DAS related developments

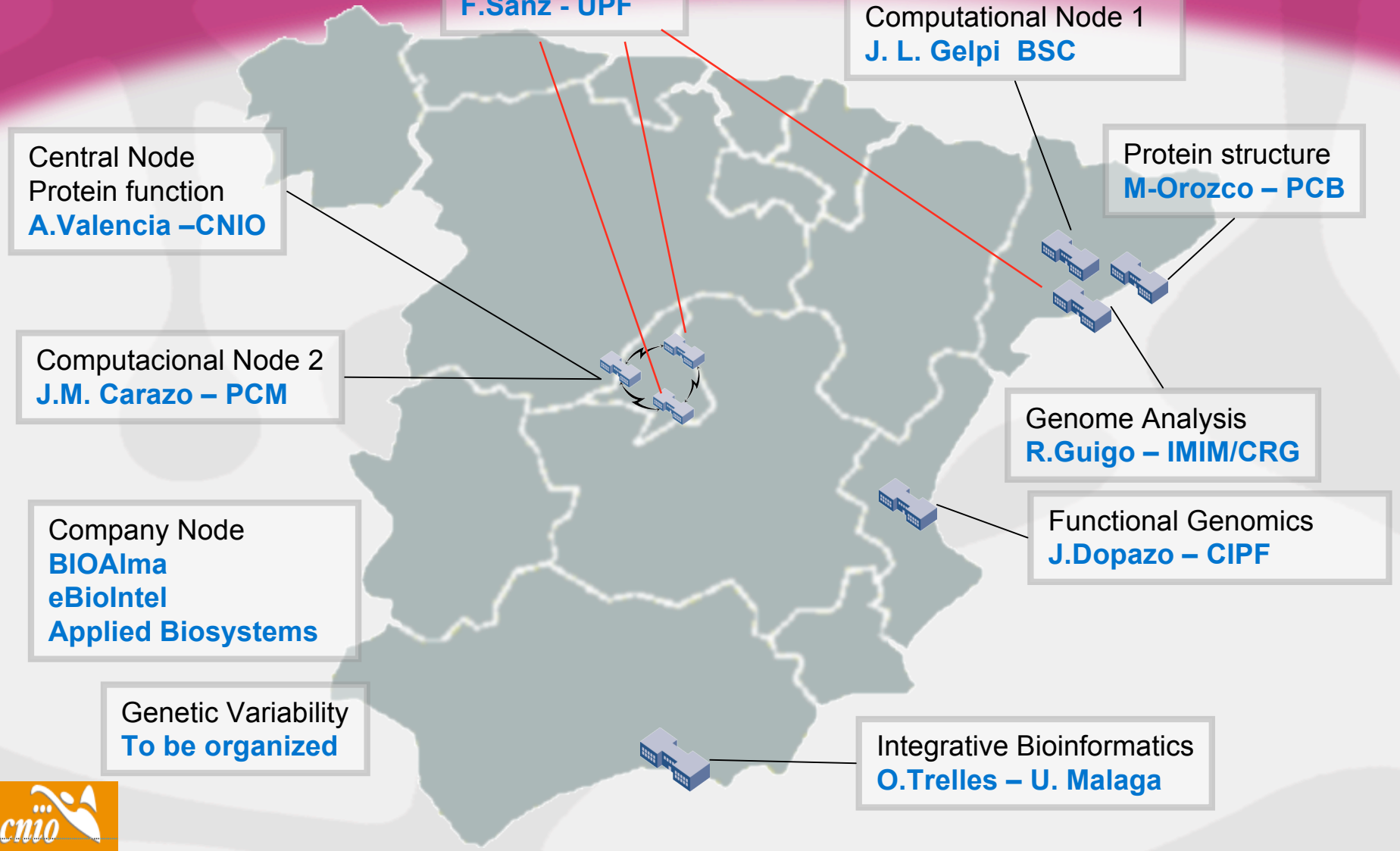
Alfonso Valencia
Structural and Computational Biology Programme
Spanish National Cancer Research Centre
CNIO, Madrid

Sanger
Feb. 2007

A Virtual Institute

INB

INSTITUTO NACIONAL DE BIOINFORMÁTICA



cnio

Alfonso Valencia

INB

INSTITUTO NACIONAL DE BIOINFORMÁTICA

INB initial resource structure

BioMoby

DAS

Instituto Nacional de Bioinformatica

Horizontal
BSC / PCM
J L Gelpi / JM Carazo
Computational resources

BioMoby Server

Users

Genomics Platforms
Genomics Projects

Databases

Software

Gene Finding

Function Prediction

DNA arrays

3Ds

Vertical 1
R Guigo
IMIM
CRG

Vertical 2
A Valencia
CNIO

Vertical 3
J Dopazo
CIPF

Vertical 4
M Orozco
PCB

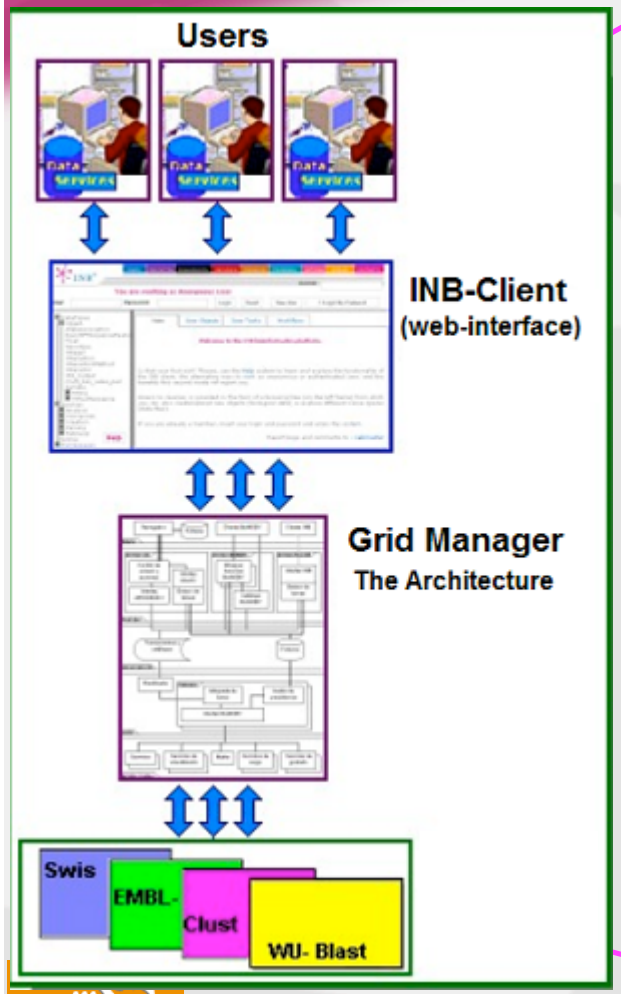
Vertical 5
O Trelles
U. Malaga

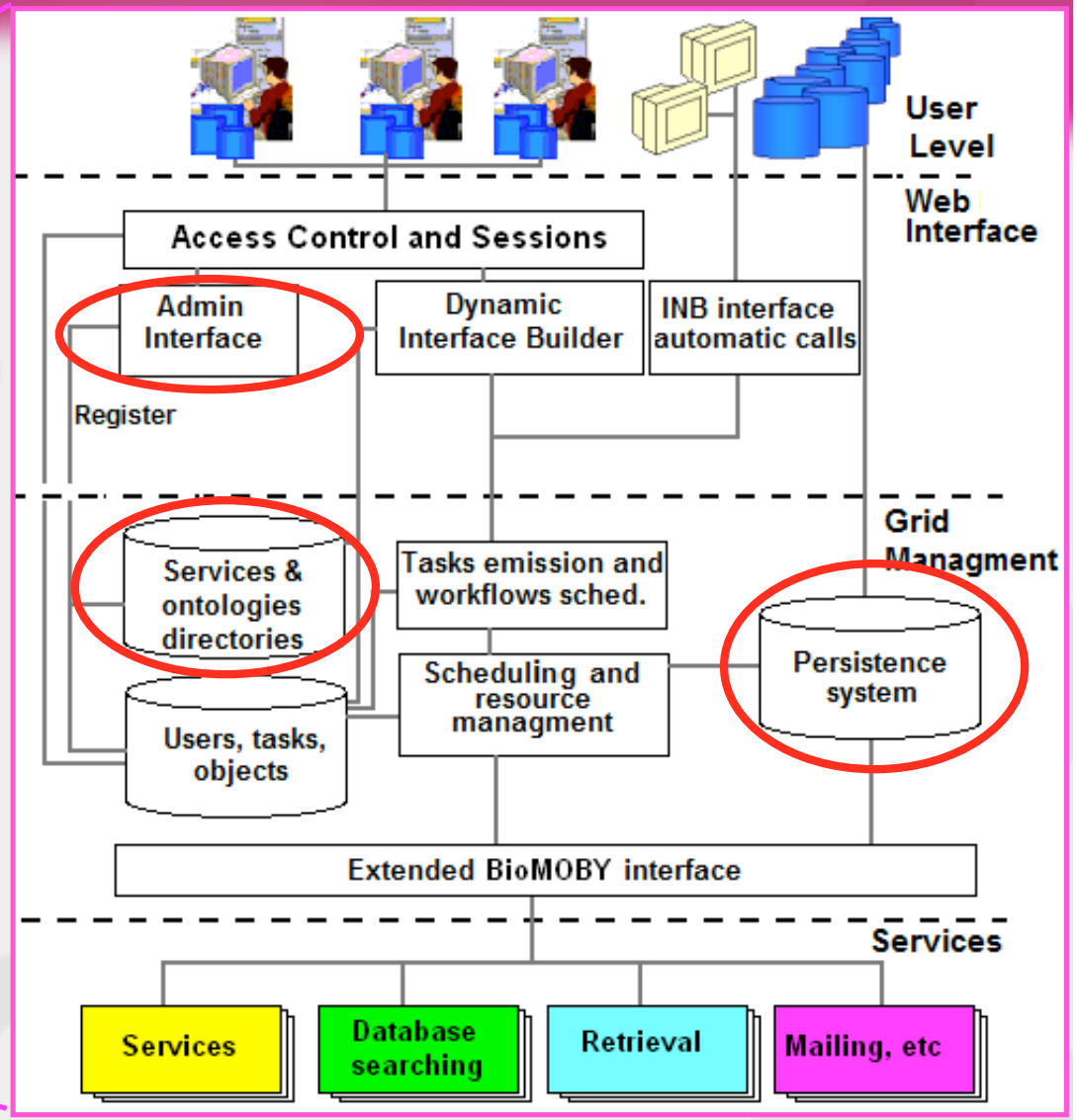
cnio

Alfonso Valencia

INB Moby architecture and Client.
> than 200 webservices under control

INB
INSTITUTO NACIONAL DE BIOINFORMÁTICA





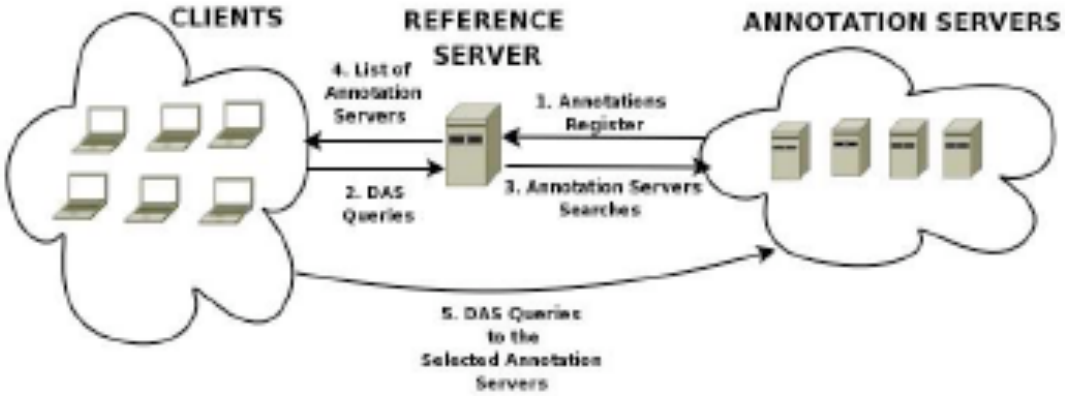
cnio
Alfonso Valencia

Oswaldo Trelles, UMalaga

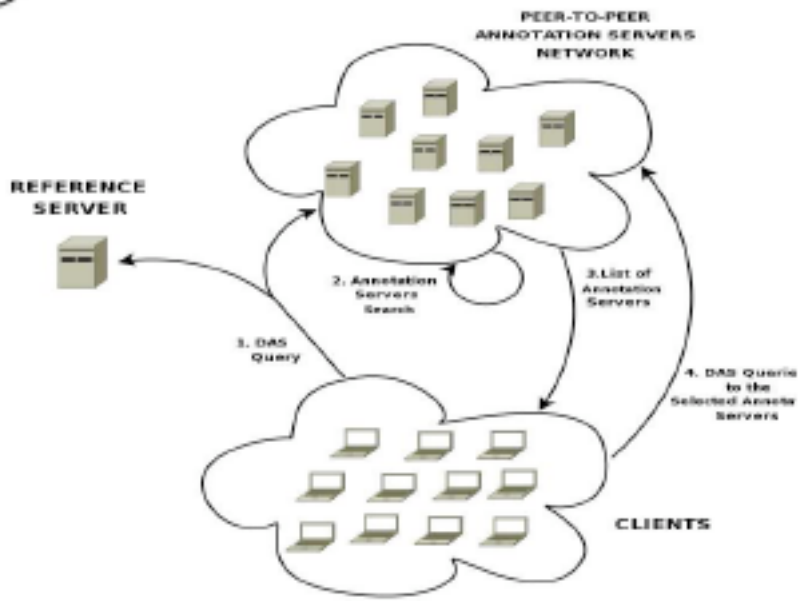
PEER-TO-PEER SOLUTIONS ON BIOINFORMATICS APPLICATIONS

C. Guerrero, R. Cuevas, I. Martinez, A. Azcorra
Universidad Carlos III Madrid

" …. possible advantages derived from the usage of peer-to-peer (p2p) technologies on DAS environments"



Current versus mix model

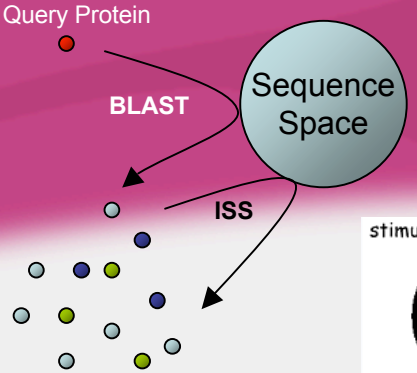


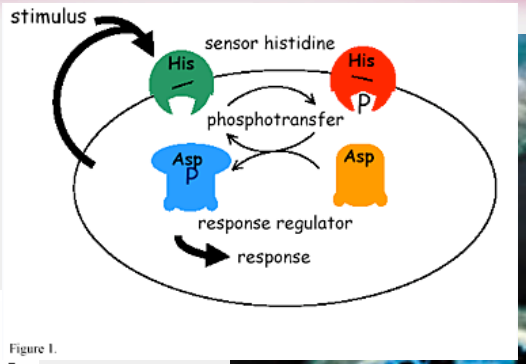
INB

INSTITUTO NACIONAL DE BIOINFORMÁTICA

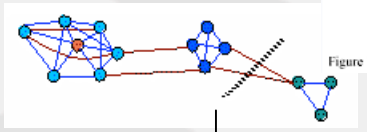
Two component system
Streptomyces coelicolor
JF Martin (NV2)

A. Homology Search

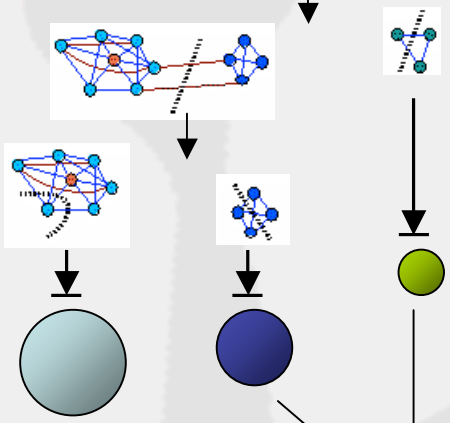




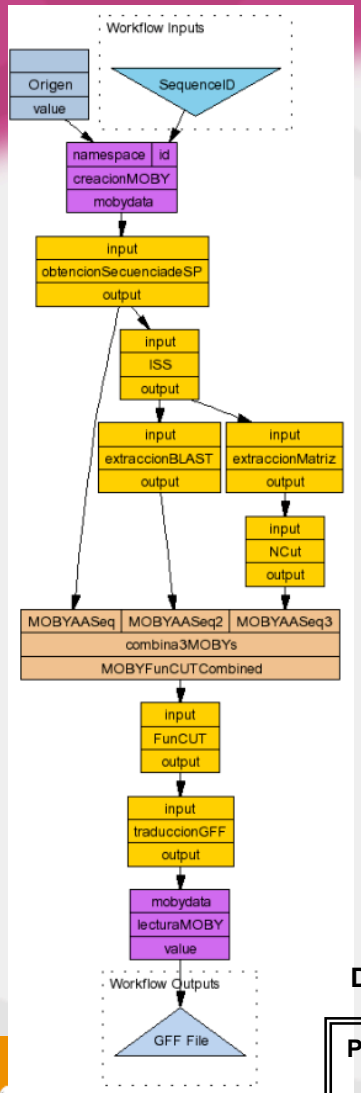
B. Sequence Space Mapping

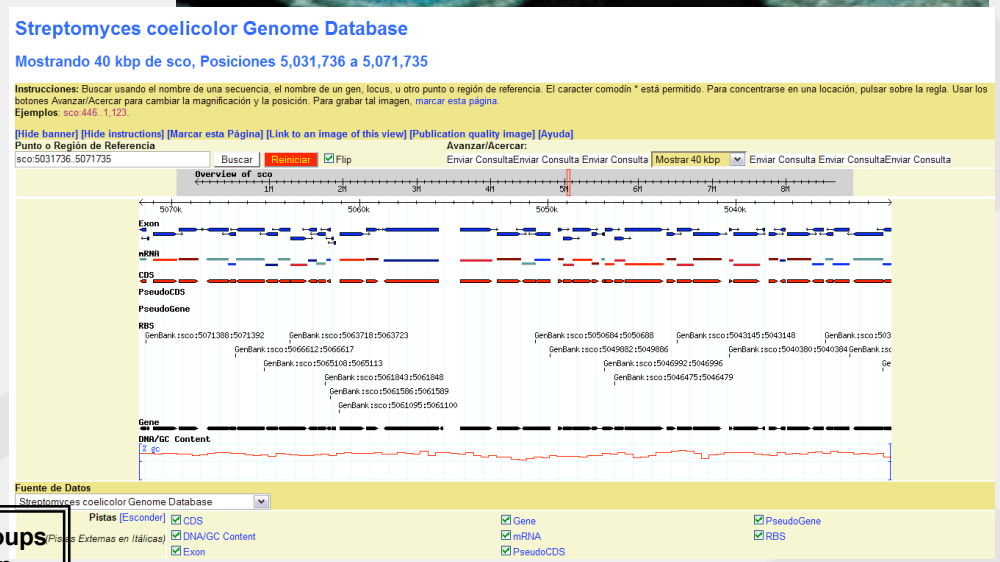


C. Automatic identification of groups. Clustering.



D. Annotation





Problem Protein Annotation

Neighbor Groups Annotation

Alfonso Valencia

CNB-CSIC / U. Leon

by Osvaldo Graña

Streptomyces coelicolor computational workflow

SCO genome, 7825

INB

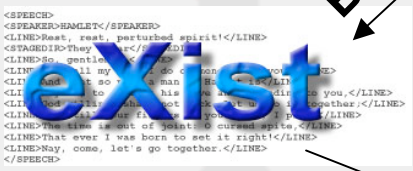
FUNcut functional annotation

INSTITUTO NACIONAL DE B

3,5 weeks later

Blast homologs

XML db



STRING
Interactions

DAS server

Sco.org.uk
Manual annotations

GBrowse

GenBank

INTERNET

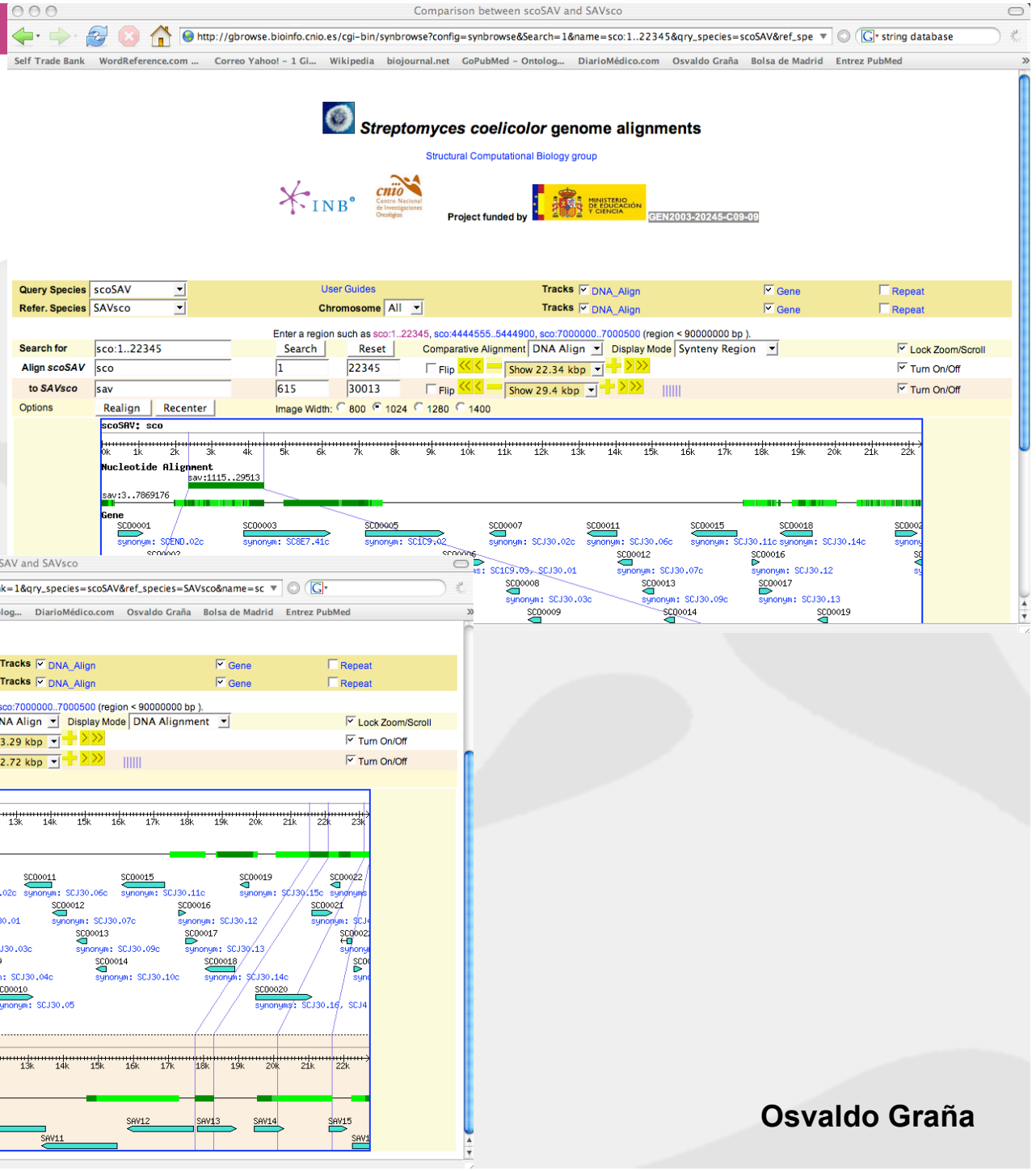
S. coelicolor GeneDB
Physicochemical properties

Pfam
Domains

Metabolic pathways
KEGG Kyoto Encyclopedia of Genes and Genomes
ScoCyc

Osvaldo Graña

SynBrowse alignment viewer



Osvaldo Graña

Hoffmann Valencia Nat Genet 2004

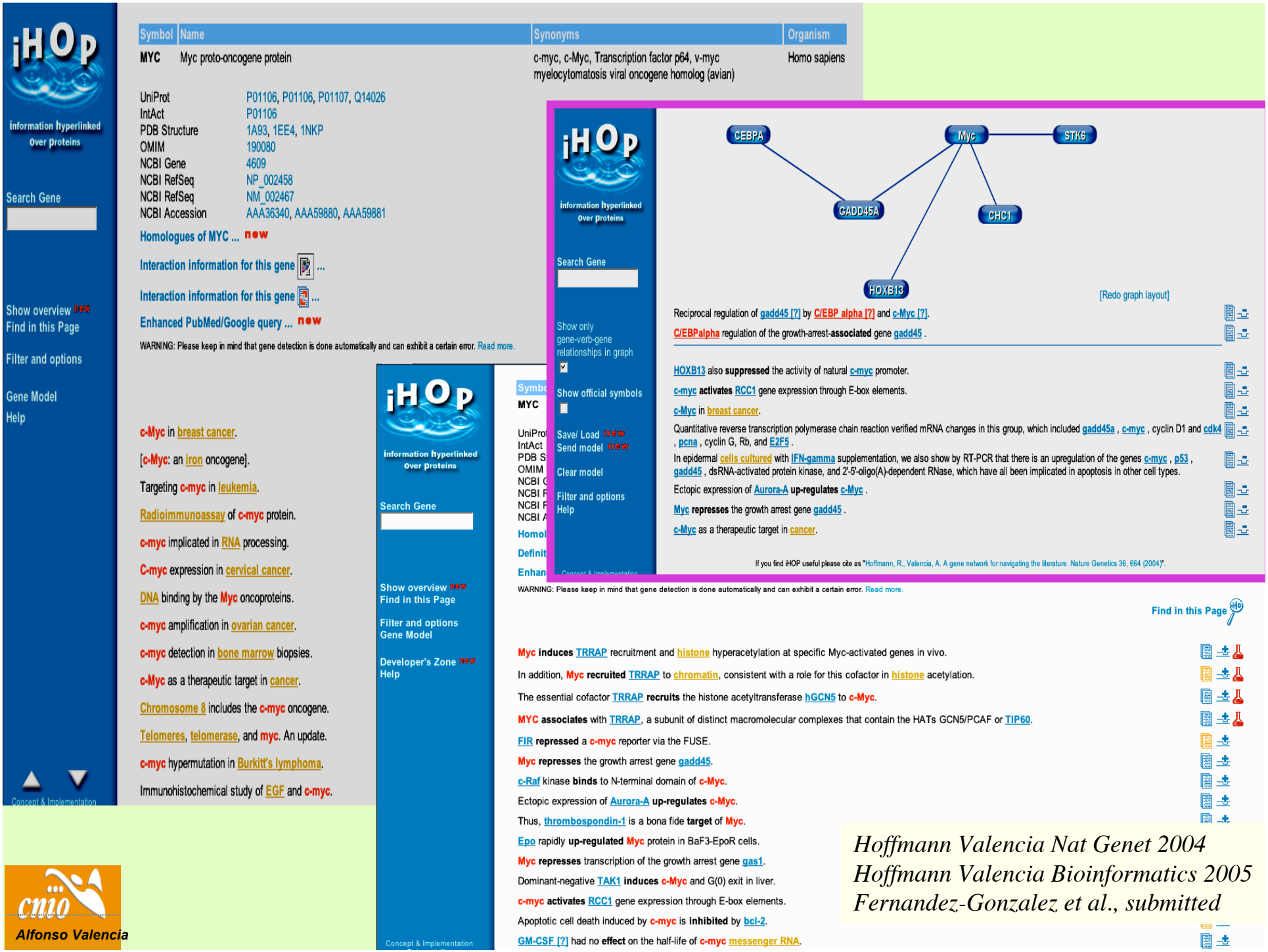
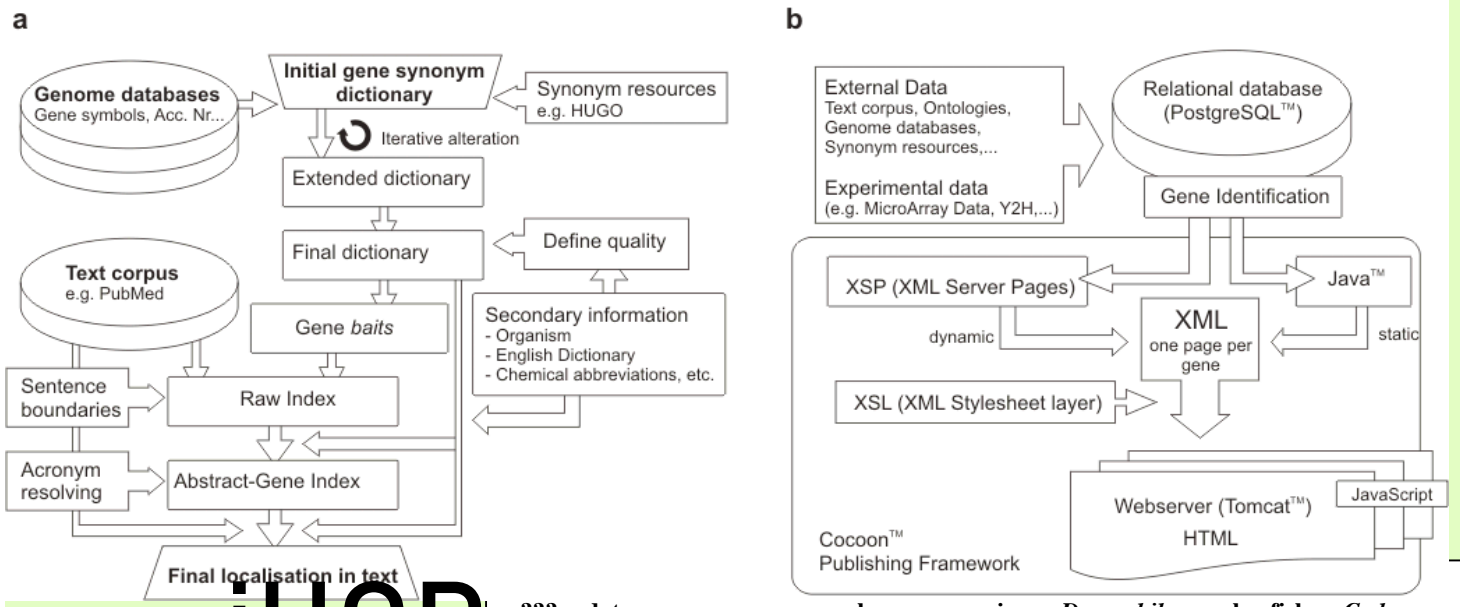

iHOp
information hyperlinked over proteins
Search Gene
Show overview new
Find in this Page
Filter and options
Gene Model
Help
Concept & Implementation
Symbol
Name
Synonyms
Organism
MYC
Myc proto-oncogene protein
c-myc, c-Myc, Transcription factor p64, v-myc myelocytomatosis viral oncogene homolog (avian)
Homo sapiens
UniProt P01106, P01106, P01107, Q14026
IntAct P01106
PDB Structure 1A93, 1EE4, 1NKP
OMIM 190080
NCBI Gene 4609
NCBI RefSeq NP_002458
NCBI RefSeq NM_002467
NCBI Accession AAA36340, AAA59880, AAA59881
Homologues of MYC ... new
Interaction information for this gene ...
Interaction information for this gene ...
Enhanced PubMed/Google query ... new
WARNING: Please keep in mind that gene detection is done automatically and can exhibit a certain error. Read more.
c-Myc in breast cancer.
[c-Myc: an iron oncogene].
Targeting c-myc in leukemia.
Radioimmunoassay of c-myc protein.
c-myc implicated in RNA processing.
C-myc expression in cervical cancer.
DNA binding by the Myc oncoproteins.
c-myc amplification in ovarian cancer.
c-myc detection in bone marrow biopsies.
c-Myc as a therapeutic target in cancer.
Chromosome 8 includes the c-myc oncogene.
Telomeres, telomerase, and myc. An update.
c-myc hypermutation in Burkitt's lymphoma.
Immunohistochemical study of EGF and c-myc.
Show only gene-verb-gene relationships in graph
Show official symbols
Save/ Load new
Send model new
Clear model
Filter and options
Help
CEBPA
Myc
STK6
GADD45A
CHC1
HOXB13
[Redo graph layout]
Reciprocal regulation of gadd45 [?] by C/EBP alpha [?] and c-Myc [?].
C/EBPalpha regulation of the growth-arrest-associated gene gadd45 .
HOXB13 also suppressed the activity of natural c-myc promoter.
c-myc activates RCC1 gene expression through E-box elements.
c-Myc in breast cancer.
Quantitative reverse transcription polymerase chain reaction verified mRNA changes in this group, which included gadd45a , c-myc , cyclin D1 and cdk4 , pcna , cyclin G, Rb, and E2F5 .
In epidermal cells cultured with IFN-gamma supplementation, we also show by RT-PCR that there is an upregulation of the genes c-myc , p53 , gadd45 , dsRNA-activated protein kinase, and 2'-5'-oligo(A)-dependent RNase, which have all been implicated in apoptosis in other cell types.
Ectopic expression of Aurora-A up-regulates c-Myc .
Myc represses the growth arrest gene gadd45 .
c-Myc as a therapeutic target in cancer.
If you find iHOP useful please cite as "Hoffmann, R., Valencia, A. A gene network for navigating the literature. Nature Genetics 36, 664 (2004)".
Show overview new
Find in this Page
Filter and options
Gene Model
Developer's Zone new
Help
Concept & Implementation
MYC
WARNING: Please keep in mind that gene detection is done automatically and can exhibit a certain error. Read more.
Find in this Page
Myc induces TRRAP recruitment and histone hyperacetylation at specific Myc-activated genes in vivo.
In addition, Myc recruited TRRAP to chromatin, consistent with a role for this cofactor in histone acetylation.
The essential cofactor TRRAP recruits the histone acetyltransferase hGCN5 to c-Myc.
MYC associates with TRRAP, a subunit of distinct macromolecular complexes that contain the HATs GCN5/PCAF or TIP60.
FIR repressed a c-myc reporter via the FUSE.
Myc represses the growth arrest gene gadd45.
c-Raf kinase binds to N-terminal domain of c-Myc.
Ectopic expression of Aurora-A up-regulates c-Myc.
Thus, thrombospondin-1 is a bona fide target of Myc.
Epo rapidly up-regulated Myc protein in BaF3-EpoR cells.
Myc represses transcription of the growth arrest gene gas1.
Dominant-negative TAK1 induces c-Myc and G(0) exit in liver.
c-myc activates RCC1 gene expression through E-box elements.
Apoptotic cell death induced by c-myc is inhibited by bcl-2.
GM-CSF [?] had no effect on the half-life of c-myc messenger RNA.
cnio
Alfonso Valencia
Hoffmann Valencia Nat Genet 2004
Hoffmann Valencia Bioinformatics 2005
Fernandez-Gonzalez et al., submitted

Figure 4

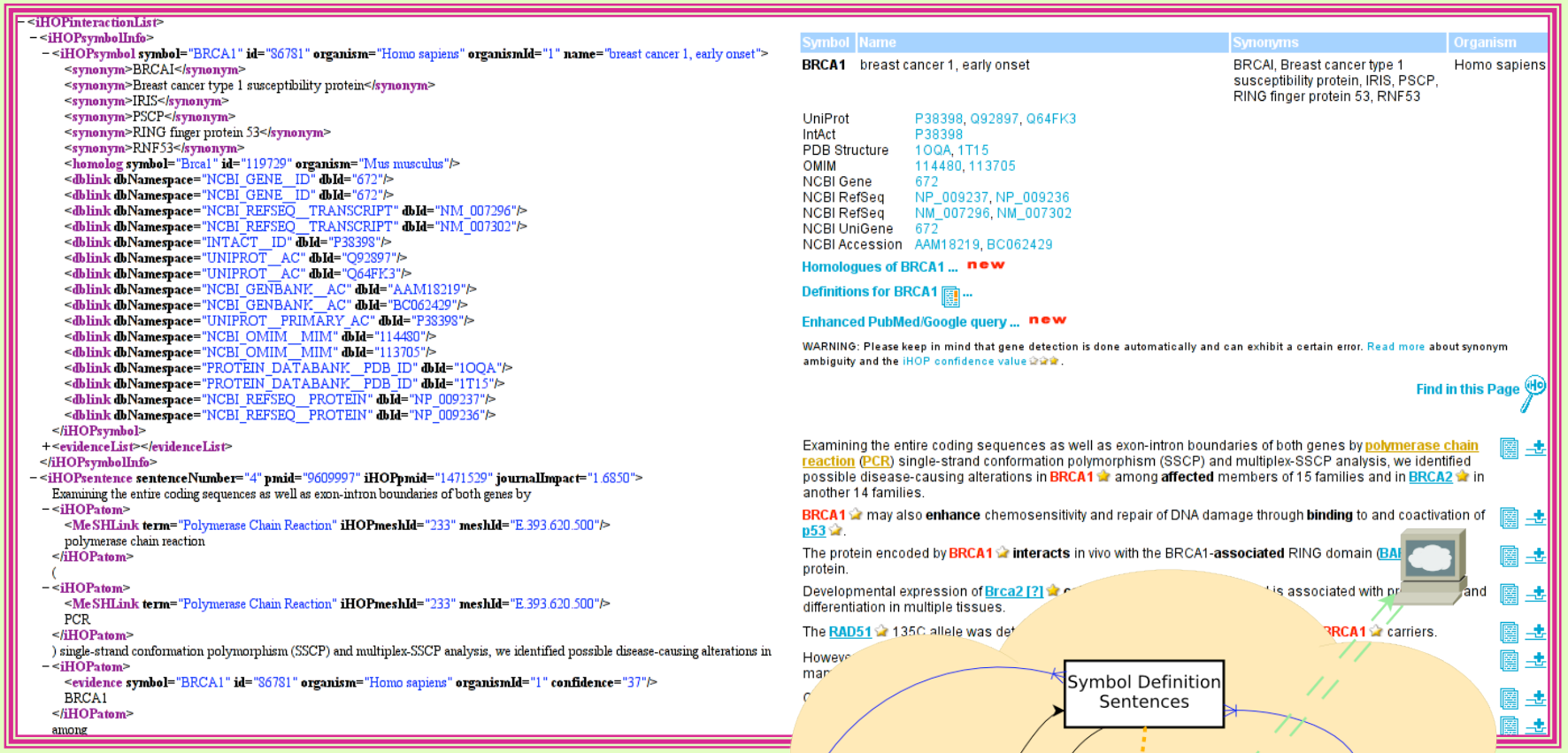


iHOP inside

Hoffmann Valencia
Bioinform. 2005

???update	humans	mice	*Drosophila*	zebrafish	*C.elegans*	*Arabidopsis*	yeast	*E.coli*	Average
Relevant docs (RD) (*Goldstd.* LocusLink)	30760	7357	12711	119	170	-	-	-	10223.4
RD: Recall (%)	84.3	76.6	79.6	74	95.3				81.96
Nr of gene-article references (GA) (*Goldstd.* LocusLink)	36816	8471	27994	147	279	-	-	-	14741.4
GA: Recall	26327	5639	14495	105	251	-	-	-	9363.4
call (%)	71.5	66.6	51.8	71.4	90	-	-	-	70.26
ene localisations (GL) (*Goldstd.* manual)	403	381	438	360	597	295	536	476	435.75
GL: True positives (*Goldstd.* manual)	351	354	409	352	584	271	533	442	412
GL: Precision, exact localisation (%)	87.1	92.9	93.4	97.8	97.8	91.9	99.4	92.9	94.15
F-measure (%)	78.5	77.6	66.6	82.5	93.7	-	-	-	80

cnio
Alfonso Valencia

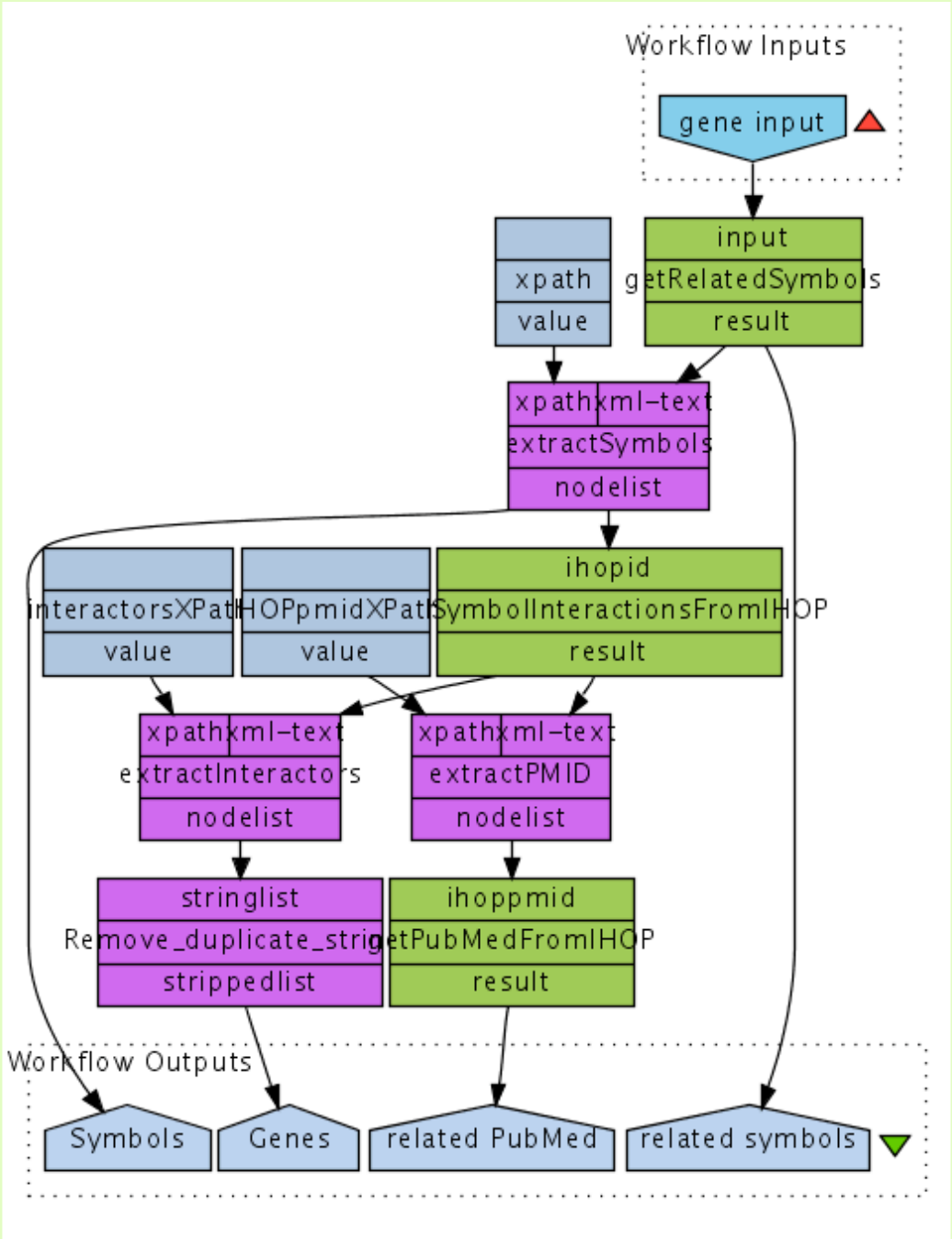


iHOP web services

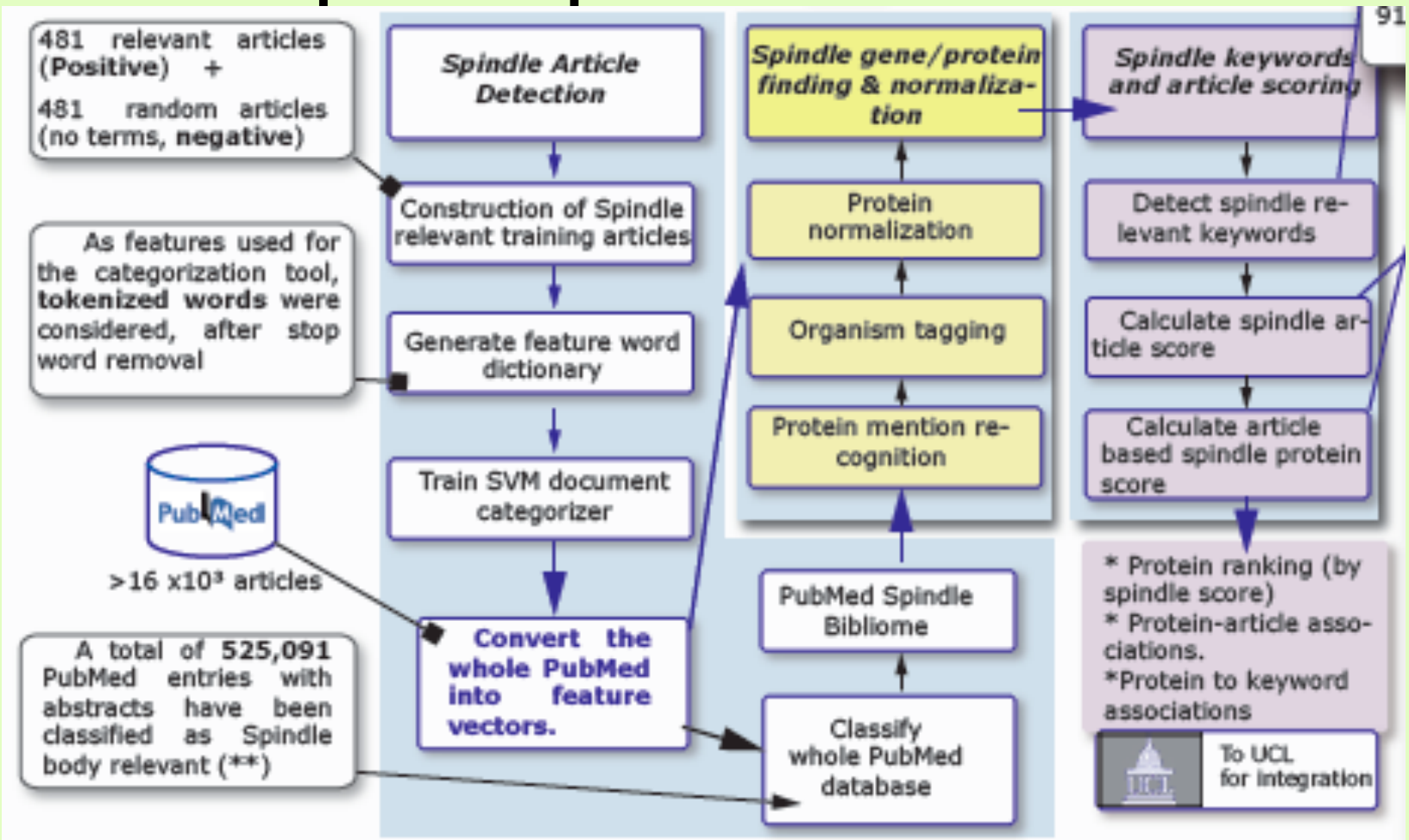
http://www. ihop-net.org/UniPub/iHOP/webservices/.

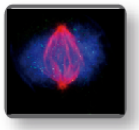
cnio
Alfonso Valencia

JM Fernández, R Hoffmann


Workflow Inputs
gene input
input
getRelatedSymbols
result
xpath
value
xpathxml-text
extractSymbols
nodelist
ihopid
SymbolInteractionsFromIHOP
result
interactorsXPath
value
IHOPpmidXPath
value
xpathxml-text
extractInteractors
nodelist
xpathxml-text
extractPMID
nodelist
stringlist
Remove_duplicate_strings
strippedlist
ihoppmid
getPubMedFromIHOP
result
Workflow Outputs
Symbols
Genes
related PubMed
related symbols

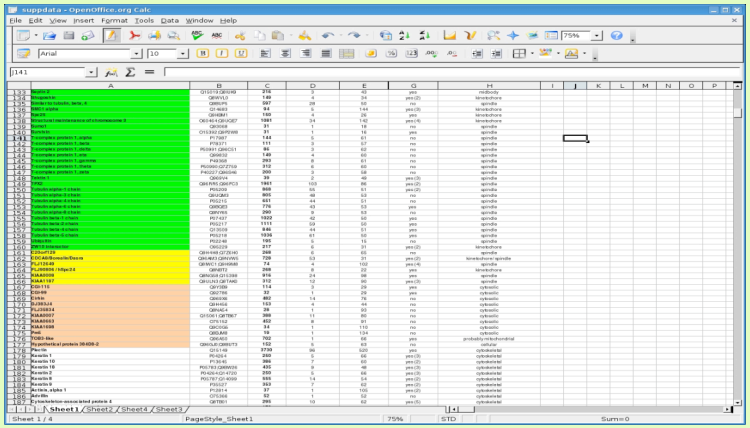
Spindle protein classification



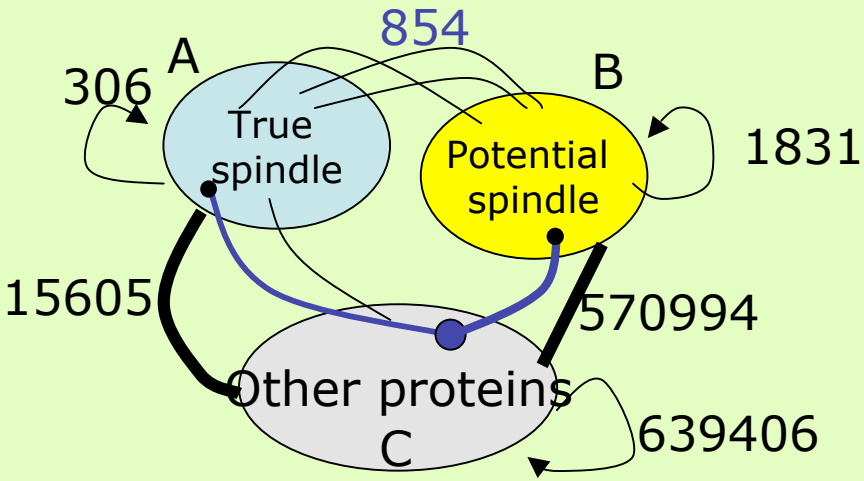


Sauer et al. Proteome analyses of the human mitotic spindle. *Mol Cell Proteomics.(2005)*

Original data



iHOP interactions: *interactions extracted From literature*



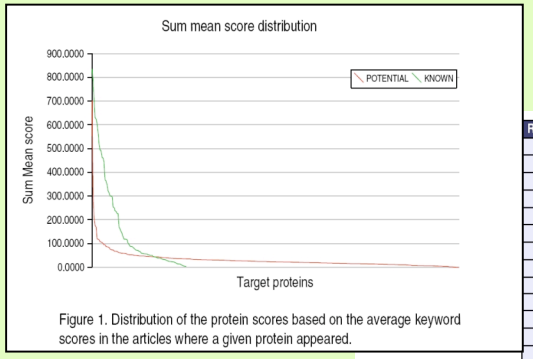
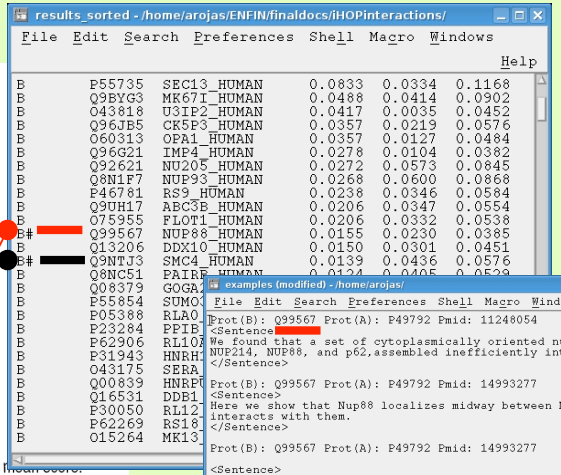
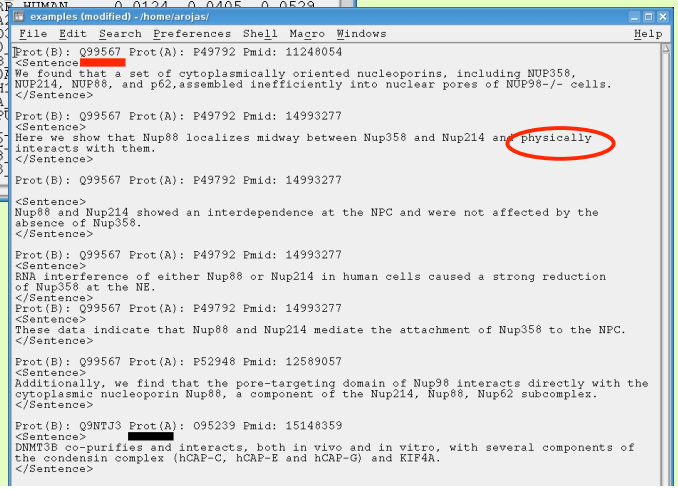


Figure 1. Distribution of the protein scores based on the average keyword scores in the articles where a given protein appeared.

RANK	SWISSPROT_ID	SCORE	REVISION
1	GCP6_HUMAN	701.74	Y
2	CASC5_HUMAN	312.84	Y
3	DLG7_HUMAN	209.03	Y
4	NOC3L_HUMAN	177.47	Y
5	SMC4_HUMAN	169.77	Y
6	CEBPZ_HUMAN	121.06	Y
7	FRAP_HUMAN	117.92	Y
8	MGN_HUMAN	115.46	Y
9	SMC2_HUMAN	107.45	Y
10	RNPC2_HUMAN	106.27	Y
11	IPO7_HUMAN	103.91	Y
12	CTRO_HUMAN	100.32	Y
13	LAP2A_HUMAN	94.15	Y
14	LAP2B_HUMAN	94.15	Y
15	MPP10_HUMAN	87.21	Y
16	PWP1_HUMAN	85.55	Y
17	PRP19_HUMAN	85.23	Y
18	NOG1_HUMAN	83.58	U
19	NUP88_HUMAN	76.51	Y
20	SAM68_HUMAN	74.83	Y

Table 1. Ranked list of the top scoring potential spindle proteins based on the sum mean score. The last column corresponds to the manual classification of the top ranking proteins according to a revision of the corresponding SwissProt records. Y: records provides positive evidence in favor of having an association to spindle proteins, U: neither positive nor negative evidence found, N: negative evidence found of not being a spindle related protein.





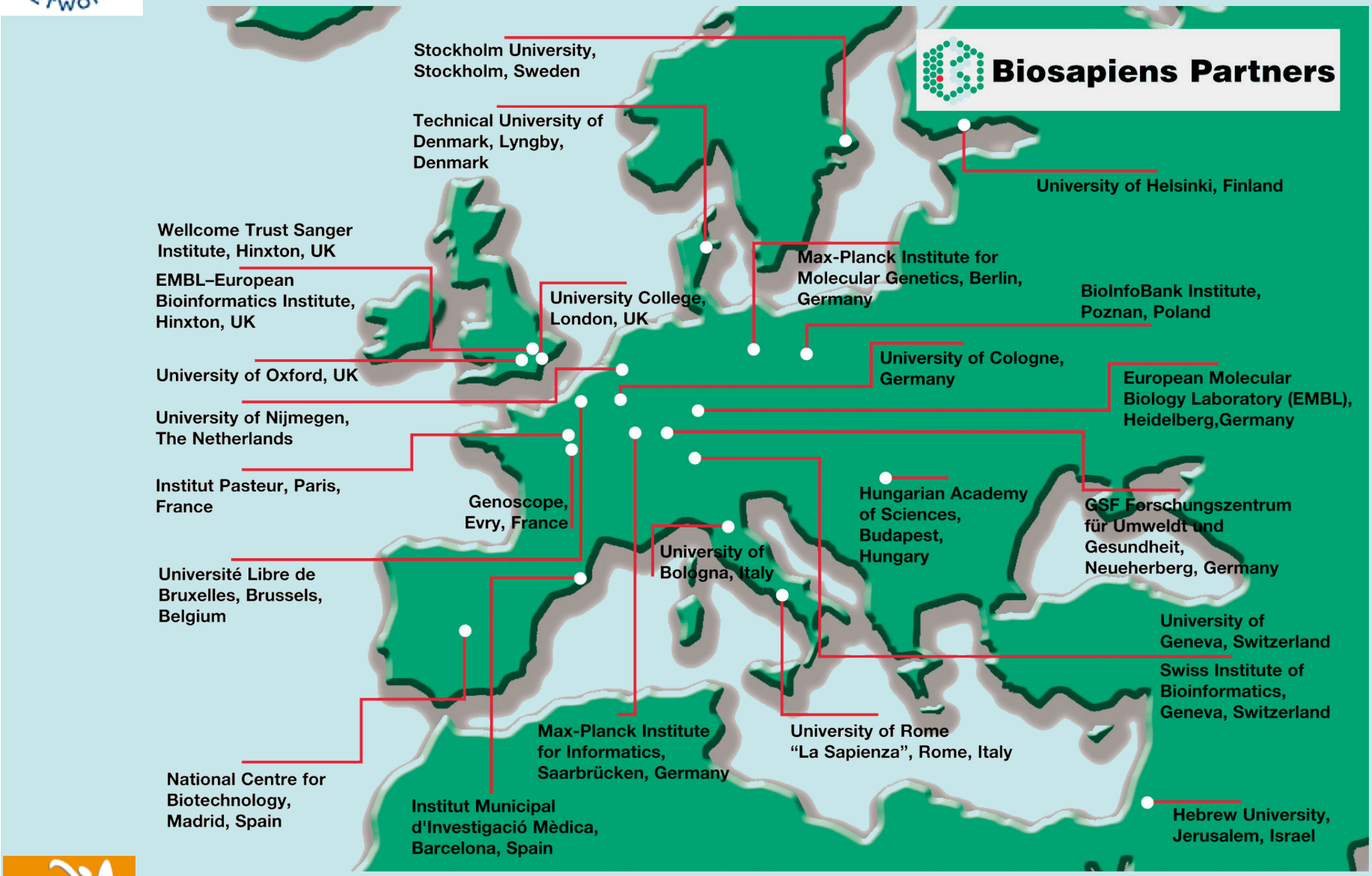
SVM scoring based on Keyword scoring

By Martin Krallinger

By Edu Leon
David de Juan
Jose M Fernandez

cnio
Alfonso Valencia

BioSapiens



Alfonso Valencia

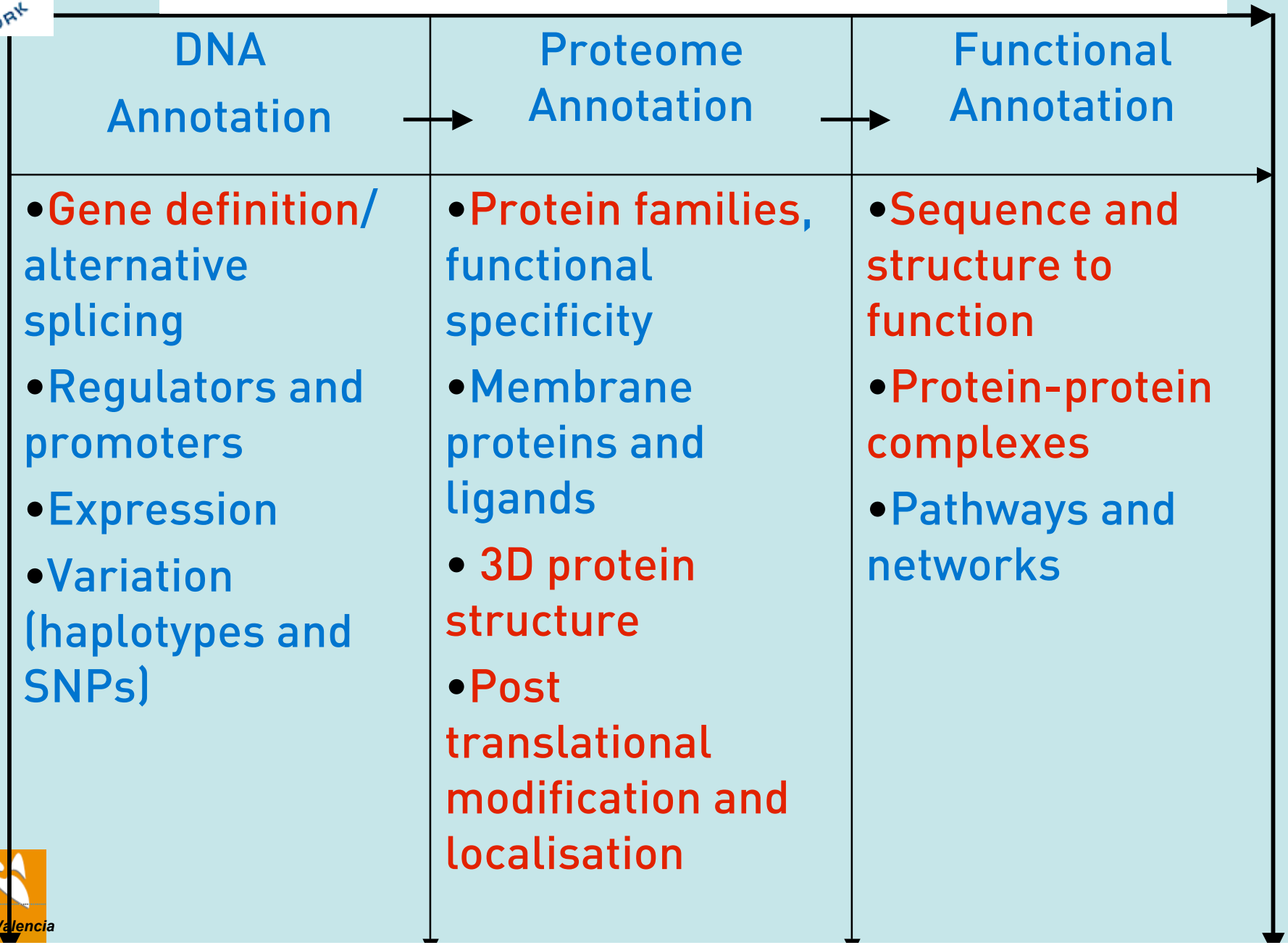
BioSapiens

European Bioinformatics Institute – Janet Thornton

University of Rome “La Sapienza” - Anna Tramontano
Consejo Superior de Investigaciones Científicas - Alfonso Valencia
Technical University of Denmark - Søren Brunak

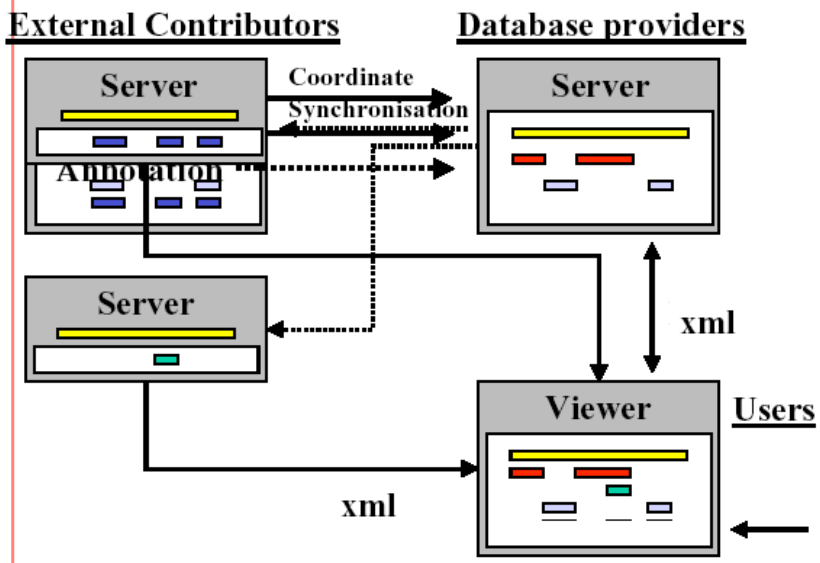
European Molecular Biology Laboratory – Peer Bork
German National Centre for Health and Environment – Dmitrij Frishman
Université Libre de Bruxelles - Shoshona Wodak
Institut Municipal d'Investigació Mèdica - Roderic Guigó
The Wellcome Trust Sanger Institute - Tim Hubbard
Max-Planck Institute for Informatics - Thomas Lengauer
The Hebrew University of Jerusalem - Michal Linial
Stockholms Universitet - Gunnar von Heijne
University of Oxford - Richard Mott
University College London - Christine Orengo
Stichting Katholieke Universiteit - Gert Vriend
Swiss Institute of Bioinformatics - Anne-Lise Veuthey
University of Helsinki - Esko Ukkonen
University of Geneva - Stylianos Antonarakis
Institute of Enzymology (Hungarian Academy of Sciences) - László Patthy
Universität zu Köln - Dietmar Schomburg
Institut Pasteur - Antoine Danchin
BioInfoBank Institute - Leszek Rychlewski
Max-Planck Institute for Molecular Genetics – Martin Vingron
Genoscope – Vincent Shachter
University of Bologna – Rita Casadio

A European Virtual Institute for Genome Annotation



Alfonso Valencia

Distributed Annotation



Why DAS in BioSapiens?

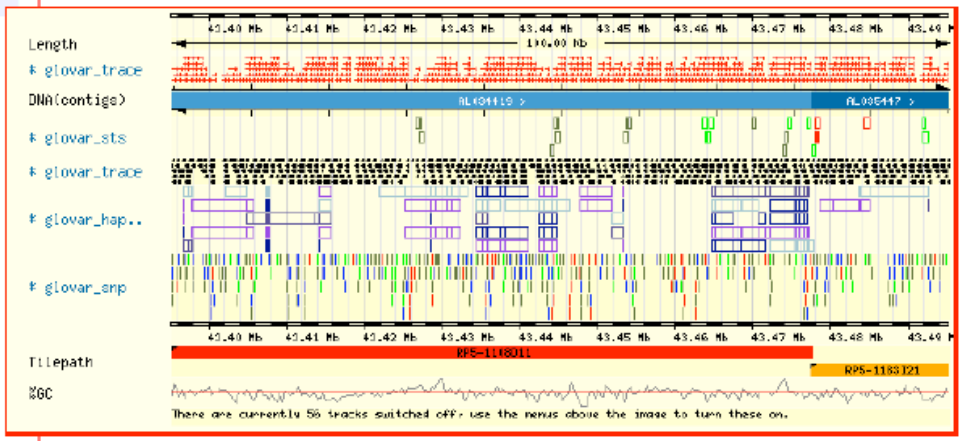
- Positives:
- Established and working now
- Simple
- Good fit with many annotation sites (you) and fewer display/viewing sites
- No need for highly centralised control

- Negatives:
- Not semantically rich
  - Can extend with DAS/2 or other mechanisms

DAS extensions

- Protein DAS
  - Provide features in protein coordinates
  - Can handle Uniprot/RefSeq coordinates rather than Ensembl coordinates
- GeneDAS
  - Provide text on geneview pages

Genomic DAS

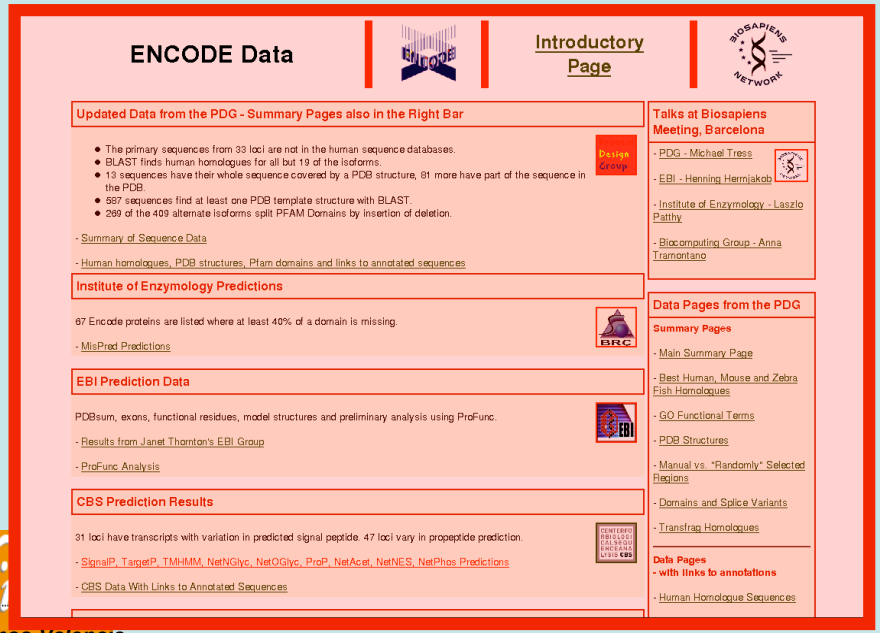


By Ewan Birney

BioSapiens

The BioSapiens-sponsored project concentrated on the protein coding loci and in particular on the alternatively spliced products.

This work is part of the BioSapiens efforts for the annotation of the human genome (www.biosapiens.info).



ENCODE Data

Introductory Page

Updated Data from the PDG - Summary Pages also in the Right Bar

- The primary sequences from 33 loci are not in the human sequence databases.
- BLAST finds human homologues for all but 19 of the isoforms.
- 13 sequences have their whole sequence covered by a PDB structure, 61 more have part of the sequence in the PDB.
- 587 sequences find at least one PDB template structure with BLAST.
- 269 of the 409 alternate isoforms split PFAM Domains by insertion of deletion.

- Summary of Sequence Data
- Human homologues, PDB structures, Pfam domains and links to annotated sequences

Institute of Enzymology Predictions

67 Encode proteins are listed where at least 40% of a domain is missing.

- MisPred Predictions

EBI Prediction Data

PDBsum, exons, functional residues, model structures and preliminary analysis using ProFunc.

- Results from Janet Thornton's EBI Group
- ProFunc Analysis

CBS Prediction Results

31 loci have transcripts with variation in predicted signal peptide. 47 loci vary in propeptide prediction.

- SignalP, TargetP, TMHMM, NetNGlyc, NetOGlyc, ProP, NetAcet, NetNES, NetPhos Predictions
- CBS Data With Links to Annotated Sequences

Talks at Biosapiens Meeting, Barcelona

- PDG - Michael Tress
- EBI - Henning Hermjakob
- Institute of Enzymology - Laszlo Patthy
- Biocomputing Group - Anna Tramontano

Data Pages from the PDG

Summary Pages

- Main Summary Page
- Best Human, Mouse and Zebra Fish Homologues
- GO Functional Terms
- PDB Structures
- Manual vs. "Randomly" Selected Regions
- Domains and Splice Variants
- Transfrag Homologues

Data Pages - with links to annotations

- Human Homologue Sequences

Splice variants were annotated by the Havana Group as part of the GENCODE consortium.

This complete set of splice variants allows us to pose interesting questions, particularly whether the gene products would fold into functional proteins.

Funded by | Coordinated by

A European Virtual Institute for Genome Annotation

Structure Function Pipeline.

ProFunc | Protein Design Group | CATH Protein Structure Classification | STRING

From this page you can submit your own structure or analyse an existing PDB entry.

Enter Details:

Email Address:

NCBI Taxon ID: OR Species Name:

If you want to see the NCBI tax IDs click here

Select Option:

Upload PDB-format file: Browse...

or

Get existing PDB file* PDB code:

RUN

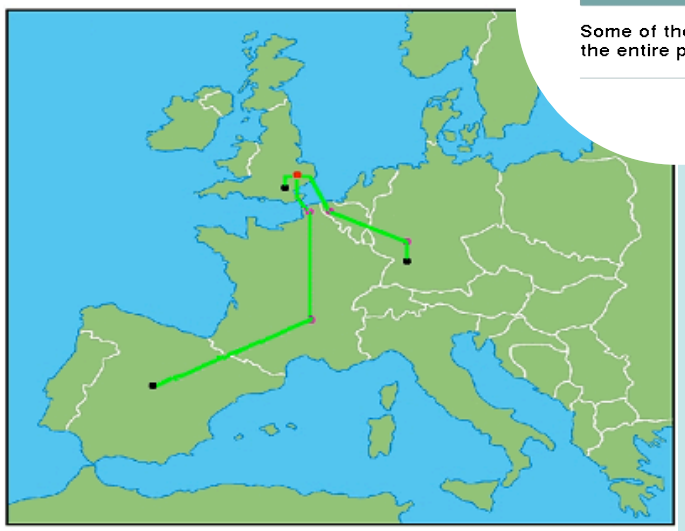
Some of the methods take minutes to run; others take hours. You will be notified by e-mail when the entire process is complete, but can check on preliminary results as they become available.

E-mail to:- *James Watson* watson@ebi.ac.uk

Coordinated by EBI | Funded by

Key to map:

QUEUED RUNNING COMPLETED FAILE



- **Utilises SOAP interfaces to simultaneously access:**
  - **ProFunc (EBI, Hinxton, UK)**
  - **CATH (UCL, London, UK)**
  - **FUNcut (CNB, Madrid, Spain)**
  - **STRING (EMBL, Heidelberg, Germany)**

Structure to Function Pipeline - running

cnio
Alfonso Valencia

Structure to Function Pipeline - results

A European Virtual Institute for Genome Annotation

Pipeline results for Uploaded structure

Profunc Results

Sequence motifs

InterPro scan for sequence motifs. Chain A

8 motifs matched in scan against PROSITE, PRINTS, PFam-A, TIGRFAM, PROFILES and PRODOM motifs

	Type	Motif	Name
1.	???	G3D.3.40.50.300	no description
2.	ScanRegExp	PS00039	DEAD_ATP_HELICASE
3.	superfamily	SSF52540	P-loop containing nucleoside triphosphate hydrolases
4.	HMMSmart	SM00487	no description
5.	HMMPfam	PF00270	DEAD

... plus others

InterPro scan for sequence motifs. Chain B

10 motifs matched in scan against PROSITE, PRINTS, PFam-A, TIGRFAM, PROFILES and PRODOM motifs

	Type	Motif	Name
1.	HMMSmart	SM00487	no description
2.	HMMSmart	SM00490	no description
3.	ScanRegExp	PS00039	DEAD_ATP_HELICASE
4.	???	G3D.3.40.50.300x2	no description
5.	HMMPfam	PF00270	DEAD

... plus others

Matches to the Superfamily HMM library

Search of sequence vs Superfamily HMM library. Chain A

1 motif matched in scan against the Superfamily HMM library

	Residue range(s)	Motif	Superfamily name
1.	20-221	52540	P-loop containing nucleoside triphosphate hydrolases

DNA-binding templates.

No hits obtained from any of the 1386 DNA-binding templates.

Reverse template comparison *vs* structures in PDB.

20 significant hits out of 1190 auto-generated templates.

	Score	Template	PDB	Name
1.	1060.000	TMP01059	1fuu	Yeast initiation factor 4a
2.	989.000	TMP01059	1qde	Crystal structure of the atpase domain of translation initiation factor 4a from saccharomyces cerevisiae-the prototype of the dead box protein family
3.	897.000	TMP00259	1fuk	Crystal structure of the carboxy terminal domain of yeast eif4a
4.	562.688	TMP00676	1t5i	Crystal structure of thE C-terminal domain of uap56
5.	648.875	TMP00038	1hv8	Crystal structure of a dead box protein from the hyperthermophile methanococcus jannaschii

... plus others

Generated on Fri 20 Jan 2006

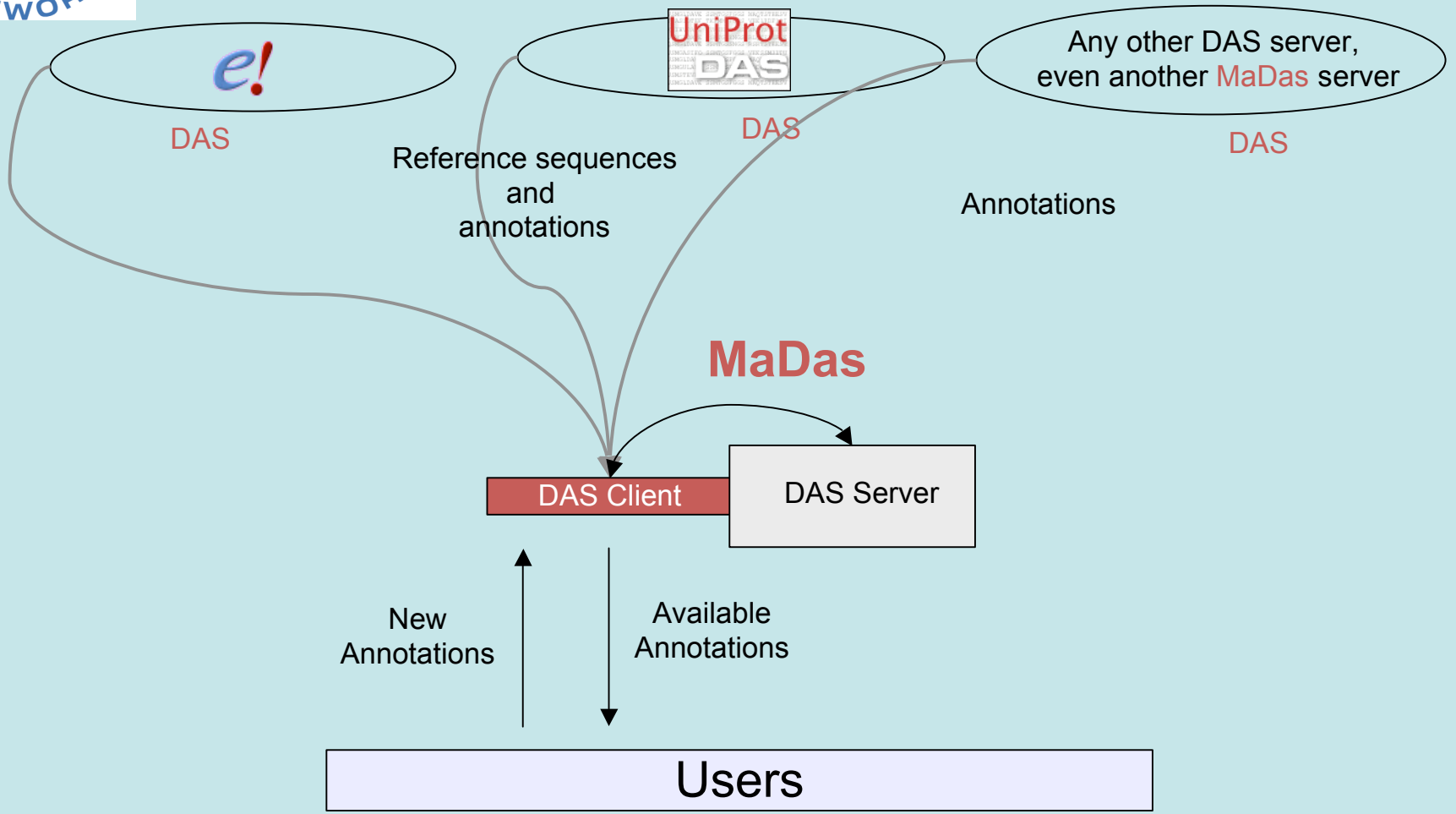
CATH database

CATH results for blast scan:

PDB Code of Match	Description	Sequence Identity	Evalue of the Match	
1qde	Gene regulation Translation initiation factor 4a. Chain: a. Fragment: n-terminal domain, residues 9-232. Synonym: eif4a. Engineered: yes	81.517	2e-65	5e-65
1fuk	Translation Eukaryotic initiation factor 4a. Chain: a. Fragment: carboxy terminal domain (residues 230-394). Synonym: yeast initiation factor 4a, eif4a. Engineered: yes	90.323	1e-61	
1hv8	RNA binding protein Putative atp-dependent RNA helicase mj0669. Chain: a, b. Synonym: dead box helicase. Engineered: yes	35.106	1e-18	5e-19
1q0u	RNA binding protein Bstdead. Chain: a, b. Fragment: n-terminal domain. Engineered: yes	34.444	4e-11	9e-11
1b73	Isomerase Glutamate racemase. Chain: a. Ec: 5.1.1.3	25.773	0.22	
1ayx	Hydrolase Glucoamylase. Chain: null. Engineered: yes	26.984	0.67	1.4
1ff9	Oxidoreductase Saccharopine reductase. Chain: a. Engineered: yes	28.571	0.88	1.8
1ng6	Structural genomics, unknown function Hypothetical protein ygey. Chain: a. Engineered: yes	23.529	1.1	
1q5y	Transferase Sr protein kinase. Chain: a, b. Fragment: sky1pdeltan(137)deltas. Engineered: yes	45.833	2.0	4.1
1ckv	Hydroxylase regulatory protein Protein b. Chain: null. Biological_unit: monomer	29.787	3.3	7.0
1d7u	Lyase 2,2-dialkylglycine decarboxylase (pyruvate). Chain: a. Synonym: dgd. Engineered: yes	52.381	5.7	
1j5p	Oxidoreductase Aspartate dehydrogenase. Chain: a. Engineered: yes	33.333	5.7	
1auu	Transcription regulation Sacy. Chain: a, b. Fragment: RNA binding domain, residues 1 - 55. Engineered: yes. Biological_unit: dimer. Other_details: n terminal region of the sacy protein. The structure is a symmetric dimer in solution	23.404	7.0	

Alfonso Valencia

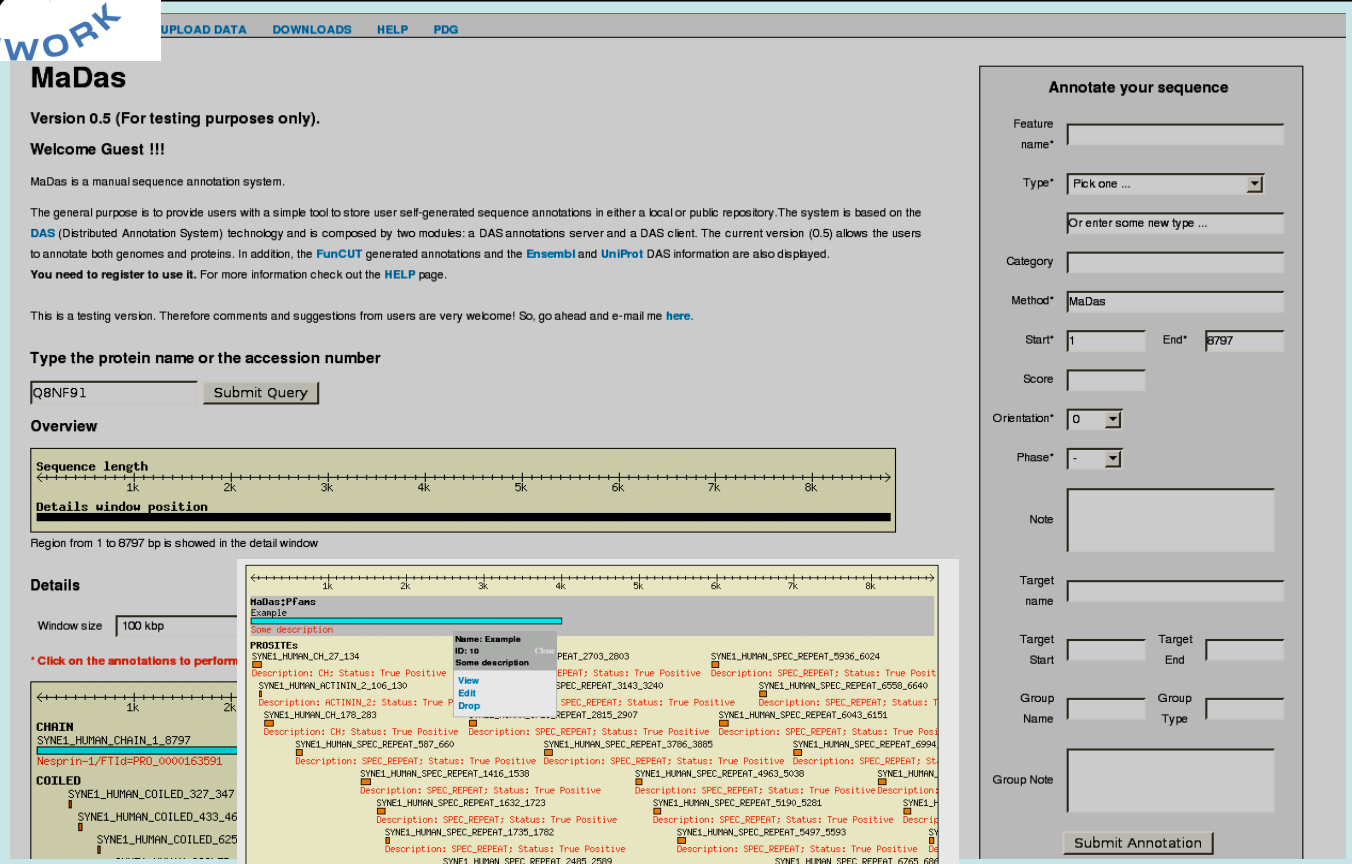
MaDas
Manual Sequence Annotation System



Developed by Victor de la Torre

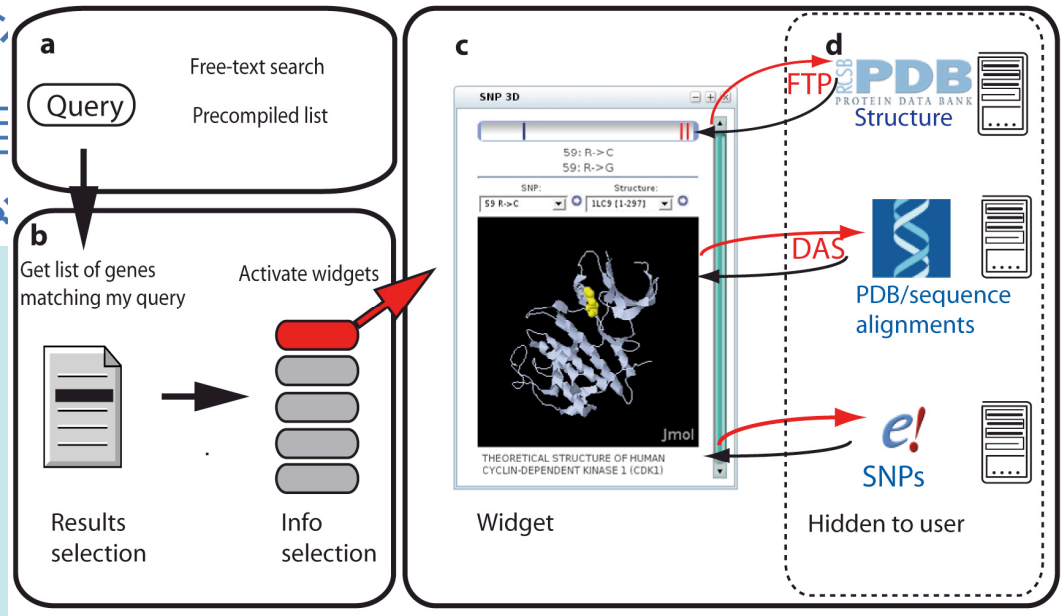
Alfonso Valencia

MaDas Web Interface



- MaDas allows users to add, edit, or remove self generated sequence annotations
- Allows to upload multiple annotations from a file.
- 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.

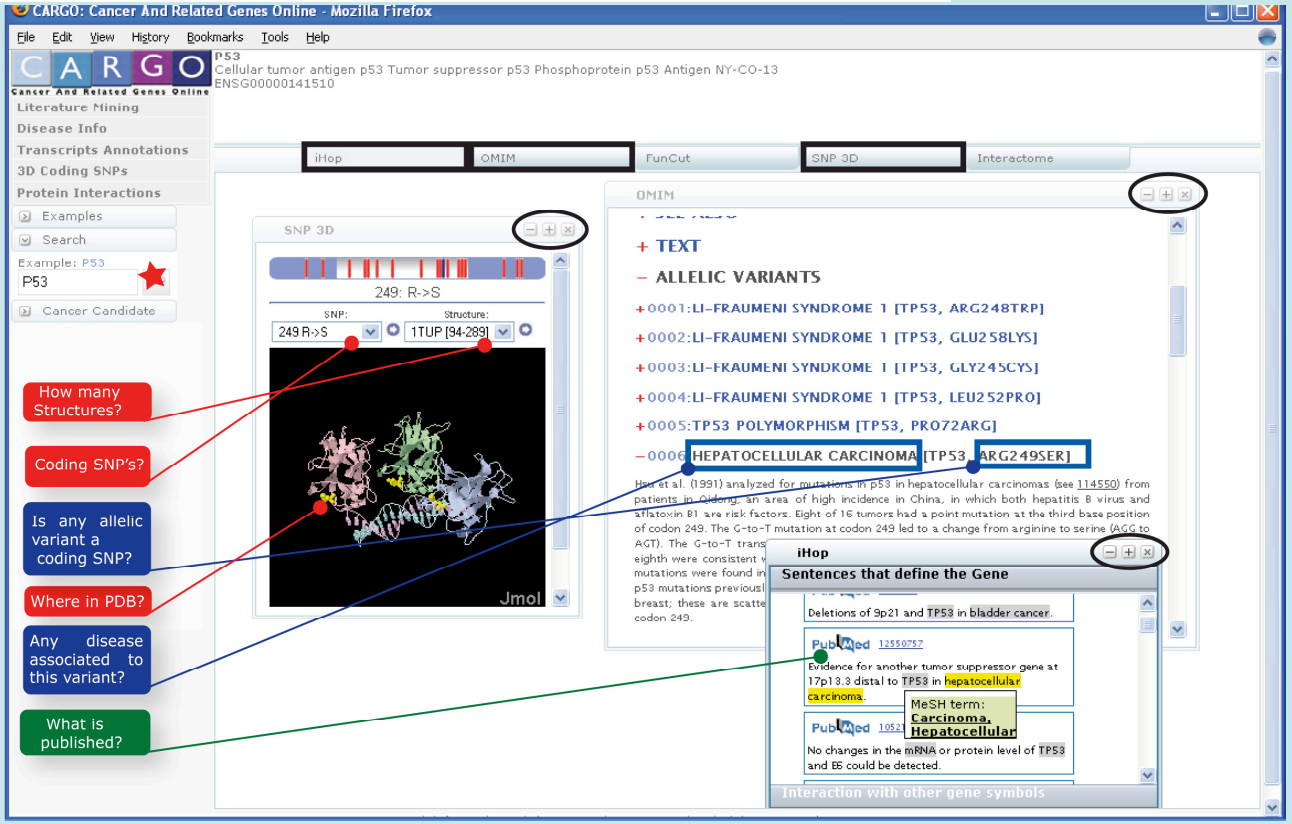
Alfonso Valencia



CARGO

bioinfo.cnio.es/cargo/

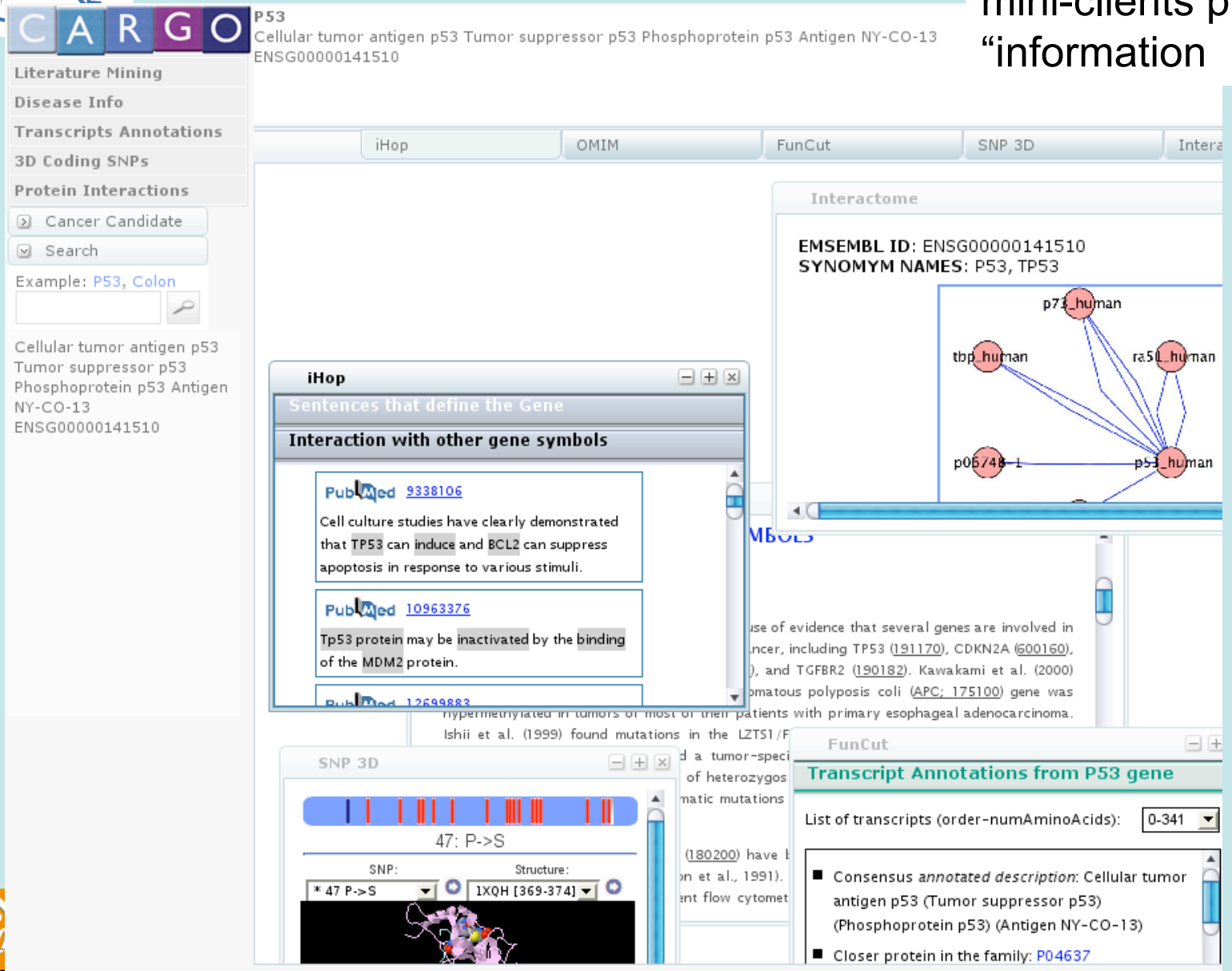
I Cases, AM Rojas, A Carro, JM Fernandez-G, JF Vera, JM Rodriguez, G Gomez, EA Leon, A Valencia, D G Pisano*



Alfonso Valencia

CARGO in the BioSapiens analysis of cancer gene list WP

Some Widget are DAS mini-clients producing "information digestion"



PDG Protein Design Group

www.cnio.es

INB

BIOSAPIENS NETWORK

ECCB

BioCreAtIvE

COMBIO

ENFIN
Enabling Systems Biology

EMERGENCE COST
CA on Synthetic Biology

www.pdg.cnb.uam.es

www.inba.org

- Osvaldo Graña
- Victor de la Torre
- Ildefonso Cases, Jose M. Fernandez-G, Jaime F. Vera, Jose M. Rodriguez, Gonzalo Gomez, Eduardo A. Leon, Angel Carro, Ana Rojas, David G. Pisano
-Robert Hoffmann, Martin Krallinger, Jose M. Fernandez

- BIOCREATIVE II. biocreative.sourceforge.net
- CASP7 predictioncenter.org

cnio
Alfonso Valencia