN THE WORKFLOW
- DETECT WHICH STEPS CAN BE HANDLED THROUGH TEXT MINING ASSISTANCE
- PRIORITIZE MOST TIME CONSUMING STEPS
- FIND SUITABLE TEXT MINING APPROACH FOR EACH PARTICULAR TASK
- EVALUATE ANNOTATION EFFICICIENCY USING TEXT MINING ASSISTANCE
- USER FEEDBACK AND POTENTIAL ITERATIVE IMPROVEMENTS

ARTICLE IDENTIFICATION:TRIAGE TASK (1)



ARTICLE IDENTIFICATION:TRIAGE TASK (2)



ARTICLE IDENTIFICATION:TRIAGE TASK (3)

- Traditionally addressed using keyword searches (e.g. Species names, interaction keywords, gene names, etc,..).
- Importance of triage task depends strongly on the annotation type and criteria used, organism source and literature volume.
- Potential text mining approaches for this task:
- More sophisticated keyword searches and Information retrieval (term weightings, Boolean queries, MeSH terms).
- Use of rules, regular expressions and pattern mining
- Document similarity (eTBLAST, vector space model)
- Machine learning and text categorization approaches (usually requires some sort of labeled text, e.g. PPI relevant articles) to learn which words

  are useful to classify articles as relevant to the topic.
- For full text articles often retrieval is done at the level of text passages
- Sometime the triage task is combined with the bio-entity identification task
- Examples: BCMS, Genomics TREC, PreBIND,…

ANNOTATION EVENT IDENTIFICATION TASK

- Often consist in extraction of some kind of biological relation, e.g. Between two proteins (PPI), proteins and genes (TF and regulated genes),
- Between gene products and functional terms (GO, phenotypes) or between proteins and compounds.
- Often require the identification of some evidential text passages for the annotation event
- Is a very complex process, often domain export knowledge inference.
- Based on interpretation of author provided articles by curator
- Often requires mapping to controlled vocabulary terms and ontologies

- Text Mining approaches for this task:
- Automatic extraction of annotations, often based on sentence co-occurence assumption
- Article, passage, sentence classifiers
- Provide ranked collection of evidence passages
- Some approaches use patterns (trigger words), regular expressions or syntactic relations.

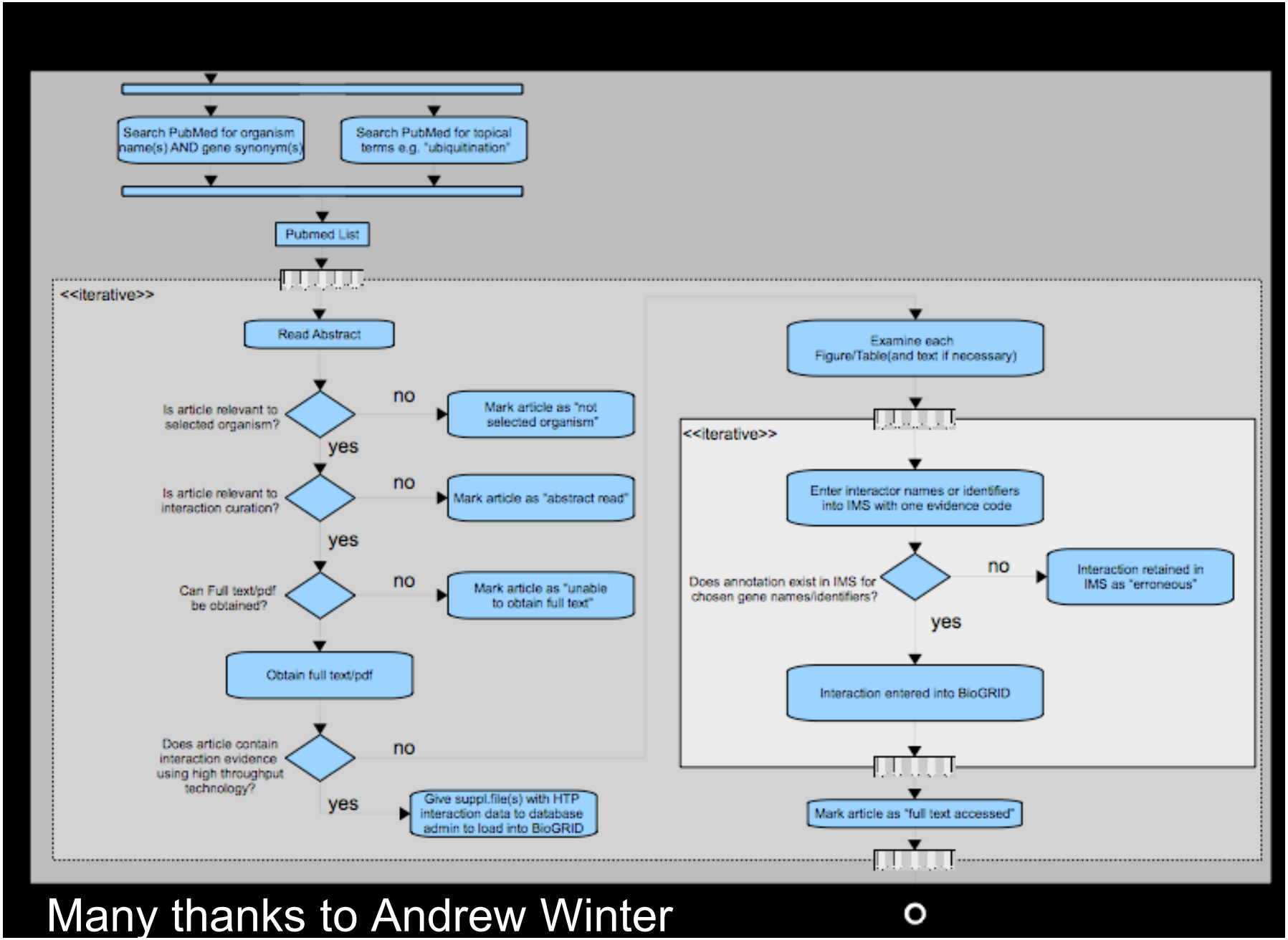
EVIDENTIAL QUALIFIER IDENTIFICATION TASK

- Evidential support for a given annotation important for interpretation.
- Indicative of the reliability of a given annotation and useful also for bioinformatics analysis
- Examples: GO evidence codes, PSI-MI interaction detection methods,

  Oreganno evidence codes, …

- Text mining approaches
- Either addressed as additional information for a given annotation event or through labeling the articles with evidence qualifiers
- Some NLP approaches more concerned with linguistic cues expressing uncertainty or negation
- Example: BioCreative II IMS task

PPI ANNOTATION OF BIOGRID



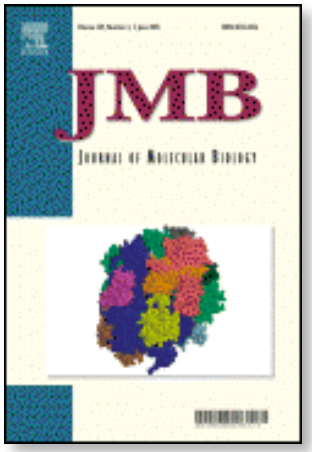
Many thanks to Andrew Winter

Pre-processing scientific articles

- Document Standardization: variety of formats (ASCII, HTML, XML, PDF, scanned PDF, SGML), convert them into a common format and encoding.
- XML /Extensible Markup language, standard way to insert tags onto a text to identify its parts
- OCR (Optical Character Recognition), used to digitalize older literature (PMC Back Issue Digitization initiative).
- Recover article Structure and content
- pdftotext, PDFLib,PDF Concerter
- Tokenization: break a stream of characters into words (tokens), e.g. white space, special chars.
- Each token is an instance of a type
- Stemming and lemmatization: standardize word tokens (e.g. Morphological analysis and
- Inflectional stemming, convert words to their corresponding root form)
- Lexical analysis of the text with the objective of treating digits, hyphens, punctuation marks, and the case of letters
- Elimination of stop-words
- Selection of index terms

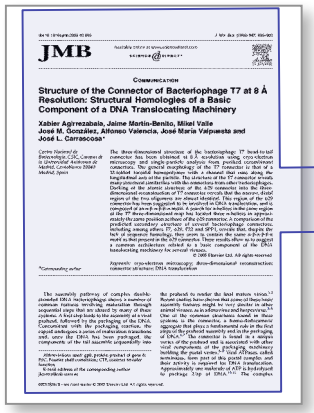
Xu et al. (2008) Improving OCR Performance in Biomedical Literature Retrieval through Preprocessing and Postprocessing. Proc SMBM 08

Basic characteristics: exploring textual data



Considerations of Journal-specific characteristics:

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

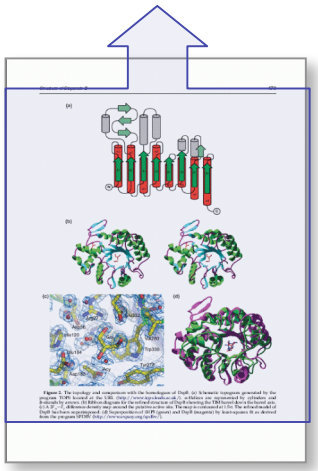


Full text:

- Title
- Authors
- Abstract
- Text Body
- References

Tables & table legends

Figures & figure legends



Processing levels of natural language texts

PROCESSING LEVEL	TASKS AND APPLICATIONS
Character & strings level	Word tokenization, sentence boundary detection, gene symbol recognition, text pattern extraction
Word token level	POS-tagging, parsing, chunking, term extraction, gene mention recognition
Sentence level	Sentence classification and retrieval and ranking, question answering, automatic summarization
Sentence window level	Anaphora resolution
Paragraph & passages level	Detection of rhetorical zones
Whole document level	Document similarity calculation
Multi-document collection level	Document clustering, multi-document summarization

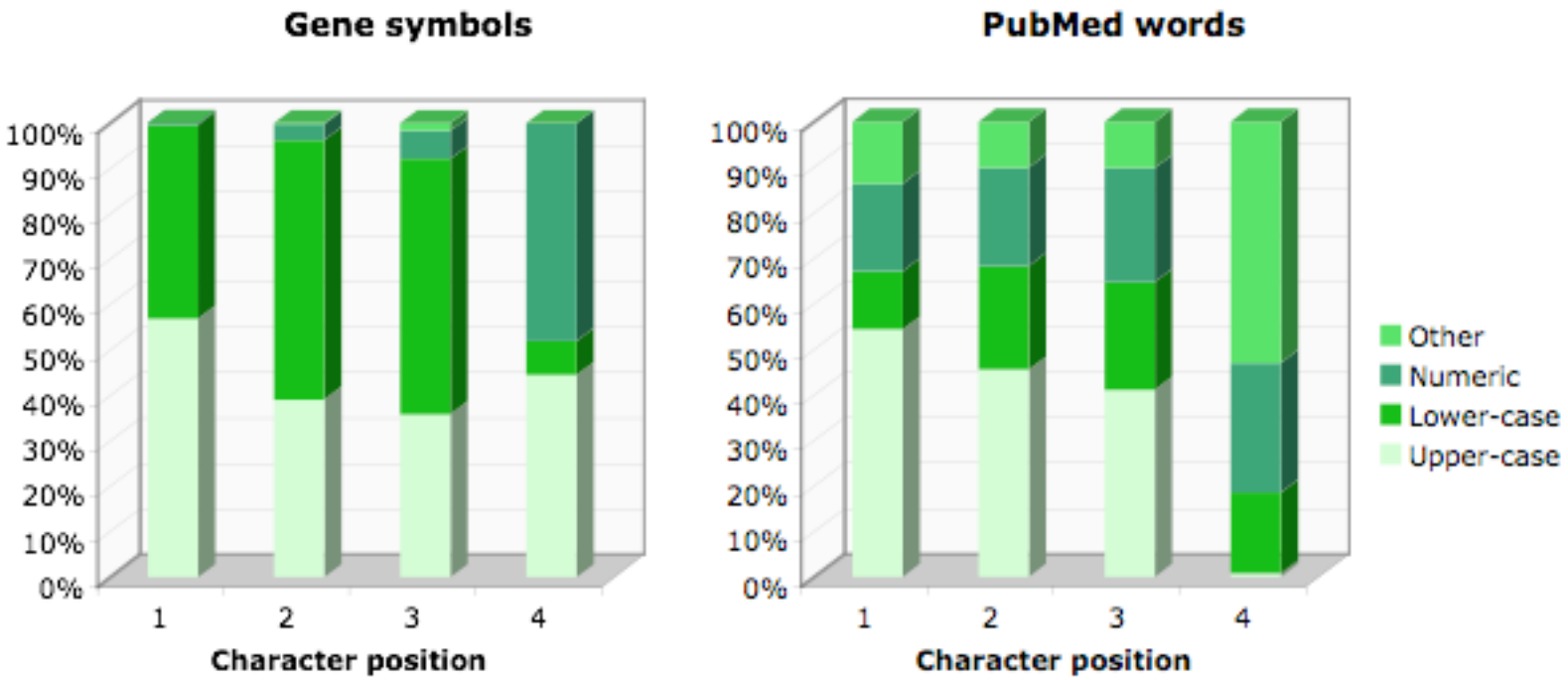
Krallinger M, et al. **Analysis of biological processes and diseases using text mining approaches**. *Methods Mol Biol*. (2009), to appear

Basic characteristics: biomedical literature

- Heavy use of domain specific terminology (12% biochemistry related technical terms*), examples: chemoattractant, fibroblasts, angiogenesis
- Polysemic words (word sense disambiguation), examples: **APC:** (1) Argon Plasma Coagulation (2) Activated Protein C; or **teashirt:** (1) a type of cloth (2) a gene name (tsh).
- Heavy use of acronyms, examples: Activated protein C (APC) , or vascular endothelial growth factor (VEGF)
- Most words with low frequency (data sparseness)

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

Word morphology and gene symbols



Krallinger M, et al. **Analysis of biological processes and diseases using text mining approaches**. *Methods Mol Biol*. (2009), to appear

Basic characteristics: biomedical literature

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

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

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

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

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

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

Variability in Biomedical language

Table 2 | Most frequently used words in various countries

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

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

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

Literature repositories for life sciences

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

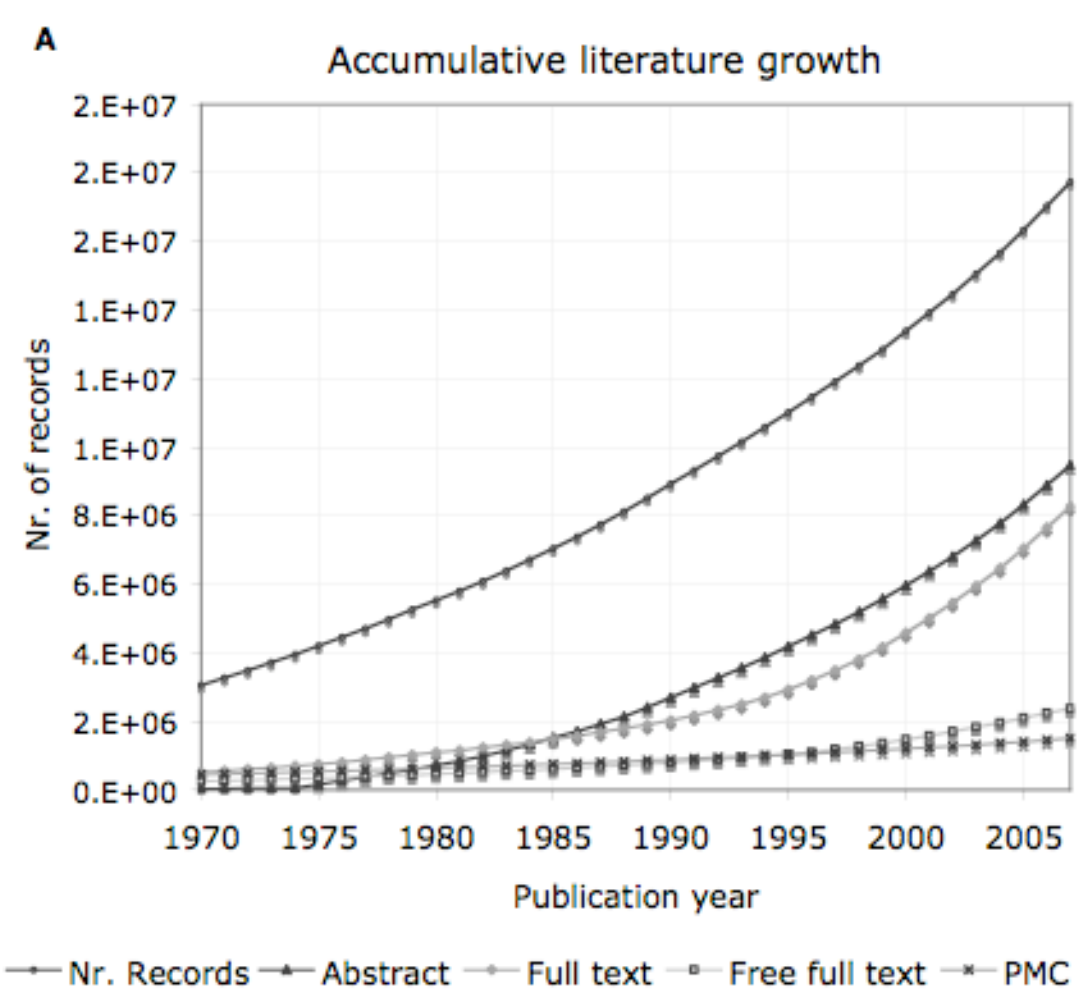
PubMed database

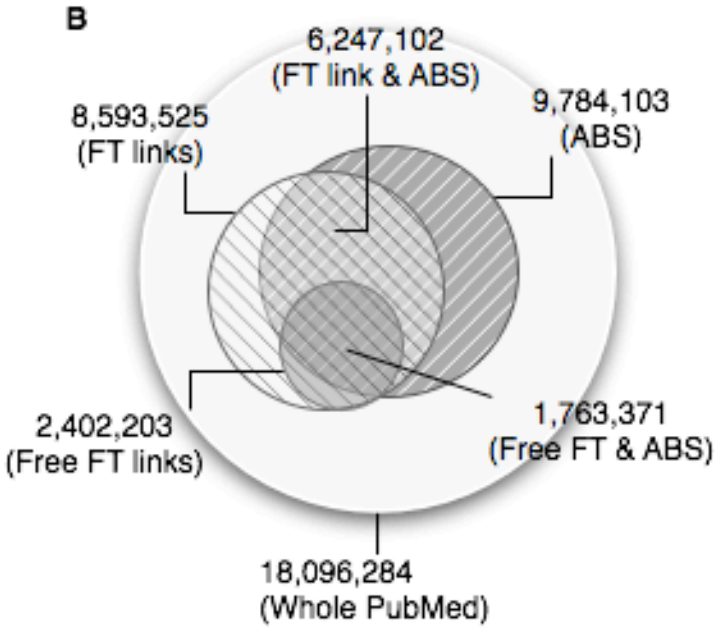
- Scientific articles: new scientific discoveries.
- Citation entries of scientific articles of all biomedical sciences, nursing, biochemistry, engineering, chemistry, environmental sciences, psychology, etc,...
- Developed at the NCBI (NIH).
- Digital library contains more than 16 million citations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

PubMed database

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

PubMed growth





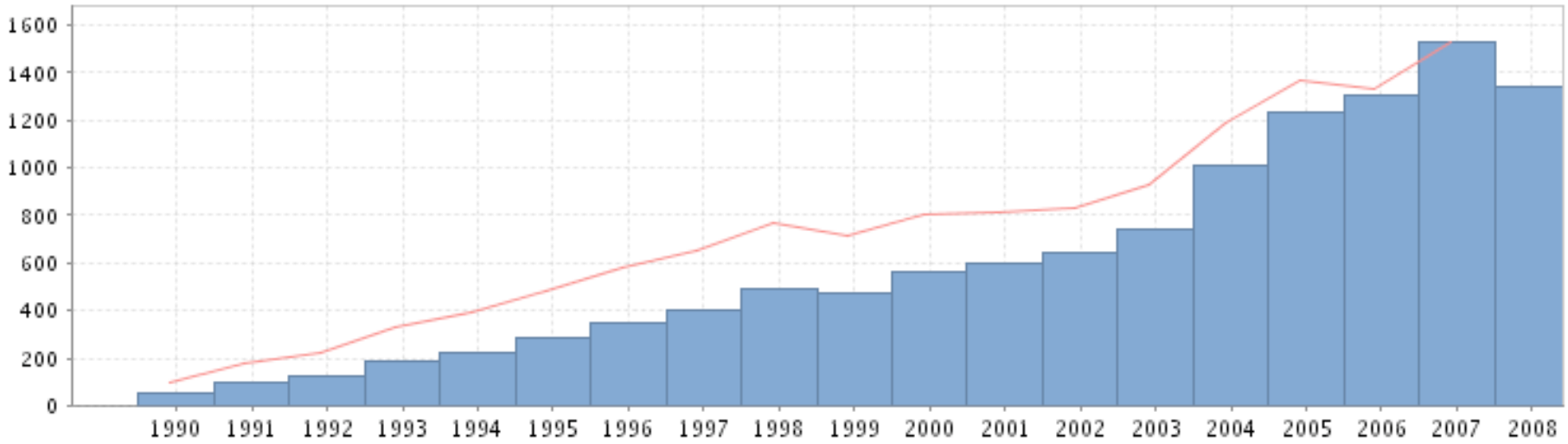
C

Language	Nr. Records
English	14,376,762
German	766,738
Russian	632,650
French	631,486
Japanese	369,915
Italian	285,432
Spanish	255,086

Krallinger M, et al. **Analysis of biological processes and diseases using text mining approaches**. *Methods Mol Biol*. (2009), to appear

PubMed is accumulating over 600,000 new entries every year

Arabidopsis articles in 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>
```

Biomedical corpora and text collections

- Medtag corpus, includes the 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 for 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.

Word	Base Form	Part-Of-Speech	Chunk	Named Entity
HAX-1	**HAX-1**	**NN**	**B-NP**	**B-protein**
associates	*associate*	*VBZ*	*B-VP*	*O*
with	with	IN	B-PP	O
cortactin	**cortactin**	**NN**	**B-NP**	**B-protein**
in	in	IN	B-PP	O
the	the	DT	B-NP	O
apical	apical	JJ	I-NP	O
membrane	membrane	NN	I-NP	O
of	of	IN	B-PP	O
hepatocytes	**hepatocyte**	**NNS**	**B-NP**	**B-cell_type**
.	.	.	O	O
Word	**Morphology**	**Grammar**	**Syntax**	**Semantics**

Krallinger M, et al **Linking genes to literature: text mining, information extraction, and retrieval applications for biology.** Genome Biol. 2008;9 Suppl 2:S8

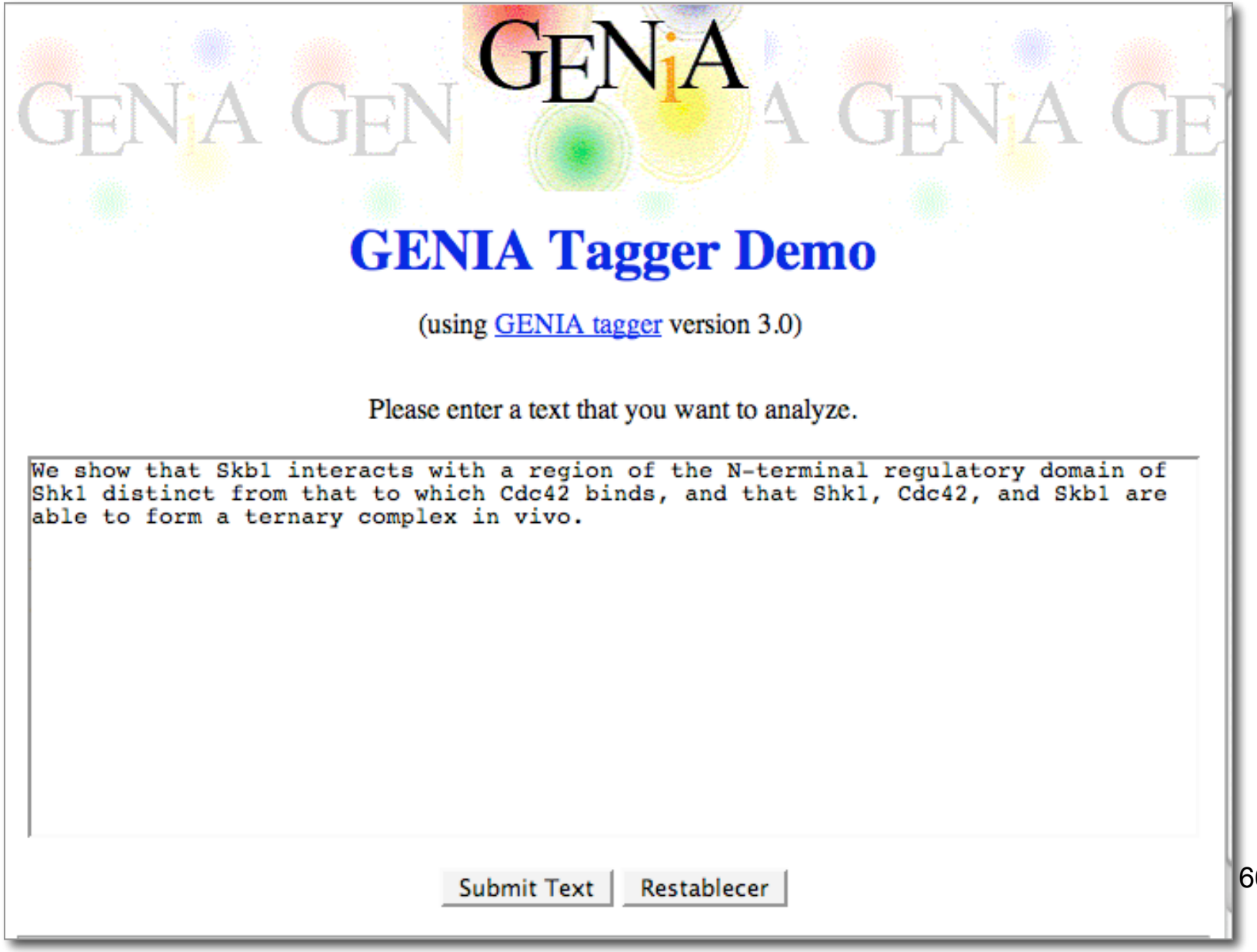
Grammatical features

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

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

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

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 .

http://text0.mib.man.ac.uk/software/geniatagger/index.html

Morphological features

- **Word structure analysis**
- **Rules of how words relate to each other.**
- **Example 1: plural formation rules, e.g.:**

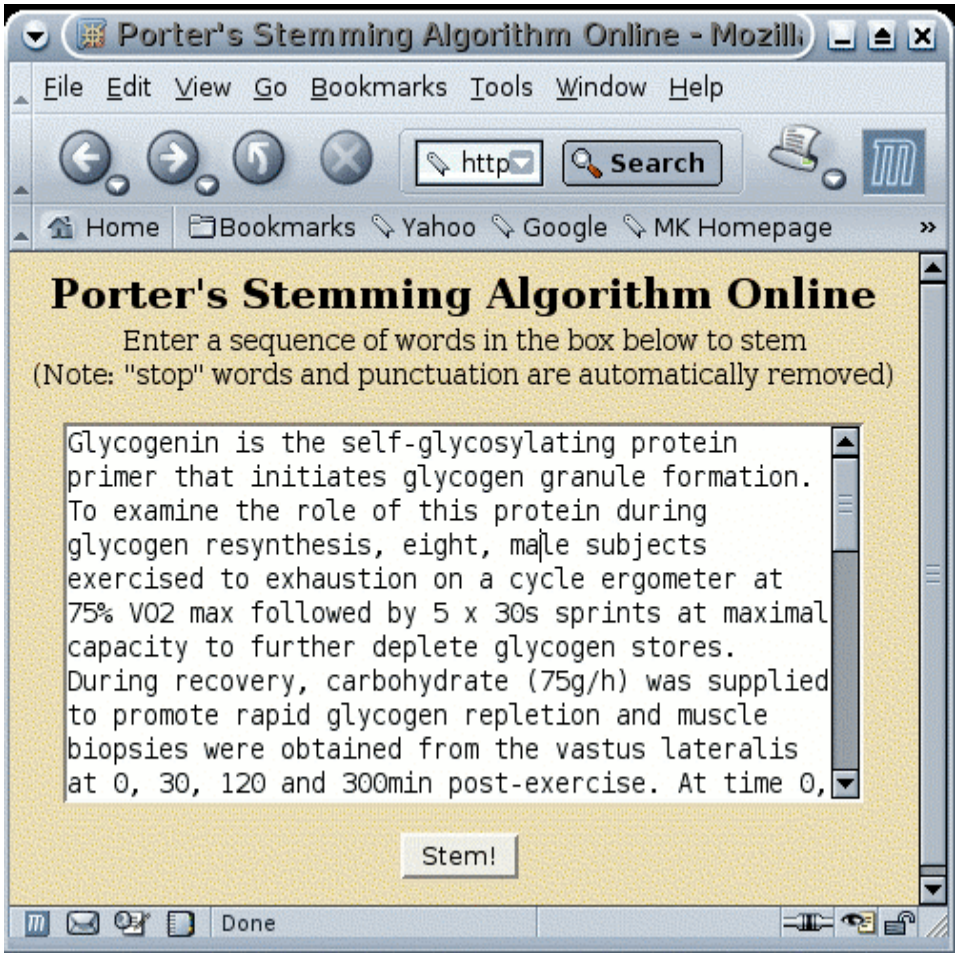
  gene and *genes* or *caspase* and *caspases*
- **Example 2: verb inflection rules, e.g.**

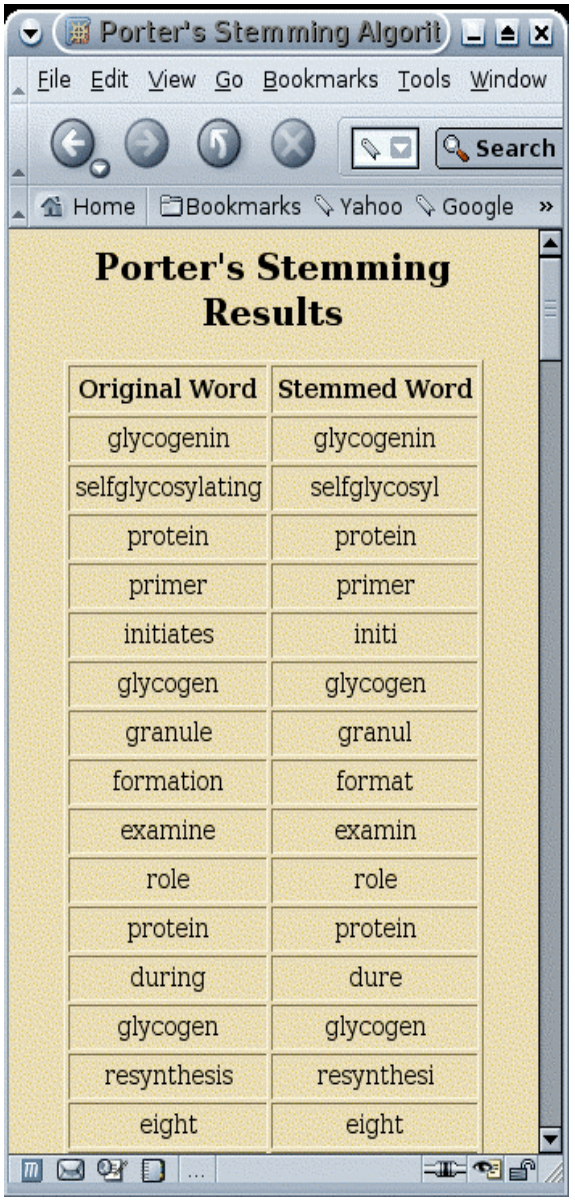
  phosphorylate, *phosphorylates* and *phosphorylating*

  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 {Porter, 1980}**
- **Problem: collapse two semantically different words, e.g:**

  gallery and gall.

Stemmer example results





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

Syntactic features

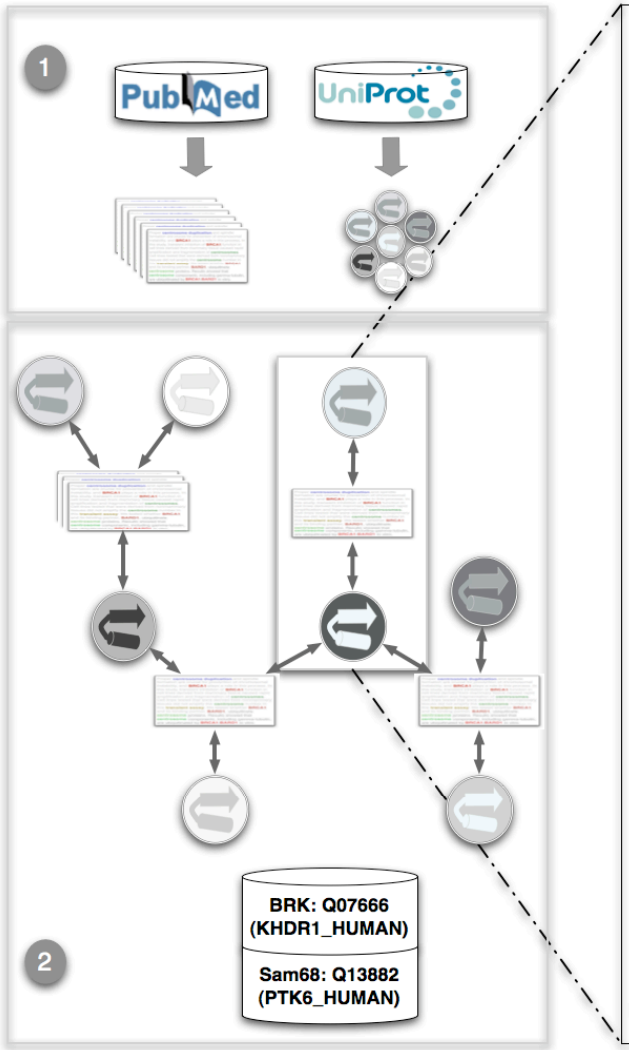
- **Relationships between words in a sentence: syntactic structure**
- **Shallow parsers analyze such relations at a coarse level, identification of phrases (groups of words which function as a syntactic unit).**

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

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

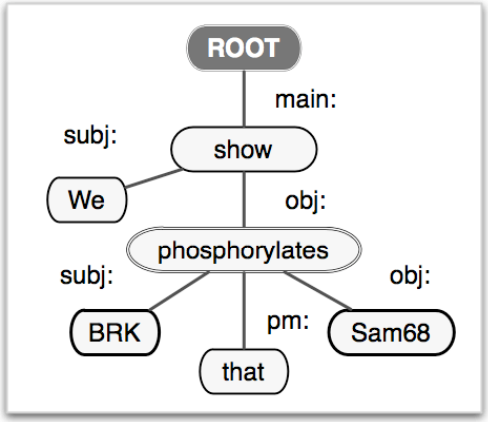
- **Word labeled to corresponding phrase.**
- **Noun phrases (head is a noun, NP) e.g. 'Caspase-3' and 'INF-gamma‘ and verbal phrases (head is a verb, VP).**

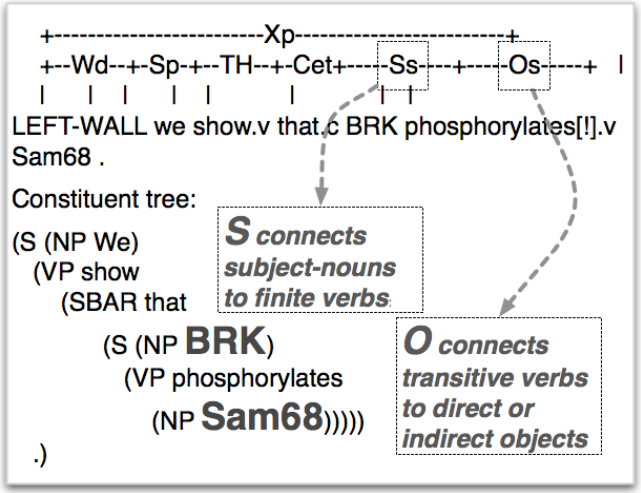
Protein interaction & Syntactic features



We show that BRK phosphorylates Sam68.

WORD	BASE FORM	POS	CHUNK	NER	Phrase syntax
We	We	*PRP*	B-NP	O	nominal head, pro-nominal
show	show	*VBP*	B-VP	O	main verb, indicative present
that	that	*IN*	B-SBAR	O	reposed marker, clause marker
BRK	BRK	*NN*	B-NP	B-protein	nominal head, proper noun, single-word noun phrase
phosphorylates	phosphorylate	*VBZ*	B-VP	O	main verb, indicative present
Sam68	Sam68	*NN*	B-NP	B-protein	nominal head, noun, single-word noun phrase, sentence
.	.	.	O	O	boundary





3

Krallinger M, et al. **Analysis of biological processes and diseases using text mining approaches**. *Methods Mol Biol*. (2009), to appear

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

PMID	Regulator	Regulated	Type	Regulation association evidence sentence				iHOP	WikiG	TAIR	Source
16055635	E2FB AT5G22220	CDKB1;1 AT3G54180	Activation	Indeed, our recent data show that **E2FB** can directly induce the promoter of the Arabidopsis **CDKB1;1** gene (Z. Magyar, unpublished results).	0	None		iHOP	WikiG	TAIR	
16055635	E2FA At2g36010	CDKB1;1 AT3G54180	Activation	It is not clear how **E2FA** could promote the expression of **CDKB1;1**, which has a separate expression window in S- and M-phases (Magyar et al., 1997, 2000).	0	None		iHOP	WikiG	TAIR	

NLP Tasks

- **Information Retrieval (IR)**
- **Text clustering**
- **Text classification**
- **Information extraction (IE)**
- **Question Answering (QA)**
- **Automatic summarization**

Main task types which have been addressed by Bio-NLP systems

Additional task types

- **Natural Language Generation**
- **Anaphora resolution**
- **Text zoning**
- **Machine translation**
- **Text proofing**
- **Speech recognition**

Information Retrieval (IR) and Search Engines

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

VECTOR SPACE MODEL

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

(1) Document indexing ,

(2) Term weighting,

(3) Similarity coefficient

w: term weight

tf: term frequency

idf: inverted document frequency

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

- **Query: a list of terms or even whole documents.**
- **Query as vectors of terms.**

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

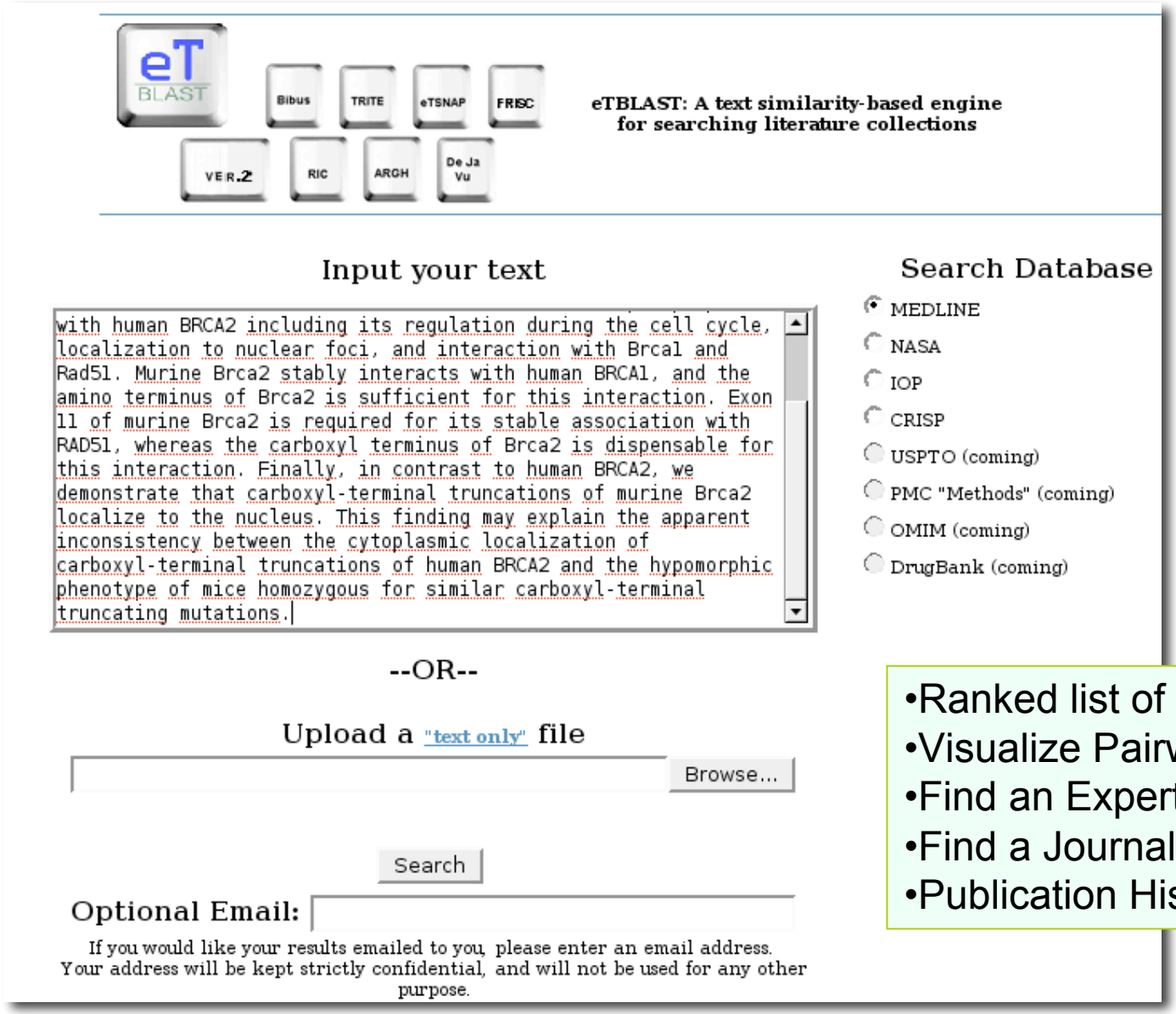
- **Term weighting (w) according to their frequency:**
  within the document (i) & within the document collection (d)
- **Widespread term weighting: tf x idf.**
- **Calculate similarity between those vectors.**
- **Cosine similarity often used.**

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

- **Return a ranked list.**
- **Example: related article search in PubMed**

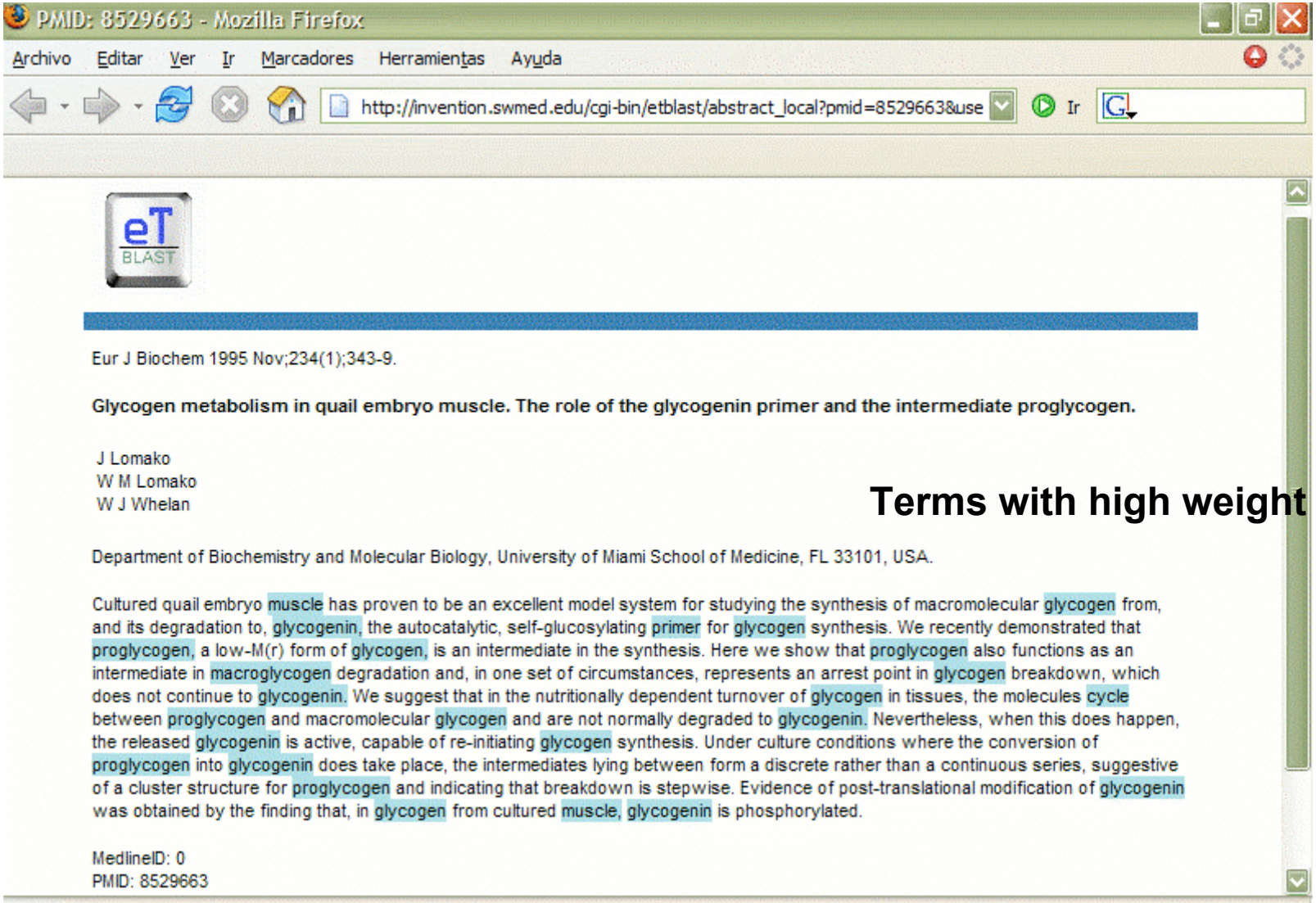
sim(Q,D): similarity between query and document

eTBLAST



- Ranked list of abstracts
- Visualize Pairwise Comparisons
- Find an Expert in this Field
- Find a Journal for your Manuscript
- Publication History of this Topic

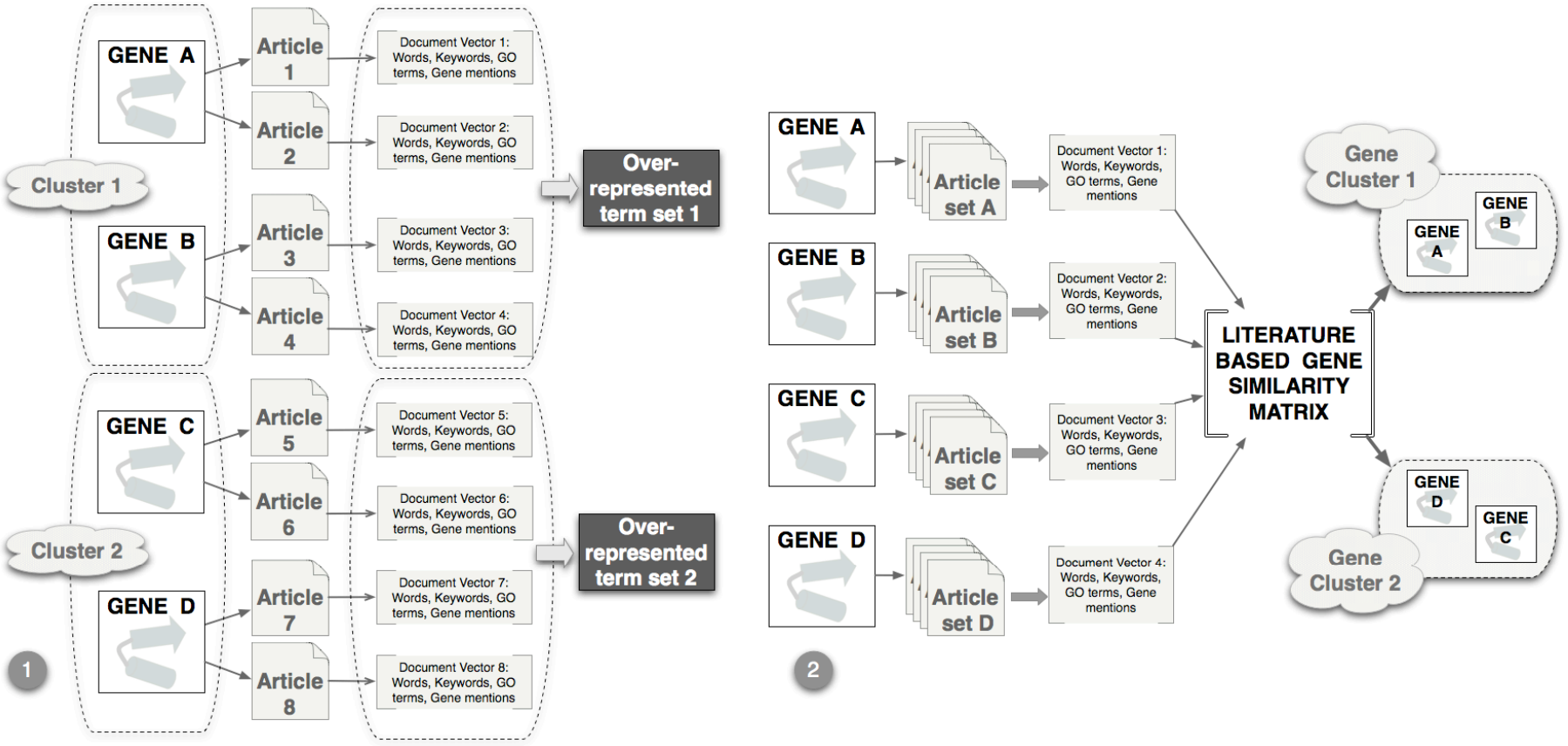
eTBLAST results: high scoring words



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.
•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.

Clustering documents, genes, terms



Krallinger M, et al. **Analysis of biological processes and diseases using text mining approaches**. *Methods Mol Biol*. (2009), to appear

Text classification

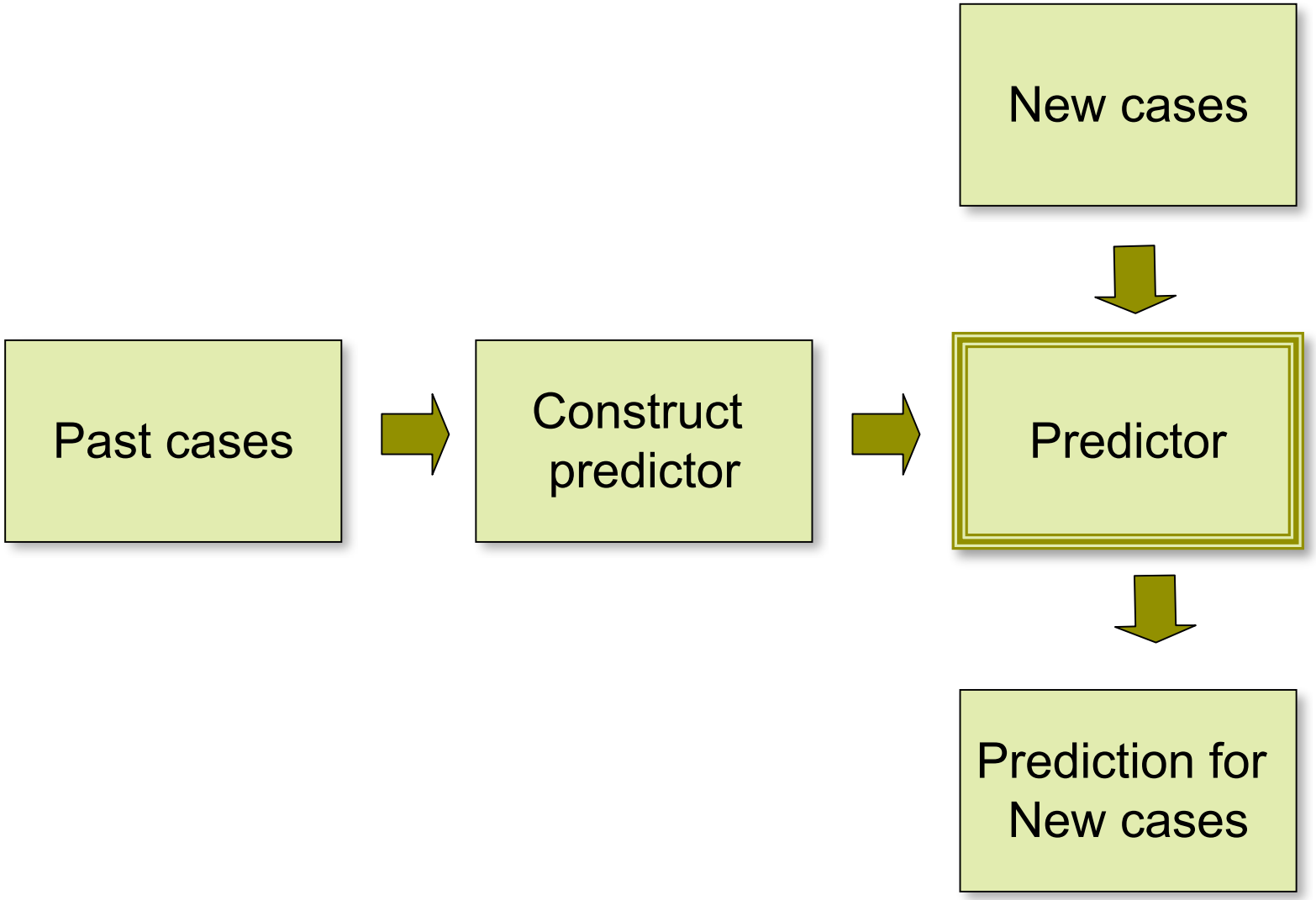
•Common problem in information science.
•Assignment of an electronic document to one or more categories, based on its contents (words).
•Can be divided into two sorts: supervised document classification where some external mechanism (such as human feedback) provides information on the correct classification for documents, and unsupervised document classification.

• Document classification techniques include:

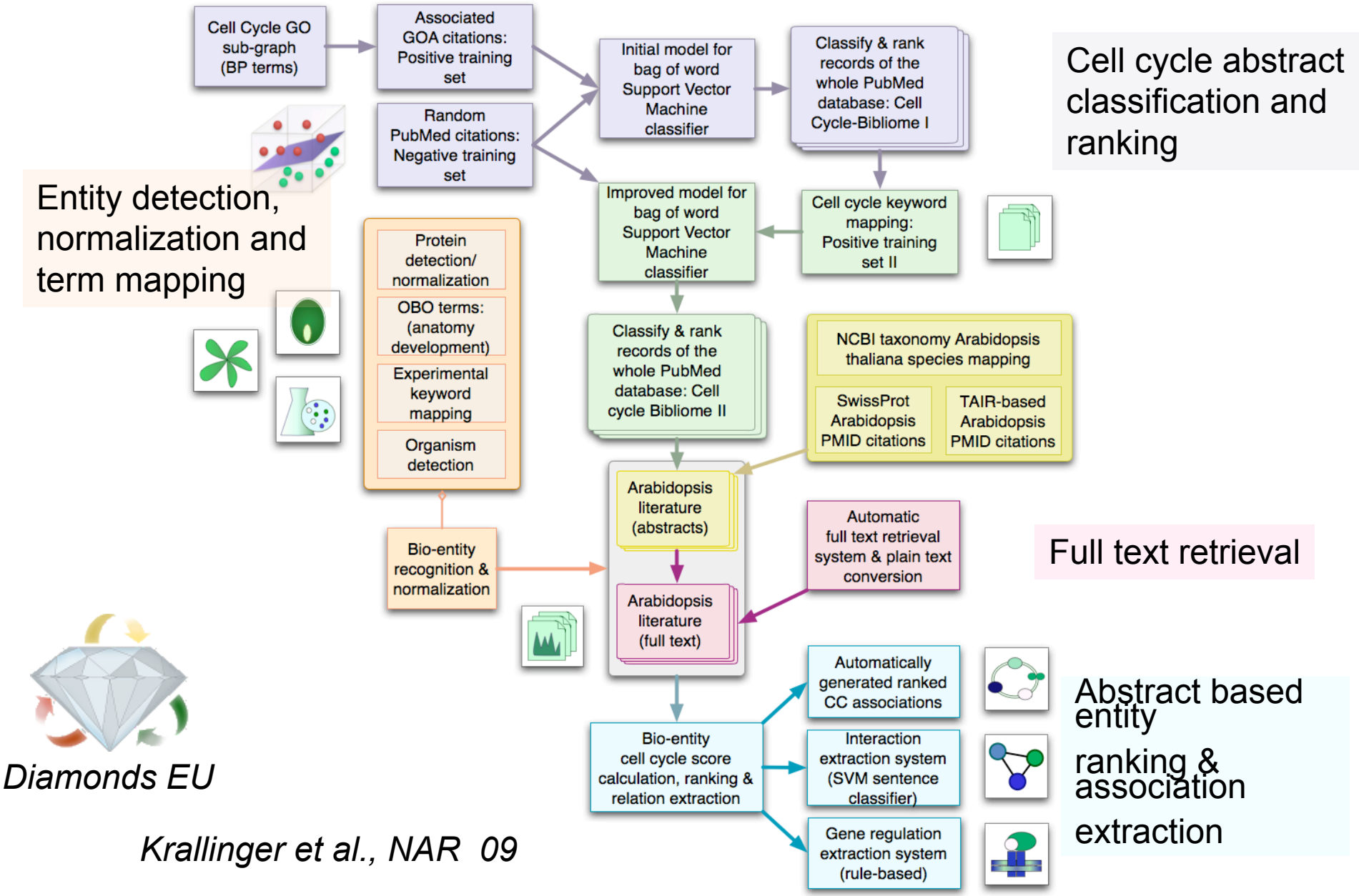
  * naive Bayes classifier
  * tf-idf
  * latent semantic indexing
  * support vector machines
  * artificial neural network
  * kNN
  * decision trees, such as ID3
  * Concept Mining

• Classification techniques have been applied to spam filtering
• Cane use the bow toolkit, SVMlight, LibSVM etc,..

Text classification & supervised learning



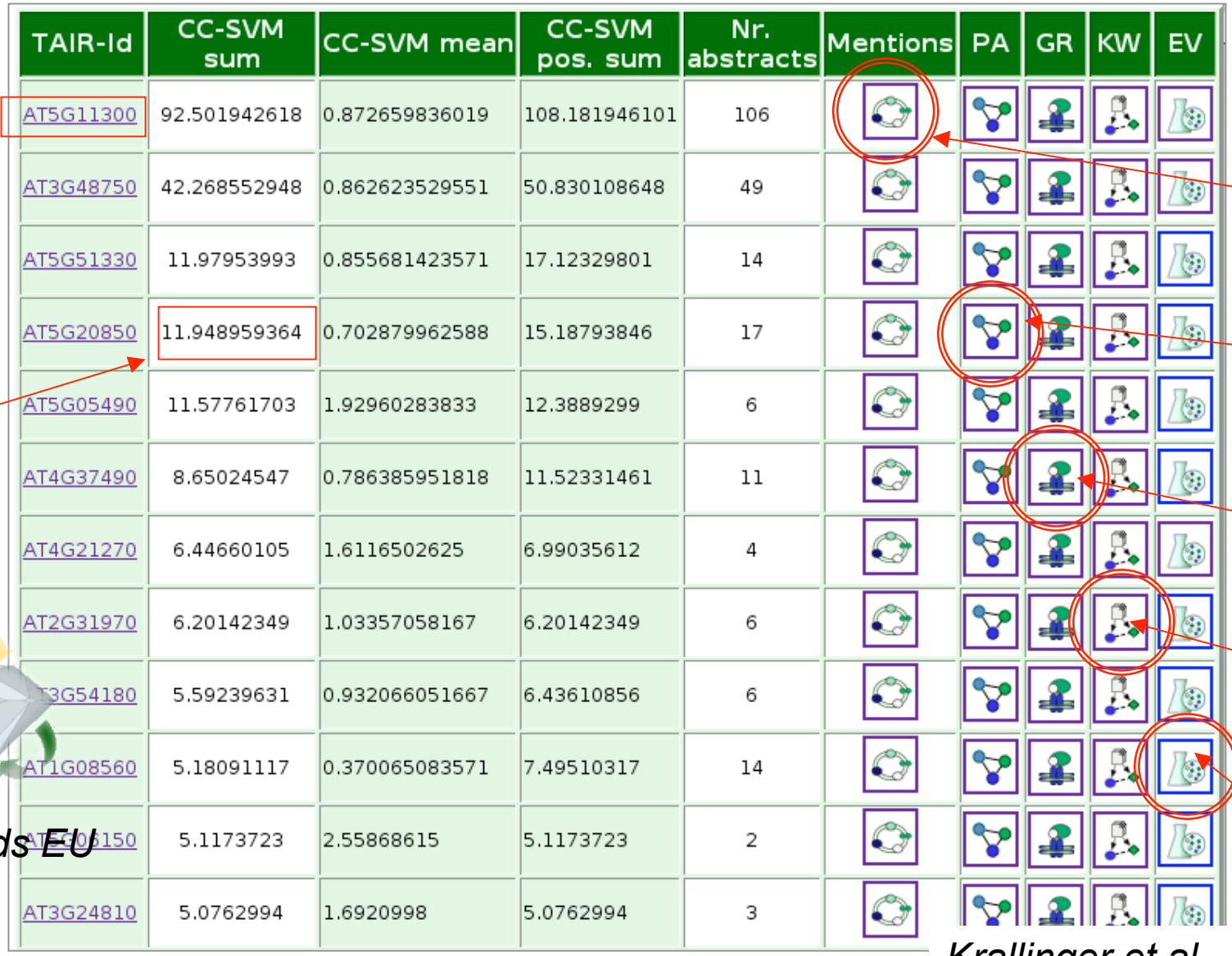
System overview



Diamonds EU

Krallinger et al., NAR 09

Cell cycle protein ranking



TAIR-Id	CC-SVM sum	CC-SVM mean	CC-SVM pos. sum	Nr. abstracts	Mentions	PA	GR	KW	EV
AT5G11300	92.501942618	0.872659836019	108.181946101	106					
AT3G48750	42.268552948	0.862623529551	50.830108648	49					
AT5G51330	11.97953993	0.855681423571	17.12329801	14					
AT5G20850	11.948959364	0.702879962588	15.18793846	17					
AT5G05490	11.57761703	1.92960283833	12.3889299	6					
AT4G37490	8.65024547	0.786385951818	11.52331461	11					
AT4G21270	6.44660105	1.6116502625	6.99035612	4					
AT2G31970	6.20142349	1.03357058167	6.20142349	6					
T3G54180	5.59239631	0.932066051667	6.43610856	6					
AT1G08560	5.18091117	0.370065083571	7.49510317	14					
AT[illegible]150	5.1173723	2.55868615	5.1173723	2					
AT3G24810	5.0762994	1.6920998	5.0762994	3					

TAIR db gene identifier

Sum of CC abstract scores

CC score ranked abstracts

Interaction sentences

Gene regulation

Keyword Co-occurrence

Experiment keywords

Diamonds EU

Krallinger et al., NAR 09

Protein abstract associations

The following results were retrieved for your query: AT3G48750

[PubMed-ID] 7523194

[TAIRID] AT3G48750

[NAMES] cdc2 # cdk2

[TITLE] **Olomoucine, an inhibitor of the cdc2/cdk2 kinases activity, blocks plant cells at the G1 to S and G2 to M cell cycle transitions.**

[ABSTRACT] The cdc2/cdk2 protein kinases play key roles in the cell cycle at two control points: the G1/S transition and the entry into mitosis. Olomoucine, a specific inhibitor of these kinases, was tested in two plant cell systems: Petunia mesophyll protoplasts induced to divide and Arabidopsis thaliana cell suspension cultures. The cell cycle status was analysed from DNA histograms or through continuous labelling of cells with 5-bromodeoxyuridine (BrdUrd) followed by double staining with bis-benzimide (Hoechst 33258) and propidium iodide (PI). Such analyses resolve cells from several generations according to the extent of their DNA replication. Olomoucine was shown to reversibly arrest differentiated Petunia cells induced to divide at G1 phase and cycling Arabidopsis cells in late G1 and G2. A comparison of the effects of aphidicolin, oryzalin and olomoucine suggests that in the Arabidopsis cell suspension culture, a cdc2/cdk2-like kinase is activated at a restriction point in late G1.

Cell cycle terms: G1 phase; mitosis; cell cycle

Species ambiguity scores: 0.666667

Cell cycle scores: 3.76766

[PubMed-ID] 9428718

[TAIRID] AT3G48750

[NAMES] CDC2 # p34cdc2

[TITLE] **Plant CDC2 is not only targeted to the pre-prophase band, but also co-localizes with the spindle, phragmoplast, and chromosomes.**

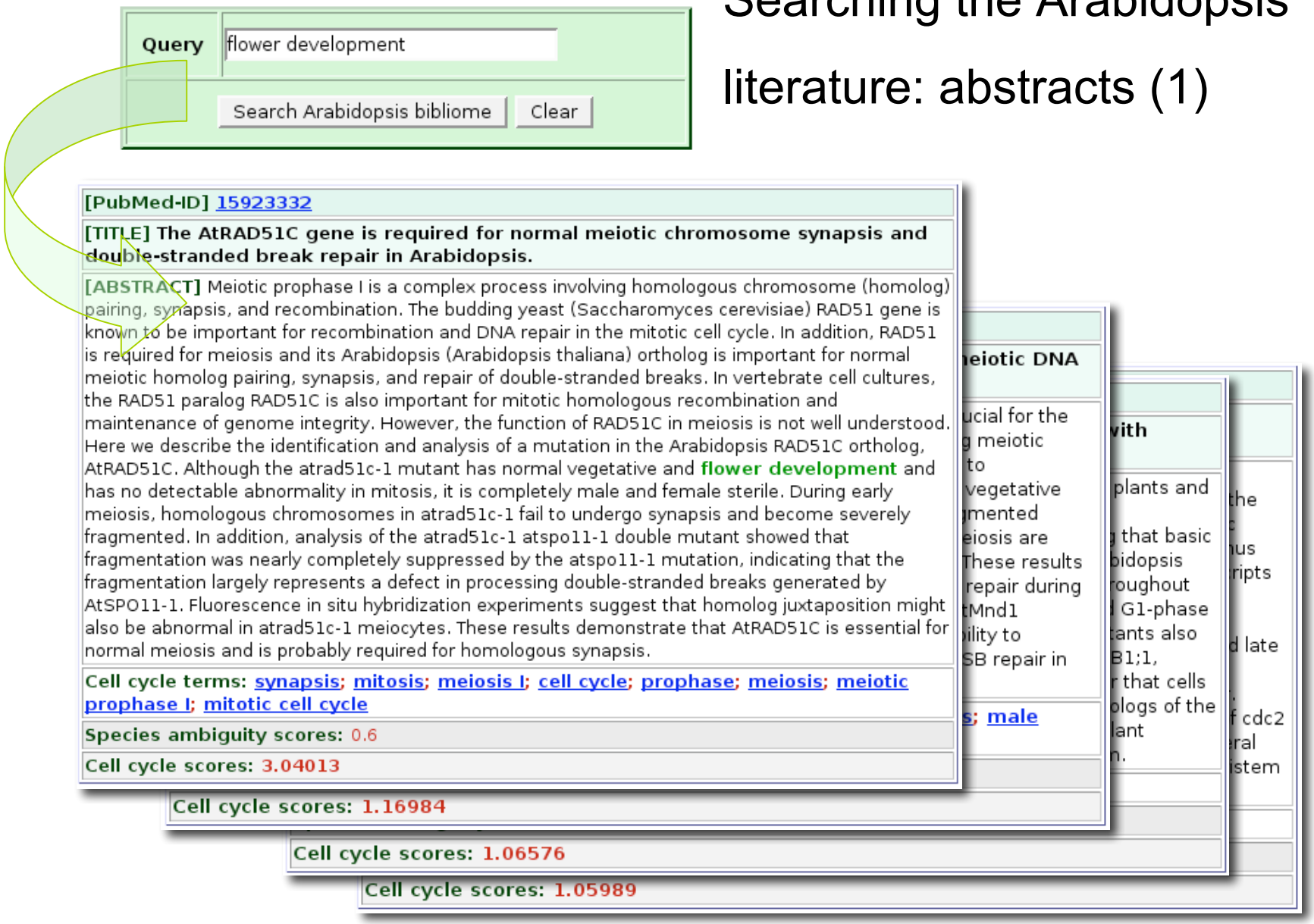
[ABSTRACT] A polyclonal antiserum against the p34cdc2 homologue of Arabidopsis thaliana, CDC2aAt, was used in parallel with a polyclonal antiserum against the PSTAIRE motif to study the subcellular localization of CDC2 during the cell cycle of isolated root tip cells of Medicago sativa. During interphase, CDC2 was located in the nucleus and in the cytoplasm. The cytoplasmic localization persisted during the complete cell cycle, whereas the nuclear signal disappeared at nuclear envelope breakdown. At the beginning of anaphase, the anti-CDC2aAt antibody transiently co-localized with condensed chromosomes. The chromosomal co-localization disappeared as anaphase continued and remained excluded from the separated chromosomes until cytokinesis, when CDC2 re-located to the newly forming nuclei. We also observed a co-localization of CDC2 with three microtubular structures, the pre-prophase band, the spindle, and the phragmoplast.

Cell cycle terms: cell cycle; interphase; prophase; anaphase

Species ambiguity scores: 0.5

Cell cycle scores: 3.4199

Searching the Arabidopsis literature: abstracts (1)



Mitotic spindle relevance protein ranking

481 (P/N)

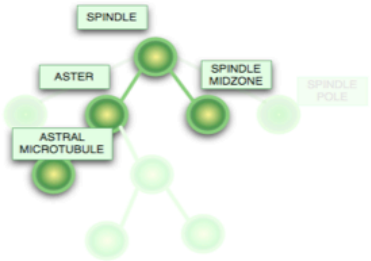
Spindle GO sub-graph (Spindle terms)

Associated GOA citations: Positive training set

Random PubMed citations: Negative training set

Initial model for bag of word Support Vector Machine classifier

Classify & rank records of the whole PubMed database: Spindle-Bibliome I



Spindle keyword mapping: Positive training set II

3498 (P/N)

Improved model for bag of word Support Vector Machine classifier

Experimental keyword mapping

Taxonomy mapping

Protein detection/ normalization

OBO term finding (GO, disease)

Bio-entity recognition and normalization

Classify & rank records of the whole PubMed database: Spindle-Bibliome II

Bio-entity Spindle score calculation, ranking & relation extraction

Automatically generated ranked associations

Spindle Literature Information Architecture

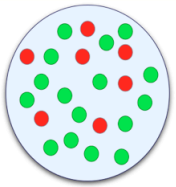
Experimental characterization

Manual curation and annotation

Integration with Bioinformatics approaches

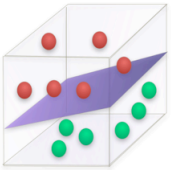
- 123,816 Abstracts
- 1,029,552 Sentences

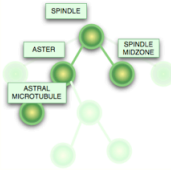
3. OVERVIEW



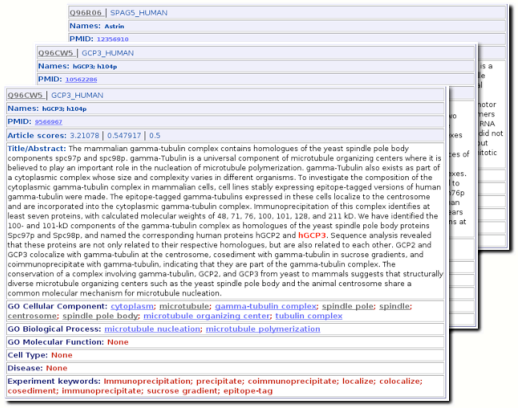
- Small collection of known spindle-relevant proteins
- Large set of potentially spindle-relevant proteins (from large scale experiments)
- Experimentally verified new spindle-relevant proteins

Large scale experiments often result in a list of potentially relevant genes or proteins which require further more detailed characterization.

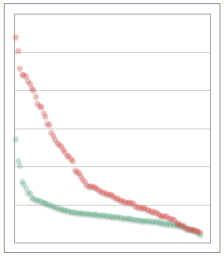




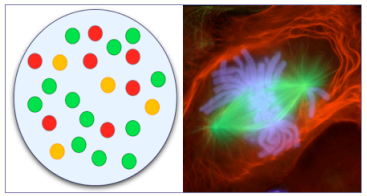
PubMed



Using supervised learning techniques, such as Support Vector Machines it is possible to automatically categorize large collections of textual data. Combined wit protein normalization and keyword mapping, it is possible to both, rank proteins in terms of their contextual association to spindle poles as well as provide ranked pointers to the corresponding textual evidences.



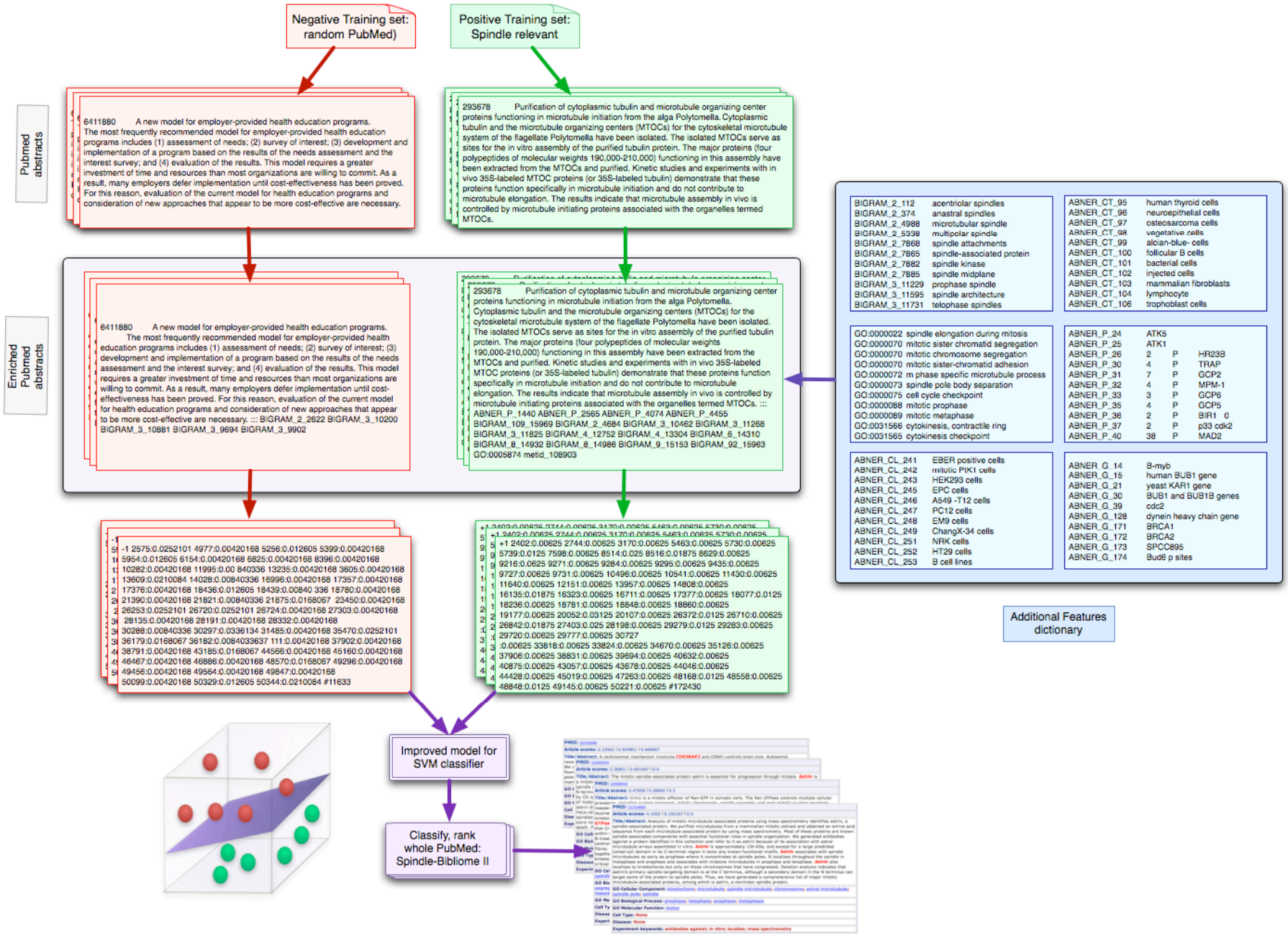
Based on the average document spindle scores Proteins can be ranked according to their spindle associations



The prioritized proteins can then be Experimentally characterized in more detail.

Type	Accession	UniProt-Id	Score
K	Q9BVA0	KTNB1_HUMAN	4.58482
K	O75330	HMMR_HUMAN	3.31712
K	Q96R06	SPAG5_HUMAN	3.23685
K	Q96CW5	GCP3_HUMAN	2.77059
K	Q14980	NUMA1_HUMAN	2.63128
K	Q9UGJ1	GCP4_HUMAN	2.55049
K	Q9BSJ2	GCP2_HUMAN	2.54923
K	Q7Z460	CLAP1_HUMAN	2.52104
P	Q9BZE4	NOG1_HUMAN	2.3877
K	Q96SN8	CK5P2_HUMAN	2.23562
K	O95613	PCNT_HUMAN	2.02624
K	Q13257	MD2L1_HUMAN	2.02318
K	Q9NQS7	INCE_HUMAN	1.92961
K	O43264	ZW10_HUMAN	1.90836
K	Q02224	CENPE_HUMAN	1.89177
K	P46060	RGP1_HUMAN	1.86112
P	O95347	SMC2_HUMAN	1.85301
P	Q9NTJ3	SMC4_HUMAN	1.8333
K	O14965	STK6_HUMAN	1.8002
K	O43683	BUB1_HUMAN	1.79463
K	O95229	ZWINT_HUMAN	1.75663
K	Q9H410	CT172_HUMAN	1.74528
K	P53350	PLK1_HUMAN	1.6669
K	P07900	HS90A_HUMAN	1.64744
K	O43684	BUB3_HUMAN	1.64454
K	O15021	CND1_HUMAN	1.61889
P	P39023	RL3_HUMAN	1.60029
P	O14137	BOP1_HUMAN	1.60029
K	O60566	BUB1B_HUMAN	1.59503

For each protein links to their associated articles as well as keywords, e.g. Spindle and Cell Cycle GO terms and experimental keywords are provided


Negative Training set: random PubMed)
Positive Training set: Spindle relevant
Pubmed abstracts
6411880 A new model for employer-provided health education programs. The most frequently recommended model for employer-provided health education programs includes (1) assessment of needs; (2) survey of interest; (3) development and implementation of a program based on the results of the needs assessment and the interest survey; and (4) evaluation of the results. This model requires a greater investment of time and resources than most organizations are willing to commit. As a result, many employers defer implementation until cost-effectiveness has been proved. For this reason, evaluation of the current model for health education programs and consideration of new approaches that appear to be more cost-effective are necessary.
293678 Purification of cytoplasmic tubulin and microtubule organizing center proteins functioning in microtubule initiation from the alga Polytomella. Cytoplasmic tubulin and the microtubule organizing centers (MTOCs) for the cytoskeletal microtubule system of the flagellate Polytomella have been isolated. The isolated MTOCs serve as sites for the in vitro assembly of the purified tubulin protein. The major proteins (four polypeptides of molecular weights 190,000-210,000) functioning in this assembly have been extracted from the MTOCs and purified. Kinetic studies and experiments with in vivo 35S-labeled MTOC proteins (or 35S-labeled tubulin) demonstrate that these proteins function specifically in microtubule initiation and do not contribute to microtubule elongation. The results indicate that microtubule assembly in vivo is controlled by microtubule initiating proteins associated with the organelles termed MTOCs.
Enriched Pubmed abstracts
6411880 A new model for employer-provided health education programs. The most frequently recommended model for employer-provided health education programs includes (1) assessment of needs; (2) survey of interest; (3) development and implementation of a program based on the results of the needs assessment and the interest survey; and (4) evaluation of the results. This model requires a greater investment of time and resources than most organizations are willing to commit. As a result, many employers defer implementation until cost-effectiveness has been proved. For this reason, evaluation of the current model for health education programs and consideration of new approaches that appear to be more cost-effective are necessary. ::: BIGRAM_2_2622 BIGRAM_3_10200 BIGRAM_3_10881 BIGRAM_3_9694 BIGRAM_3_9902
293678 Purification of cytoplasmic tubulin and microtubule organizing center proteins functioning in microtubule initiation from the alga Polytomella. Cytoplasmic tubulin and the microtubule organizing centers (MTOCs) for the cytoskeletal microtubule system of the flagellate Polytomella have been isolated. The isolated MTOCs serve as sites for the in vitro assembly of the purified tubulin protein. The major proteins (four polypeptides of molecular weights 190,000-210,000) functioning in this assembly have been extracted from the MTOCs and purified. Kinetic studies and experiments with in vivo 35S-labeled MTOC proteins (or 35S-labeled tubulin) demonstrate that these proteins function specifically in microtubule initiation and do not contribute to microtubule elongation. The results indicate that microtubule assembly in vivo is controlled by microtubule initiating proteins associated with the organelles termed MTOCs. ::: ABNER_P_1440 ABNER_P_2565 ABNER_P_4074 ABNER_P_4455 BIGRAM_109_15969 BIGRAM_2_4684 BIGRAM_3_10462 BIGRAM_3_11268 BIGRAM_3_11825 BIGRAM_4_12752 BIGRAM_4_13304 BIGRAM_6_14310 BIGRAM_8_14932 BIGRAM_8_14986 BIGRAM_9_15153 BIGRAM_92_15963 GO:0005874 metid_108903
BIGRAM_2_112 acentriolar spindles
BIGRAM_2_374 anastral spindles
BIGRAM_2_4988 microtubular spindle
BIGRAM_2_5338 multipolar spindle
BIGRAM_2_7868 spindle attachments
BIGRAM_2_7865 spindle-associated protein
BIGRAM_2_7882 spindle kinase
BIGRAM_2_7885 spindle midplane
BIGRAM_3_11229 prophase spindle
BIGRAM_3_11595 spindle architecture
BIGRAM_3_11731 telophase spindles
ABNER_CT_95 human thyroid cells
ABNER_CT_96 neuroepithelial cells
ABNER_CT_97 osteosarcoma cells
ABNER_CT_98 vegetative cells
ABNER_CT_99 alcian-blue- cells
ABNER_CT_100 follicular B cells
ABNER_CT_101 bacterial cells
ABNER_CT_102 injected cells
ABNER_CT_103 mammalian fibroblasts
ABNER_CT_104 lymphocyte
ABNER_CT_106 trophoblast cells
GO:0000022 spindle elongation during mitosis
GO:0000070 mitotic sister chromatid segregation
GO:0000070 mitotic chromosome segregation
GO:0000070 mitotic sister-chromatid adhesion
GO:0000072 m phase specific microtubule process
GO:0000073 spindle pole body separation
GO:0000075 cell cycle checkpoint
GO:0000088 mitotic prophase
GO:0000089 mitotic metaphase
GO:0031566 cytokinesis, contractile ring
GO:0031565 cytokinesis checkpoint
ABNER_P_24 ATK5
ABNER_P_25 ATK1
ABNER_P_26 2 P HR23B
ABNER_P_30 4 P TRAP
ABNER_P_31 7 P GCP2
ABNER_P_32 4 P MPM-1
ABNER_P_33 3 P GCP6
ABNER_P_35 4 P GCP5
ABNER_P_36 2 P BIR1 0
ABNER_P_37 2 P p33 cdk2
ABNER_P_40 38 P MAD2
ABNER_CL_241 EBER positive cells
ABNER_CL_242 mitotic PtK1 cells
ABNER_CL_243 HEK293 cells
ABNER_CL_245 EPC cells
ABNER_CL_246 A549 -T12 cells
ABNER_CL_247 PC12 cells
ABNER_CL_248 EM9 cells
ABNER_CL_249 ChangX-34 cells
ABNER_CL_251 NRK cells
ABNER_CL_252 HT29 cells
ABNER_CL_253 B cell lines
ABNER_G_14 B-myb
ABNER_G_15 human BUB1 gene
ABNER_G_21 yeast KAR1 gene
ABNER_G_30 BUB1 and BUB1B genes
ABNER_G_39 cdc2
ABNER_G_128 dynein heavy chain gene
ABNER_G_171 BRCA1
ABNER_G_172 BRCA2
ABNER_G_173 SPCC895
ABNER_G_174 Bud6 p sites
Additional Features dictionary
-1 2575:0.0252101 4977:0.00420168 5256:0.012605 5399:0.00420168 5954:0.012605 6154:0.00420168 6825:0.00420168 8396:0.00420168 10282:0.00420168 11995:0.00 840336 13235:0.00420168 3605:0.00420168 13609:0.0210084 14028:0.00840336 16996:0.00420168 17357:0.00420168 17376:0.00420168 18436:0.012605 18439:0.00840 336 18780:0.00420168 21390:0.00420168 21821:0.00840336 21875:0.0168067 23450:0.00420168 26253:0.0252101 26720:0.0252101 26724:0.00420168 27303:0.00420168 28135:0.00420168 28191:0.00420168 28332:0.00420168 30288:0.00840336 30297:0.0336134 31485:0.00420168 35470:0.0252101 36179:0.0168067 36182:0.0084033637 111:0.00420168 37902:0.00420168 38791:0.00420168 43185:0.0168067 44566:0.00420168 45160:0.00420168 46467:0.00420168 46886:0.00420168 48570:0.0168067 49296:0.00420168 49456:0.00420168 49564:0.00420168 49847:0.00420168 50099:0.00420168 50329:0.012605 50344:0.0210084 #11633
+1 2402:0.00625 2744:0.00625 3170:0.00625 5463:0.00625 5730:0.00625 5739:0.0125 7598:0.00625 8514:0.025 8516:0.01875 8629:0.00625 9216:0.0625 9271:0.00625 9284:0.00625 9295:0.00625 9435:0.00625 9727:0.00625 9731:0.00625 10496:0.00625 10541:0.00625 11430:0.00625 11640:0.00625 12151:0.00625 13957:0.00625 14808:0.00625 16135:0.01875 16323:0.00625 16711:0.00625 17377:0.00625 18077:0.0125 18236:0.00625 18781:0.00625 18848:0.00625 18860:0.00625 19177:0.00625 20052:0.03125 20107:0.00625 26372:0.0125 26710:0.00625 26842:0.01875 27403:0.025 28198:0.00625 29279:0.0125 29283:0.00625 29720:0.00625 29777:0.00625 30727 :0.00625 33818:0.00625 33824:0.00625 34670:0.00625 35126:0.00625 37906:0.00625 38831:0.00625 39694:0.00625 40632:0.00625 40875:0.00625 43057:0.00625 43678:0.00625 44046:0.00625 44428:0.00625 45019:0.00625 47263:0.00625 48168:0.0125 48558:0.00625 48848:0.0125 49145:0.00625 50221:0.00625 #172430
Improved model for SVM classifier
Classify, rank whole PubMed: Spindle-Bibliome II

Information Extraction

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

Krallinger M, et al **Linking genes to literature: text mining, information extraction, and retrieval applications for biology.** Genome Biol. 2008;9 Suppl 2:S8

TAGGING BIO-ENTITIES IN TEXT

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

SOME TRICKY CASES OF GENE TAGGING

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

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

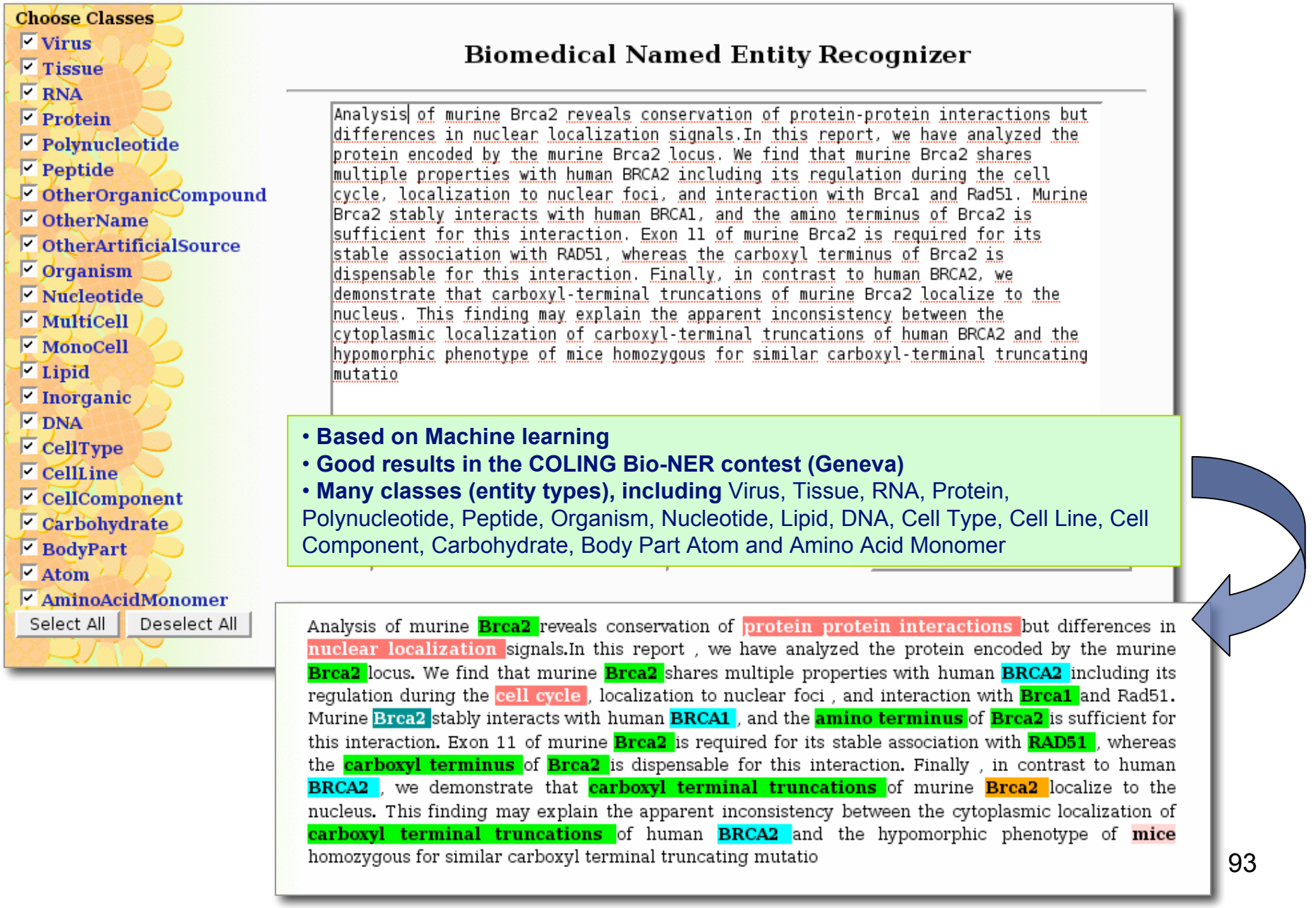
(3) The Drosophila takeout gene is regulated by the somatic sex-determination pathway and affects male courtship behavior. [PMID:12435630]

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

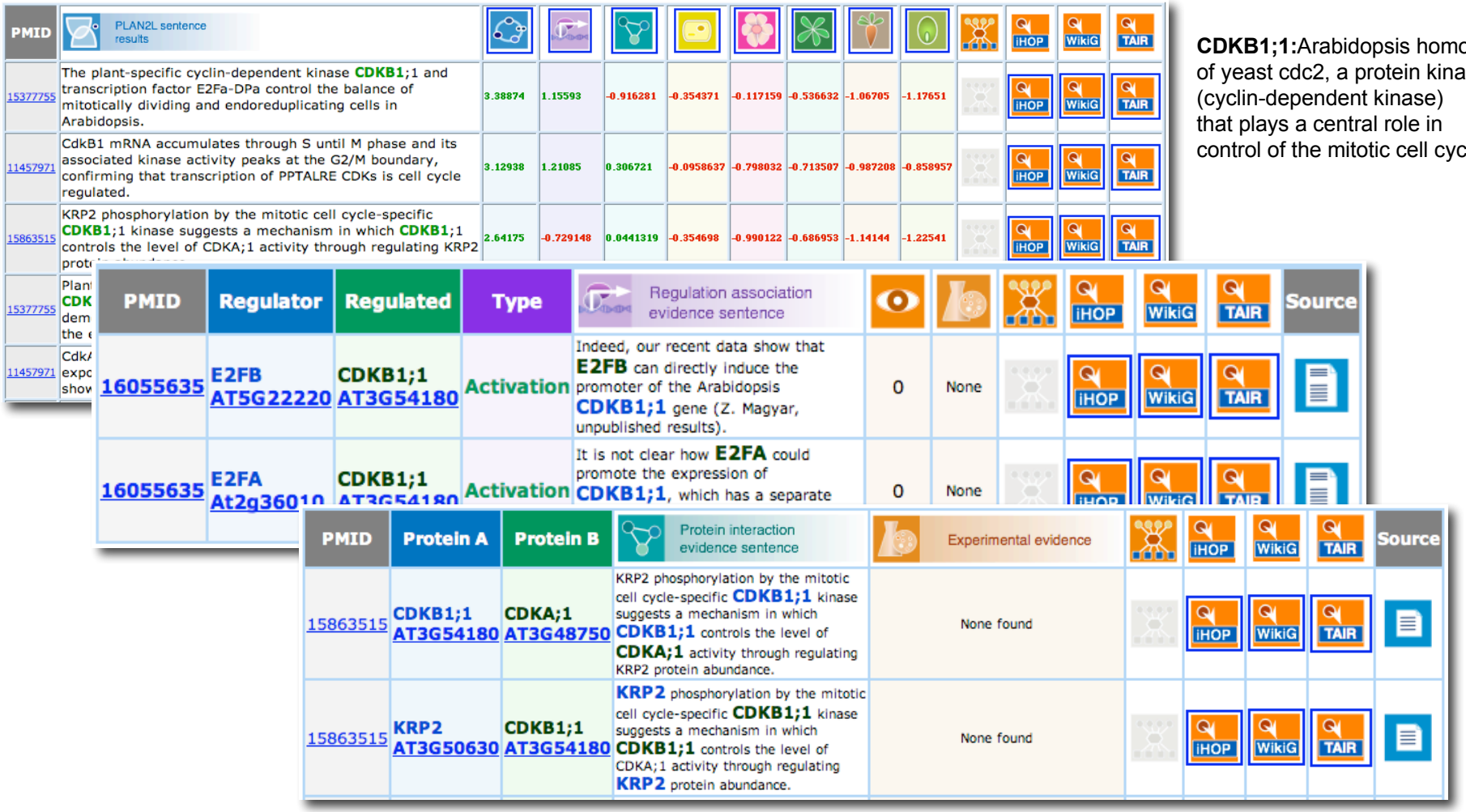
(5) A new longevity gene, Indy (for I'm not dead yet), which doubles the average …. [PMID:12391301]

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

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


Choose Classes
Virus
Tissue
RNA
Protein
Polynucleotide
Peptide
OtherOrganicCompound
OtherName
OtherArtificialSource
Organism
Nucleotide
MultiCell
MonoCell
Lipid
Inorganic
DNA
CellType
CellLine
CellComponent
Carbohydrate
BodyPart
Atom
AminoAcidMonomer
Select All
Deselect All
Biomedical Named Entity Recognizer
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 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 mutatio
• Based on Machine learning
• Good results in the COLING Bio-NER contest (Geneva)
• Many classes (entity types), including Virus, Tissue, RNA, Protein, Polynucleotide, Peptide, Organism, Nucleotide, Lipid, DNA, Cell Type, Cell Line, Cell Component, Carbohydrate, Body Part Atom and Amino Acid Monomer
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 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 mutatio

PLAN2L: a web tool for integrated text mining & literature-based bioentity relation extraction



PMID	PLAN2L sentence results									iHOP	WikiG	TAIR
15377755	The plant-specific cyclin-dependent kinase **CDKB1**;1 and transcription factor E2Fa-DPa control the balance of mitotically dividing and endoreduplicating cells in Arabidopsis.	3.38874	1.15593	-0.916281	-0.354371	-0.117159	-0.536632	-1.06705	-1.17651			
11457971	CdkB1 mRNA accumulates through S until M phase and its associated kinase activity peaks at the G2/M boundary, confirming that transcription of PPTALRE CDKs is cell cycle regulated.	3.12938	1.21085	0.306721	-0.0958637	-0.798032	-0.713507	-0.987208	-0.858957			
15863515	KRP2 phosphorylation by the mitotic cell cycle-specific **CDKB1**;1 kinase suggests a mechanism in which **CDKB1**;1 controls the level of CDKA;1 activity through regulating KRP2	2.64175	-0.729148	0.0441319	-0.354698	-0.990122	-0.686953	-1.14144	-1.22541			

PMID	Regulator	Regulated	Type	Regulation association evidence sentence			iHOP	WikiG	TAIR	Source
16055635	E2FB AT5G22220	CDKB1;1 AT3G54180	Activation	Indeed, our recent data show that **E2FB** can directly induce the promoter of the Arabidopsis **CDKB1;1** gene (Z. Magyar, unpublished results).	0	None				
16055635	E2FA At2g36010	CDKB1;1 AT3G54180	Activation	It is not clear how **E2FA** could promote the expression of **CDKB1;1**, which has a separate	0	None				

PMID	Protein A	Protein B	Protein interaction evidence sentence	Experimental evidence		iHOP	WikiG	TAIR	Source
15863515	CDKB1;1 AT3G54180	CDKA;1 AT3G48750	KRP2 phosphorylation by the mitotic cell cycle-specific **CDKB1;1** kinase suggests a mechanism in which **CDKB1;1** controls the level of **CDKA;1** activity through regulating KRP2 protein abundance.	None found					
15863515	KRP2 AT3G50630	CDKB1;1 AT3G54180	**KRP2** phosphorylation by the mitotic cell cycle-specific **CDKB1;1** kinase suggests a mechanism in which **CDKB1;1** controls the level of CDKA;1 activity through regulating **KRP2** protein abundance.	None found					

CDKB1;1: Arabidopsis homolog of yeast cdc2, a protein kinase (cyclin-dependent kinase) that plays a central role in control of the mitotic cell cycle.

http://zope.bioinfo.cnio.es/plan2l

Krallinger, M. et al . **PLAN2L: a web tool for integrated text mining and literature-based bioentity relation extraction**. To appear *in Nucl. Acids Res.*, Web Server Issue, 2009.

PLAN2L

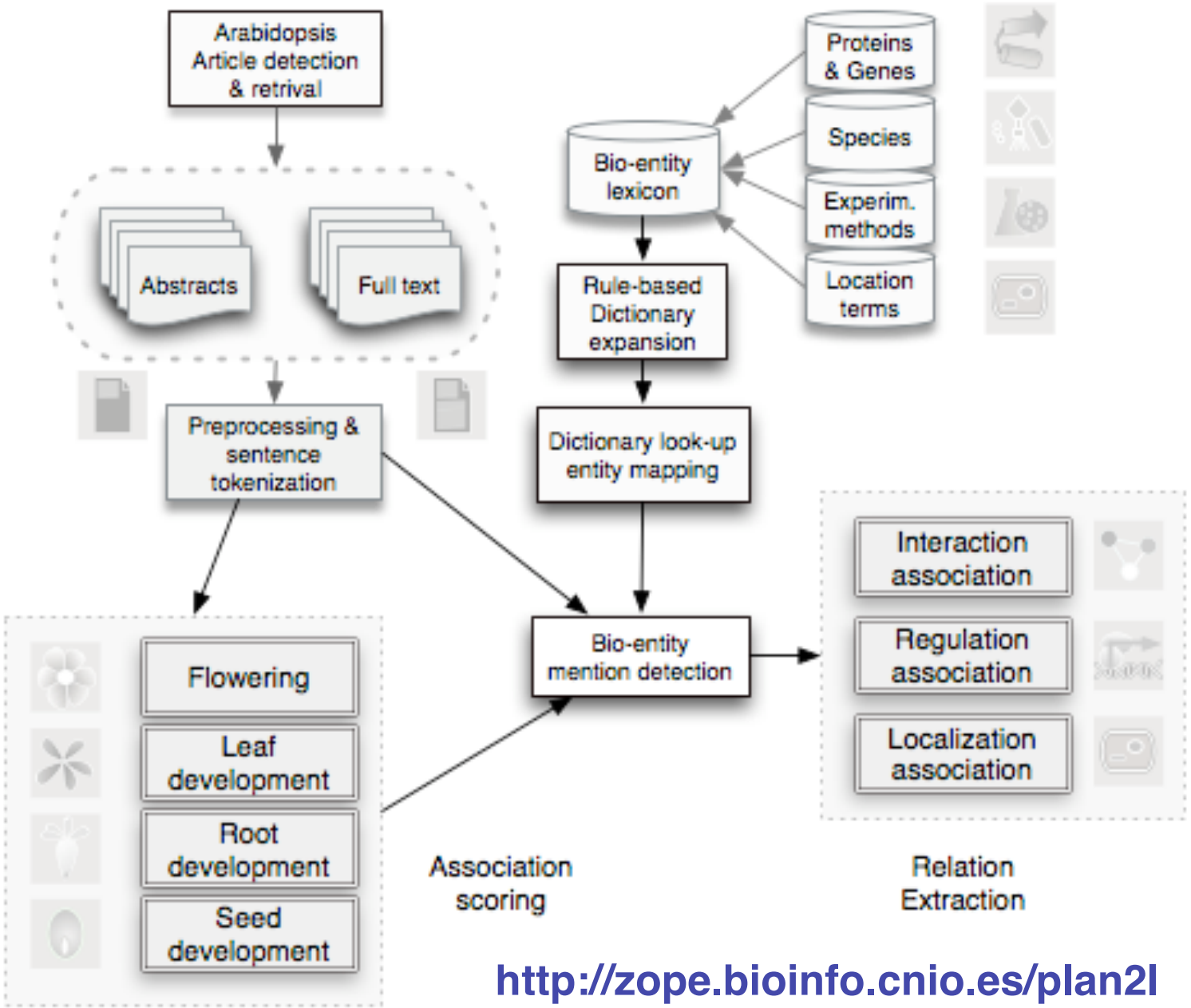
PMID	Regulator	Regulated	Type	Regulation association evidence sentence				iHOP	WikiG	TAIR	Source
16055635	E2FB AT5G22220	CDKB1;1 AT3G54180	Activation	Indeed, our recent data show that **E2FB** can directly induce the promoter of the Arabidopsis **CDKB1;1** gene (Z. Magyar, unpublished results).	0	None		iHOP	WikiG	TAIR	
16055635	E2FA At2g36010	CDKB1;1 AT3G54180	Activation	It is not clear how **E2FA** could promote the expression of **CDKB1;1**, which has a separate expression window in S- and M-phases (Magyar et al., 1997, 2000).	0	None		iHOP	WikiG	TAIR	

PMID	Proteins	Location description evidence sentence	Location terms	Location words		iHOP
11058164	LEUNIG	The **nuclear** localization of **LEUNIG**-GFP is consistent with a role of **LEUNIG** as a transcriptional regulator.	nuclear	localization		iHOP

PMID	PLAN2L Protein association evidence								
11041883	Additional analyses indicate that the absence of marginal tissues in leunig aintegumenta double mutants is not mediated by ectopic **AGAMOUS**.	-0.384139	-1.26458	-0.121035	-0.430636	2.57862	-0.0047048	-1.25884	-0.937231
16625397	SEUSS and **LEUNIG** encode components of a putative transcriptional regulatory complex that controls organ identity specification through the repression of the floral organ identity gene **AGAMOUS**.	-1.5747	1.60754	-0.190263	-0.104091	3.33405	-0.294006	-1.00271	-0.170982
11782418	The effects of seuss mutations are most striking when combined with mutations in **LEUNIG**, a previously identified repressor of **AGAMOUS**.	-0.885257	-0.754272	-0.239062	-1.4516	1.33617	-0.847629	-1.31295	-0.751337
11058164	**LEUNIG**, a putative transcriptional corepressor that regulates **AGAMOUS** expression during flower development.	-1.3353	3.11992	-0.338721	-0.641414	5.46814	-0.307351	-0.943231	-0.865053
16854969	Previously, we identified and isolated two Arabidopsis transcription co-repressors **LEUNIG** (LUG) and SEUSS (SEU) that function together in a putative co-repressor complex to prevent ectopic **AGAMOUS** (AG) transcription in flowers.	-0.726197	2.60201	-0.39487	0.303502	1.99957	-0.326139	-1.42551	-1.02715

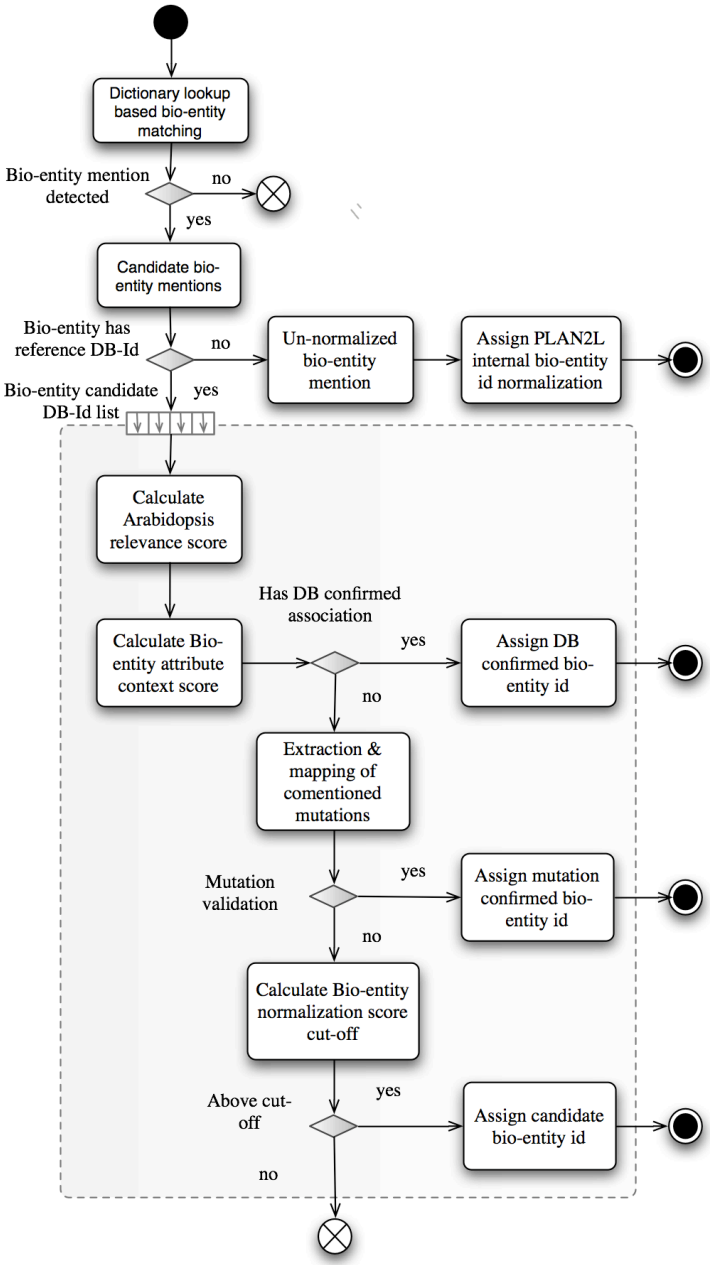
http://zope.bioinfo.cnio.es/plan2l

PLAN2L flowchart

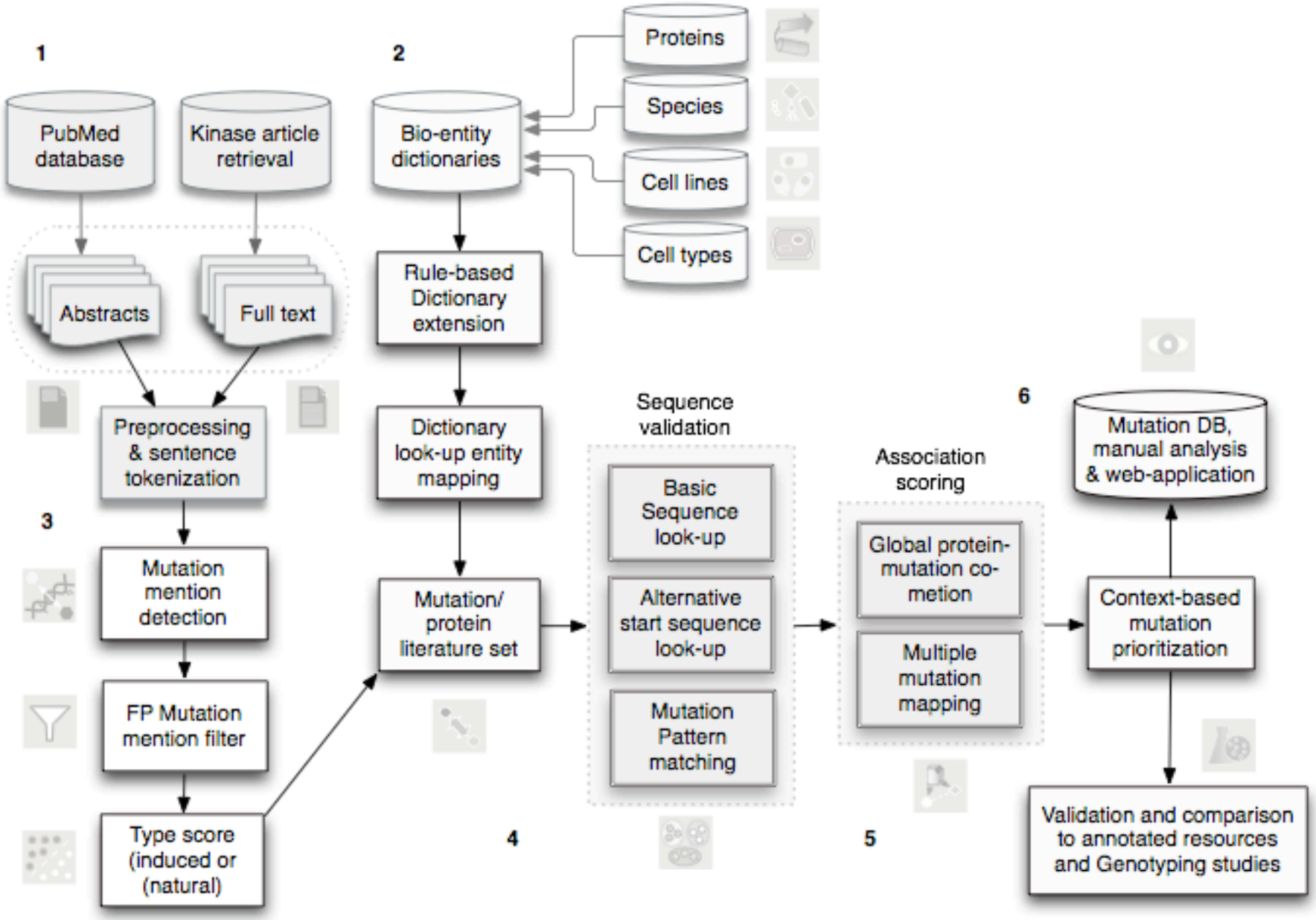


http://zope.bioinfo.cnio.es/plan2l

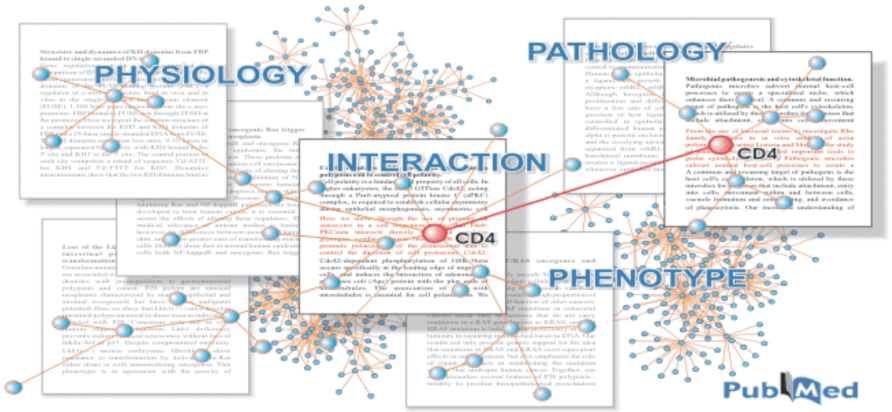
PLAN2L protein mention normalization

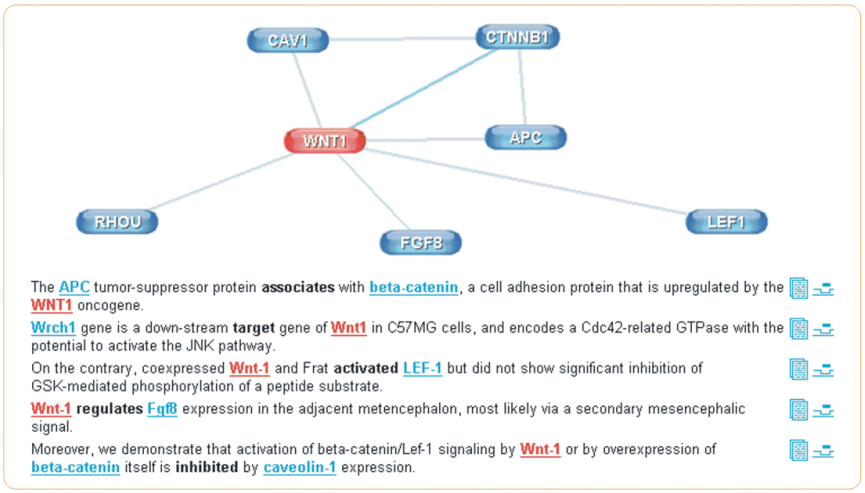


PLAN2L mutation extraction



iHOP system





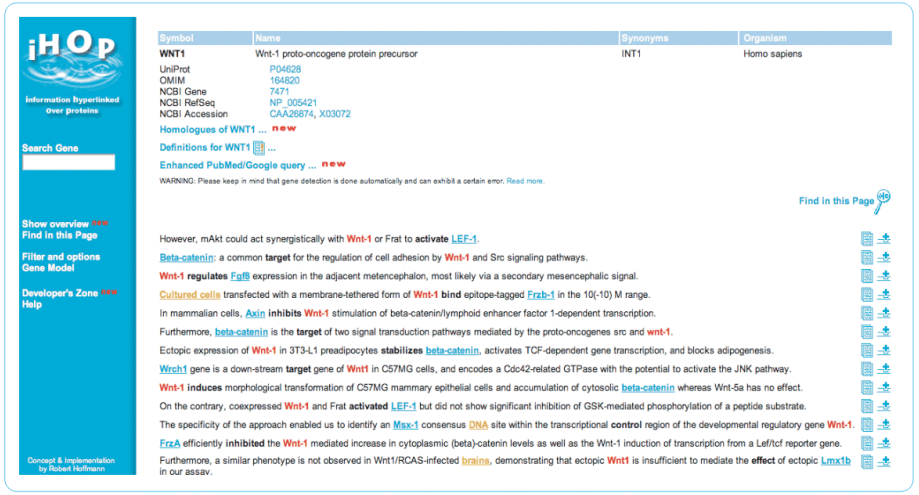
The APC tumor-suppressor protein **associates** with beta-catenin, a cell adhesion protein that is upregulated by the WNT1 oncogene.

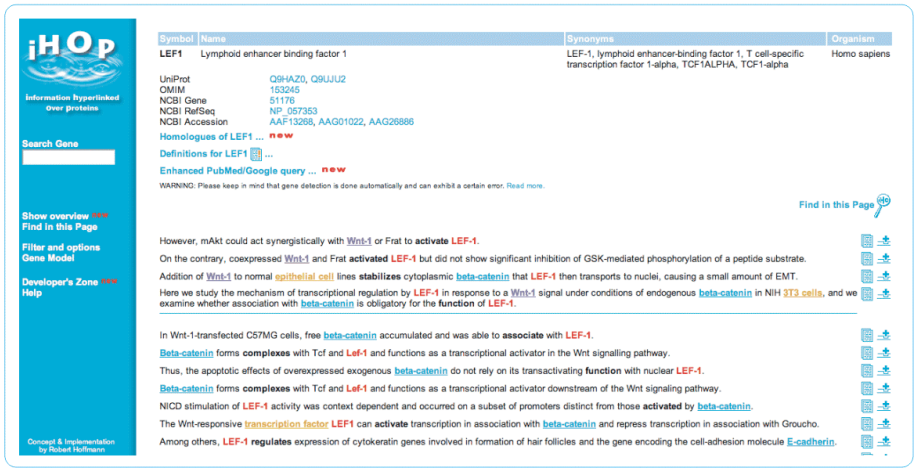
Wrch1 gene is a down-stream **target** gene of Wnt1 in C57MG cells, and encodes a Cdc42-related GTPase with the potential to activate the JNK pathway.

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

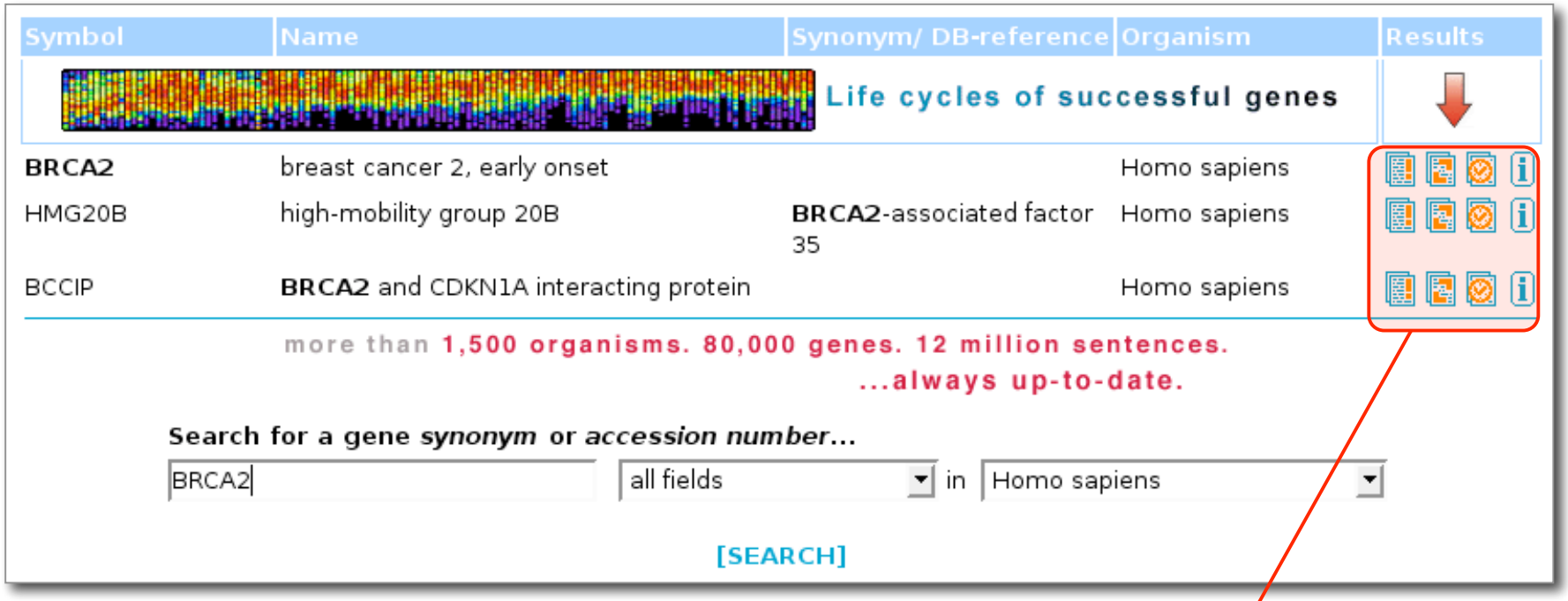
Wnt-1 **regulates** Fgf8 expression in the adjacent metencephalon, most likely via a secondary mesencephalic signal.

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.



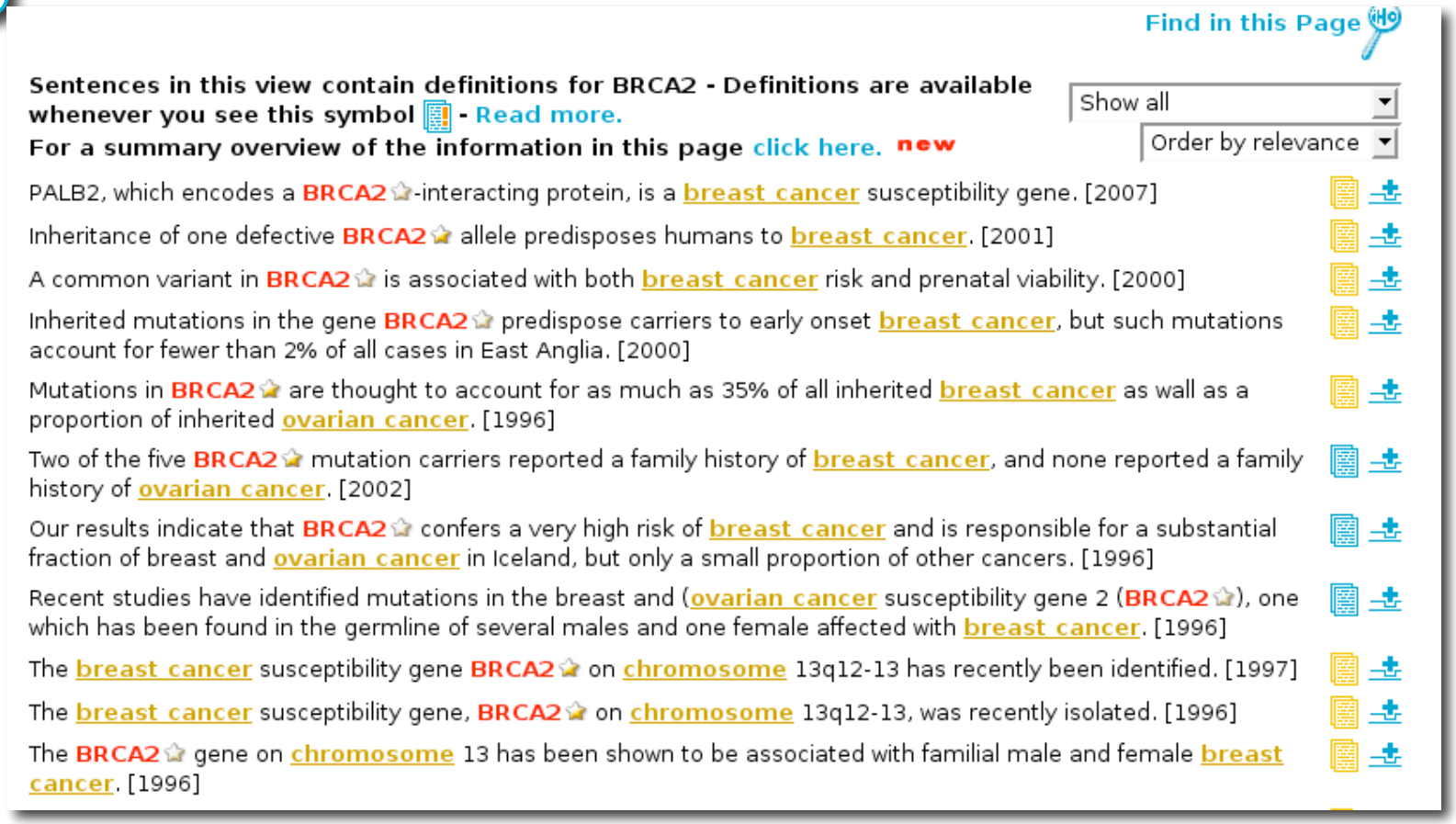


iHOP system: query to DB record



Results options

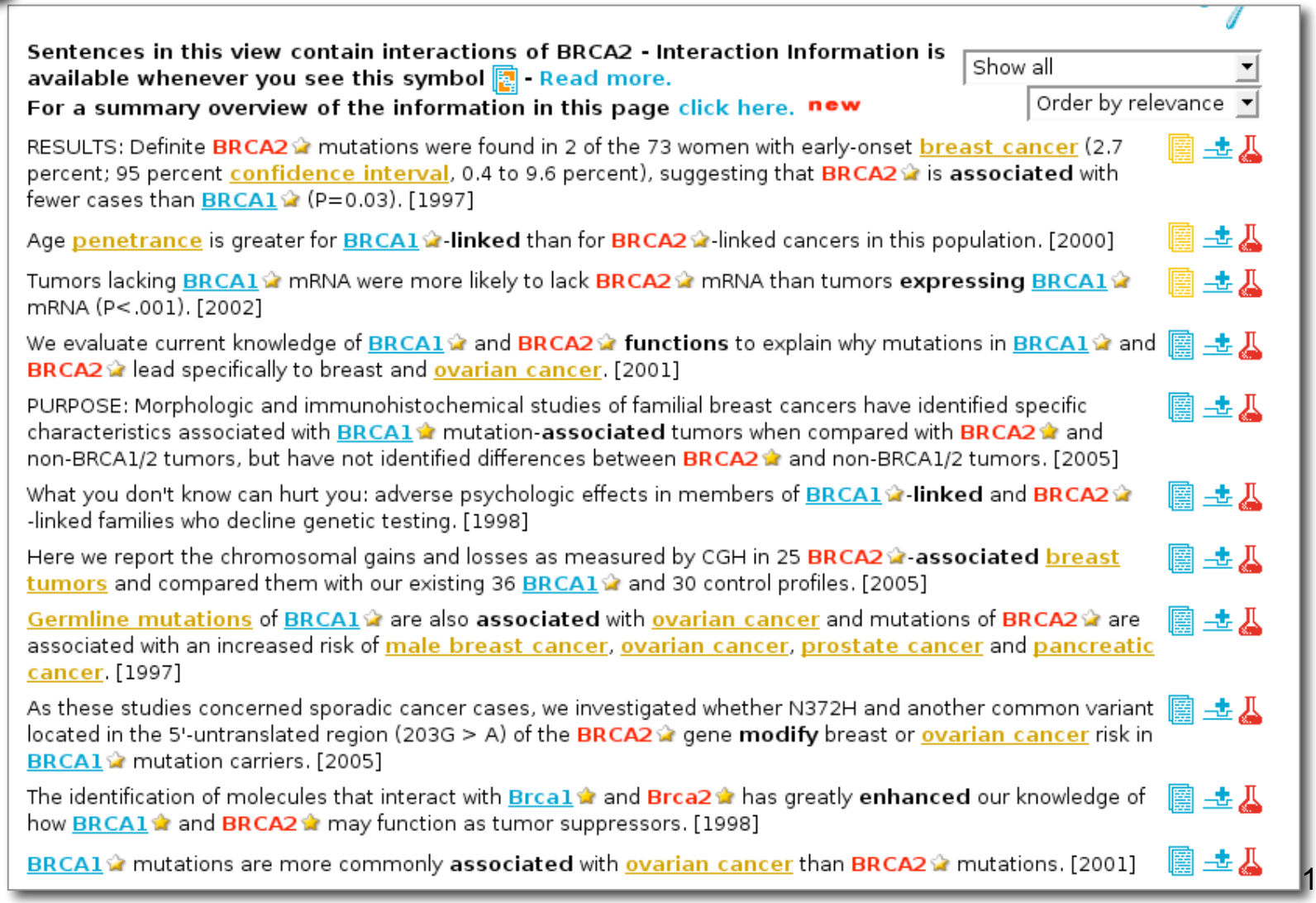
iHOP system: Defining information



Colour legend
- Main gene
- Associated genes
- Relevant Biomedical terms
- Compounds

Defining Information for this Gene

iHOP system: 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

Show all
Order by relevance

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]

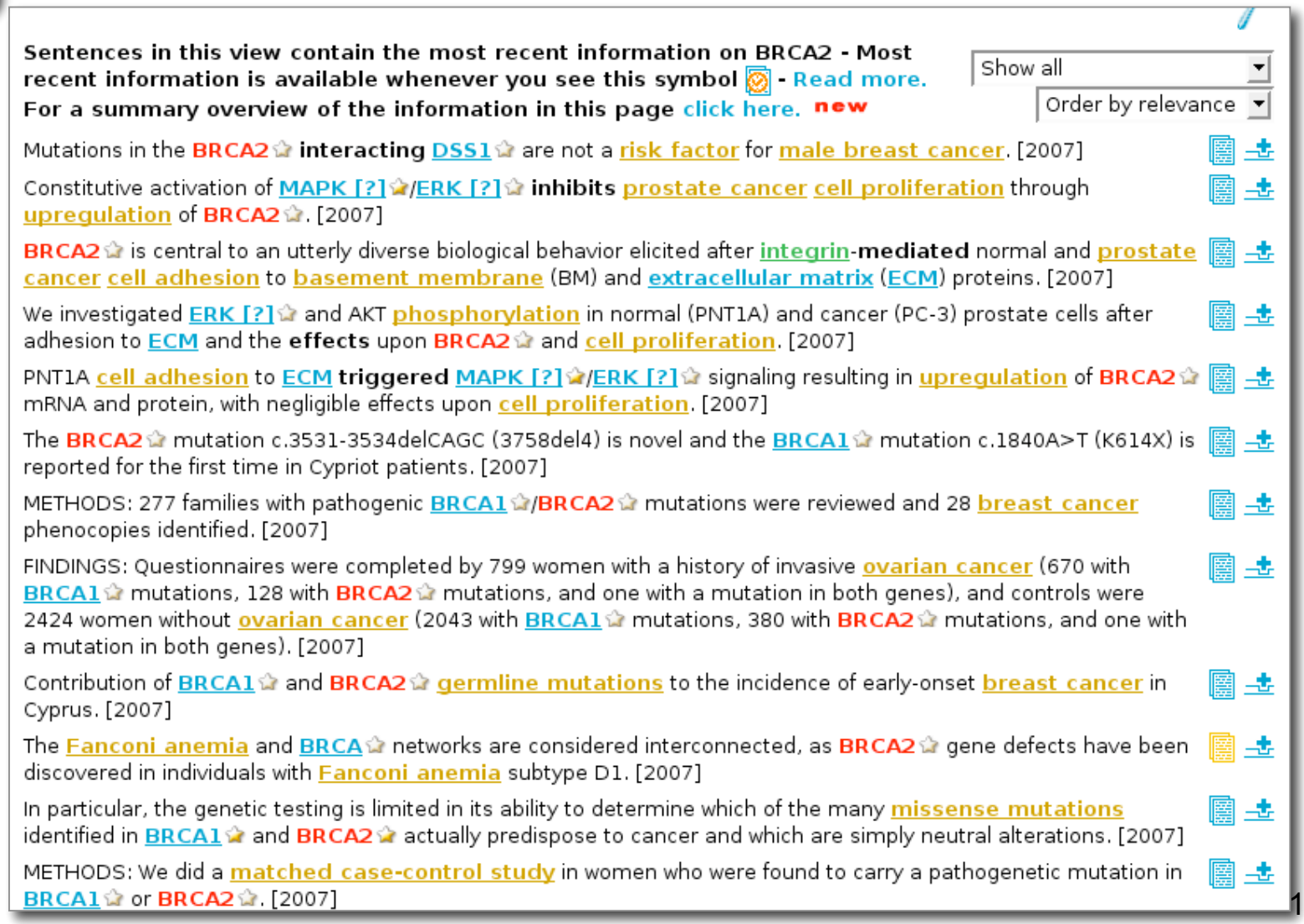
Germline mutations of BRCA1 are also **associated** with ovarian cancer and mutations of BRCA2 are associated with an increased risk of male breast cancer, ovarian cancer, prostate cancer and pancreatic cancer. [1997]

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]

iHOP system: 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

Show all

Order by relevance

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]

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

Homologues of BRCA2 ...

Interaction information for BRCA2 ...

Most recent information for BRCA2 ... **new**

Enhanced PubMed/Google query ...

WARNING: Please keep in mind that gene detection is done automatically and can exhil
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.

Gene model is a interactive graph where you can add interesting sentences and interactions.

iHOP system: confidence

The synomnym ambiguity limitation

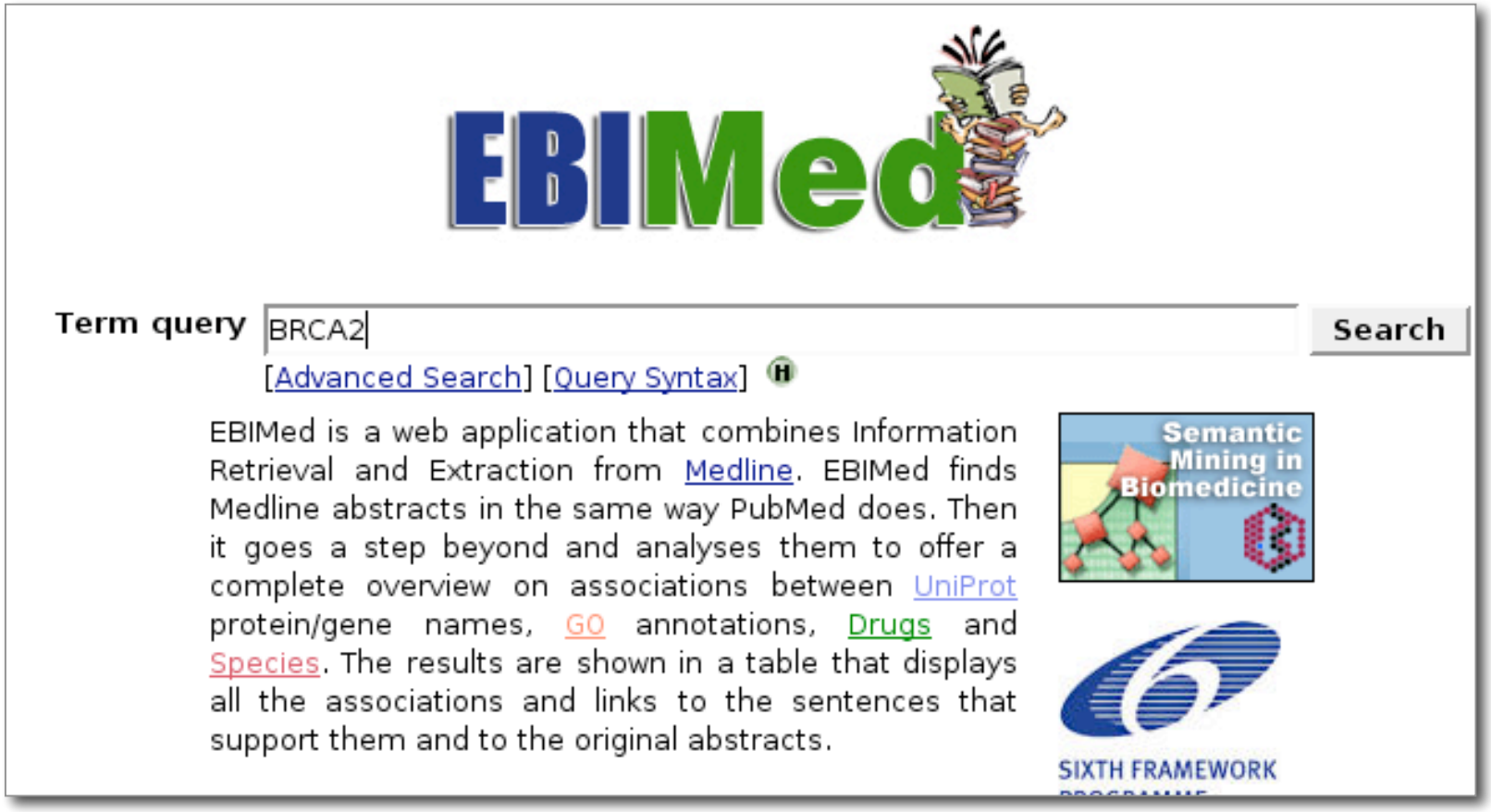
Many gene or protein synonyms are ambiguous, thus one and the same synonym is often used for different genes.
Even human experts can have difficulties to resolve such ambiguities and automatic systems, like iHOP,
will therefore always exhibit certain errors.

The iHOP confidence value

Although no definite solution for the problem of synonym ambiguity is in sight, it is possible to put an automatically derived confidence value to specific gene references.
This iHOP confidence value is illustrated through the colour intensity of a star

The absence of a star does not mean that a certain term could not be a gene,
but simply that supporting evidence is not available.

EBIMed


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

- *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).*

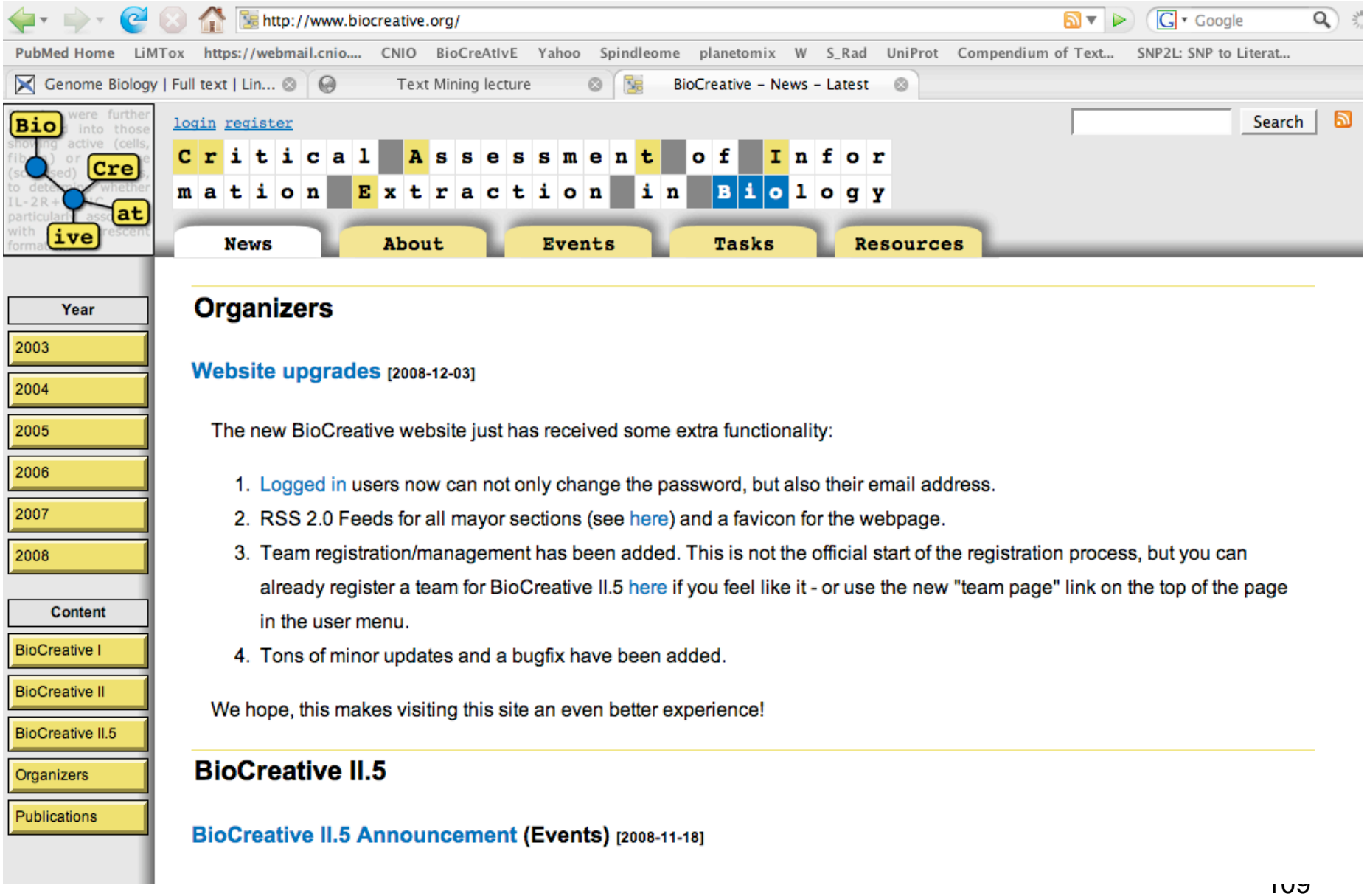
first << 1/40 >> last

Protein/Gene	Protein/Gene	Cellular component	Biological process	Molecular function	Drug	Species
BRCA2 *or* FANCD1 ***(score: 1603)***	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 (3/6)	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)	

GOPubMed

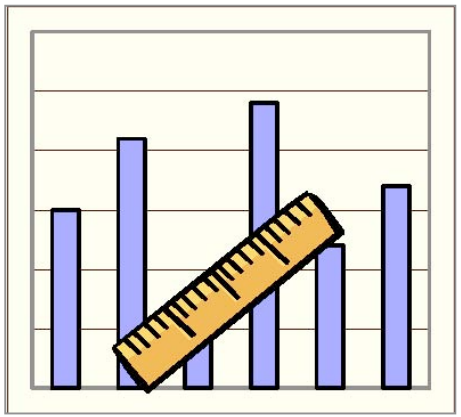

what
Top 5 categories
Mutation [766]
Breast Neoplasms [7
Humans [588]
Genes [751]
Genes, BRCA2 [487]
Top categories of GO
Anatomy [366]
Biological Sciences [9
biological_process [4
cellular_component [1
Chemicals and Drugs
Diseases [941]
Health Care [796]
molecular_function [1
Named Groups [627]
Natural Sciences [791
Organisms [641]
Psychiatry and Psych
Techniques and Equip
Technology, Industry,
Found categories of G
Find related categorie
My last 5 queries
BRCA2
who
where
BRCA2
find it!
gopubme
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

BioCreative

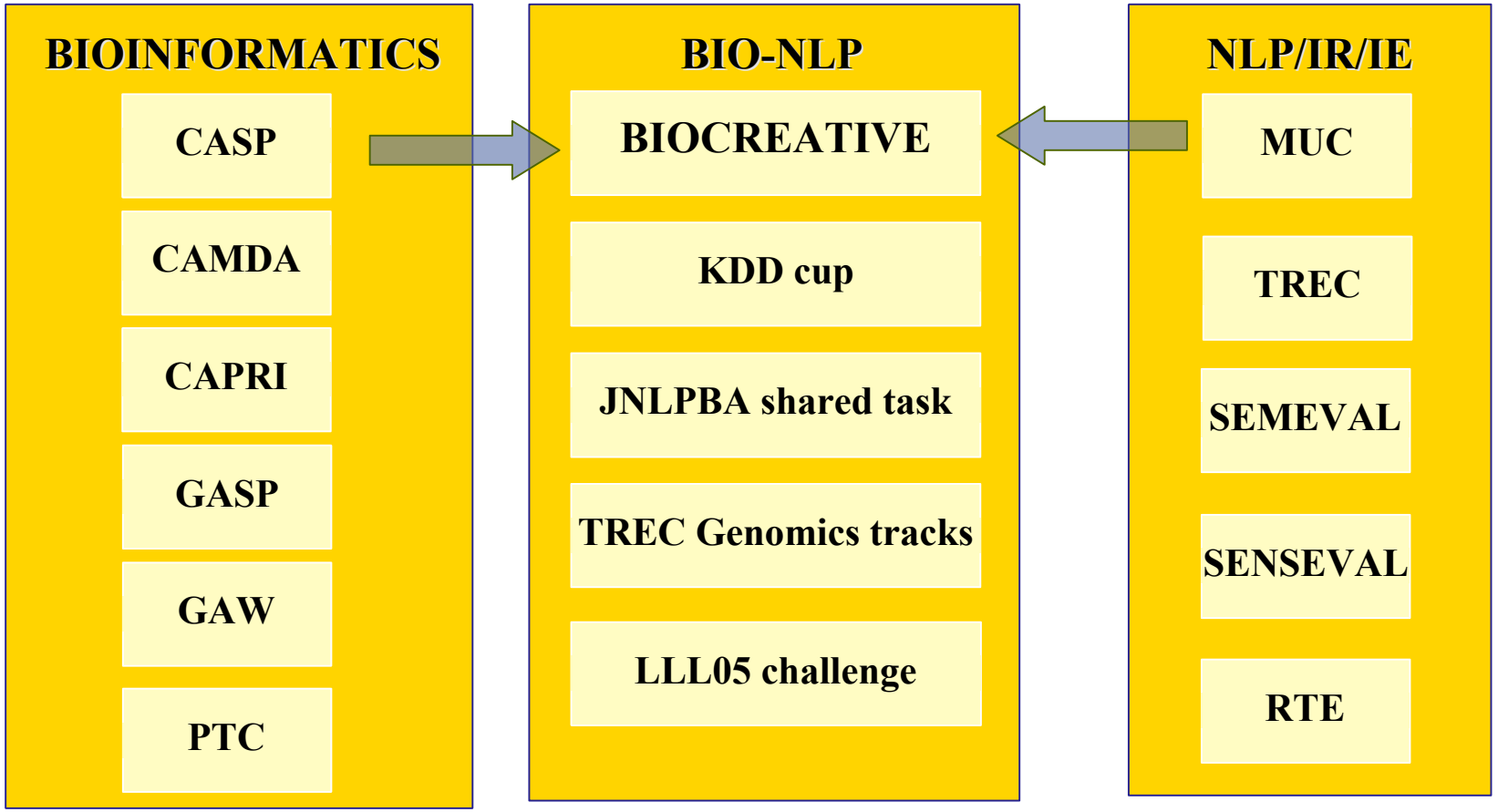


Why community assessments?

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

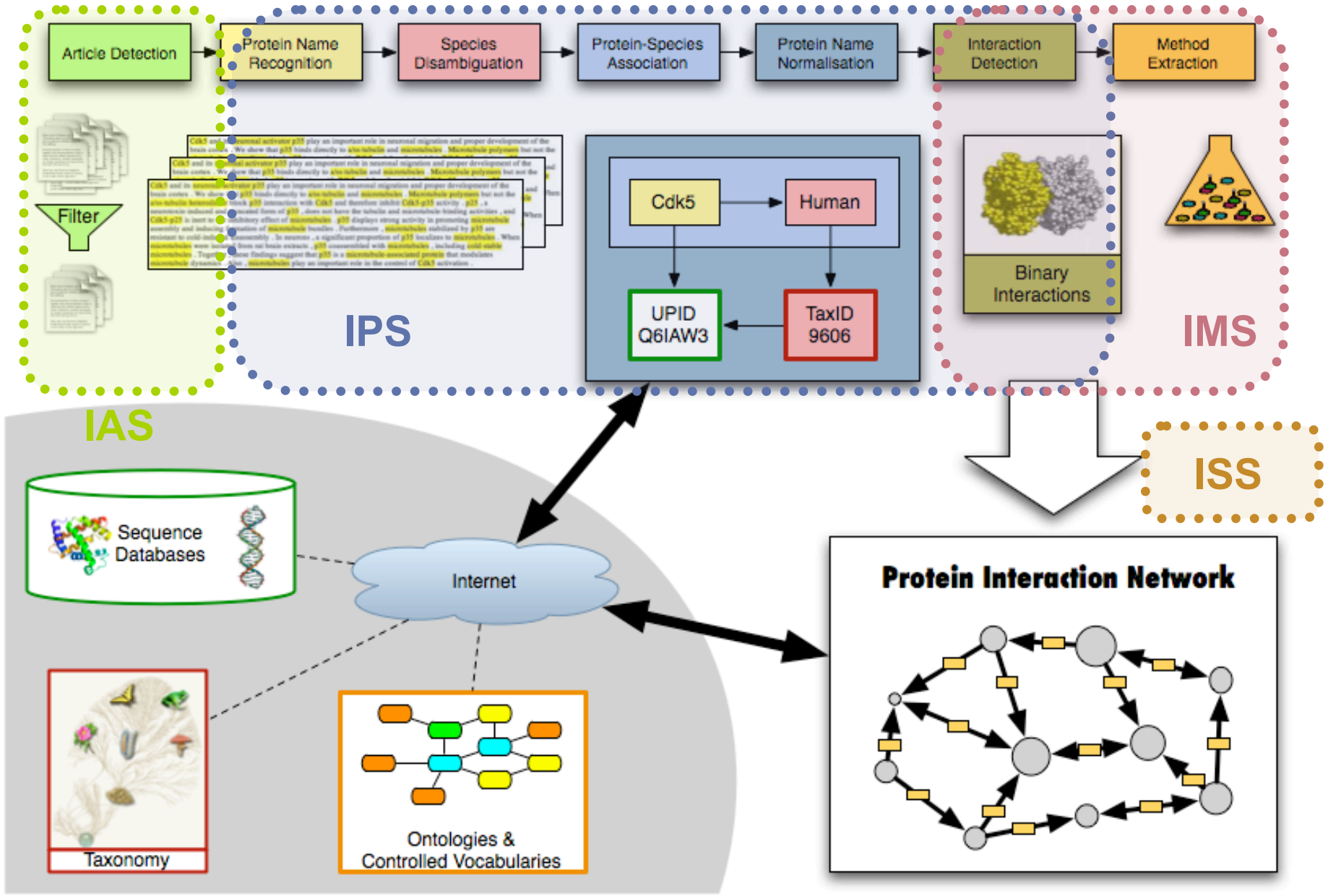


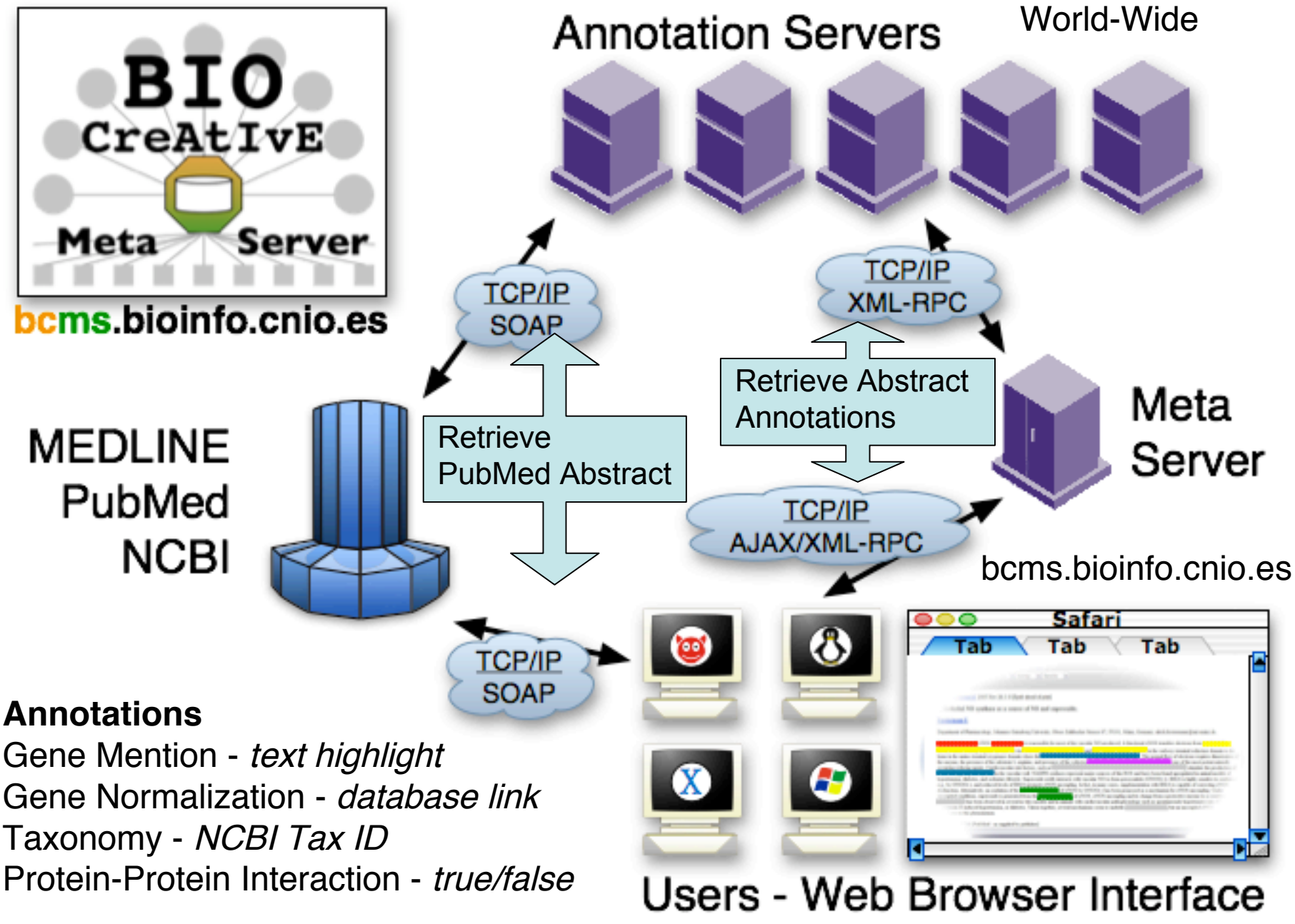
Community assessments



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

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


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



Annotations

Gene Mention - *text highlight*
Gene Normalization - *database link*
Taxonomy - *NCBI Tax ID*
Protein-Protein Interaction - *true/false*

Participants - Annotation Servers

- Alias I, New York, Bob Carpenter
- Georgetown University, Hongfang Liu
- Humboldt Univ., Berlin, Jörg Hakenberg
- Inst. of Biomed. Inf., Taiwan, Cheng-Ju Kuo
- Inst. of Inform. Sci., Taiwan, Richard Tsai
- Jena Univ., Germany, Kathrin Tomanek
- Milwaukee Marquette Univ., Craig Struble
- National Inst. of Health, William Lau
- Norweg. Univ. of Sci. and Tech., Janny Chen
- Seoul National University, Sun Kim
- Univ. of Colorado, William Baumgartner
- University of Edinburgh, Barry Haddow
- University of Geneva, Patrick Ruch
- University of Michigan, Arzucan Ozgur
- Univ. of Pennsylvania, Kuzman Ganchev
- Yale University, ThaiBinh Luong

Main advantages of BCMS

- **Data Integration:** multi-site annotations
- **Simplicity of usage:** single API with many annotations
- **User-oriented:** TM & biologist
- **Novel/ unique:** first system in biomedical text mining
- **Scalability:** additional systems
- **Extensibility:** additional annotation types
- **Flexibility:** additional input text types, e.g. full-text articles

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	0.886
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

Acknowledgements

Prof. Alfonso Valencia & Structural Computational Biology group at CNIO.