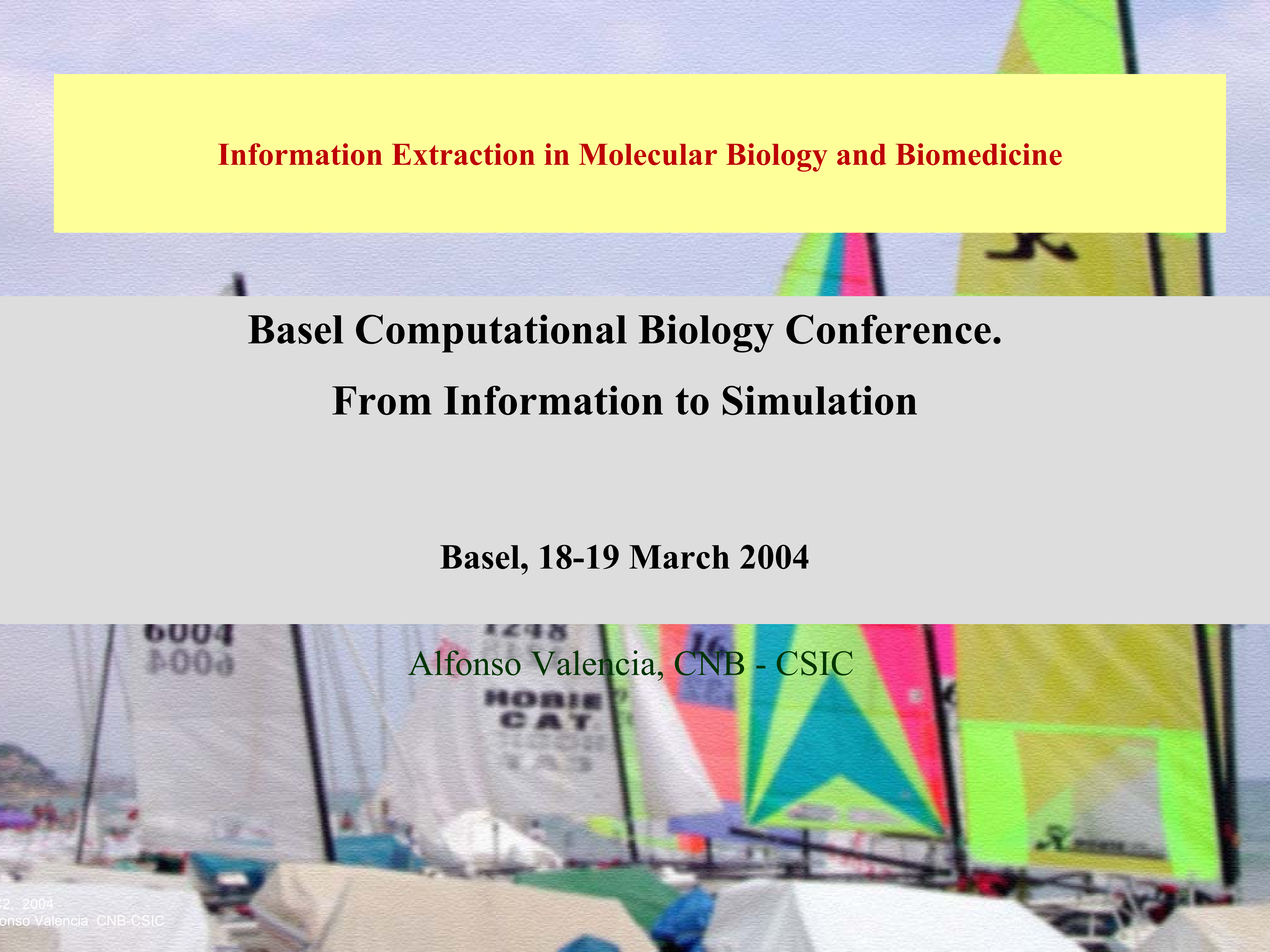


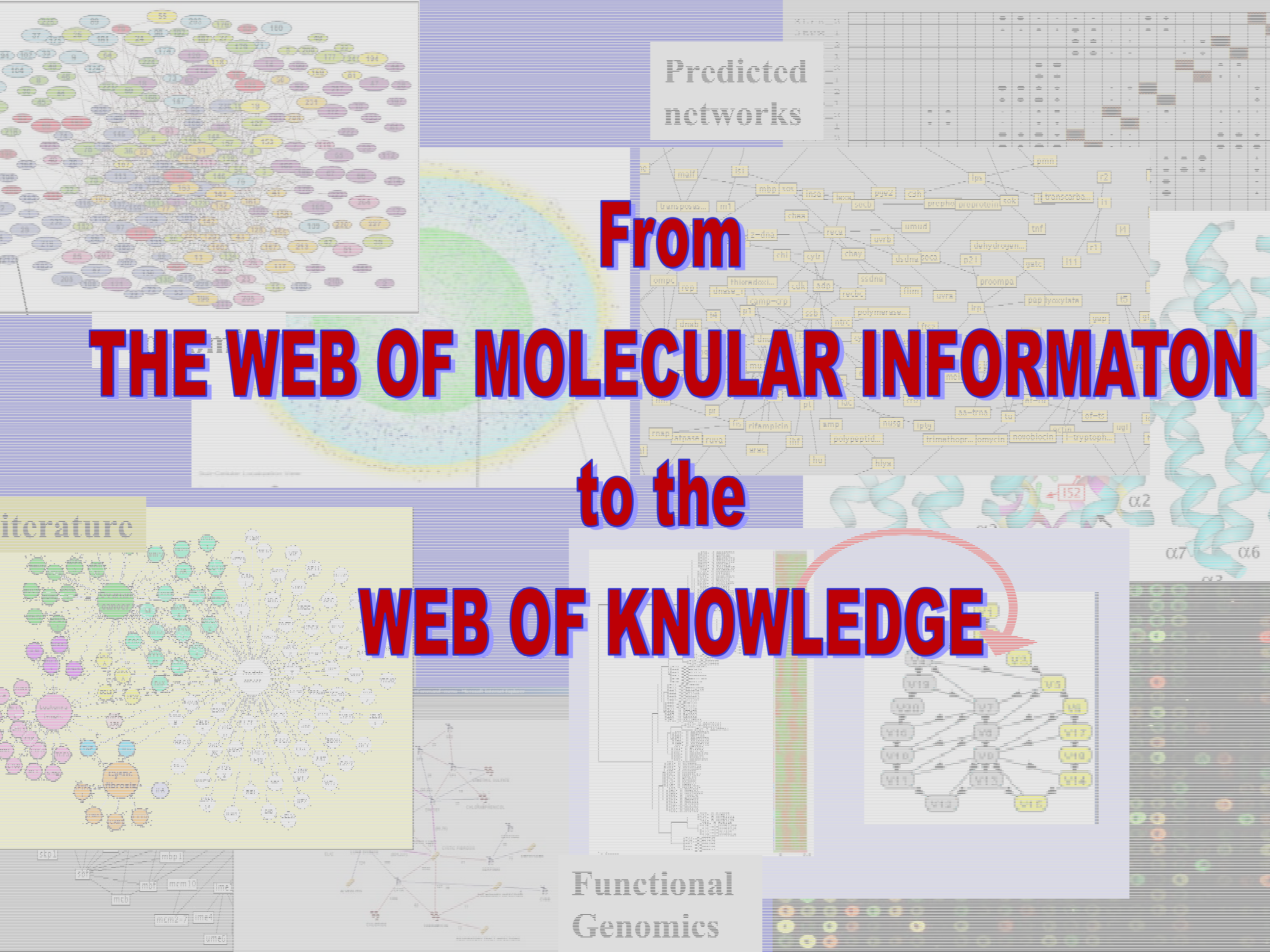
Information Extraction in Molecular Biology and Biomedicine

Basel Computational Biology Conference.

From Information to Simulation

Basel, 18-19 March 2004

Alfonso Valencia, CNB - CSIC



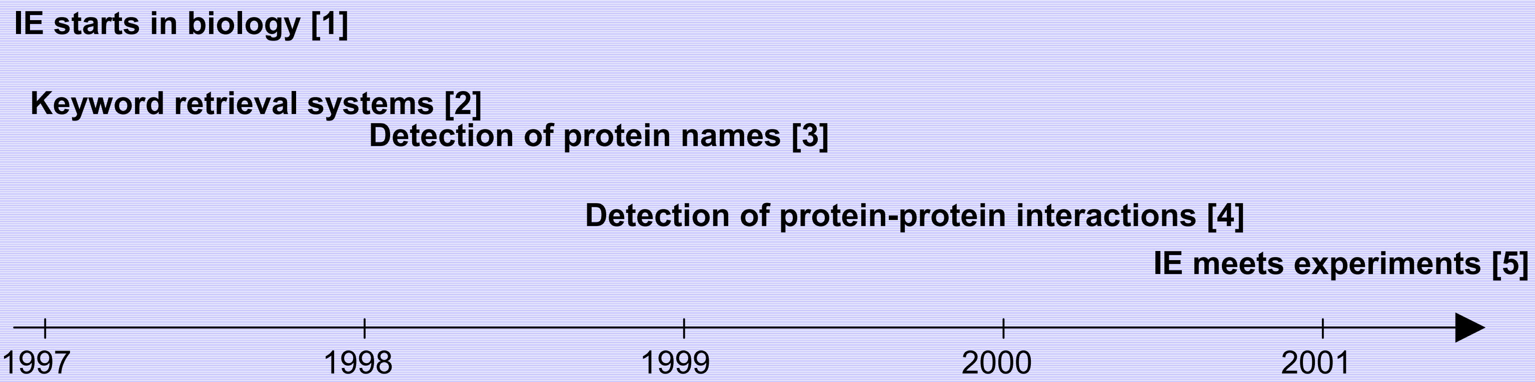
From

THE WEB OF MOLECULAR INFORMATON

to the

WEB OF KNOWLEDGE

Information Extraction in Molecular Biology



IE starts in biology [1]

Keyword retrieval systems [2]

Detection of protein names [3]

Detection of protein-protein interactions [4]

IE meets experiments [5]

Ohta et al. (1997). "Automatic construction of Knowledge Bases form Biological Papers"

Andrade and Valencia (1997). "Automatic annotation for biological sequences by extraction of keywords from MEDLINE abstracts"

Fukuda et al. (1998). "Information Extraction: Identifying Protein Names from Biological Papers"

Proux et al. (1998). "Detecting Gene Symbols and Names in Biological Texts: a first step ..."

Blaschke et al. (1999). "Automatic Extraction of Biological Information ...: Protein-Protein Interactions"

Park et al. (2001). "Incremental Parsing for Automatic Pathway Identification with Combinatorial Categorical Grammar"

Proux et al. (2000). "... Information Extraction Strategy for gathering Data on Genetic Interactions"

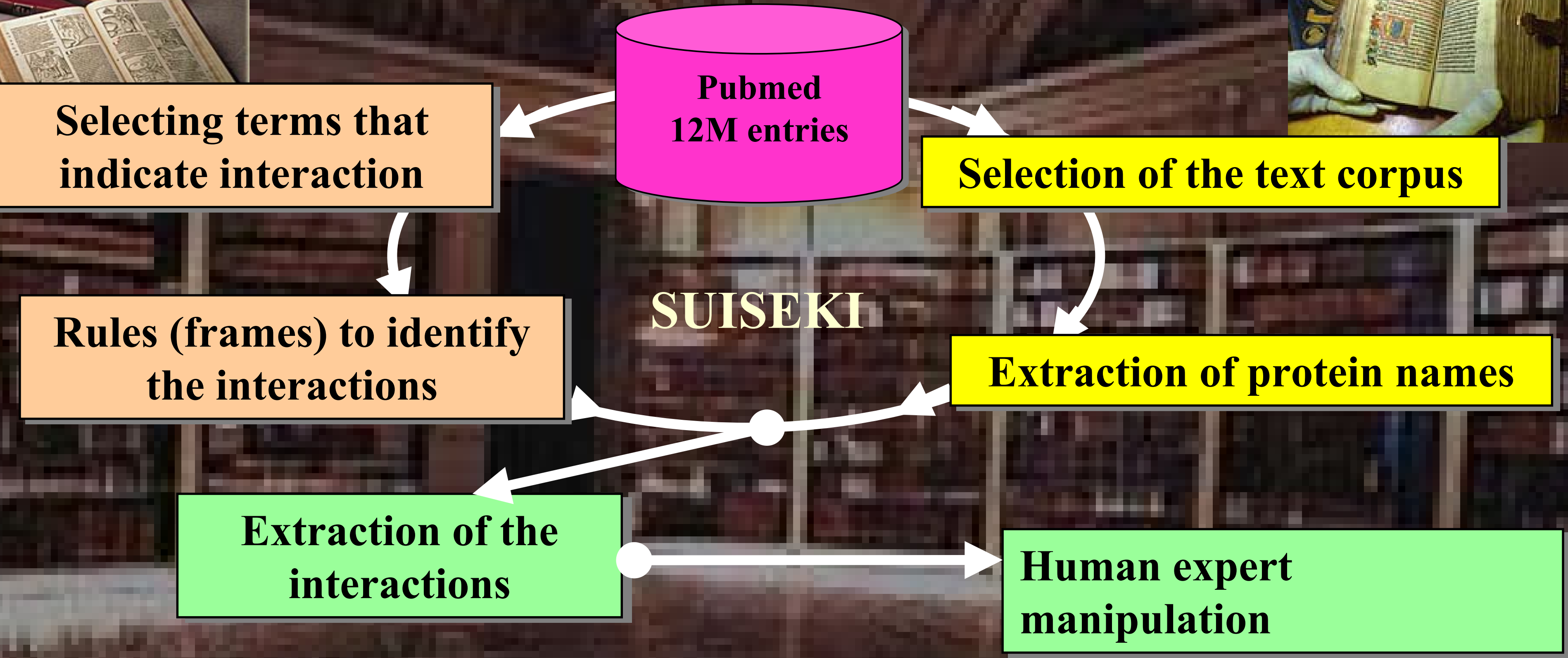
Rindflesch et al. (2000). "EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature"

Sekimizu et al. (1998). "Identifying the Interaction between Genes and Gene Products based on frequently seen Verbs in Medline stracts"

Thomas et al. (2000). "Automatic Extraction of Protein Interactions from Scientific Abstracts"

Blaschke et al. (2001). "Mining functional information associated with expression arrays"

Jenssen et al. (2001). "A literature network of human genes for high-throughput analysis of gene expression"



Pubmed
12M entries
Selecting terms that indicate interaction
Selection of the text corpus
SUISEKI
Rules (frames) to identify the interactions
Extraction of protein names
Extraction of the interactions
Human expert manipulation

Action words are for example:

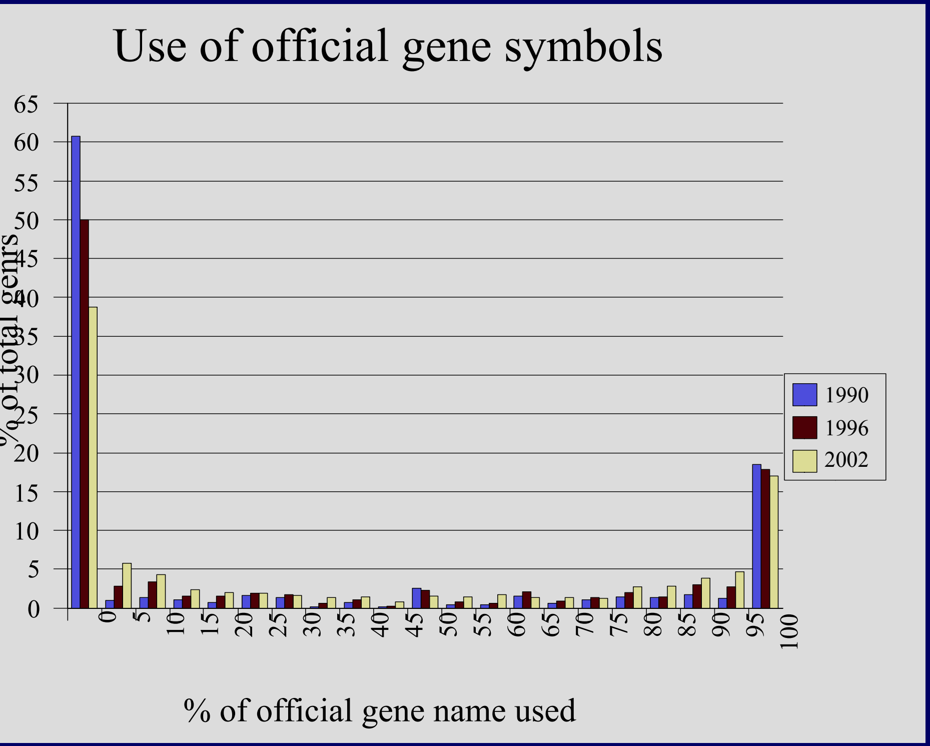
activate, associated with, bind, interact, phosphorylate, regulate

* [protein A] ... verb indicating an action ... [protein B]

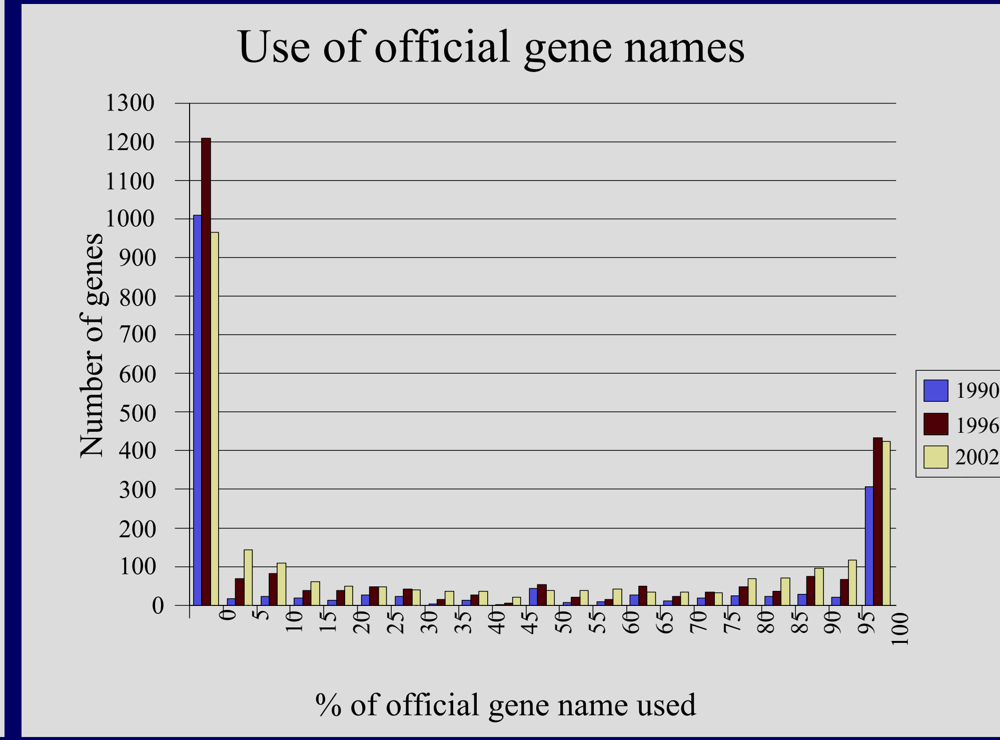
"After extensive purification, Cdk2 was still bound to cyclin D1"

Basic problem: identify Bio-entities in text

• Genes and proteins	cdk2, interleukin-8
• Chemicals, metabolites	acrilamide, fructose 6-phosphate
• Drugs	aspirin, prozac
• Diseases	Diabetes mellitus, Angelman syndrome
• Pathways, processes	Pentose phosphate pathway, DNA replication
• Species, tissues	*Saccharomyces cerevisiae*, vertebrates, brain
• Cell types, cell lines, mutations	macrophages, cd4+, liver, 95arg->trp
• Experimental techniques	2D electrophoresis, NMR



Use of official gene symbols



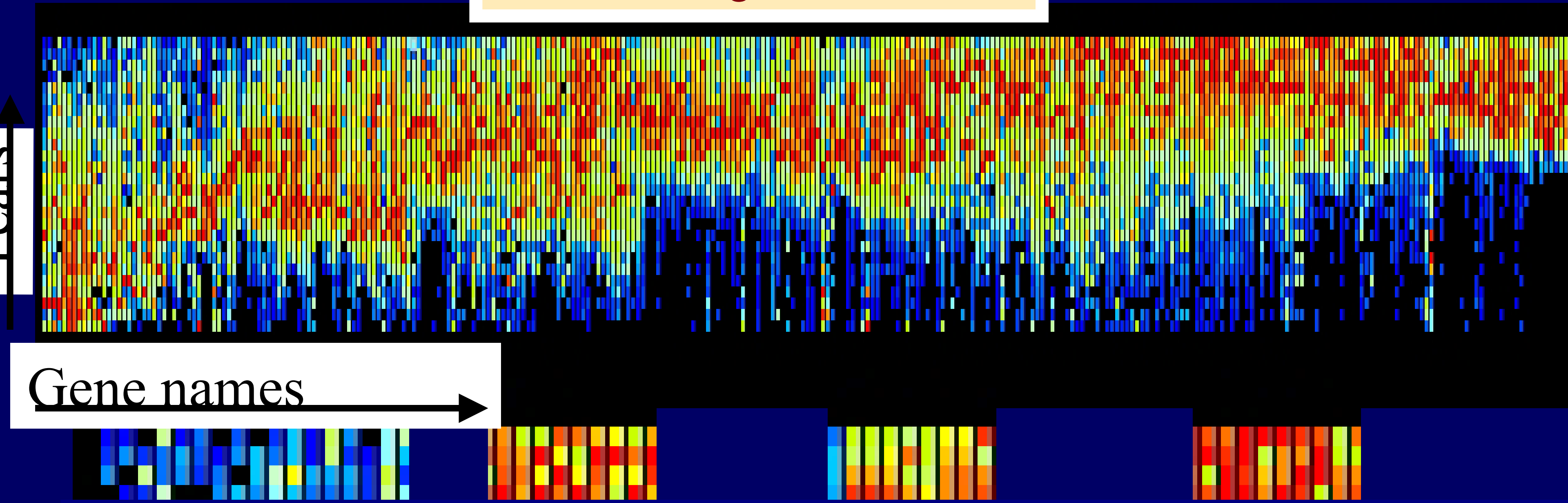
Use of official gene names

OFFICIAL	**62542**	**44.46 %**
ALIAS	**51749**	**36.79 %**
PROTEIN	**26363**	**18.74 %**

The 2492 selected genes in the year 2002 were cited 140654 time

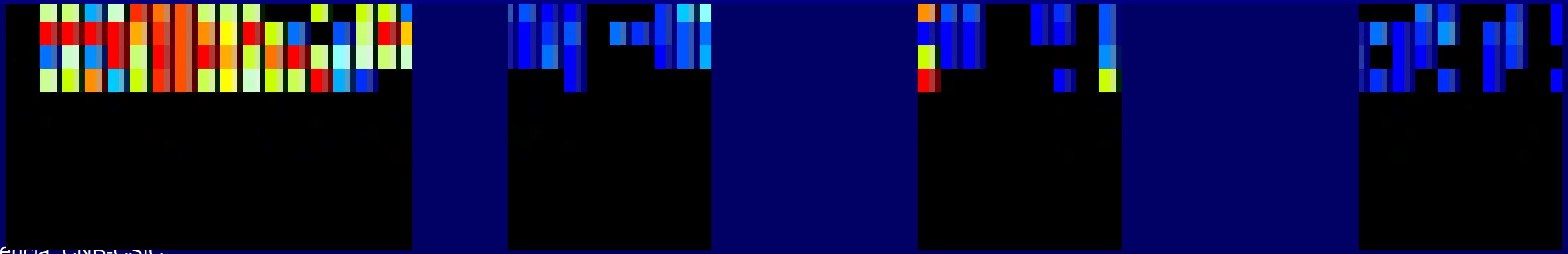
Tamames et al., 2

Evolution of gene names



The evolution of gene names over time is a “scale free” proces

- **“critical state” system**
- **the evolution of a gene name cannot be predicted**
- **some gene name act as attractors of other names**



Hoffmann, Valencia TIG

Example of annotation of a PubMed article

[Lipid-poor apolipoproteins]$_{\text{prot}}$ remove cellular [cholesterol]$_{\text{chem}}$ and [phospholipids]$_{\text{chem}}$ by an [active transport pathway]$_{\text{proc}}$ controlled by an [ATP binding cassette transporter]$_{\text{prot}}$ called [ABCA1]$_{\text{prot}}$. Mutations in [ABCA1]$_{\text{prot}}$ cause [Tangier disease]$_{\text{dis}}$, a [severe HDL deficiency syndrome]$_{\text{dis}}$ characterized by a rapid turnover of plasma [apolipoprotein A-I]$_{\text{prot}}$, accumulation of [sterol]$_{\text{chem}}$ in tissue [macrophages]$_{\text{cell}}$, and prevalent [atherosclerosis]$_{\text{dis}}$

- 🟦 Genes/proteins
- 🟨 Chemicals
- 🟩 Diseases
- 🟥 Pathways/processes
- 🟪 Cell types

Descriptions for RHO gene

Visual pigments are the light-absorbing molecules that mediate vision. They consist of an apoprotein, opsin, covalently linked to cis-retinal.

Defects in RHO are one of the causes of autosomal dominant retinitis pigmentosa.

Tissue specificity: Rod shaped photoreceptor cells which mediates vision in a dim light.

Keywords

cis-retinal

light-absorbing

opsin

photoreceptor

pigment

pigmentosa

retinitis

rod

Tamames et al., 2

Second step: Label the words in the article according to the definitions

Sentence: **The subunit alpha of DNA polymerase is a key component of the replication machine**

Definitions:

295	POLA	DNA polymerase, alpha catalitic subunit
297	POLB	DNA polymerase, subunit beta
298	POLD	DNA polymerase, delta subunit

Labelling

The subunit alpha of DNA polymerase is a key

295	**295**	**295**	**295**
297		**297**	**297**
298		**298**	**298**

component of the replication machinery.

Tamames et al., 2

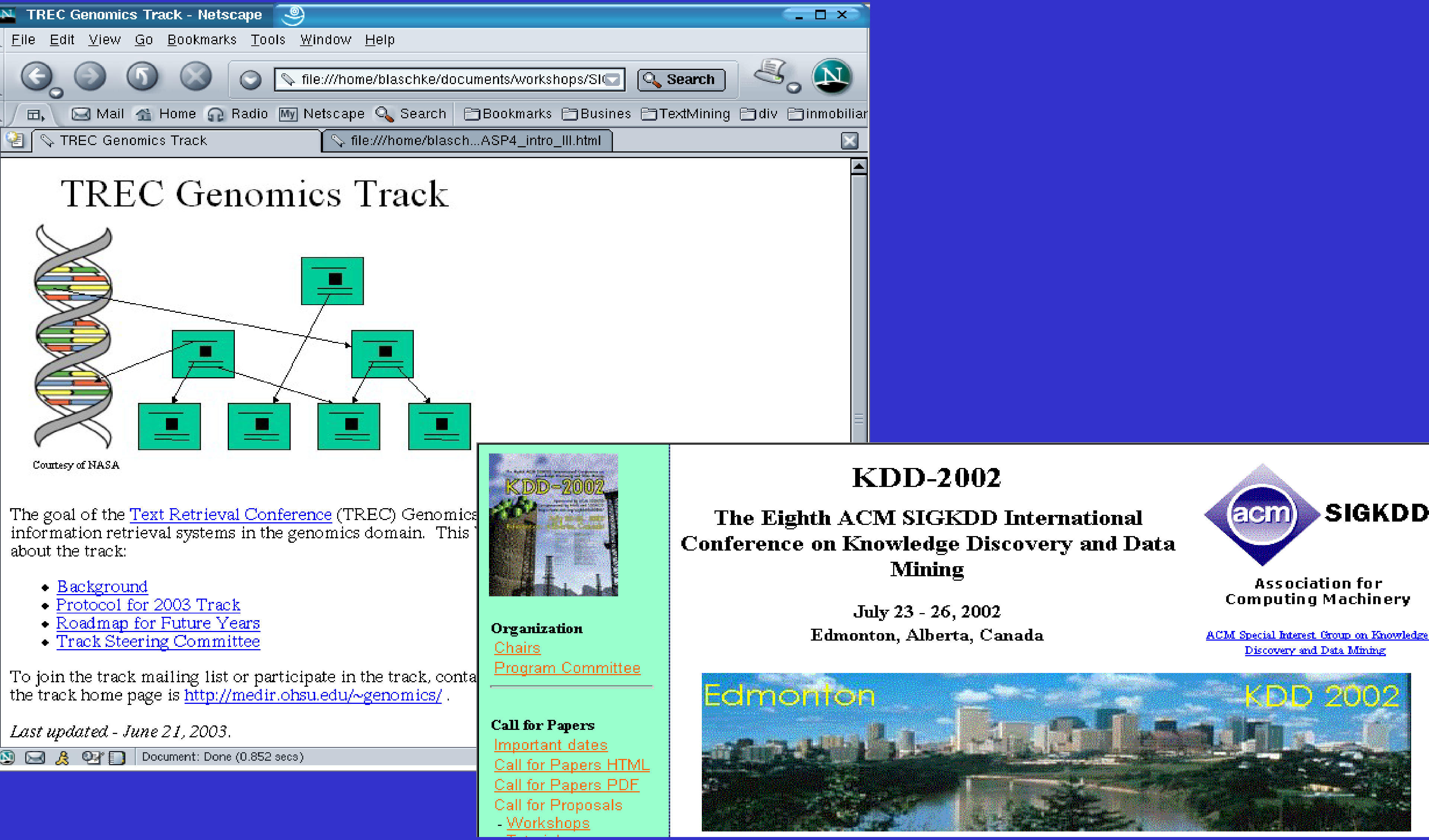
RESULTS of TEXT DETECTIVE

	Possible genes	True pos	Precision (PubMed)	Selected/ Correct	Recall	Precision
Curated set of articles	2173	612	28%	648 / 575	93%	88%

Selected difficult cases

Symbol	Total in PubMed	True pos	Precision (PubMed)	Selected/ Correct	Recall	Precision
HK1 (Hexokinase 1)	101	36	35.6%	42 / 32	89%	76%
LHB (Luteinizing hormone beta)	113	11	9.7%	7 / 6	55%	86%
RSN (Restin)	41	3	7%	1 / 1	33%	100%
SCT (Secretin) (Year 2001 only)	158	1	0.6%	1 / 1	100%	100%

TREC – Genomics Track and KDD



TREC Genomics Track

The goal of the Text Retrieval Conference (TREC) Genomics
information retrieval systems in the genomics domain. This
about the track:

- Background
- Protocol for 2003 Track
- Roadmap for Future Years
- Track Steering Committee

To join the track mailing list or participate in the track, conta
the track home page is http://medir.ohsu.edu/~genomics/ .

Last updated - June 21, 2003.

KDD-2002

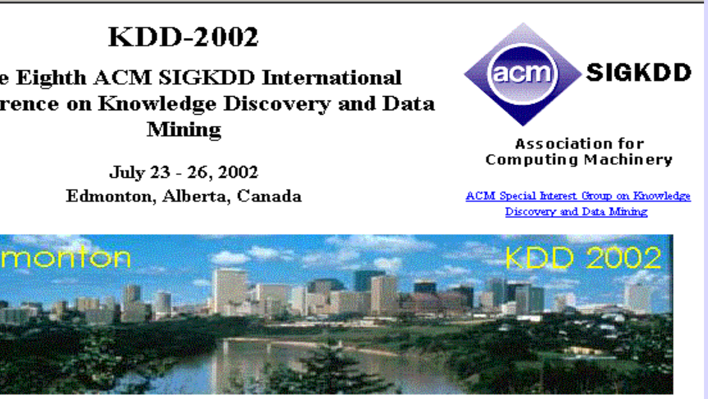
The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining

July 23 - 26, 2002
Edmonton, Alberta, Canada

Association for Computing Machinery

ACM Special Interest Group on Knowledge Discovery and Data Mining

By Alexander Ye



KDD-2002

e Eighth ACM SIGKDD International
rence on Knowledge Discovery and Data
Mining

July 23 - 26, 2002
Edmonton, Alberta, Canada

acm
SIGKDD
Association for
Computing Machinery
ACM Special Interest Group on Knowledge Discovery and Data Mining
monton
KDD 2002

Does that paper contain any curatable gene product information (Yes/No)?

For each gene mentioned in the paper, does that paper have experimental results for

- Transcript(s) of that gene (Yes/No)?
- Protein(s) of that gene (Yes/No)?

- Also produce a ranked list of the papers

Training set (6 weeks) 862 full papers and list of genes

Test set (2 weeks) 213 papers

	Best	**Median**
Ranked-list:	**84%**	**69%**
Yes/No curate paper:	**78%**	**58%**
Yes/No gene products:	**67%**	**35%**

learForest - Celera team used manually generated rules and patterns

A. Yeh, MITRE

BioCreAtIvE

Many groups are now working in the area of text mining. However, despite increased activity in this area, there are no common standards or shared evaluation criteria to enable comparison among the different approaches. Therefore the BioLINK group Biological Literature, Information and Knowledge, [BioLINK]) is organizing a CASP-like evaluation for the text data mining community applied to biology: BioCreAtIvE - Critical Assessment of Information Extraction systems in Biology. Following the experience of CASP, the emphasis will be more on the comparison of methods and the community assessment of scientific progress, rather than on the purely competitive aspects.

http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html

Thurs Nov 13: Test data available (for all tasks/sub-tasks) Nov 19: System submissions are due
Wed Dec 31: Results back to participating groups Feb 04 Submission for workshop
March 30: EMBO Evaluation Workshop, Grenada, Spain

Biocreative team: Swiss-Prot/EBI MITRE CNB/Protein Design Group

Database curators: R. Apweiler, E. Camon, C. O'Donnovan SWISS-PROT C. Wu: PIR J. Blake: MGI I. Donaldson: BIND

Text mining researchers: A. Valencia and C. Blaschke: CNB L. Hirschman and A. Yeh: MITRE L. Hunter, U. of Colorado S-K Ng, Institute for Infocomm Research, Singapore C. Friedman, Columbia

Help from L. Brivell EMBO J. Wilbur and L. Tanabe NCBI

Full-text access: HighWire Press

BioCreAtIvE

Training

674 unique GOs and 1907 in total (i.e. each GO appears about 3 times)
636 papers released for training + 150 Nature journals (Nat. Gen, Nat. Med and Oncogene)

Testing

About 200 proteins. same number of papers and maybe twice as much GO annotations

Task 1: entity extraction

The goal in defining this task was to provide a way of assessing the ability of an automated system to identify the genes (or proteins, where there is ambiguity) mentioned in text.

The "natural language processing" or MUC version of this task has required that a system identify each mention of a gene-or-protein in the text.

http://www.mitre.org/public/biocreative/

Task 2: functional annotation of gene products

The second task will address the assignment of GO annotations to human proteins [GOA]

1. 'Recover' text that provides evidence for the GO annotation
2. Provide GO annotation for human proteins
3. Selection of relevant papers

http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html

BioCreAtIvE

Task 1: Entity Extraction Task

1. Gene list annotation (Creating a list of genes mentioned in an abstract)

   Useful for indexing

2. Gene name mentions (Using data provided by John Wilbur and Lorrie Tanabe, NCBI)

   Corresponds to “named entity” task in the natural language processing

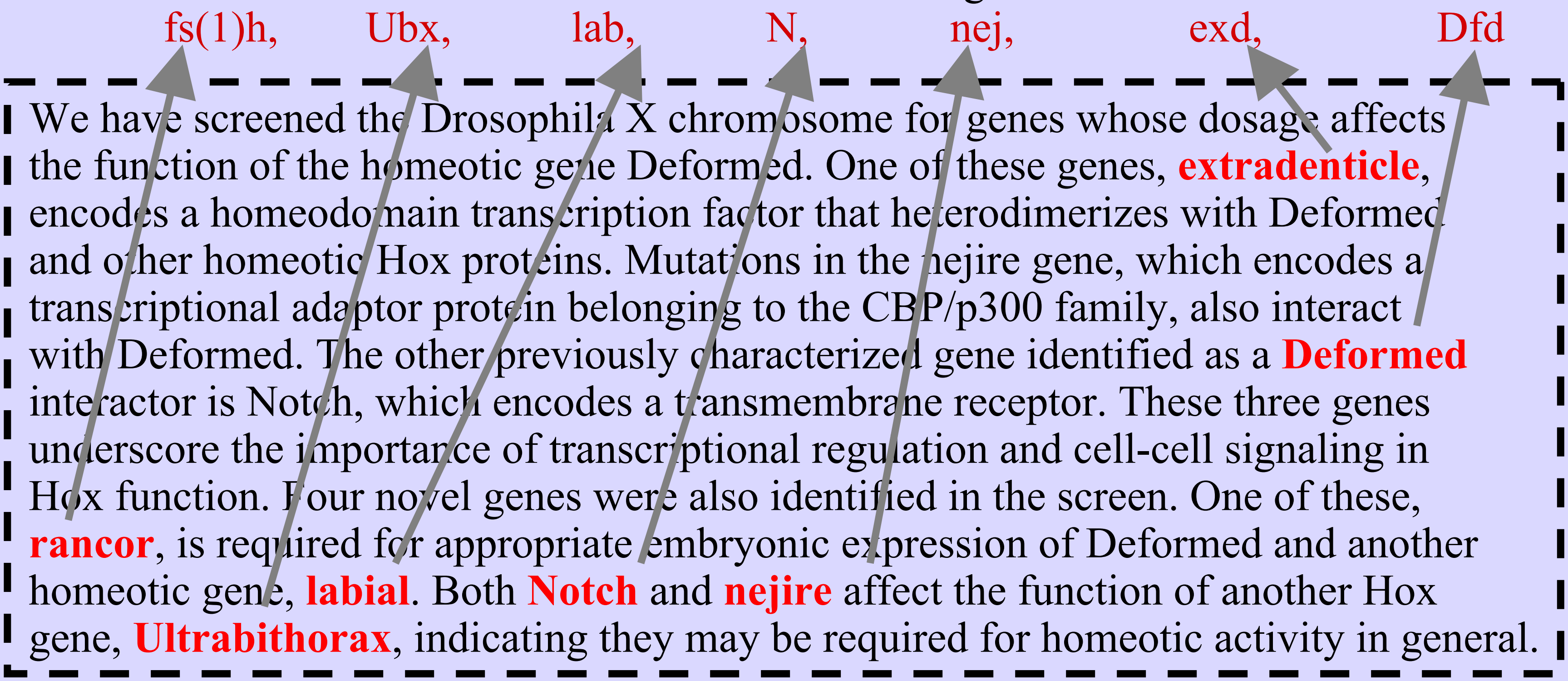
3. Gene references (flagging all references to a named gene in a text)

   Useful as a building block for capturing relations

A. Yeh, MITRE

Entity Extraction Part 1: What a Contestant's System Should Return

- Return a list of the standardized names of the genes mentioned in each abstract:



fs(1)h, Ubx, lab, N, nej, exd, Dfd

We have screened the Drosophila X chromosome for genes whose dosage affects the function of the homeotic gene Deformed. One of these genes, **extradenticle**, encodes a homeodomain transcription factor that heterodimerizes with Deformed and other homeotic Hox proteins. Mutations in the nejire gene, which encodes a transcriptional adaptor protein belonging to the CBP/p300 family, also interact with Deformed. The other previously characterized gene identified as a **Deformed** interactor is Notch, which encodes a transmembrane receptor. These three genes underscore the importance of transcriptional regulation and cell-cell signaling in Hox function. Four novel genes were also identified in the screen. One of these, **rancor**, is required for appropriate embryonic expression of Deformed and another homeotic gene, **labial**. Both **Notch** and **nejire** affect the function of another Hox gene, **Ultrabithorax**, indicating they may be required for homeotic activity in general.

- Also mark 1 text mention of each gene in a list
  - Indicate which gene a mention is for

A. Yeh, MITRE

Part 1: Data Availability (Noisy)

PubMed/Medline abstracts

Papers for various model organism databases (Drosophila, mouse, yeast) and lists of genes (standardized names)

Databases synonym lists

- Training set (1000’s of Abstracts/Organism):
  - correct answer **fs(1)h, Ubx, lab, N, nej, exd, Dfd**
  - Genes in which a known name or synonym appears in the abstract **fs(1)h, Ubx, N, nej, exd, Dfd**
  - Other genes in the list **Cg25C, cnc, kis, stout, apt,** that do not appear in the abstract

- Training set (1000’s of Abstracts/Organism):
  - Test set of 400 abstracts manually tagged

A. Yeh, MITRE

Entity Extraction Part 2: Gene Name Mention

- Data provided by John Wilbur & Lorrie Tanabe, NCBI
  - 10,000 sentences manually annotated for genes
  - Separate development training and test sets
- occurrences of gene-related mentions and text spans

Mutation of **TTF-1-binding sites** (**TBE**) 1, 3, and 4 in combination markedly decreased transcriptional activity of **SP-A promoter-chloramphenicol acetyltransferase** constructs containing **SP-A** gene sequences from -256 to +45.

A. Yeh, MITRE

Task 3: Gene references (flagging all references to a named gene in a text)

- Find every explicit gene name mention and not so explicit references
- Smaller training set (fewer number of abstracts)
  - Information is not available in the databases and more work is involved to annotate one abstract
- Related with relation extraction (examine every mention of every gene)

A. Yeh, MITRE

Task 2. GO annotation

.'Recover' text that proves the GO annotation:

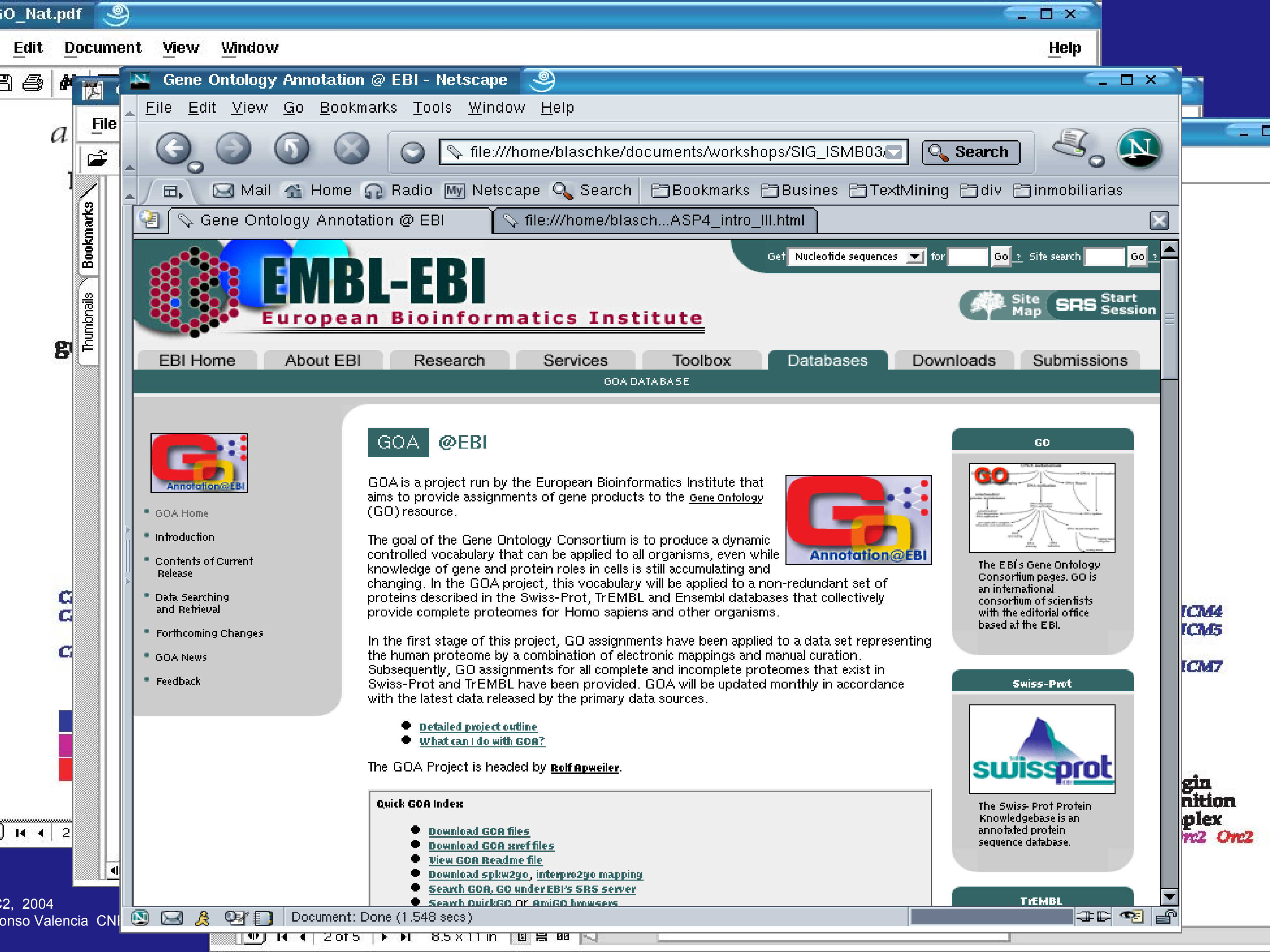
Protein, GO annotation, associated publication >>> provide a part of the document that would (to a human xpert) prove the original annotation.

.Provide GO annotation for human proteins:

rotein, associated publication >>> 'annotate' automatically the GO class and provide a part of the locument to prove the annotation.

.Selection of relevant papers:

rotein and a large number of papers (many irrelevant) >> relevant papers with information suitable to derive GO annotation and parts of the papers useful for annotation.


GO_Nat.pdf
Edit Document View Window Help
Gene Ontology Annotation @ EBI - Netscape
File Edit View Go Bookmarks Tools Window Help
file:///home/blaschke/documents/workshops/SIG_ISMB03/
Search
Mail Home Radio Netscape Search Bookmarks Busines TextMining div inmobiliarias
Gene Ontology Annotation @ EBI
file:///home/blasch...ASP4_intro_III.html
Bookmarks
Thumbnails
EMBL-EBI
European Bioinformatics Institute
Get Nucleotide sequences for Go Site search Go
Site Map
SRS Start Session
EBI Home About EBI Research Services Toolbox Databases Downloads Submissions
GOA DATABASE
Annotation@EBI
GO
Swiss-Prot
swissprot
TrEMBL
2 of 5
8.5 x 11 in
C2, 2004
onso Valencia CNI

- GOA Home
- Introduction
- Contents of Current Release
- Data Searching and Retrieval
- Forthcoming Changes
- GOA News
- Feedback

GOA @EBI

GOA is a project run by the European Bioinformatics Institute that aims to provide assignments of gene products to the Gene Ontology (GO) resource.

The goal of the Gene Ontology Consortium is to produce a dynamic controlled vocabulary that can be applied to all organisms, even while knowledge of gene and protein roles in cells is still accumulating and changing. In the GOA project, this vocabulary will be applied to a non-redundant set of proteins described in the Swiss-Prot, TrEMBL and Ensembl databases that collectively provide complete proteomes for Homo sapiens and other organisms.

In the first stage of this project, GO assignments have been applied to a data set representing the human proteome by a combination of electronic mappings and manual curation. Subsequently, GO assignments for all complete and incomplete proteomes that exist in Swiss-Prot and TrEMBL have been provided. GOA will be updated monthly in accordance with the latest data released by the primary data sources.

- Detailed project outline
- What can I do with GOA?

The GOA Project is headed by Rolf Apweiler.

Quick GOA Index

- Download GOA files
- Download GOA xref files
- View GOA Readme file
- Download spkw2go, interpro2go mapping
- Search GOA, GO under EBI's SRS server
- Search QuickGO or AmiGO browsers

The EBI's Gene Ontology Consortium pages. GO is an international consortium of scientists with the editorial office based at the EBI.

The Swiss-Prot Protein Knowledgebase is an annotated protein sequence database.

Document: Done (1.548 secs)

Challenges

- ***We do not provide protein name dictionaries, i.e. the name of a protein in the GOA file may not be used in the associated documents but a synonym that may be found in Swiss-Prot or in other databases. It is the responsibility of the participants to collect synonyms lists to detect the protein names correctly in the documents.***

- ***GO consists of three (non overlapping) parts (biochemical function, molecular process, cellular component) that are treated separately***

- ***One protein can have many functions (be part of many processes, be localized in different places in the cell) and can therefore appear many times in the corresponding parts of GO***

- ***The function of a protein (its molecular processes, cellular components) can be described in many different articles and in different WAYS***

- ***The GO codes have to be predicted precisely***

- ***One article can describe different functions (processes, components) of the same protein AND/OR mention a number of proteins of which all or just a subset are relevant in our evaluation task***

- ***Full-text articles are long and in general only a (small) section of the whole paper is relevant for classification of a certain protein (maybe a paragraph or two)***

Examples

1.- RGS4 GO:0005516 calmodulin binding activity PMID 10747990

'Indeed, Ca2+/calmodulin binds a complex of RGS4 and a transition state analog of Galpha i1-GDP-AlF4-'

2.- p21waf/cip1 GO: 0008285 negative regulation of cell proliferation PMID 10692450

'The p21waf/cip1 protein is a universal inhibitor of cyclin kinases and plays an important role in inhibiting cell proliferation'

3.- Thrombin GO:0006915 apoptosis PMID 10692450

'Induction of Apoptosis by Thrombin'

4.- RGS1,RGS2,RGS4,RGS16 GO: 0008277 regulation of G-protein coupled receptor protein signaling pathway PMID 10747990

'We report that calmodulin binds in a Ca2+-dependent manner to all RGS proteins we tested, including RGS1, RGS2, RGS4, RGS10, RGS16, and GAIP' and later in the text 'To investigate the role of Ca2+ in feedback regulation of G protein signaling by RGS proteins, we characterized ...'.

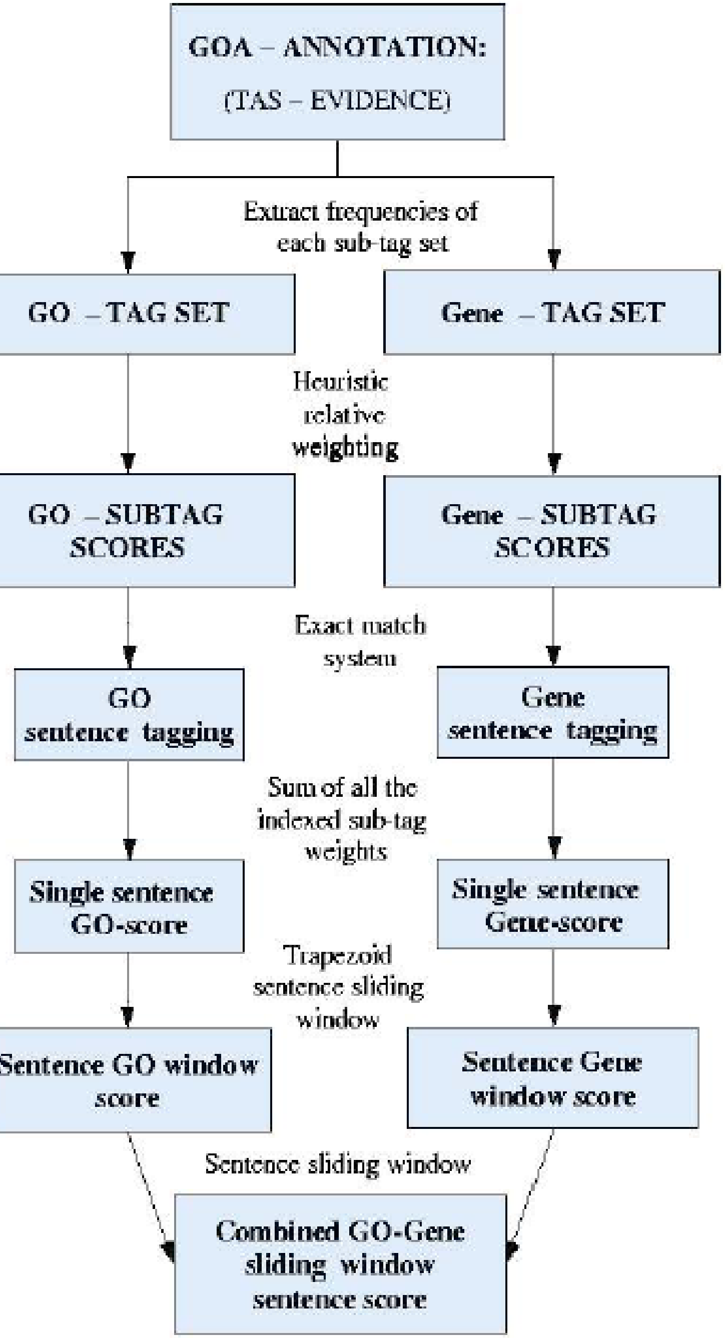
establish first a the relation between the individual proteins and the fact that they are all RGS proteins and then interpret from the second sentence later in the text that these proteins are related to G protein signaling

5.- MIP-1alpha GO:0007186 G-protein coupled receptor protein signaling pathway PMID 10734056

'Taken together, these results indicate that CCR1-mediated responses are regulated at several steps in the signaling pathway, by receptor phosphorylation at the level of receptor/G protein coupling and by an unknown mechanism at the level of phospholipase C activation' and later 'In this study, the CCR1 receptor, which binds RANTES, MIP-1alpha , MCP-2, and MCP-3 with high affinity'.

The first sentence establishes that CCR1 is related to a G-protein coupled receptor pathway and the second sentence states that MIP-1alpha binds to this receptor and it can be deduced that it is therefore also related to this process.

CNB Text to GOA



GOA – ANNOTATION: (TAS – EVIDENCE)

Extract frequencies of each sub-tag set

GO – TAG SET

Gene – TAG SET

Heuristic relative weighting

GO – SUBTAG SCORES

Gene – SUBTAG SCORES

Exact match system

GO sentence tagging

Gene sentence tagging

Sum of all the indexed sub-tag weights

Single sentence GO-score

Single sentence Gene-score

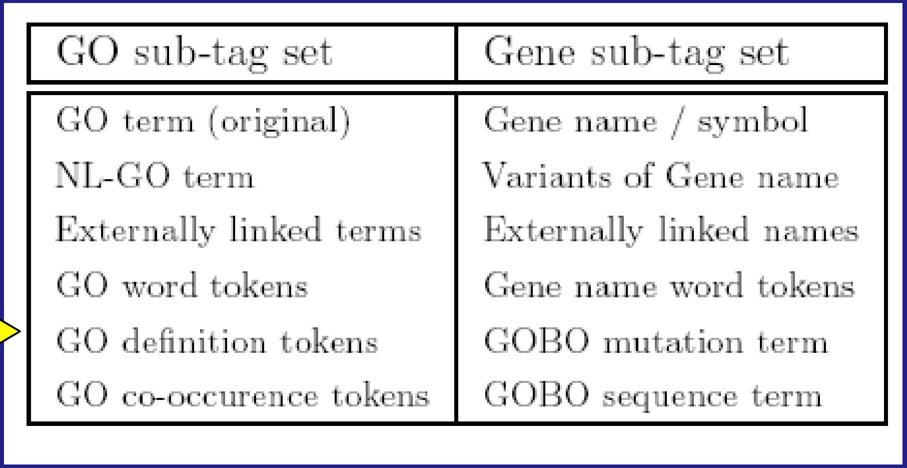
Trapezoid sentence sliding window

Sentence GO window score

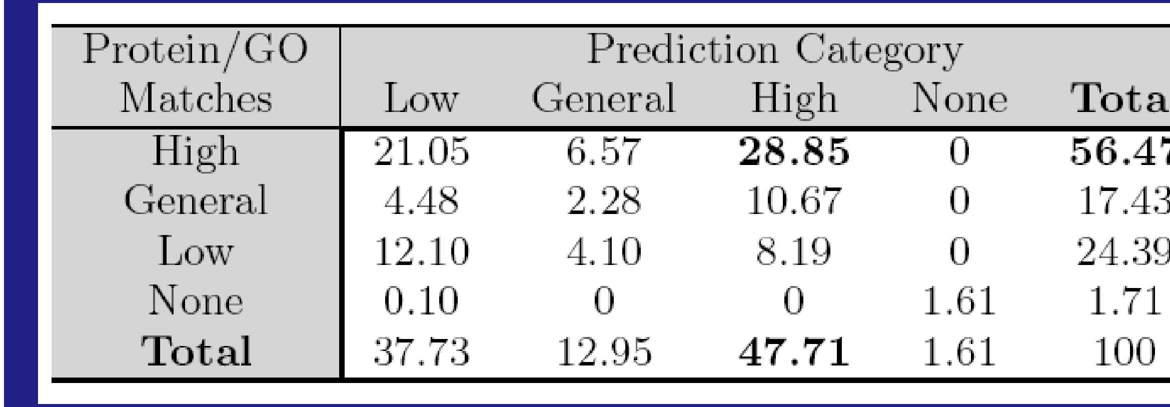
Sentence Gene window score

Sentence sliding window

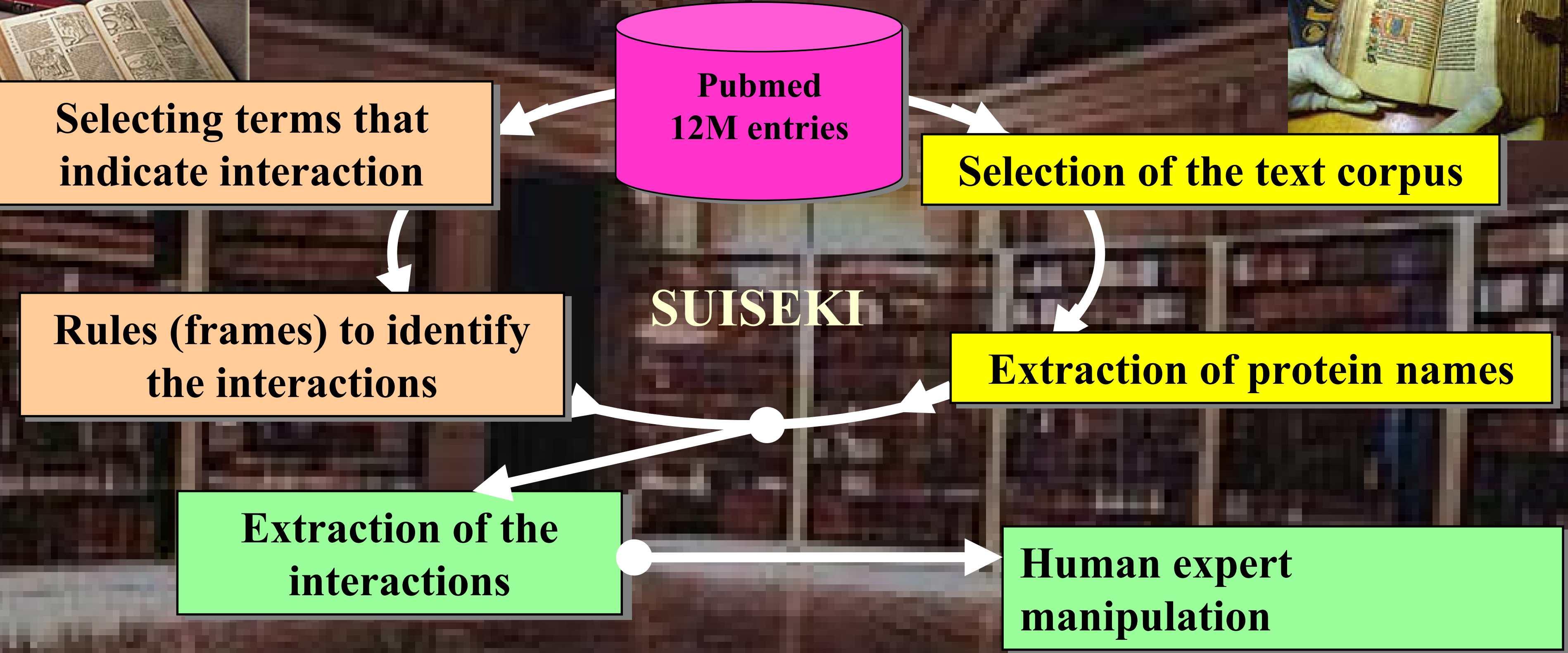
Combined GO-Gene sliding window sentence score



GO sub-tag set	Gene sub-tag set
GO term (original)	Gene name / symbol
NL-GO term	Variants of Gene name
Externally linked terms	Externally linked names
GO word tokens	Gene name word tokens
GO definition tokens	GOBO mutation term
GO co-occurence tokens	GOBO sequence term



Protein/GO	Prediction Category				
Matches	Low	General	High	None	**Tota**
High	21.05	6.57	**28.85**	0	**56.47**
General	4.48	2.28	10.67	0	17.43
Low	12.10	4.10	8.19	0	24.39
None	0.10	0	0	1.61	1.71
Total	37.73	12.95	**47.71**	1.61	100

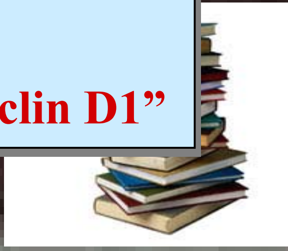

Pubmed
12M entries
Selecting terms that indicate interaction
Selection of the text corpus
Rules (frames) to identify the interactions
SUISEKI
Extraction of protein names
Extraction of the interactions
Human expert manipulation

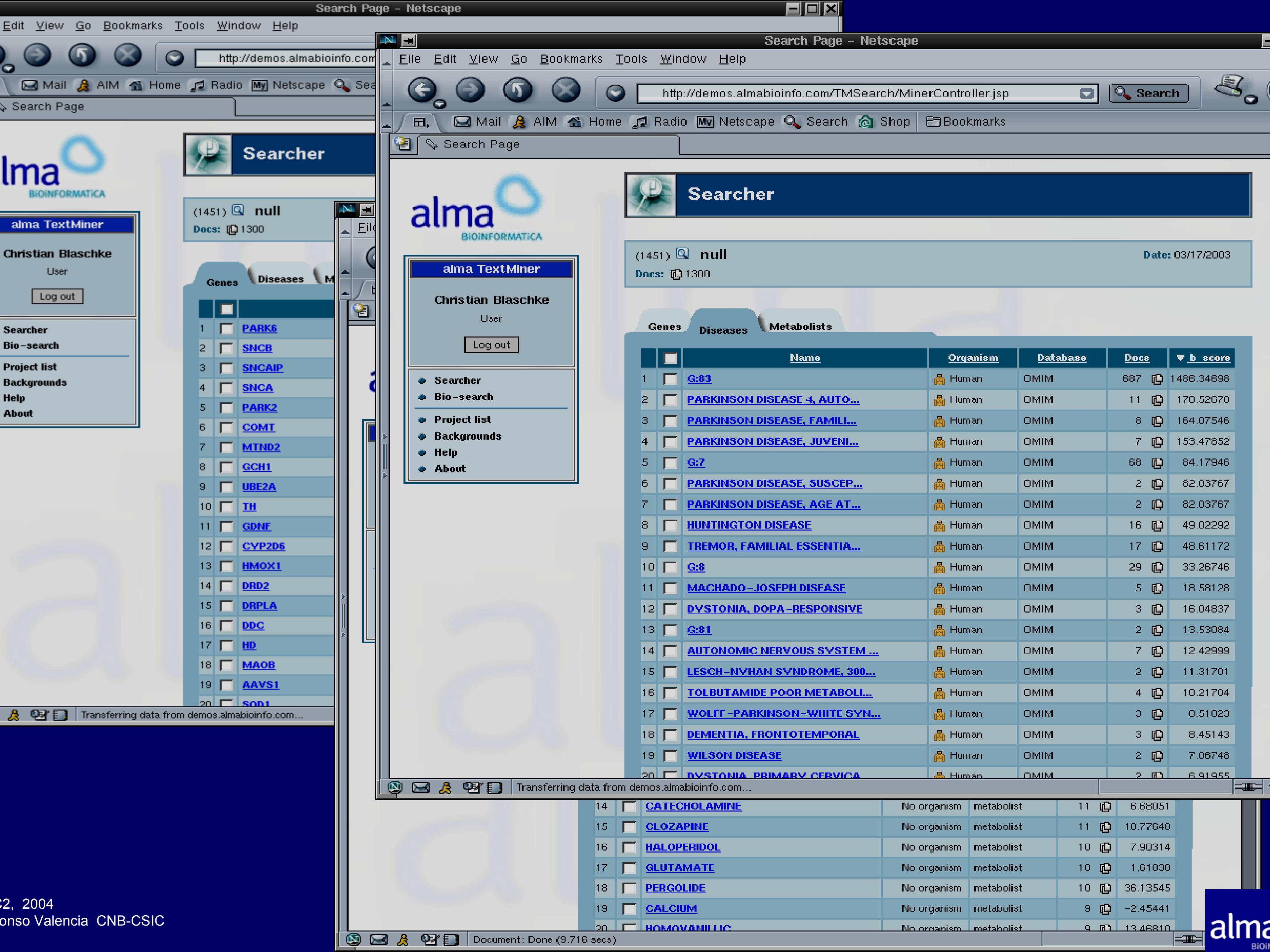
Action words are for example:

activate, associated with, bind, interact, phosphorylate, regulate

* [protein A] ... verb indicating an action ... [protein B]

"After extensive purification, Cdk2 was still bound to cyclin D1"




Search Page - Netscape
File Edit View Go Bookmarks Tools Window Help
http://demos.almabioinfo.com/TMSearch/MinerController.jsp
Search
Mail AIM Home Radio Netscape Search Shop Bookmarks
Search Page
alma BiOiNFORMATiCA
Searcher
alma TextMiner
Log out
Searcher
Bio-search
Project list
Backgrounds
Help
About
Genes
Diseases
Metabolists
Transferring data from demos.almabioinfo.com...
Document: Done (9.716 secs)

Christian Blaschke
User

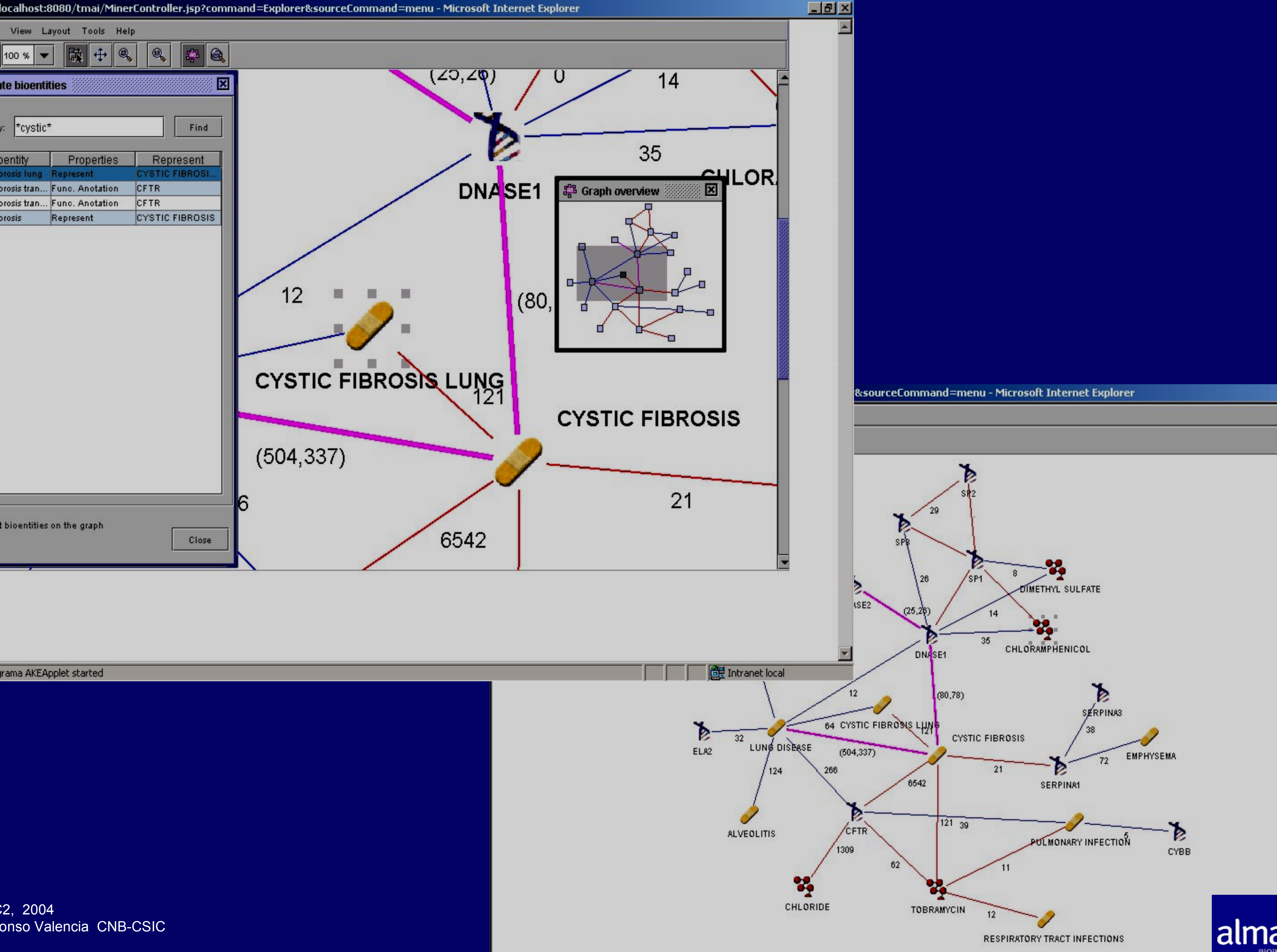
(1451) null Date: 03/17/2003
Docs: 1300

	□	Name	Organism	Database	Docs	▼ b_score
1	□	G:83	Human	OMIM	687	1486.34698
2	□	PARKINSON DISEASE 4, AUTO...	Human	OMIM	11	170.52670
3	□	PARKINSON DISEASE, FAMILI...	Human	OMIM	8	164.07546
4	□	PARKINSON DISEASE, JUVENI...	Human	OMIM	7	153.47852
5	□	G:7	Human	OMIM	68	84.17946
6	□	PARKINSON DISEASE, SUSCEP...	Human	OMIM	2	82.03767
7	□	PARKINSON DISEASE, AGE AT...	Human	OMIM	2	82.03767
8	□	HUNTINGTON DISEASE	Human	OMIM	16	49.02292
9	□	TREMOR, FAMILIAL ESSENTIA...	Human	OMIM	17	48.61172
10	□	G:8	Human	OMIM	29	33.26746
11	□	MACHADO-JOSEPH DISEASE	Human	OMIM	5	18.58128
12	□	DYSTONIA, DOPA-RESPONSIVE	Human	OMIM	3	16.04837
13	□	G:81	Human	OMIM	2	13.53084
14	□	AUTONOMIC NERVOUS SYSTEM ...	Human	OMIM	7	12.42999
15	□	LESCH-NYHAN SYNDROME, 300...	Human	OMIM	2	11.31701
16	□	TOLBUTAMIDE POOR METABOLI...	Human	OMIM	4	10.21704
17	□	WOLFF-PARKINSON-WHITE SYN...	Human	OMIM	3	8.51023
18	□	DEMENTIA, FRONTOTEMPORAL	Human	OMIM	3	8.45143
19	□	WILSON DISEASE	Human	OMIM	2	7.06748
20	□	DYSTONIA, PRIMARY CERVICA	Human	OMIM	2	6.91955

	□	
1	□	PARK6
2	□	SNCB
3	□	SNCAIP
4	□	SNCA
5	□	PARK2
6	□	COMT
7	□	MTND2
8	□	GCH1
9	□	UBE2A
10	□	TH
11	□	GDNF
12	□	CYP2D6
13	□	HMOX1
14	□	DRD2
15	□	DRPLA
16	□	DDC
17	□	HD
18	□	MAOB
19	□	AAVS1
20	□	SOD1

	□					
14	□	CATECHOLAMINE	No organism	metabolist	11	6.68051
15	□	CLOZAPINE	No organism	metabolist	11	10.77648
16	□	HALOPERIDOL	No organism	metabolist	10	7.90314
17	□	GLUTAMATE	No organism	metabolist	10	1.61838
18	□	PERGOLIDE	No organism	metabolist	10	36.13545
19	□	CALCIUM	No organism	metabolist	9	-2.45441
20	□	HOMOVANILLIC	No organism	metabolist	9	13.46810

C2, 2004
onso Valencia CNB-CSIC



localhost:8080/tmai/MinerController.jsp?command=Explorer&sourceCommand=menu - Microsoft Internet Explorer

te bioentities

y: *cystic* Find

entity	Properties	Represent
rosis lung	Represent	CYSTIC FIBROSI...
rosis tran...	Func. Anotation	CFTR
rosis tran...	Func. Anotation	CFTR
rosis	Represent	CYSTIC FIBROSIS

t bioentities on the graph

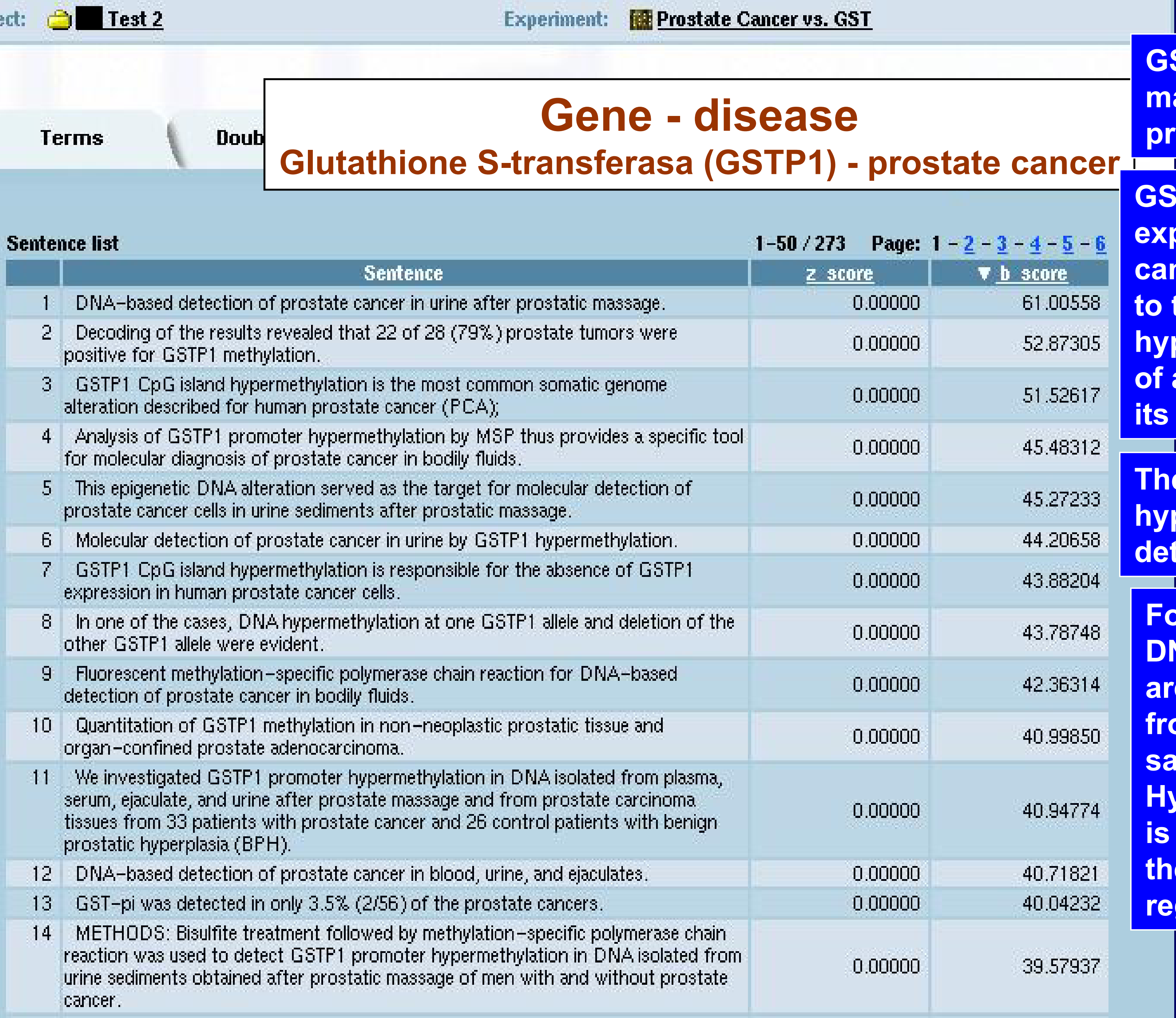
rama AKEApplet started

&sourceCommand=menu - Microsoft Internet Explorer

C2, 2004
onso Valencia CNB-CSIC

Gene - disease

Glutathione S-transferasa (GSTP1) - prostate cancer



Sentence list 1-50 / 273 Page: 1 - 2 - 3 - 4 - 5 - 6

	Sentence	z score	▼ b score
1	DNA-based detection of prostate cancer in urine after prostatic massage.	0.00000	61.00558
2	Decoding of the results revealed that 22 of 28 (79%) prostate tumors were positive for GSTP1 methylation.	0.00000	52.87305
3	GSTP1 CpG island hypermethylation is the most common somatic genome alteration described for human prostate cancer (PCA);	0.00000	51.52617
4	Analysis of GSTP1 promoter hypermethylation by MSP thus provides a specific tool for molecular diagnosis of prostate cancer in bodily fluids.	0.00000	45.48312
5	This epigenetic DNA alteration served as the target for molecular detection of prostate cancer cells in urine sediments after prostatic massage.	0.00000	45.27233
6	Molecular detection of prostate cancer in urine by GSTP1 hypermethylation.	0.00000	44.20658
7	GSTP1 CpG island hypermethylation is responsible for the absence of GSTP1 expression in human prostate cancer cells.	0.00000	43.88204
8	In one of the cases, DNA hypermethylation at one GSTP1 allele and deletion of the other GSTP1 allele were evident.	0.00000	43.78748
9	Fluorescent methylation-specific polymerase chain reaction for DNA-based detection of prostate cancer in bodily fluids.	0.00000	42.36314
10	Quantitation of GSTP1 methylation in non-neoplastic prostatic tissue and organ-confined prostate adenocarcinoma.	0.00000	40.99850
11	We investigated GSTP1 promoter hypermethylation in DNA isolated from plasma, serum, ejaculate, and urine after prostate massage and from prostate carcinoma tissues from 33 patients with prostate cancer and 26 control patients with benign prostatic hyperplasia (BPH).	0.00000	40.94774
12	DNA-based detection of prostate cancer in blood, urine, and ejaculates.	0.00000	40.71821
13	GST-pi was detected in only 3.5% (2/56) of the prostate cancers.	0.00000	40.04232
14	METHODS: Bisulfite treatment followed by methylation-specific polymerase chain reaction was used to detect GSTP1 promoter hypermethylation in DNA isolated from urine sediments obtained after prostatic massage of men with and without prostate cancer.	0.00000	39.57937

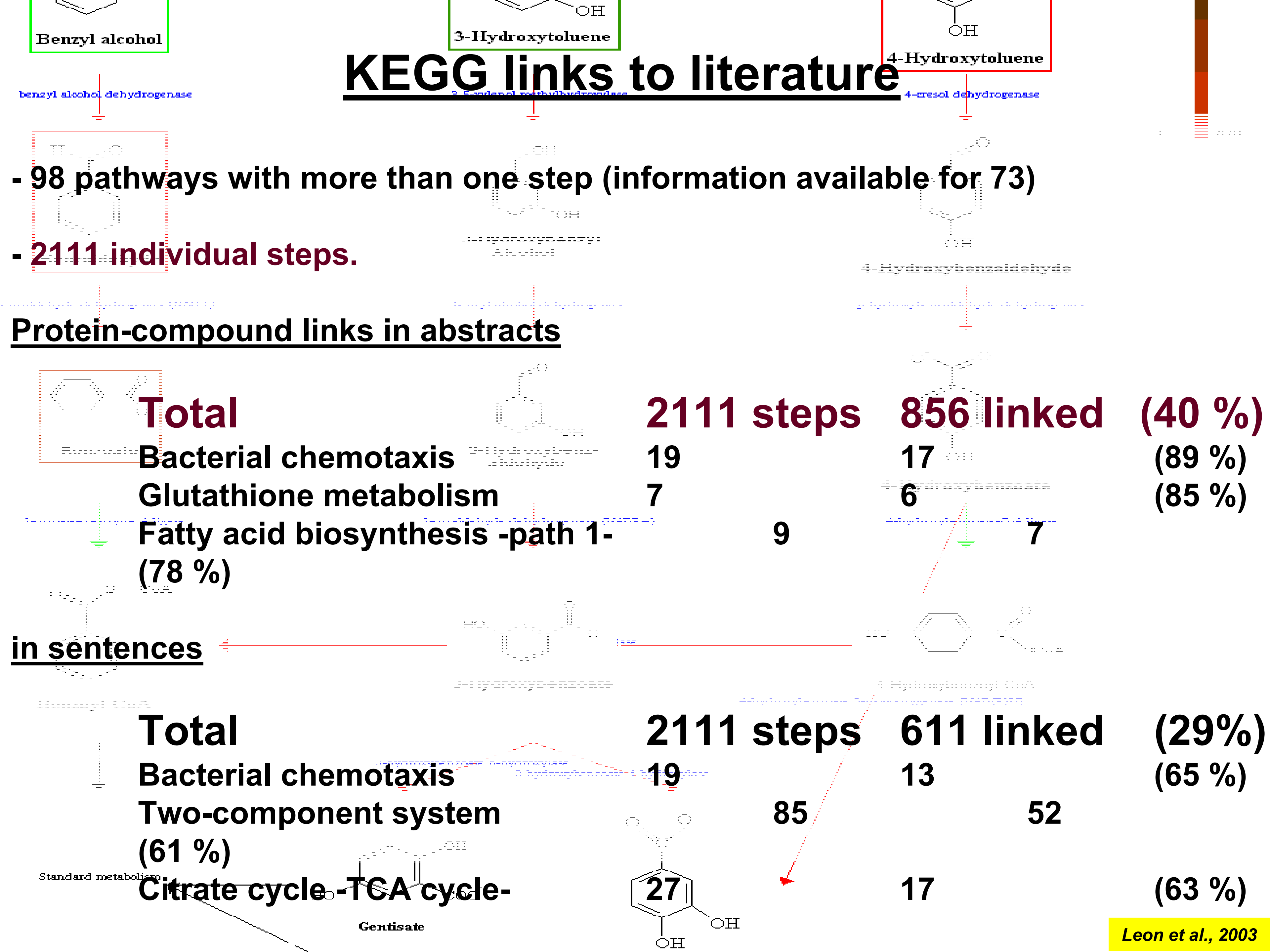
GST used as marker of prostate cance

GST is not expressed in cancer cells du to the hypermetylatio of a CpG island its promotor

The hypemetylation detected by PC

For diagnosis, DNA samples are extracted from urine samples. Hypermetylatio is analyzed on the promoter regions

alma



Benzyl alcohol
3-Hydroxytoluene
4-Hydroxytoluene
benzyl alcohol dehydrogenase
3,5-xylenol methylhydroxylase
4-cresol dehydrogenase
Benzaldehyde
3-Hydroxybenzyl Alcohol
4-Hydroxybenzaldehyde
benzaldehyde dehydrogenase (NAD+)
benzyl alcohol dehydrogenase
p-hydroxybenzaldehyde dehydrogenase
Benzoate
3-Hydroxybenzaldehyde
4-Hydroxybenzoate
benzoate-coenzyme A ligase
benzaldehyde dehydrogenase (NADP+)
4-hydroxybenzoate-CoA ligase
Benzoyl-CoA
3-Hydroxybenzoate
4-Hydroxybenzoyl-CoA
4-hydroxybenzoate 3-monooxygenase (NAD(P)H)
3-hydroxybenzoate 6-hydroxylase
3-hydroxybenzoate 4-hydroxylase
Standard metabolism
Gentisate

KEGG links to literature

- 98 pathways with more than one step (information available for 73)
- 2111 individual steps.

Protein-compound links in abstracts

Total	2111 steps	856 linked	(40 %)
Bacterial chemotaxis	19	17	(89 %)
Glutathione metabolism	7	6	(85 %)
Fatty acid biosynthesis -path 1-	9	7	(78 %)

in sentences

Total	2111 steps	611 linked	(29%)
Bacterial chemotaxis	19	13	(65 %)
Two-component system	85	52	(61 %)
Citrate cycle -TCA cycle-	27	17	(63 %)

Leon et al., 2003

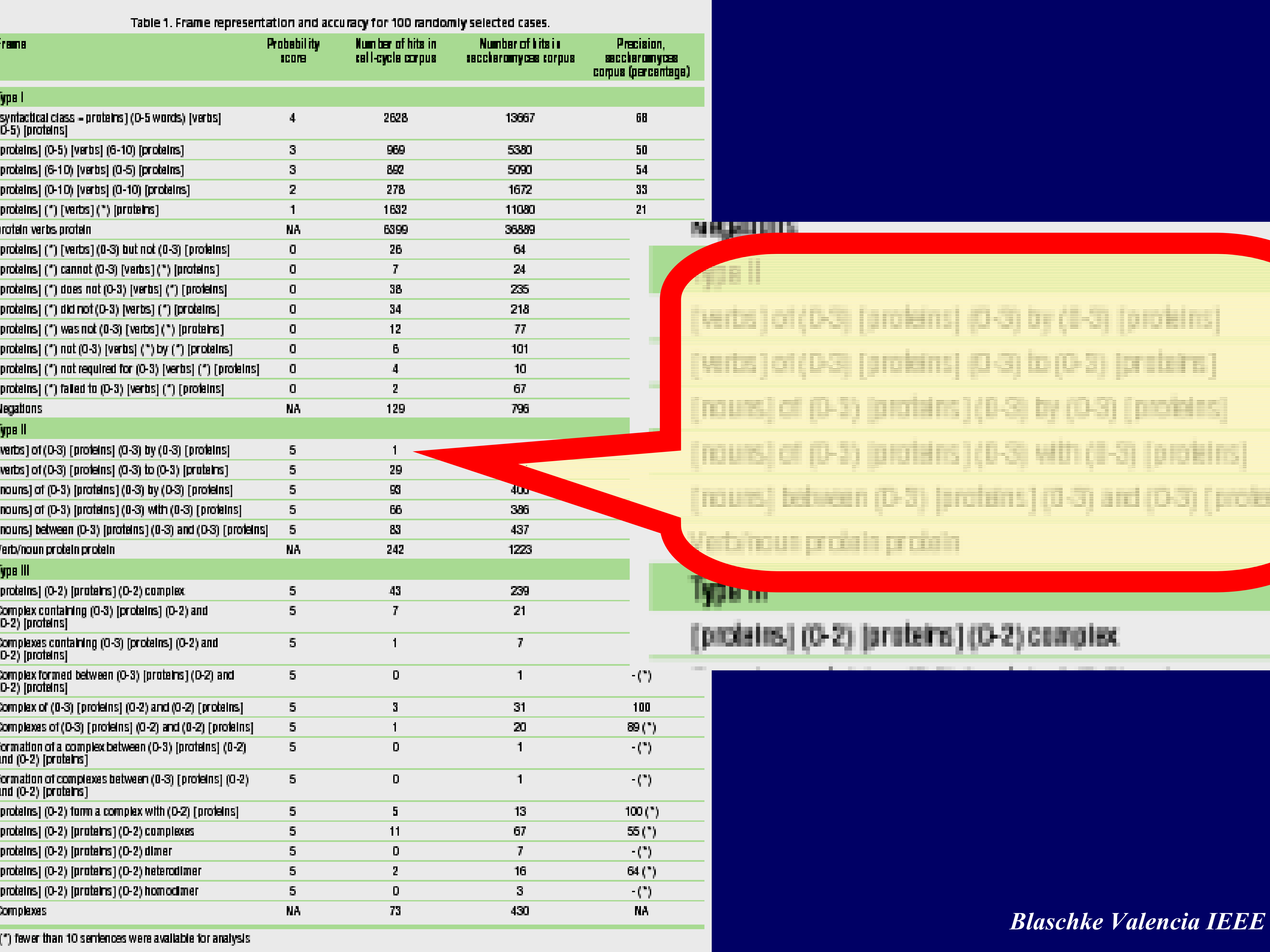


Table 1. Frame representation and accuracy for 100 randomly selected cases.

rame	Probability score	Number of hits in cell-cycle corpus	Number of hits in saccharomyces corpus	Precision, saccharomyces corpus (percentage)
ype I				
syntactical class = proteins] (0-5 words) [verbs] 0-5) [proteins]	4	2628	13667	68
proteins] (0-5) [verbs] (6-10) [proteins]	3	969	5380	50
proteins] (6-10) [verbs] (0-5) [proteins]	3	892	5090	54
proteins] (0-10) [verbs] (0-10) [proteins]	2	278	1672	33
proteins] (*) [verbs] (*) [proteins]	1	1632	11080	21
rotein verbs protein	NA	6399	36889	
proteins] (*) [verbs] (0-3) but not (0-3) [proteins]	0	26	64	
proteins] (*) cannot (0-3) [verbs] (*) [proteins]	0	7	24	
proteins] (*) does not (0-3) [verbs] (*) [proteins]	0	38	235	
proteins] (*) did not (0-3) [verbs] (*) [proteins]	0	34	218	
proteins] (*) was not (0-3) [verbs] (*) [proteins]	0	12	77	
proteins] (*) not (0-3) [verbs] (*) by (*) [proteins]	0	6	101	
proteins] (*) not required for (0-3) [verbs] (*) [proteins]	0	4	10	
proteins] (*) failed to (0-3) [verbs] (*) [proteins]	0	2	67	
legations	NA	129	796	
ype II				
verbs] of (0-3) [proteins] (0-3) by (0-3) [proteins]	5	1		
verbs] of (0-3) [proteins] (0-3) to (0-3) [proteins]	5	29		
nouns] of (0-3) [proteins] (0-3) by (0-3) [proteins]	5	93		
nouns] of (0-3) [proteins] (0-3) with (0-3) [proteins]	5	66	386	
nouns] between (0-3) [proteins] (0-3) and (0-3) [proteins]	5	83	437	
'erb/noun protein protein	NA	242	1223	
ype III				
proteins] (0-2) [proteins] (0-2) complex	5	43	239	
omplex containing (0-3) [proteins] (0-2) and 0-2) [proteins]	5	7	21	
omplexes containing (0-3) [proteins] (0-2) and 0-2) [proteins]	5	1	7	
omplex formed between (0-3) [proteins] (0-2) and 0-2) [proteins]	5	0	1	- (*)
omplex of (0-3) [proteins] (0-2) and (0-2) [proteins]	5	3	31	100
omplexes of (0-3) [proteins] (0-2) and (0-2) [proteins]	5	1	20	89 (*)
ormation of a complex between (0-3) [proteins] (0-2) nd (0-2) [proteins]	5	0	1	- (*)
ormation of complexes between (0-3) [proteins] (0-2) nd (0-2) [proteins]	5	0	1	- (*)
proteins] (0-2) form a complex with (0-2) [proteins]	5	5	13	100 (*)
proteins] (0-2) [proteins] (0-2) complexes	5	11	67	55 (*)
proteins] (0-2) [proteins] (0-2) dimer	5	0	7	- (*)
proteins] (0-2) [proteins] (0-2) heterodimer	5	2	16	64 (*)
proteins] (0-2) [proteins] (0-2) homodimer	5	0	3	- (*)
omplexes	NA	73	430	NA

(*) fewer than 10 sentences were available for analysis

Negations

Type II

[verbs] of (0-3) [proteins] (0-3) by (0-3) [proteins]

[verbs] of (0-3) [proteins] (0-3) to (0-3) [proteins]

[nouns] of (0-3) [proteins] (0-3) by (0-3) [proteins]

[nouns] of (0-3) [proteins] (0-3) with (0-3) [proteins]

[nouns] between (0-3) [proteins] (0-3) and (0-3) [prote

Verb/noun protein protein

Type III

[proteins] (0-2) [proteins] (0-2) complex

Blaschke Valencia IEEE

Suiseki (motivation)

“There are advantages to each of these approaches [grammar or pattern matching]. Generally, the less syntax is used, the more domain-specific the system is. This allows you to construct a robust system relatively quickly, but many subtleties may be lost in the interpretation of the sentence.

… In some applications, however, the domain-dependent pattern-matching approach may be the only way to attain reasonable performance in the foreseeable future”

Allen, J. (1995). Natural language understanding.

Evaluation of the system

COVERAGE interactions		coverage words		coverage	
corpus	Abstracts	identified	unique	identified	unique
cell cycle	5,283	69,193	18,942	6,778	4,657
yeast	43,417	500,943	100,729	39,126	25,988

tein names in yeast cell cycle corpus

	instances	Recall (*)	Precision (*)
ected names	1387		
racted names	1766		
rect detection	1331	96.0 %	75.4 %
mpletely rect detection	1201	86.6 %	68.0 %

aluation in 100 randomly chosen genes

Relation between accuracy and number of insta

	Names	Sentences	
	% correct names	% correct interactions	mean score
first 25%	76	80	8.5
second quarter	71	69	4.0
third quarter	60	63	3.2
last quarter	52	42	1.5

Blaschke , Valencia, E

iOp

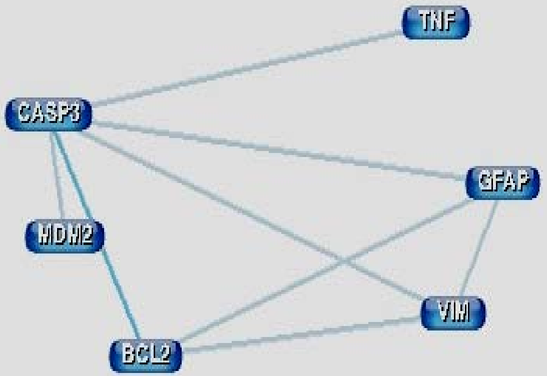
tion hyperlinked
er proteins

Gene

odel

rsion

n the graph
nd to genes
cur in at least
2 ● 3 phrases.



MDM2 is cleaved by Caspase 3 (CPP32) during apoptosis after aspartic acid-361, generating a 60 kd fragment.

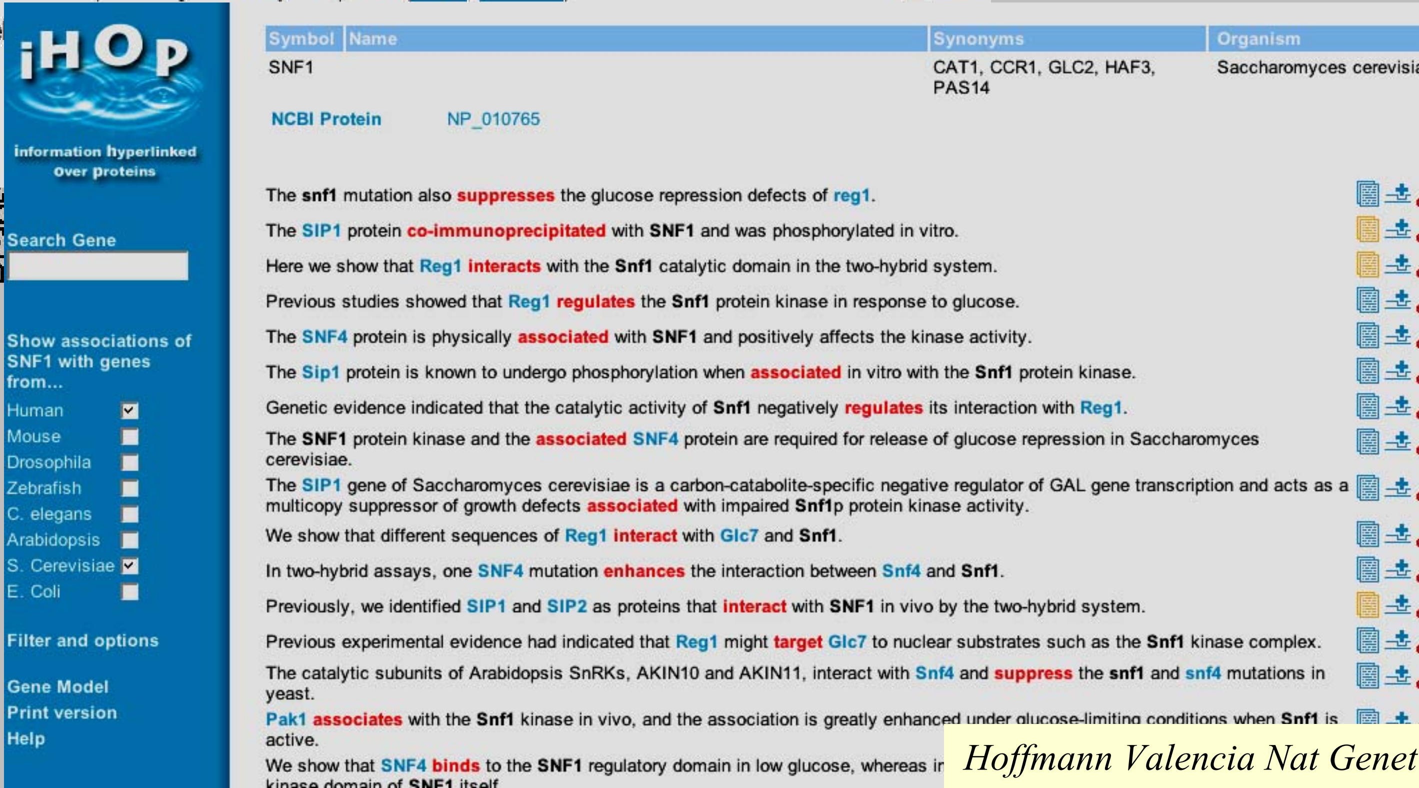
These findings indicate that IR-induced apoptosis involves activation of CPP32 and that this CrmA-insensitive apoptotic pathway is distinct from those induced by TNF and certain other stimuli.

At the end of the experiment, the whole CNS of each animal was collected for histopathology and immunohistochemistry for apoptotic markers (BAX, BCl2 and CPP32) and for glial fibrillary acid protein (GFAP; vimentin).

Fas-induced activation of the ce

MeSH Terms

🟦	level 1-3
🟩	level 4-6
🟪	level 7-9
🟥	Genes



iHOP

information hyperlinked over proteins

Search Gene

Show associations of SNF1 with genes from...

- Human ☑
- Mouse □
- Drosophila □
- Zebrafish □
- C. elegans □
- Arabidopsis □
- S. Cerevisiae ☑
- E. Coli □

Filter and options

Gene Model
Print version
Help

Symbol	Name	Synonyms	Organism
SNF1		CAT1, CCR1, GLC2, HAF3, PAS14	Saccharomyces cerevisia

NCBI Protein NP_010765

The snf1 mutation also suppresses the glucose repression defects of reg1.

The SIP1 protein co-immunoprecipitated with SNF1 and was phosphorylated in vitro.

Here we show that Reg1 interacts with the Snf1 catalytic domain in the two-hybrid system.

Previous studies showed that Reg1 regulates the Snf1 protein kinase in response to glucose.

The SNF4 protein is physically associated with SNF1 and positively affects the kinase activity.

The Sip1 protein is known to undergo phosphorylation when associated in vitro with the Snf1 protein kinase.

Genetic evidence indicated that the catalytic activity of Snf1 negatively regulates its interaction with Reg1.

The SNF1 protein kinase and the associated SNF4 protein are required for release of glucose repression in Saccharomyces cerevisiae.

The SIP1 gene of Saccharomyces cerevisiae is a carbon-catabolite-specific negative regulator of GAL gene transcription and acts as a multicopy suppressor of growth defects associated with impaired Snf1p protein kinase activity.

We show that different sequences of Reg1 interact with Glc7 and Snf1.

In two-hybrid assays, one SNF4 mutation enhances the interaction between Snf4 and Snf1.

Previously, we identified SIP1 and SIP2 as proteins that interact with SNF1 in vivo by the two-hybrid system.

Previous experimental evidence had indicated that Reg1 might target Glc7 to nuclear substrates such as the Snf1 kinase complex.

The catalytic subunits of Arabidopsis SnRKs, AKIN10 and AKIN11, interact with Snf4 and suppress the snf1 and snf4 mutations in yeast.

Pak1 associates with the Snf1 kinase in vivo, and the association is greatly enhanced under glucose-limiting conditions when Snf1 is active.

We show that SNF4 binds to the SNF1 regulatory domain in low glucose, whereas ir
kinase domain of SNF1 itself.

C2, 2004
onso Valencia CNB-CSIC

Hoffmann Valencia Nat Genet

Datamining for Chromosomal Aberrations

examples from H-CAD

Source

Source: MEDLINE, PMID=10862084

The mutational spectrum consisted of 25 nonsense, 12 frameshift, 19 splice mutations, six missense and/or small in-frame deletions, one deletion of the entire NF1 gene, and a translocation t(14;17)(q32;q11.2). Our data suggest that exons 10a-10c and 37 are mutation-rich regions...

Source: MEDLINE, PMID=2562822

Fine structure DNA mapping studies of the chromosomal region harboring the genetic defect in neurofibromatosis type I. To better map the location of the von Recklinghausen neurofibromatosis (NF1) gene, we have characterized a somatic cell hybrid designated 7AE-11... The panel included a hybrid (NF13) carrying a der(22) chromosome that was isolated from an NF1 patient with a balanced translocation, t(17;22) (q11.2;q11.2). Fifty-three of the cosmids map into a region spanning the NF13

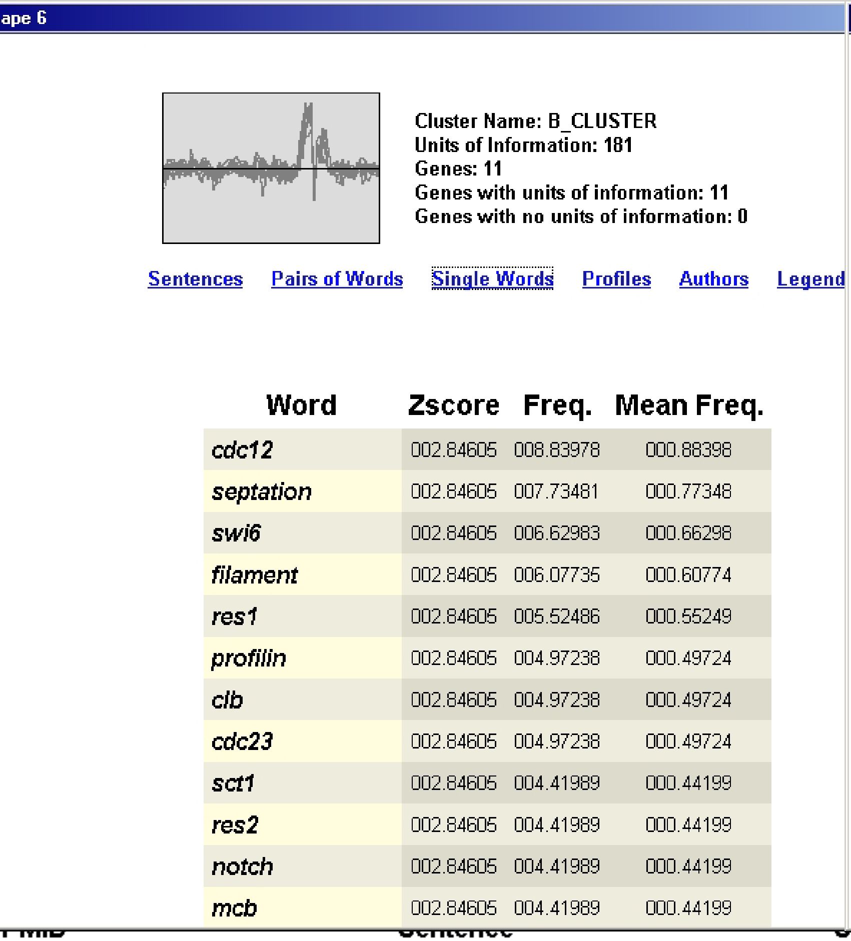
Information Extracted

translocation
t(14;17)(q32;q11.2)
t(17;22) (q11.2;q11.2)

breakpoint
17q11.2

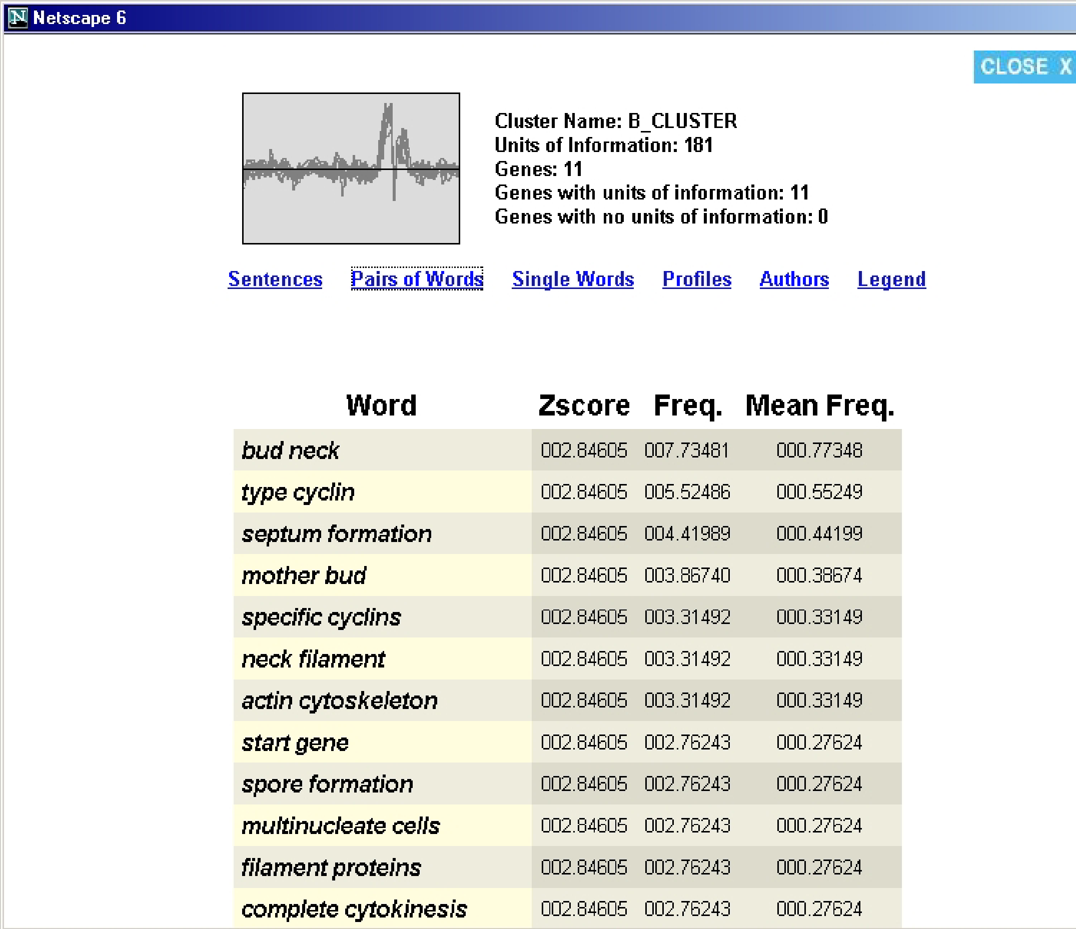
phenotype
neurofibromatosis type I

gene
NF1 gene



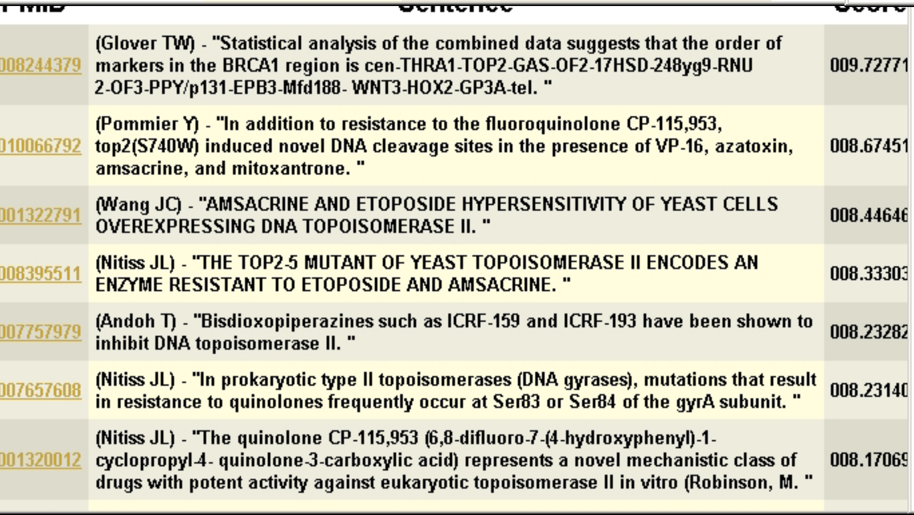
Cluster Name: B_CLUSTER
Units of Information: 181
Genes: 11
Genes with units of information: 11
Genes with no units of information: 0

Word	Zscore	Freq.	Mean Freq.
cdc12	002.84605	008.83978	000.88398
septation	002.84605	007.73481	000.77348
swi6	002.84605	006.62983	000.66298
filament	002.84605	006.07735	000.60774
res1	002.84605	005.52486	000.55249
profilin	002.84605	004.97238	000.49724
clb	002.84605	004.97238	000.49724
cdc23	002.84605	004.97238	000.49724
sct1	002.84605	004.41989	000.44199
res2	002.84605	004.41989	000.44199
notch	002.84605	004.41989	000.44199
mcb	002.84605	004.41989	000.44199

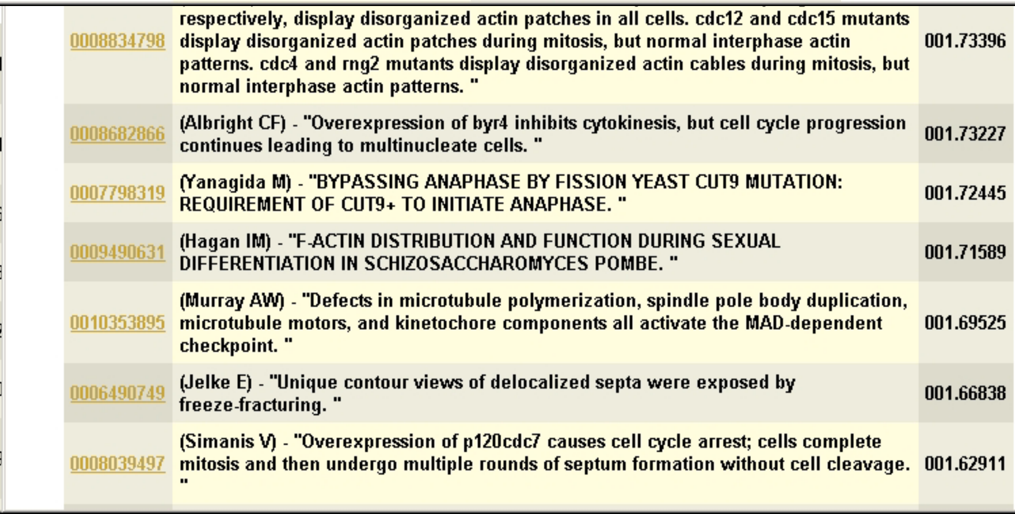


Cluster Name: B_CLUSTER
Units of Information: 181
Genes: 11
Genes with units of information: 11
Genes with no units of information: 0

Word	Zscore	Freq.	Mean Freq.
bud neck	002.84605	007.73481	000.77348
type cyclin	002.84605	005.52486	000.55249
septum formation	002.84605	004.41989	000.44199
mother bud	002.84605	003.86740	000.38674
specific cyclins	002.84605	003.31492	000.33149
neck filament	002.84605	003.31492	000.33149
actin cytoskeleton	002.84605	003.31492	000.33149
start gene	002.84605	002.76243	000.27624
spore formation	002.84605	002.76243	000.27624
multinucleate cells	002.84605	002.76243	000.27624
filament proteins	002.84605	002.76243	000.27624
complete cytokinesis	002.84605	002.76243	000.27624



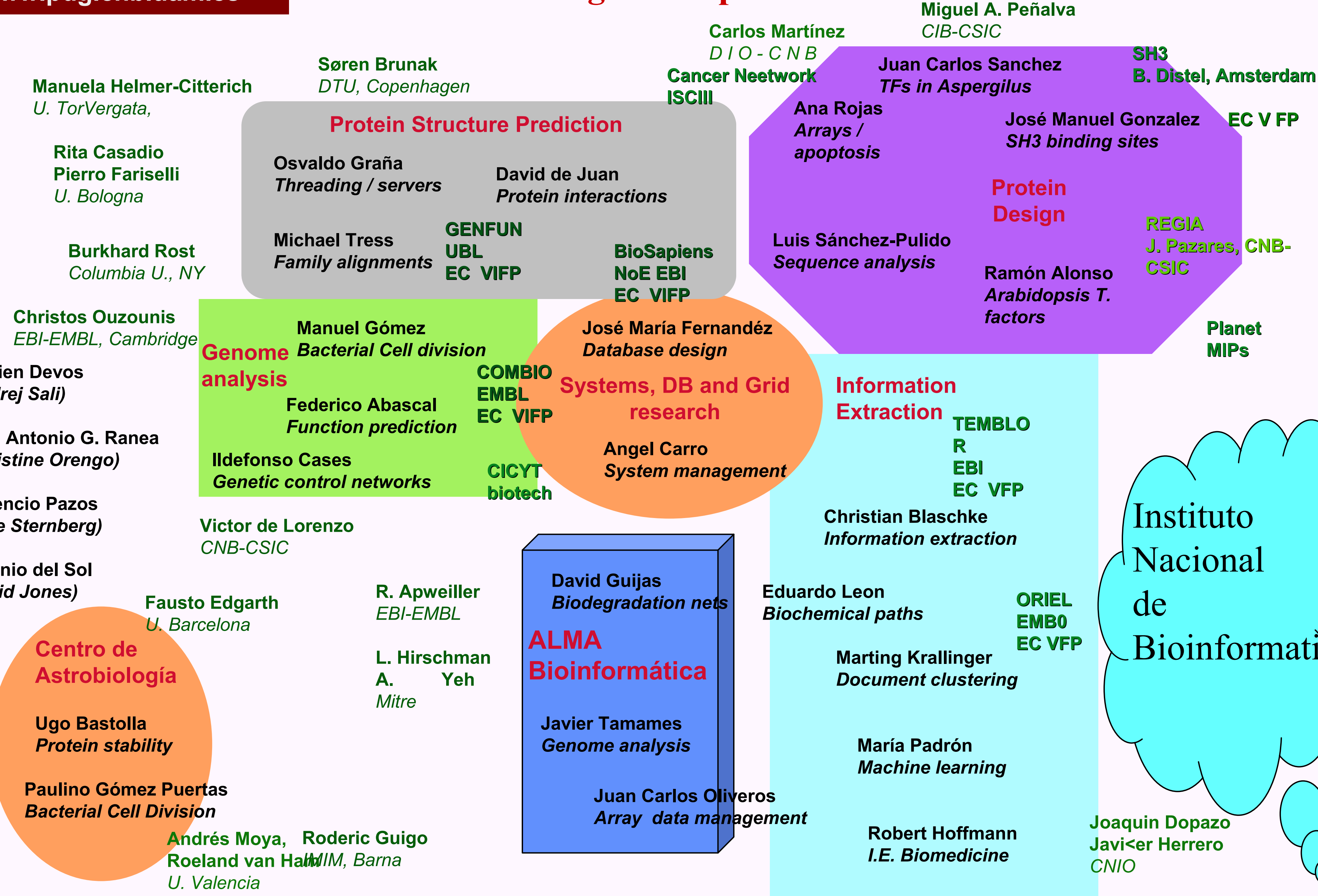
PMID	Sentence	Score
008244379	(Glover TW) - "Statistical analysis of the combined data suggests that the order of markers in the BRCA1 region is cen-THRA1-TOP2-GAS-OF2-17HSD-248yg9-RNU 2-OF3-PPY/p131-EPB3-Mfd188- WNT3-HOX2-GP3A-tel. "	009.72771
010066792	(Pommier Y) - "In addition to resistance to the fluoroquinolone CP-115,953, top2(S740W) induced novel DNA cleavage sites in the presence of VP-16, azatoxin, amsacrine, and mitoxantrone. "	008.67451
001322791	(Wang JC) - "AMSACRINE AND ETOPOSIDE HYPERSENSITIVITY OF YEAST CELLS OVEREXPRESSING DNA TOPOISOMERASE II. "	008.44646
008395511	(Nitiss JL) - "THE TOP2-5 MUTANT OF YEAST TOPOISOMERASE II ENCODES AN ENZYME RESISTANT TO ETOPOSIDE AND AMSACRINE. "	008.33303
007757979	(Andoh T) - "Bisdioxopiperazines such as ICRF-159 and ICRF-193 have been shown to inhibit DNA topoisomerase II. "	008.23282
007657608	(Nitiss JL) - "In prokaryotic type II topoisomerases (DNA gyrases), mutations that result in resistance to quinolones frequently occur at Ser83 or Ser84 of the gyrA subunit. "	008.23140
001320012	(Nitiss JL) - "The quinolone CP-115,953 (6,8-difluoro-7-(4-hydroxyphenyl)-1-cyclopropyl-4- quinolone-3-carboxylic acid) represents a novel mechanistic class of drugs with potent activity against eukaryotic topoisomerase II in vitro (Robinson, M. "	008.17069



0008834798	respectively, display disorganized actin patches in all cells. cdc12 and cdc15 mutants display disorganized actin patches during mitosis, but normal interphase actin patterns. cdc4 and rng2 mutants display disorganized actin cables during mitosis, but normal interphase actin patterns. "	001.73396
0008682866	(Albright CF) - "Overexpression of byr4 inhibits cytokinesis, but cell cycle progression continues leading to multinucleate cells. "	001.73227
0007798319	(Yanagida M) - "BYPASSING ANAPHASE BY FISSION YEAST CUT9 MUTATION: REQUIREMENT OF CUT9+ TO INITIATE ANAPHASE. "	001.72445
0009490631	(Hagan IM) - "F-ACTIN DISTRIBUTION AND FUNCTION DURING SEXUAL DIFFERENTIATION IN SCHIZOSACCHAROMYCES POMBE. "	001.71589
0010353895	(Murray AW) - "Defects in microtubule polymerization, spindle pole body duplication, microtubule motors, and kinetochore components all activate the MAD-dependent checkpoint. "	001.69525
0006490749	(Jelke E) - "Unique contour views of delocalized septa were exposed by freeze-fracturing. "	001.66838
0008039497	(Simanis V) - "Overexpression of p120cdc7 causes cell cycle arrest; cells complete mitosis and then undergo multiple rounds of septum formation without cell cleavage. "	001.62911

ww.pdg.cnb.uam.es

Protein Design Group CNB-CSIC



Miguel A. Peñalva
CIB-CSIC

Carlos Martínez
D I O - C N B

Søren Brunak
DTU, Copenhagen

Manuela Helmer-Citterich
U. TorVergata,

Cancer Neetwork
ISCIII

SH3
B. Distel, Amsterdam

EC V FP

Rita Casadio
Pierro Fariselli
U. Bologna

Burkhard Rost
Columbia U., NY

Protein Structure Prediction

Osvaldo Graña
Threading / servers

David de Juan
Protein interactions

Michael Tress
Family alignments

GENFUN
UBL
EC VIFP

BioSapiens
NoE EBI
EC VIFP

Protein Design

Juan Carlos Sanchez
TFs in Aspergilus

Ana Rojas
Arrays / apoptosis

José Manuel Gonzalez
SH3 binding sites

Luis Sánchez-Pulido
Sequence analysis

Ramón Alonso
Arabidopsis T. factors

REGIA
J. Pazares, CNB-CSIC

Planet
MIPs

Christos Ouzounis
EBI-EMBL, Cambridge

ien Devos
rej Sali)

Antonio G. Ranea
istine Orengo)

encio Pazos
e Sternberg)

nio del Sol
id Jones)

Genome analysis

Manuel Gómez
Bacterial Cell division

Federico Abascal
Function prediction

Ildefonso Cases
Genetic control networks

COMBIO
EMBL
EC VIFP

CICYT
biotech

Systems, DB and Grid research

José María Fernandéz
Database design

Angel Carro
System management

Information Extraction

TEMBLO
R
EBI
EC VFP

Christian Blaschke
Information extraction

Eduardo Leon
Biochemical paths

ORIEL
EMB0
EC VFP

Marting Krallinger
Document clustering

María Padrón
Machine learning

Robert Hoffmann
I.E. Biomedicine

Instituto Nacional de Bioinformati

Victor de Lorenzo
CNB-CSIC

Fausto Edgarth
U. Barcelona

R. Apweiller
EBI-EMBL

L. Hirschman
A. Yeh
Mitre

ALMA Bioinformática

David Guijas
Biodegradation nets

Javier Tamames
Genome analysis

Juan Carlos Oliveros
Array data management

Centro de Astrobiología

Ugo Bastolla
Protein stability

Paulino Gómez Puertas
Bacterial Cell Division

Andrés Moya,
Roeland van Halm
U. Valencia

Roderic Guigo
IMIM, Barna

Joaquin Dopazo
Javi<er Herrero
CNIO

www.pdg.cnb.uam.es

BO workshop: A critical assessment of text mining methods in
ecular biology, Granada 28. Mar - 01. April
v.pdg.cnb.uam.es/BioLINK/workshop_BioCreative_04/

n ISMB / 3rd ECCB Conference: Text mining and Genome function
diction SIGs,   Glasgow, July 31st-August 4th
://www.iscb.org/ismbeccb2004/

ona ESF workshop: Molecular Interactions: New frontiers for
nputational methods,   Verona, July 3-8
v.functionalgenomics.org.uk

rnational Joint Workshop on Natural Language Processing in
medicine and its Applications 2004,   28-29 August Switzerland
//www.genisis.ch/~natlang/JNLPBA04/

www.almabioinfo.c