www.orvidas.com

LA BIOINFORMATICA: Desde la Información al Conocimiento

o

Aprendiendo de la Naturaleza

José M. Carazo
Biocomputing Unit -Centro Nacional de Biotecnología

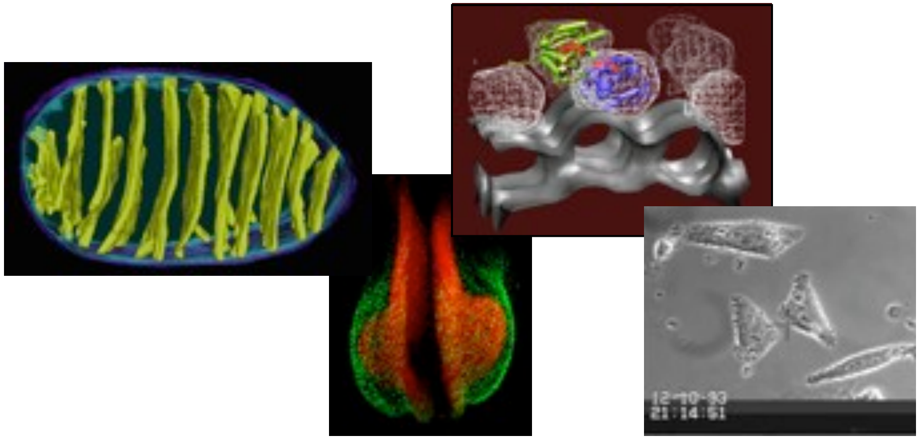
Bioinformática y Biología Computacional.

¿Por qué es tan importante?

...Porque la ingente cantidad de datos y la complejidad de sus relaciones hacen **inviable su procesamiento manual (y su reproducibilidad peligra).**

...Porque se necesita una **perspectiva global** del diseño experimental y del análisis de resultados.

...Porque la disponibilidad de archivos digitales permite generar **hipótesis verificables sobre la función/estructura de un gen o proteína** de interés por medio de la identificación de secuencias similares en organismos mejor caracterizados.



Bioinformática y Biología Computacional.

> Biology in the 21st century is being transformed from a purely lab-based science to an information science as well.
>
> ***Fuente: National Center for Biotechnology Information***

Bioinformática y Biología Computacional.

Biology in the 21st century is being transformed from a purely lab-based science to an information science as well.

Fuente: National Center for Biotechnology Information

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline.

Fuente: National Center for Biotechnology Information

Bioinformática y Biología Computacional.

Biology in the 21st century is being transformed from a purely lab-based science to an information science as well.

Fuente: National Center for Biotechnology Information

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline.

Fuente: National Center for Biotechnology Information

La "Bioinformática" ha evolucionado, de forma que ya no sólo se trata de almacenar y organizar la información sino de analizar, visualizar e interpretar mediante métodos matemáticos y computacionales

Bioinformática y Biología Computacional.

Biology in the 21st century is being transformed from a purely lab-based science to an information science as well.

Fuente: National Center for Biotechnology Information

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline.

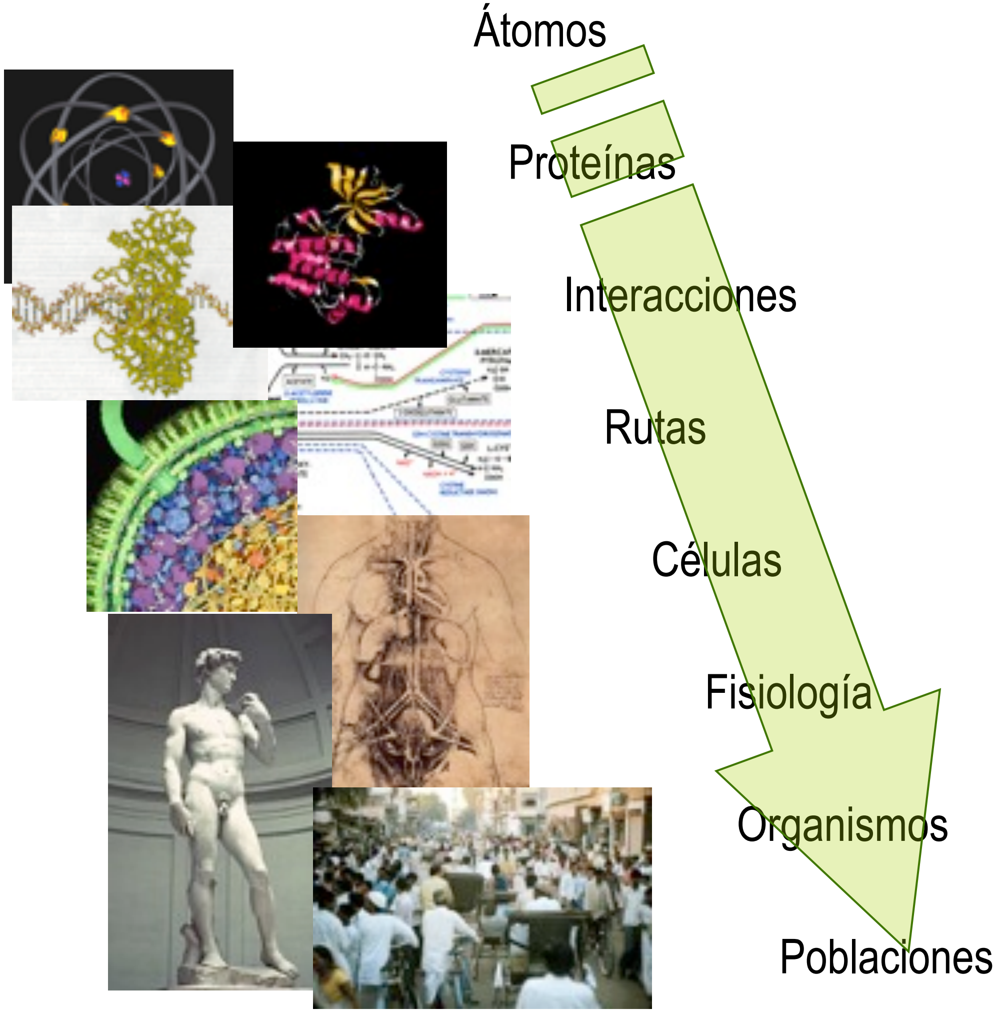
Fuente: National Center for Biotechnology Information

La "Bioinformática" ha evolucionado, de forma que ya no sólo se trata de almacenar y organizar la información sino de analizar, visualizar e interpretar mediante métodos matemáticos y computacionales

Biología Computacional.

¿Con qué tipo información tratamos en Biología y Biomedicina?

- Datos Biológicos
- Características:
  - Complejos
  - Jeráquicos
  - Heterogéneos
  - Dinámicos
  - Incompletos



Átomos

Proteínas

Interacciones

Rutas

Células

Fisiología

Organismos

Poblaciones

¿Dónde se encuentra almacenada?

PRINTS

Patent USPTO

BLOCKS

INTERPRO

PIR

PFAM

Patent PCT

GENEPEPT

PROSITEDOC

DOMO

NRL3D

LOCUS LINK

SWISSFAM

Patent JPO

TFCLASS

PROSITE

TREMBL

TFMATRIX

Medline

PRODOM

UNIGENE

EMBL

TFSITE

DSSP

DBSTS

DDBJ

TFCELL

GSDB

TIGR

SWISSPROT

Entrez

EBI

TAXONOMY

Celera

PDB

RHDB

GENBANK

GENETICCODE

HUGO

Microbial Genomes

STKE

WIT

SNP

GDB

OMIM

Fly Base

KEGG

ENZYME

dbSNP Contact

C. Elegans

FASTA

SSEARCH

Clinical DB

BLAST

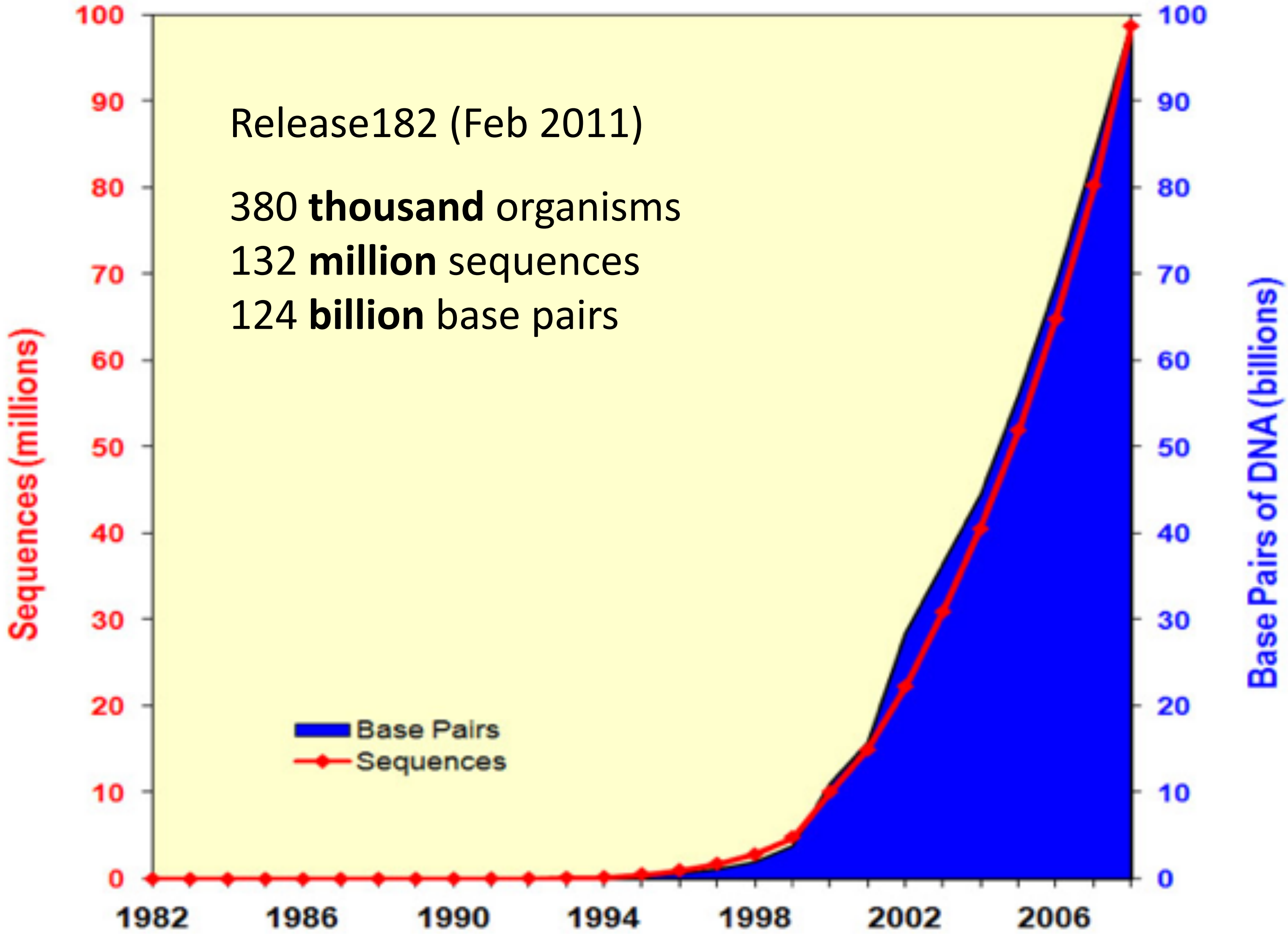
CLUSTALW

dbSNP Population

SNP Consortium

GenBank has doubled in size about every 18 months

GenBank. Benson et al. Nucleic Acids Res. 2011 Jan; 39:D32-7.

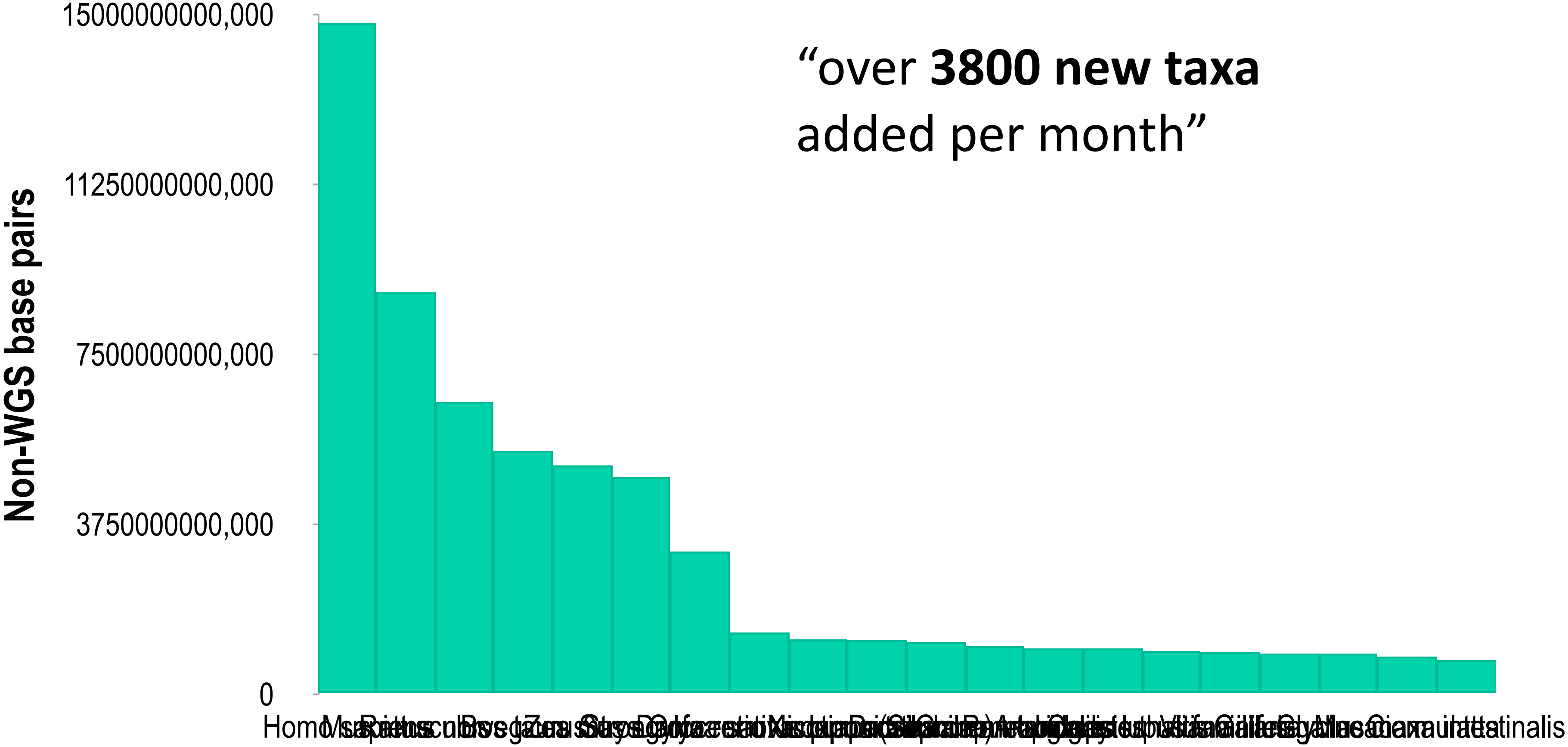


Release182 (Feb 2011)

380 **thousand** organisms
132 **million** sequences
124 **billion** base pairs

GenBank contains nucleotide sequences for more than 380.000 named ***organisms***

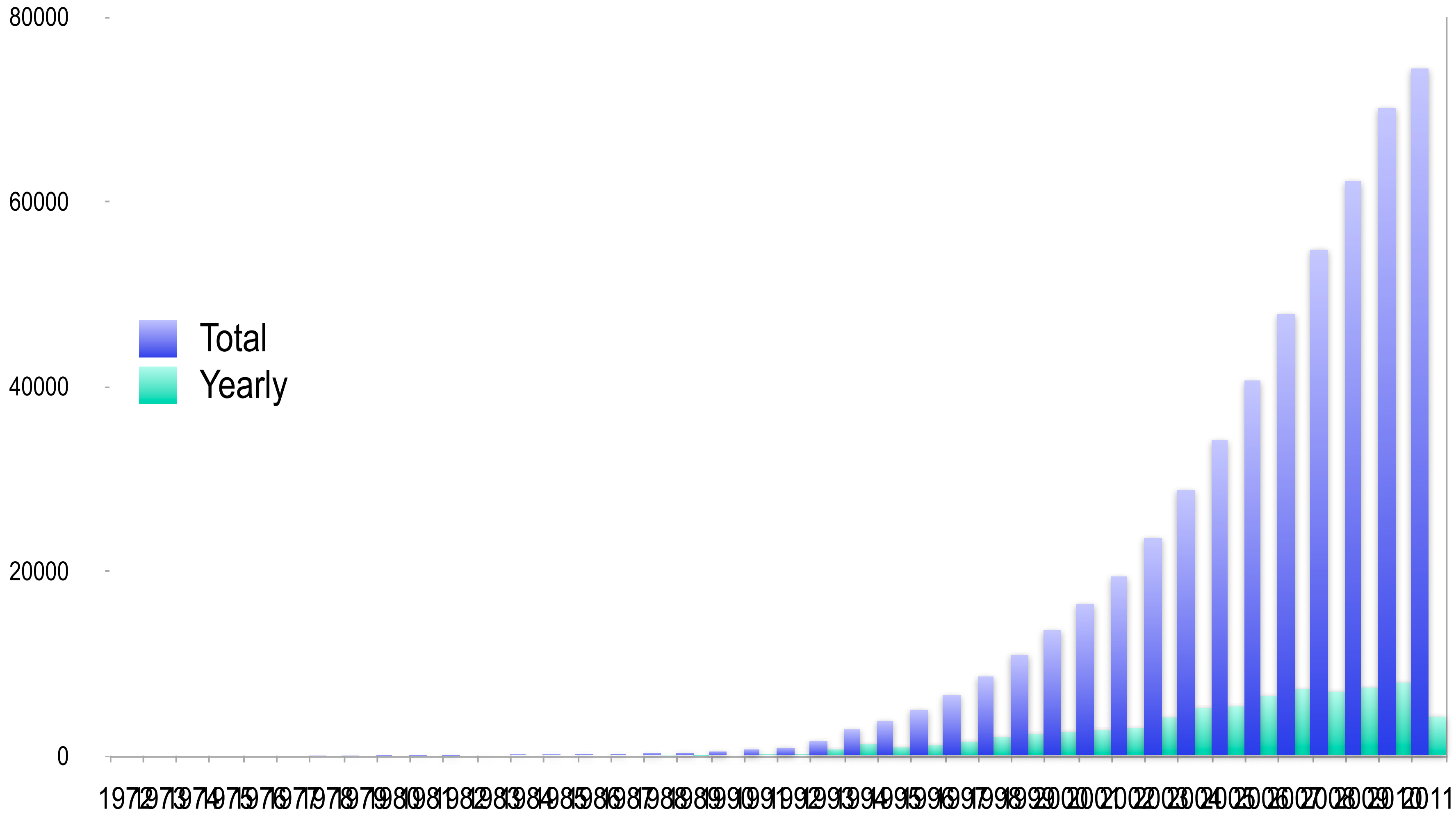
GenBank. Benson et al. Nucleic Acids Res. 2011 Jan; 39:D32-7.



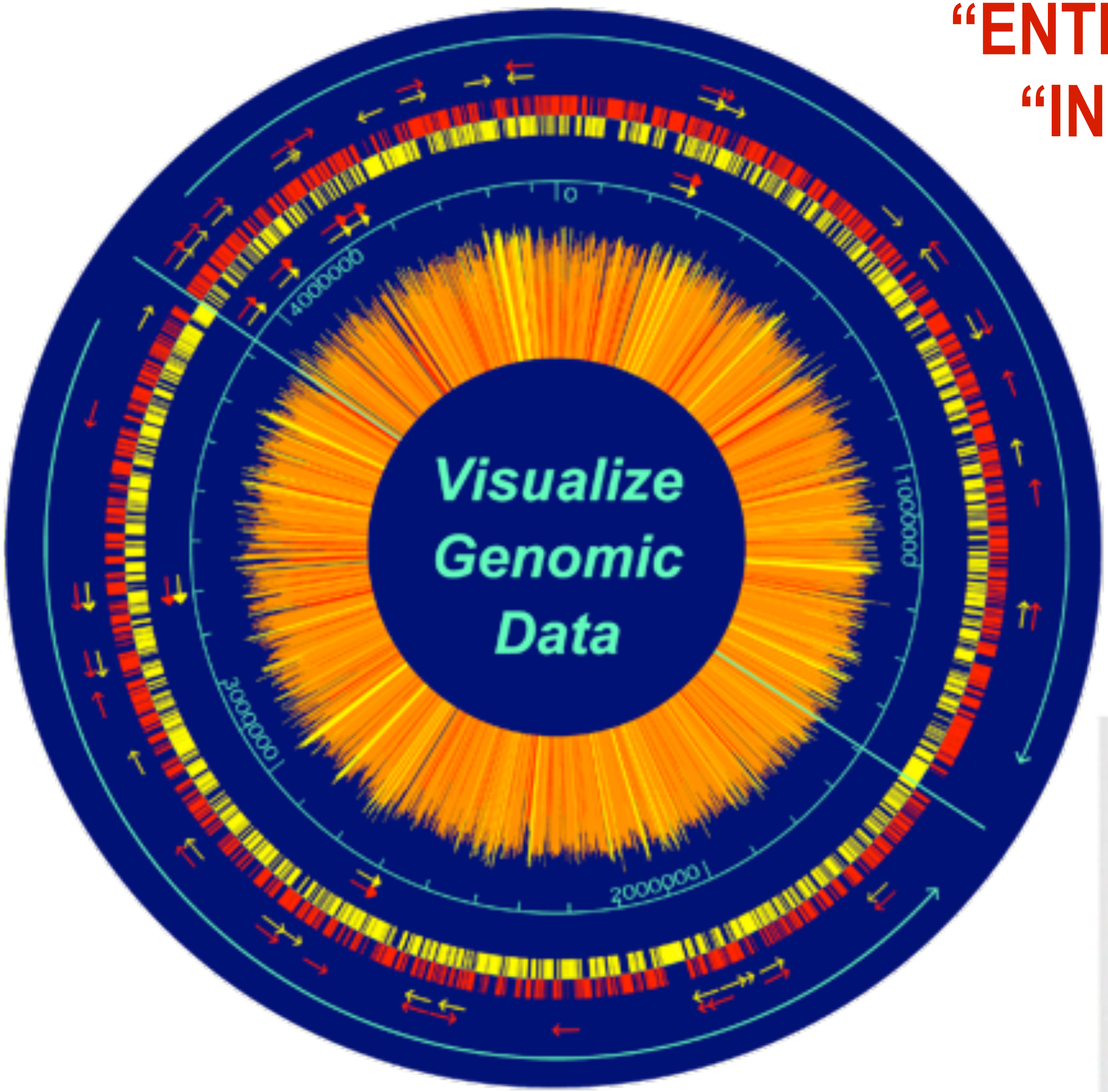
“over **3800 new taxa** added per month”

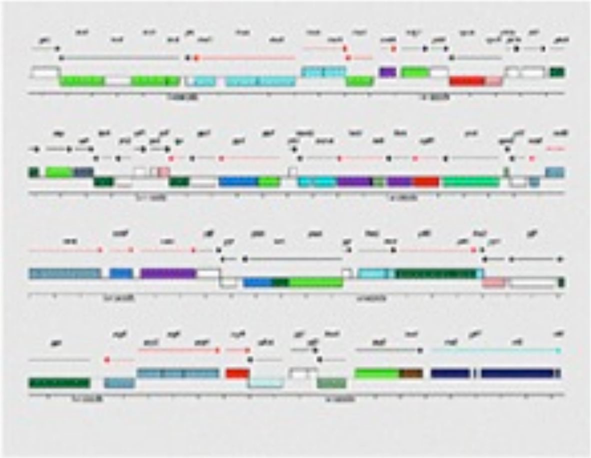
PDB Searchable structures per year

Last updated: Jul 2011 - http://www.rcsb.org/pdb/statistics/



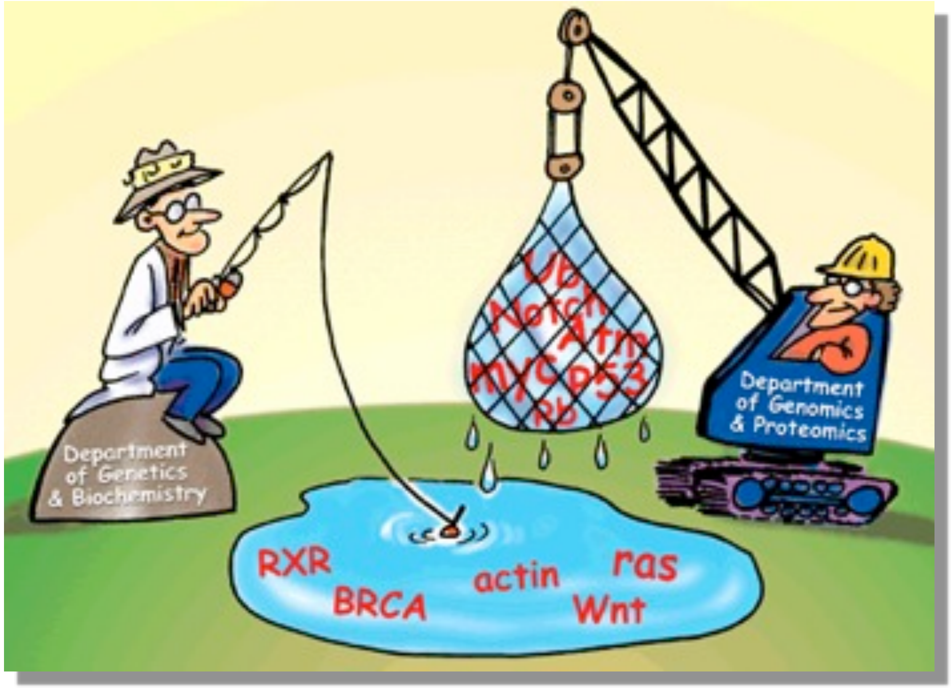
¿PODEMOS REALMENTE “ENTENDER” TODA ESTA “INFORMACION”???





Bioinformática Funcional

Desarrollo de métodos automáticos que ***ayuden*** en la interpretación funcional de los resultados experimentales


Department of Genetics & Biochemistry
Department of Genomics & Proteomics
Ub
Notch
Atm
MYC
p53
Rb
RXR
BRCA
actin
ras
Wnt

Work horse technology: Micro.arrays

Epigenetic studies

CGH

Transcriptional profiling

Genotyping

Methylation

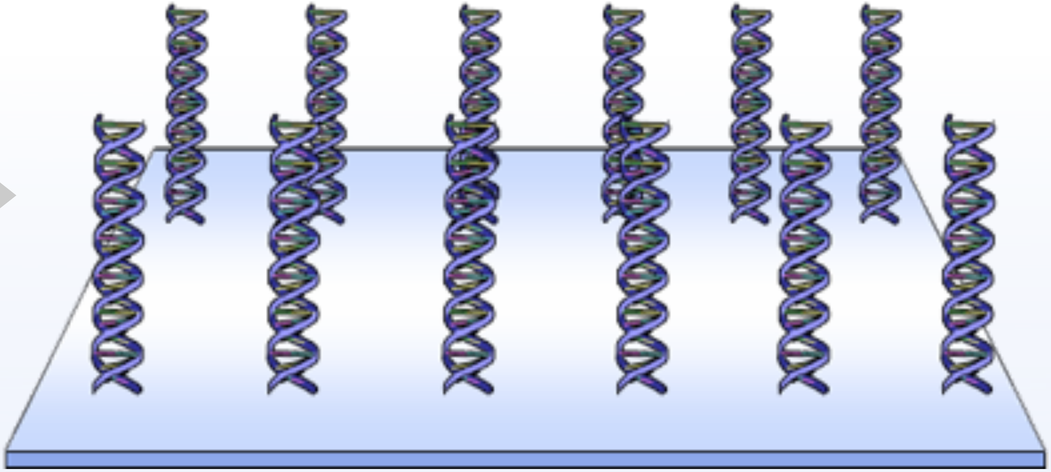
Protein Expression

Splice Variant studies

Protein selection or attachment by aptamers

Protein-ssDNA interactions

Protein-dsDNA interactions



Apart from quality issues, data interpretation is currently the main bottleneck in microarray analyses. In particular, the automated integration of complementary information in analysis algorithms is not yet well established.

Jörg D. Hoheisel: “Microarray technology: beyond transcript profiling and genotype analysis.” Nature Genetics, Vol 7. (2006)

Work horse technology: Micro.arrays

Epigenetic studies

CGH

Transcriptional profiling

Genotyping

Methylation

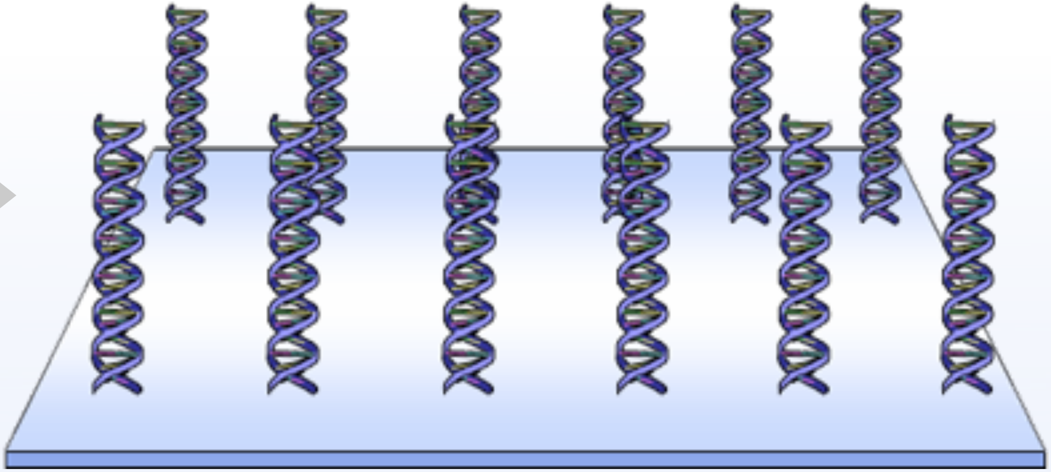
Protein Expression

Splice Variant studies

Protein selection or attachment by aptamers

Protein-ssDNA interactions

Protein-dsDNA interactions



Apart from quality issues, data interpretation is currently the main bottleneck in microarray analyses. In particular, the automated integration of complementary information in analysis algorithms is not yet well established.

Jörg D. Hoheisel: “Microarray technology: beyond transcript profiling and genotype analysis.” Nature Genetics, Vol 7. (2006)

High-throughput sequencing

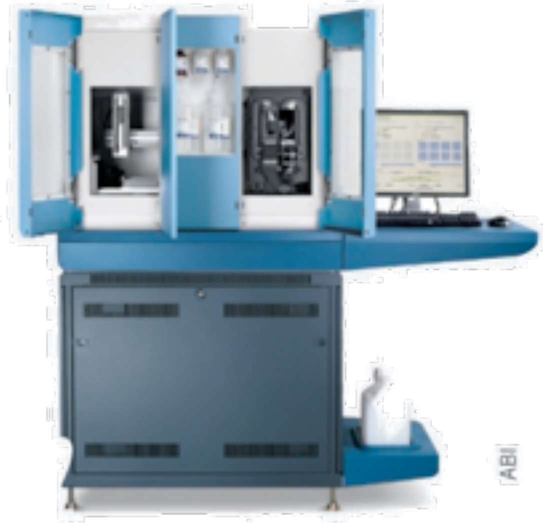
Roche’s 454

Roche

Illumina’s Solexa

illumina

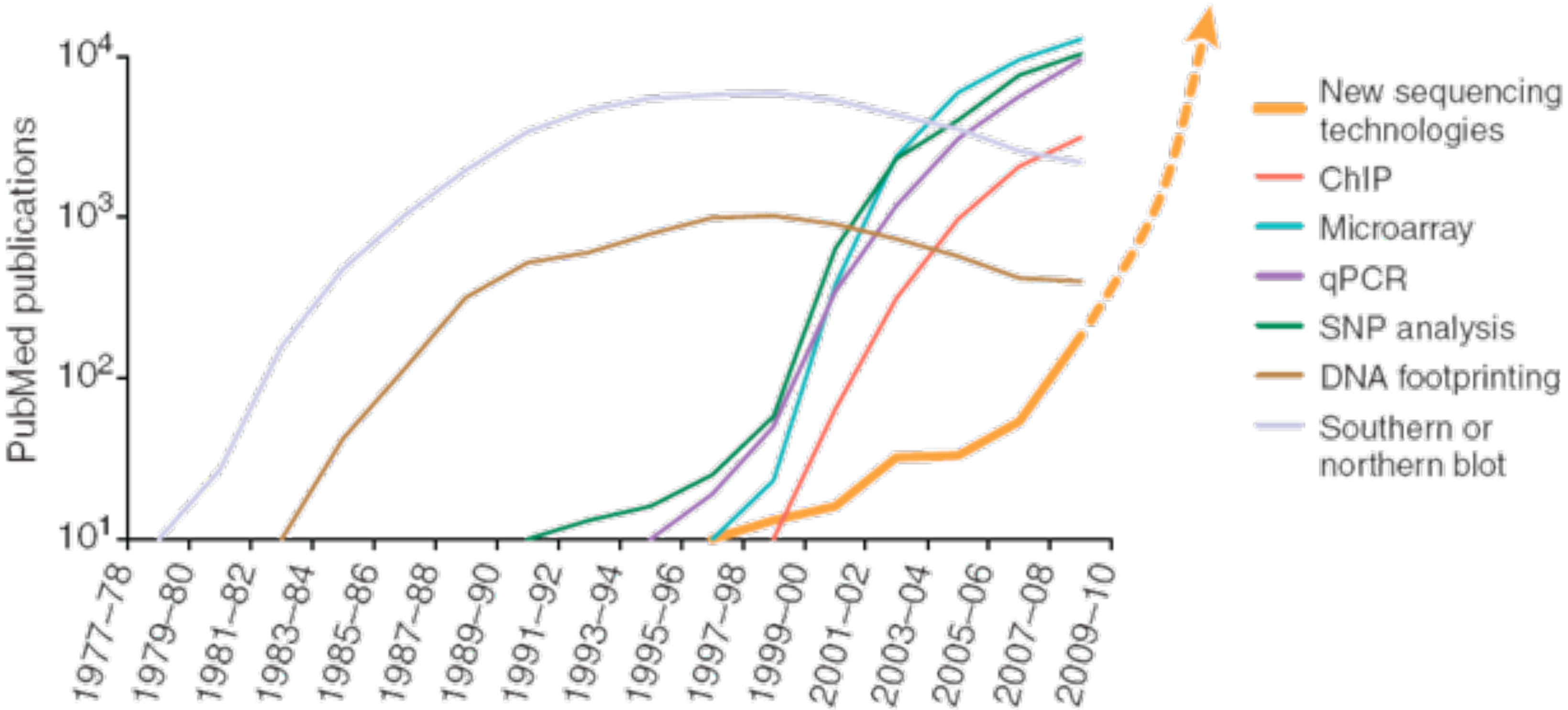
ABI’s SOLiD



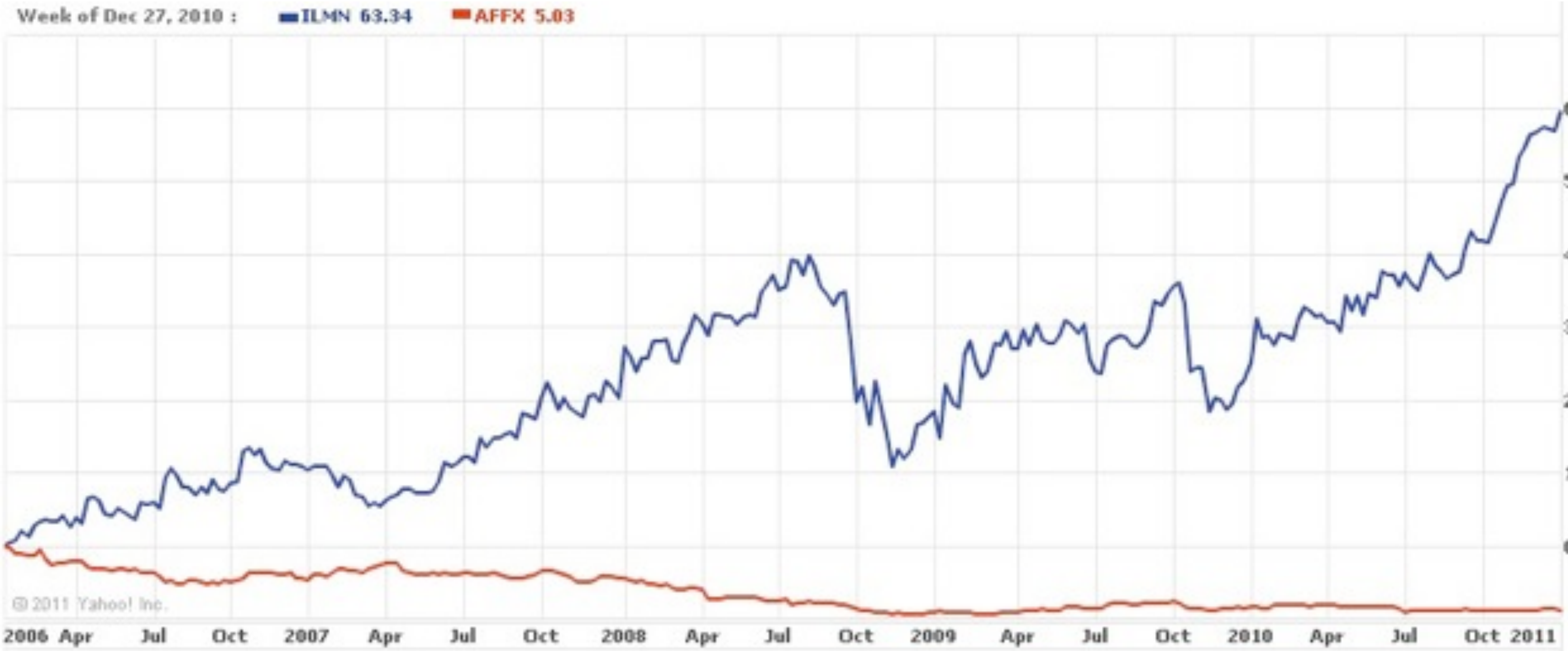
AB Applied Biosystems

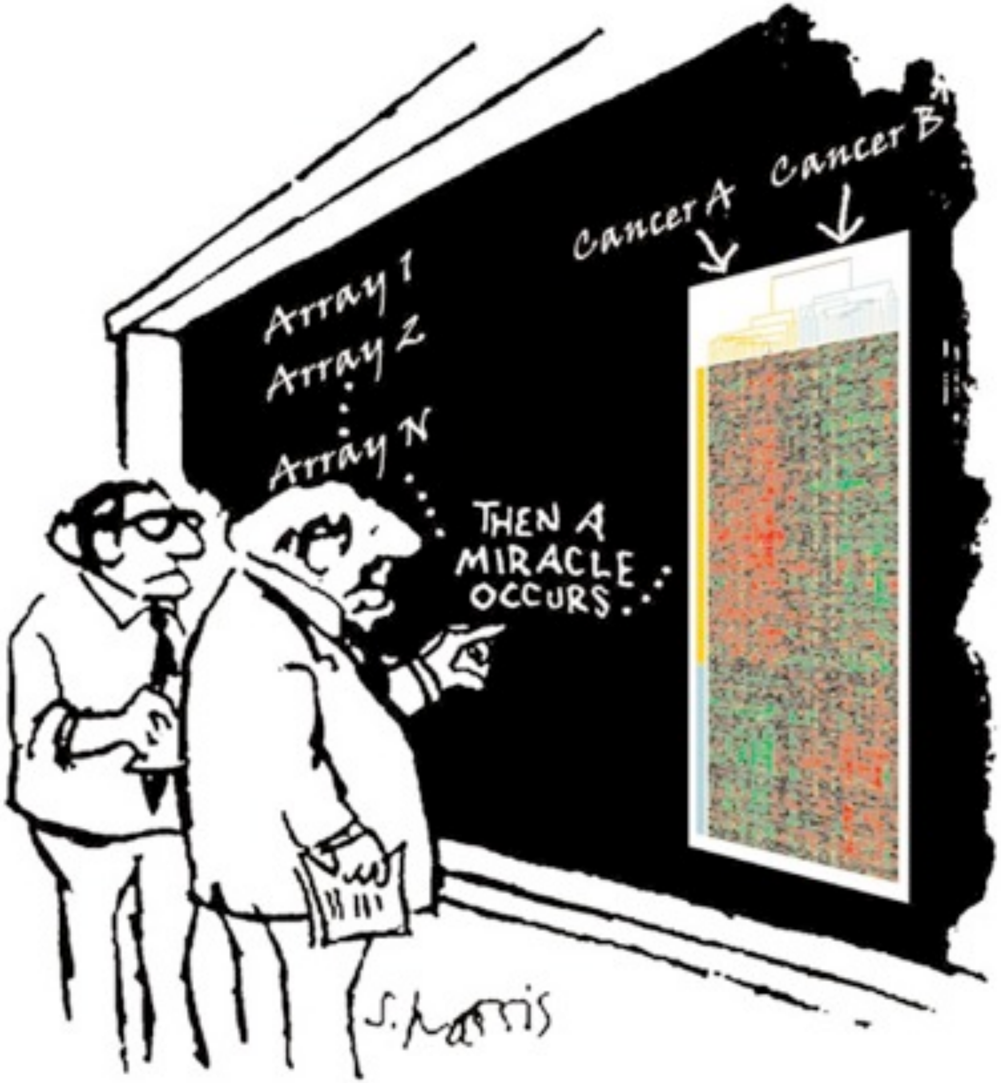
- ‘ChIP-Seq’, or genome-wide mapping of DNA-protein interactions
- ‘RNA-Seq’, analogous to expressed sequence tags (EST) or serial analysis of gene expression (SAGE))
- Full-genome re-sequencing or more targeted discovery of mutations or polymorphisms
- Mapping of structural rearrangements, including copy number variation, balanced translocation breakpoints and chromosomal inversions
- Large-scale analysis of DNA methylation
- Epigenomic state: Differences in DNA methylation patterns

Future trend



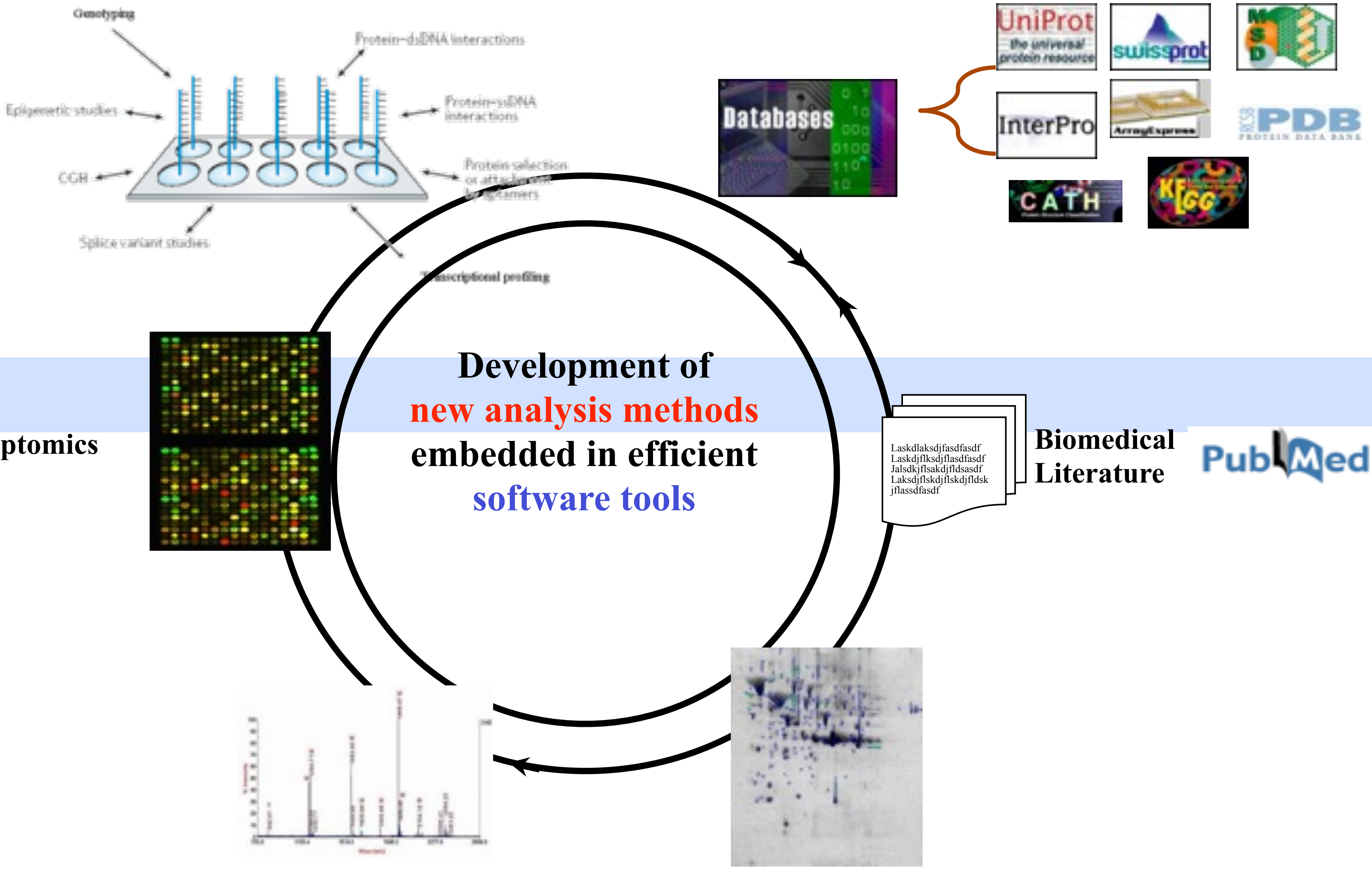
SeqSolve: NGS





"I think you should be more explicit here in step two."

Main goal of Functional Bioinformatics:



Development of
new analysis methods
embedded in efficient
software tools

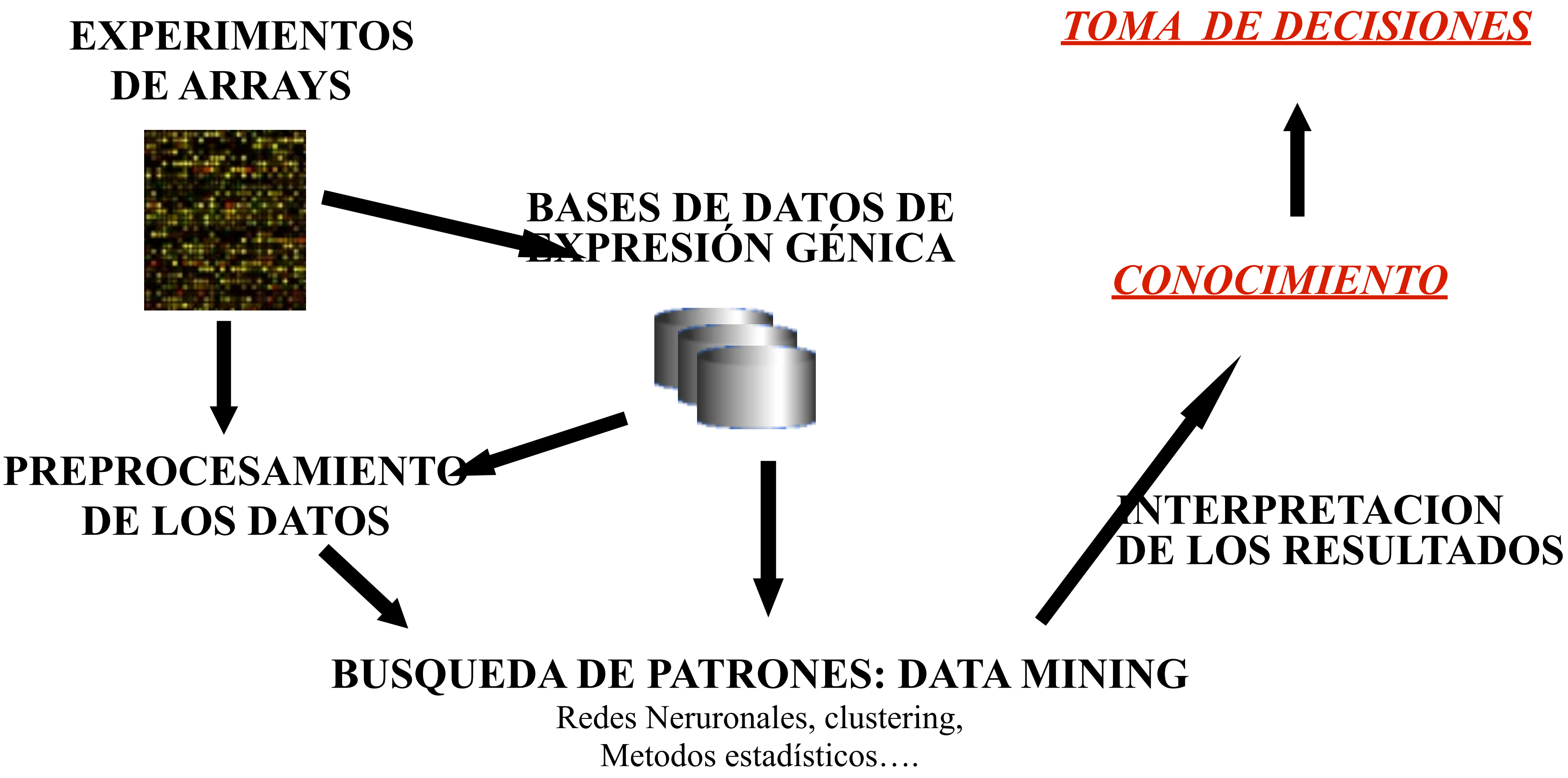
Transcriptomics

Biomedical
Literature

Metabolomic

Proteomics

DATOS EXPERIMENTALES => CONOCIMIENTO



EXPERIMENTOS
DE ARRAYS

BASES DE DATOS DE
EXPRESIÓN GÉNICA

PREPROCESAMIENTO
DE LOS DATOS

BUSQUEDA DE PATRONES: DATA MINING
Redes Neruronales, clustering,
Metodos estadísticos….

INTERPRETACION
DE LOS RESULTADOS

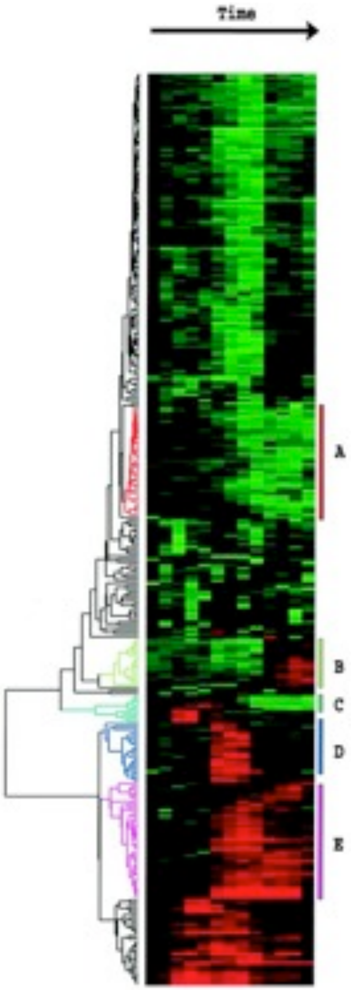
CONOCIMIENTO

TOMA DE DECISIONES

PERO…. QUE ES “DATA MINING”???

Startrek.mpg

CONOCIMIENTO BIOLOGICO A PARTIR DE LOS DATOS



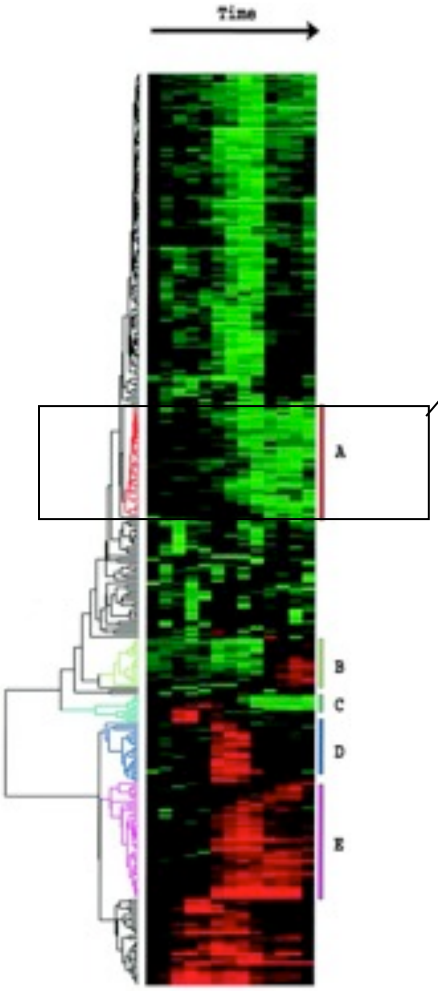
Eisen *et al*., PNAS 1998

Find group of genes sharing similar expression patterns.

Clustering algorithms remain the most popular computational approach to analyze microarray data in this line. These methods organize complex expression datasets into tractable clusters of genes sharing similar expression patterns.

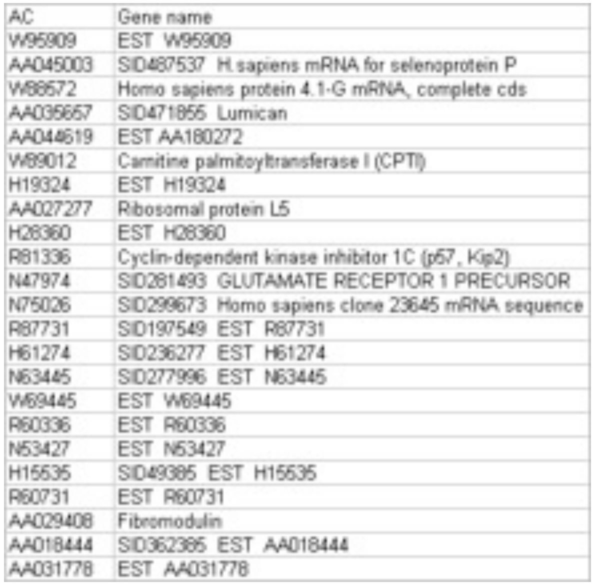
Then, we obtain a list of genes that share a similar expression pattern… but, why these genes have a similar expression pattern?

BIOLOGICAL KNOWLEDGE FROM GENE EXPRESSION DATA

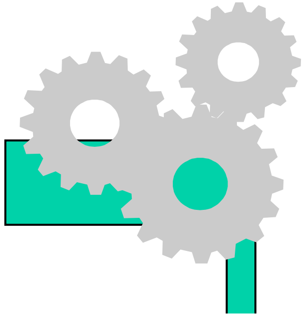


Eisen *et al*., PNAS 1998

List of genes



AC	Gene name
W95909	EST W95909
AA045003	SID487537 H.sapiens mRNA for selenoprotein P
W88572	Homo sapiens protein 4.1-G mRNA, complete cds
AA035657	SID471855 Lumican
AA044619	EST AA180272
W89012	Carnitine palmitoyltransferase I (CPTI)
H19324	EST H19324
AA027277	Ribosomal protein L5
H28360	EST H28360
R81336	Cyclin-dependent kinase inhibitor 1C (p57, Kip2)
N47974	SID281493 GLUTAMATE RECEPTOR 1 PRECURSOR
N75026	SID299673 Homo sapiens clone 23645 mRNA sequence
R87731	SID197549 EST R87731
H61274	SID236277 EST H61274
N63445	SID277996 EST N63445
W69445	EST W69445
R60336	EST R60336
N53427	EST N53427
H15535	SID49385 EST H15535
R60731	EST R60731
AA029408	Fibromodulin
AA018444	SID362385 EST AA018444
AA031778	EST AA031778



Functional relationships
Upstream sequence motifs
Literature searches
Gene Ontology …

Biological knowledge

Interpretación de datos de expresión génica:

Anotaciones y análisis de co.ocurrencia

Integrating Geo-Annotations

Data: Geographical information of a particular region

Metadata: Different types of annotations can be over lied on a ***model***

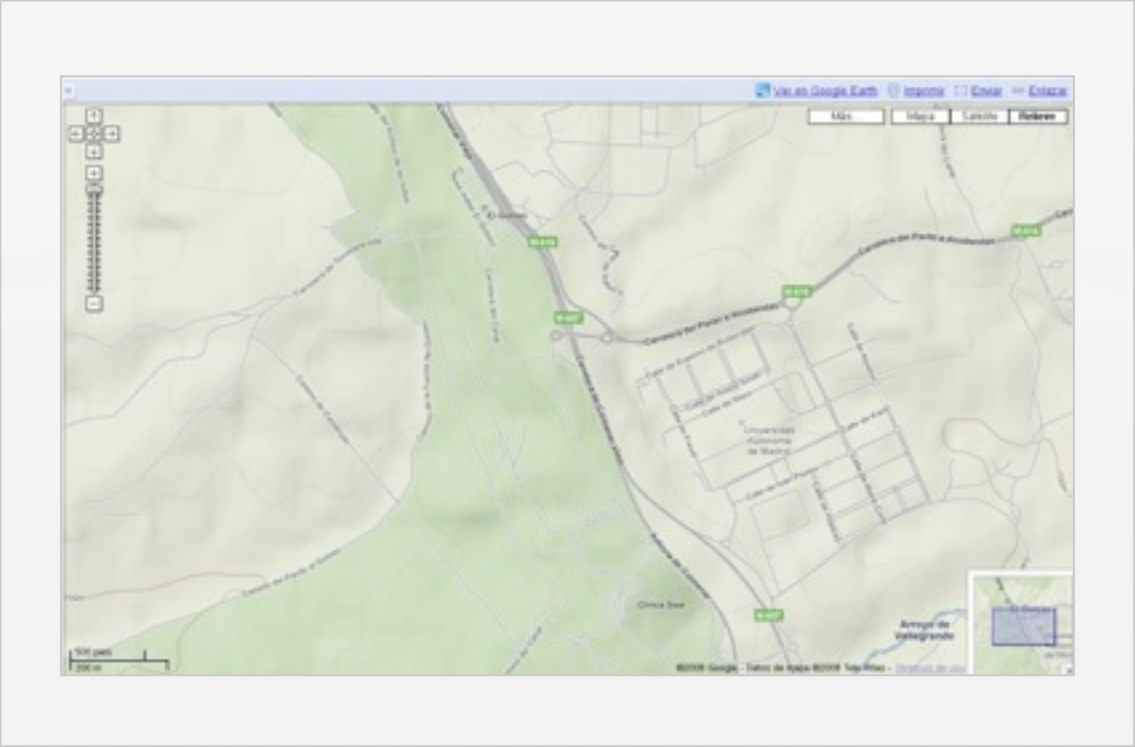
- Points of interest
- Pictures
- Etc.

Integrating Geo-Annotations

Data: Geographical information of a particular region

Metadata: Different types of annotations can be over lied on a ***model***

- Points of interest
- Pictures
- Etc.

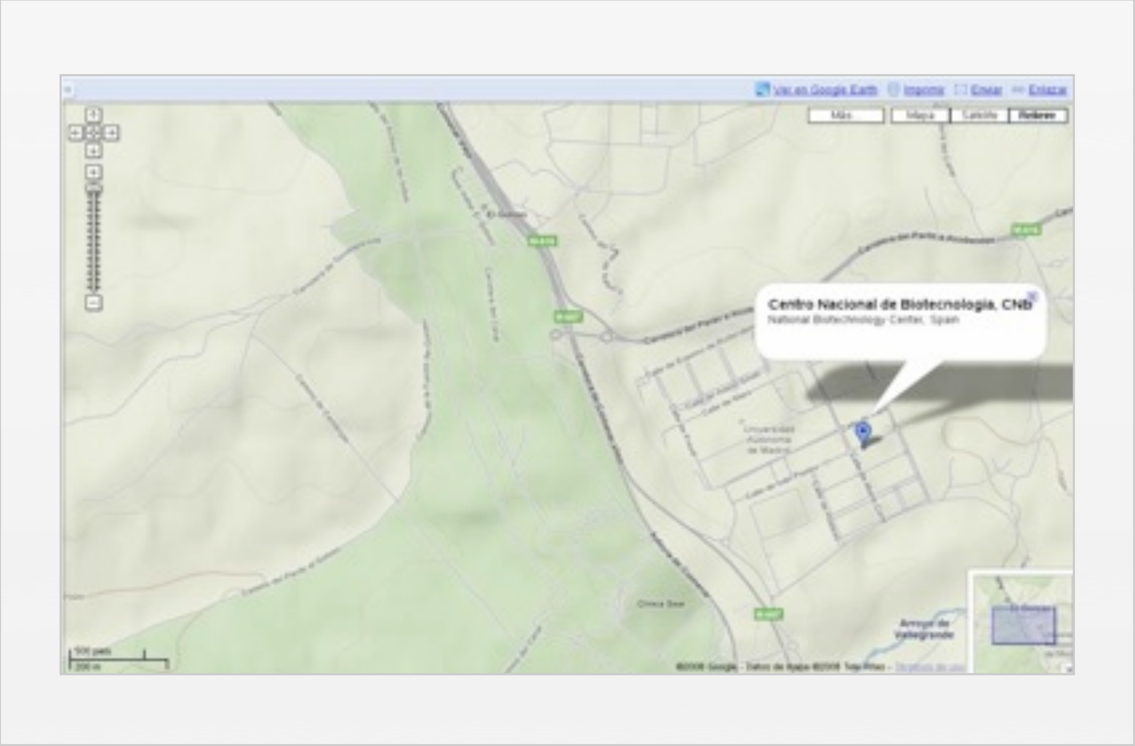


Integrating Geo-Annotations

Data: Geographical information of a particular region

Metadata: Different types of annotations can be over lied on a ***model***

- Points of interest
- Pictures
- Etc.

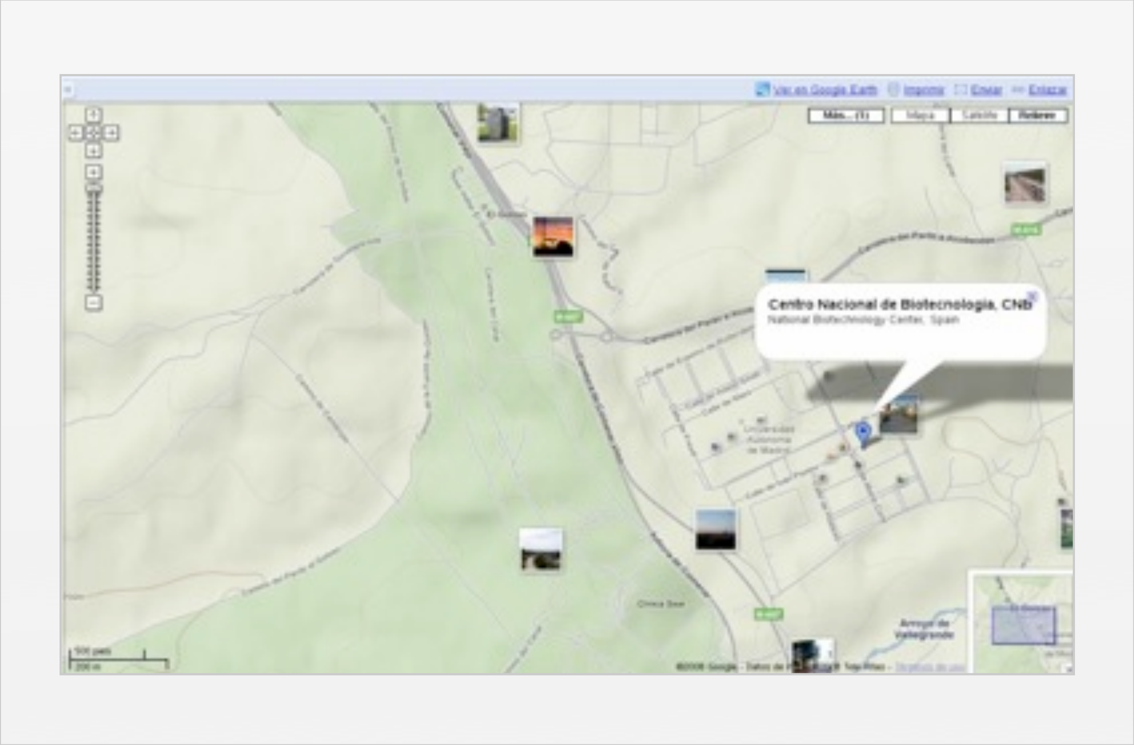

Ver en Google Earth
Imprimir
Enviar
Enlazar
Más
Mapa
Satélite
Relieve
Centro Nacional de Biotecnología, CNB
National Biotechnology Center, Spain
Universidad Autónoma de Madrid

Integrating Geo-Annotations

Data: Geographical information of a particular region

Metadata: Different types of annotations can be over lied on a ***model***

- Points of interest
- Pictures
- Etc.

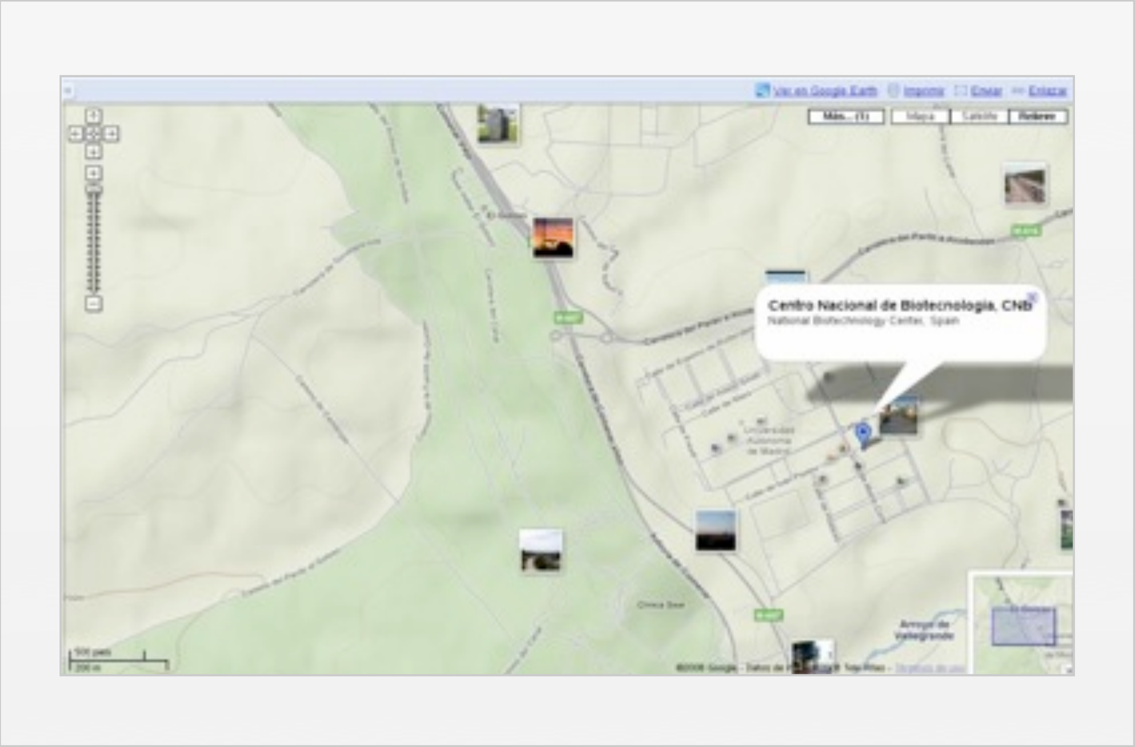

Ver en Google Earth
Imprimir
Enviar
Enlazar
Más... (1)
Mapa
Satélite
Relieve
Centro Nacional de Biotecnologia, CNB
National Biotechnology Center, Spain
Universidad Autónoma de Madrid
500 pies
200 m
©2008 Google

Integrating Geo-Annotations

Data: Geographical information of a particular region

Metadata: Different types of annotations can be over lied on a ***model***

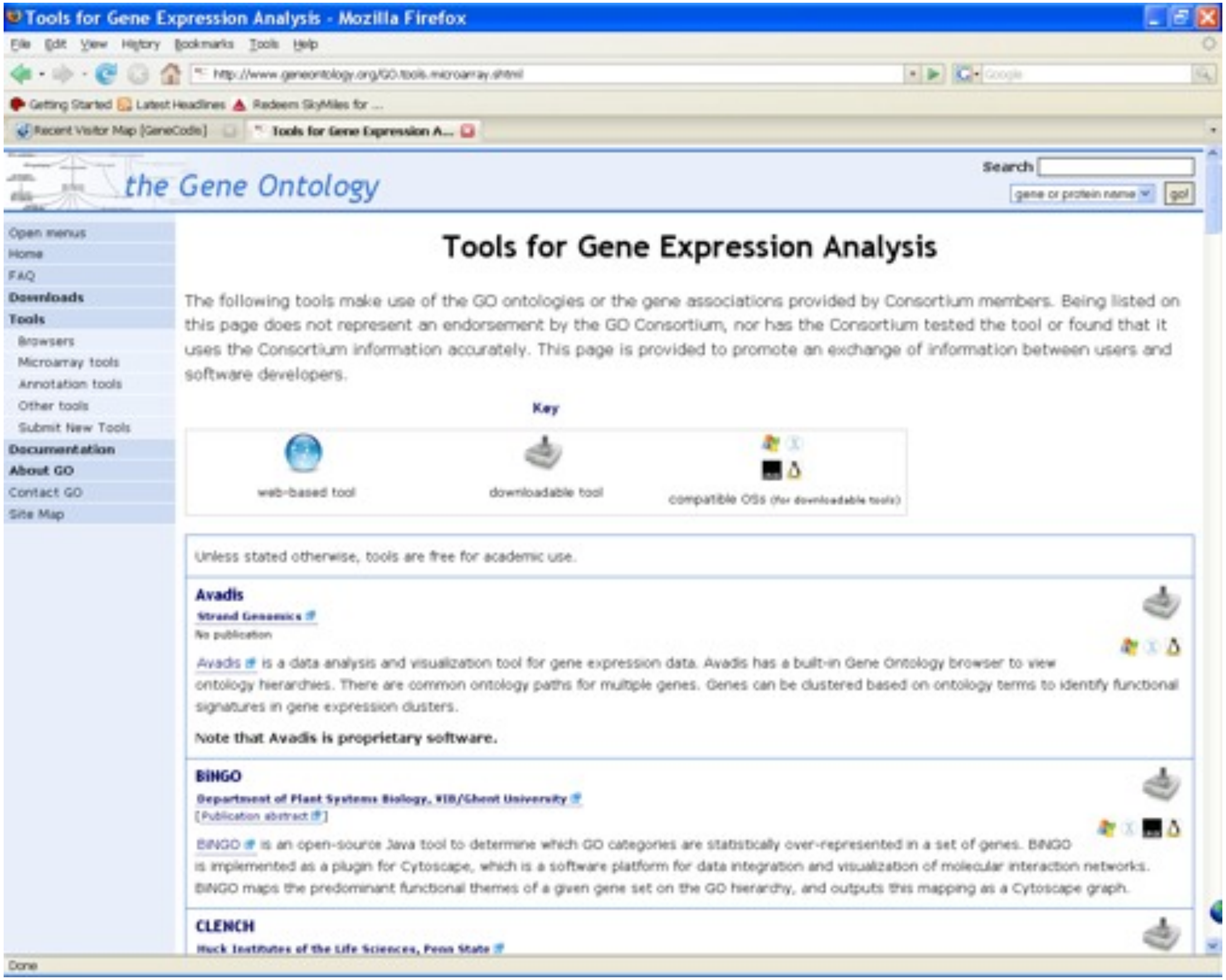
- Points of interest
- Pictures
- Etc.


Ver en Google Earth
Imprimir
Enviar
Enlazar
Más... (1)
Mapa
Satélite
Relieve
Centro Nacional de Biotecnologia, CNB
National Biotechnology Center, Spain
Universidad Autónoma de Madrid
500 pies
200 m
©2008 Google

GO Tools in microarrays:

http://www.geneontology.org/GO.tools.microarray.shtml

More than 60!


Tools for Gene Expression Analysis - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.geneontology.org/GO.tools.microarray.shtml
Google
Getting Started Latest Headlines Redeem SkyMiles for ...
Recent Visitor Map [GeneCodis]
Tools for Gene Expression A...
the Gene Ontology
Search
gene or protein name
go!
Open menus
Home
FAQ
Downloads
Tools
Browsers
Microarray tools
Annotation tools
Other tools
Submit New Tools
Documentation
About GO
Contact GO
Site Map
Done

Tools for Gene Expression Analysis

The following tools make use of the GO ontologies or the gene associations provided by Consortium members. Being listed on this page does not represent an endorsement by the GO Consortium, nor has the Consortium tested the tool or found that it uses the Consortium information accurately. This page is provided to promote an exchange of information between users and software developers.

Key

web-based tool

downloadable tool

compatible OSs (for downloadable tools)

Unless stated otherwise, tools are free for academic use.

Avadis

Strand Genomics

No publication

Avadis is a data analysis and visualization tool for gene expression data. Avadis has a built-in Gene Ontology browser to view ontology hierarchies. There are common ontology paths for multiple genes. Genes can be clustered based on ontology terms to identify functional signatures in gene expression clusters.

Note that Avadis is proprietary software.

BiNGO

Department of Plant Systems Biology, VIB/Ghent University

[Publication abstract]

BiNGO is an open-source Java tool to determine which GO categories are statistically over-represented in a set of genes. BiNGO is implemented as a plugin for Cytoscape, which is a software platform for data integration and visualization of molecular interaction networks. BiNGO maps the predominant functional themes of a given gene set on the GO hierarchy, and outputs this mapping as a Cytoscape graph.

CLENCH

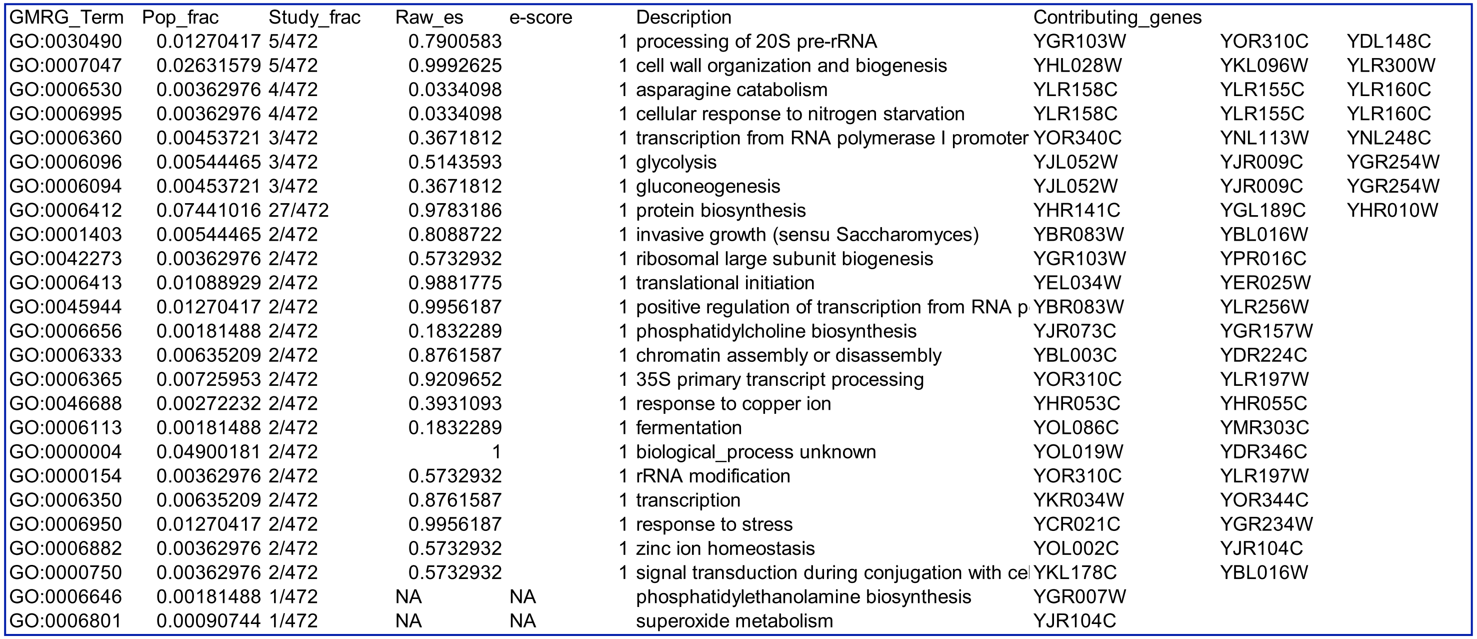
Huck Institutes of the Life Sciences, Penn State

Drawbacks of these methods....

- But all of these tools analyze an annotation independently of each other

Drawbacks of these methods….

- But all of these tools analyze an annotation independently of each other



GMRG_Term	Pop_frac	Study_frac	Raw_es	e-score	Description	Contributing_genes		
GO:0030490	0.01270417	5/472	0.7900583	1	processing of 20S pre-rRNA	YGR103W	YOR310C	YDL148C
GO:0007047	0.02631579	5/472	0.9992625	1	cell wall organization and biogenesis	YHL028W	YKL096W	YLR300W
GO:0006530	0.00362976	4/472	0.0334098	1	asparagine catabolism	YLR158C	YLR155C	YLR160C
GO:0006995	0.00362976	4/472	0.0334098	1	cellular response to nitrogen starvation	YLR158C	YLR155C	YLR160C
GO:0006360	0.00453721	3/472	0.3671812	1	transcription from RNA polymerase I promoter	YOR340C	YNL113W	YNL248C
GO:0006096	0.00544465	3/472	0.5143593	1	glycolysis	YJL052W	YJR009C	YGR254W
GO:0006094	0.00453721	3/472	0.3671812	1	gluconeogenesis	YJL052W	YJR009C	YGR254W
GO:0006412	0.07441016	27/472	0.9783186	1	protein biosynthesis	YHR141C	YGL189C	YHR010W
GO:0001403	0.00544465	2/472	0.8088722	1	invasive growth (sensu Saccharomyces)	YBR083W	YBL016W	
GO:0042273	0.00362976	2/472	0.5732932	1	ribosomal large subunit biogenesis	YGR103W	YPR016C	
GO:0006413	0.01088929	2/472	0.9881775	1	translational initiation	YEL034W	YER025W	
GO:0045944	0.01270417	2/472	0.9956187	1	positive regulation of transcription from RNA p	YBR083W	YLR256W	
GO:0006656	0.00181488	2/472	0.1832289	1	phosphatidylcholine biosynthesis	YJR073C	YGR157W	
GO:0006333	0.00635209	2/472	0.8761587	1	chromatin assembly or disassembly	YBL003C	YDR224C	
GO:0006365	0.00725953	2/472	0.9209652	1	35S primary transcript processing	YOR310C	YLR197W	
GO:0046688	0.00272232	2/472	0.3931093	1	response to copper ion	YHR053C	YHR055C	
GO:0006113	0.00181488	2/472	0.1832289	1	fermentation	YOL086C	YMR303C	
GO:0000004	0.04900181	2/472	1	1	biological_process unknown	YOL019W	YDR346C	
GO:0000154	0.00362976	2/472	0.5732932	1	rRNA modification	YOR310C	YLR197W	
GO:0006350	0.00635209	2/472	0.8761587	1	transcription	YKR034W	YOR344C	
GO:0006950	0.01270417	2/472	0.9956187	1	response to stress	YCR021C	YGR234W	
GO:0006882	0.00362976	2/472	0.5732932	1	zinc ion homeostasis	YOL002C	YJR104C	
GO:0000750	0.00362976	2/472	0.5732932	1	signal transduction during conjugation with cel	YKL178C	YBL016W	
GO:0006646	0.00181488	1/472	NA	NA	phosphatidylethanolamine biosynthesis	YGR007W		
GO:0006801	0.00090744	1/472	NA	NA	superoxide metabolism	YJR104C		

Drawbacks of these methods….

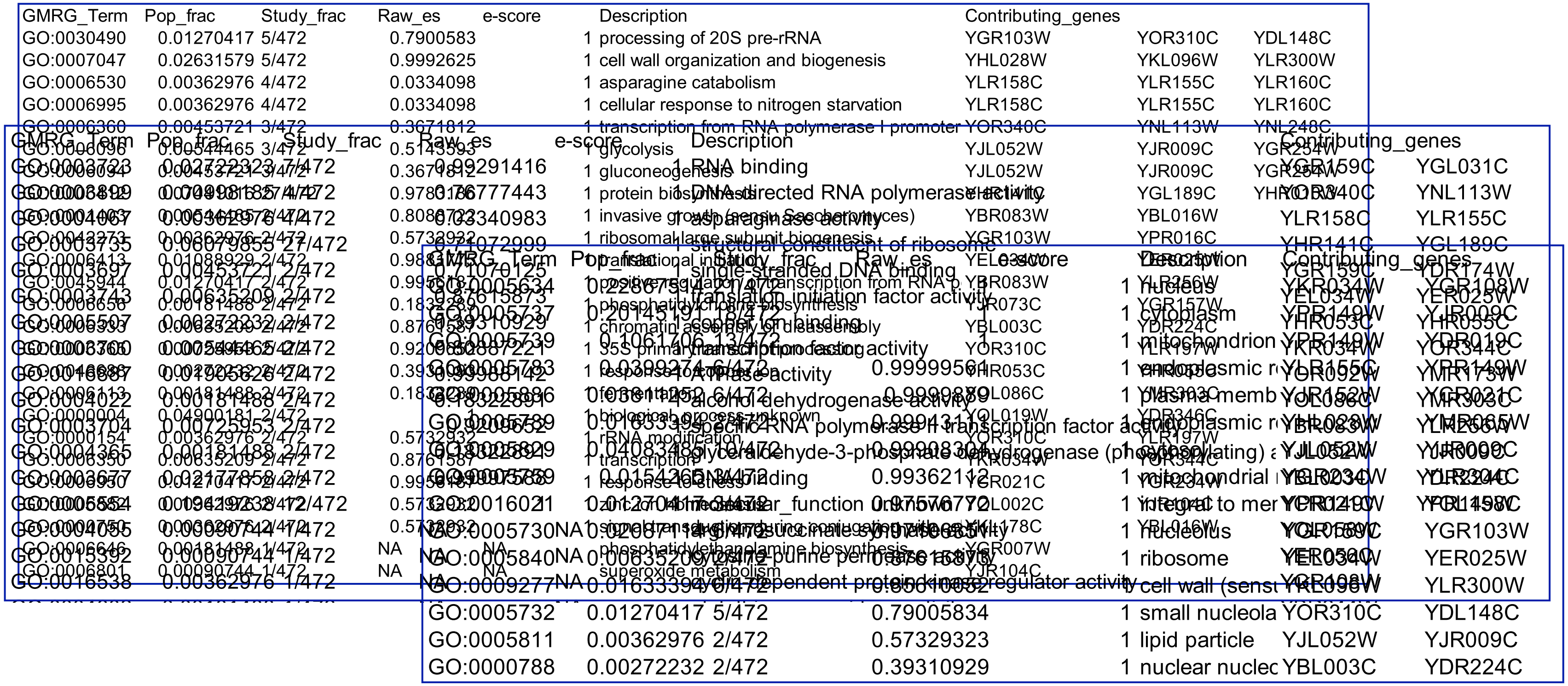
- But all of these tools analyze an annotation independently of each other

GMRG_Term	Pop_frac	Study_frac	Raw_es	e-score	Description	Contributing_genes		
GO:0030490	0.01270417	5/472	0.7900583	1	processing of 20S pre-rRNA	YGR103W	YOR310C	YDL148C
GO:0007047	0.02631579	5/472	0.9992625	1	cell wall organization and biogenesis	YHL028W	YKL096W	YLR300W
GO:0006530	0.00362976	4/472	0.0334098	1	asparagine catabolism	YLR158C	YLR155C	YLR160C
GO:0006995	0.00362976	4/472	0.0334098	1	cellular response to nitrogen starvation	YLR158C	YLR155C	YLR160C
GO:0006360	0.00453721	3/472	0.3671812	1	transcription from RNA polymerase I promoter	YOR340C	YNL113W	YNL248C
GO:0006096	0.00544465	3/472	0.5143593	1	glycolysis	YJL052W	YJR009C	YGR254W
GO:0006094	0.00453721	3/472	0.3671812	1	gluconeogenesis	YJL052W	YJR009C	YGR254W
GO:0006412	[illegible]	[illegible]	[illegible]	1	protein biosynthesis	YHR141C	YGL189C	YHR010W
GO:0001403	0.00544465	2/472	0.8088722	1	invasive growth (sensu Saccharomyces)	YBR083W	YBL016W	
GO:0042273	0.00362976	2/472	0.5732932	1	ribosomal large subunit biogenesis	YGR103W	YPR016C	
GO:0006413	0.01088929	2/472	0.9881775	1	translational initiation	YEL034W	YER025W	
GO:0045944	0.01270417	2/472	0.9956187	1	positive regulation of transcription from RNA p	YBR083W	YLR256W	
GO:0006656	0.00181488	2/472	0.1832289	1	phosphatidylcholine biosynthesis	YJR073C	YGR157W	
GO:0006333	0.00635209	2/472	0.8761587	1	chromatin assembly or disassembly	YBL003C	YDR224C	
GO:0006365	0.00725953	2/472	0.9209652	1	35S primary transcript processing	YOR310C	YLR197W	
GO:0046688	0.00272232	2/472	0.3931093	1	response to copper ion	YHR053C	YHR055C	
GO:0006113	0.00181488	2/472	0.1832289	1	fermentation	YOL086C	YMR303C	
GO:0000004	0.04900181	2/472	1	1	biological_process unknown	YOL019W	YDR346C	
GO:0000154	0.00362976	2/472	0.5732932	1	rRNA modification	YOR310C	YLR197W	
GO:0006350	0.00635209	2/472	0.8761587	1	transcription	YKR034W	YOR344C	
GO:0006950	0.01270417	2/472	0.9956187	1	response to stress	YCR021C	YGR234W	
GO:0006882	0.00362976	2/472	0.5732932	1	zinc ion homeostasis	YOL002C	YJR104C	
GO:0007750	0.00362976	2/472	0.5732932	1	signal transduction during conjugation with cel	YKL178C	YBL016W	
GO:0006646	0.00181488	1/472	NA	NA	phosphatidylethanolamine biosynthesis	YGR007W		
GO:0006801	0.00090744	1/472	NA	NA	superoxide metabolism	YJR104C		

GMRG_Term	Pop_frac	Study_frac	Raw_es	e-score	Description	Contributing_genes	
GO:0003723	0.02722323	7/472	0.99291416	1	RNA binding	YGR159C	YGL031C
GO:0003899	0.00998185	4/472	0.76777443	1	DNA-directed RNA polymerase activity	YOR340C	YNL113W
GO:0004067	0.00362976	4/472	0.03340983	1	asparaginase activity	YLR158C	YLR155C
GO:0003735	0.06079855	27/472	0.71072999	1	structural constituent of ribosome	YHR141C	YGL189C
GO:0003697	0.00453721	2/472	0.71070125	1	single-stranded DNA binding	YGR159C	YDR174W
GO:0003743	0.00635209	2/472	0.87615873	1	translation initiation factor activity	YEL034W	YER025W
GO:0005507	0.00272232	2/472	0.39310929	1	copper ion binding	YHR053C	YHR055C
GO:0003700	0.00544465	2/472	0.80887221	1	transcription factor activity	YKR034W	YOR344C
GO:0016887	0.01905626	2/472	0.99988142	1	ATPase activity	YOR092W	YMR173W
GO:0004022	0.00181488	2/472	0.18322891	1	alcohol dehydrogenase activity	YOL086C	YMR303C
GO:0003704	0.00725953	2/472	0.9209652	1	specific RNA polymerase II transcription factor activity	YBR083W	YLR256W
GO:0004365	0.00181488	2/472	0.18322891	1	glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) a	YJL052W	YJR009C
GO:0003677	0.02177858	2/472	0.99997588	1	DNA binding	YBL003C	YDR224C
GO:0005554	0.19419238	12/472	1	1	molecular_function unknown	YCR021C	YPR149W
GO:0004055	0.00090744	1/472	NA	NA	argininosuccinate synthase activity	YOL058W	
GO:0015392	0.00090744	1/472	NA	NA	cytosine-purine permease activity	YER056C	
GO:0016538	0.00362976	1/472	NA	NA	cyclin-dependent protein kinase regulator activity	YGR108W	

Drawbacks of these methods....

- But all of these tools analyze an annotation independently of each other



GMRG_Term	Pop_frac	Study_frac	Raw_es	e-score	Description	Contributing_genes		
GO:0030490	0.01270417	5/472	0.7900583	1	processing of 20S pre-rRNA	YGR103W	YOR310C	YDL148C
GO:0007047	0.02631579	5/472	0.9992625	1	cell wall organization and biogenesis	YHL028W	YKL096W	YLR300W
GO:0006530	0.00362976	4/472	0.0334098	1	asparagine catabolism	YLR158C	YLR155C	YLR160C
GO:0006995	0.00362976	4/472	0.0334098	1	cellular response to nitrogen starvation	YLR158C	YLR155C	YLR160C
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

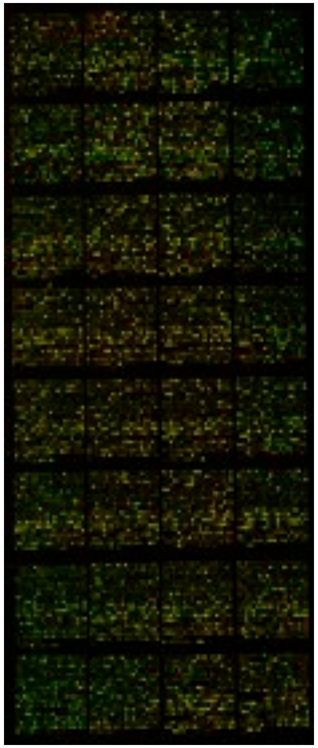
GMRG_Term	Pop_frac	Study_frac	Raw_es	e-score	Description	Contributing_genes	
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
GO:0005732	0.01270417	5/472	0.79005834	1	small nucleola	YOR310C	YDL148C
GO:0005811	0.00362976	2/472	0.57329323	1	lipid particle	YJL052W	YJR009C
GO:0000788	0.00272232	2/472	0.39310929	1	nuclear nuclec	YBL003C	YDR224C

Gene Annotation Co-ocurrence discovery

10000 (*N*) genes analyzed

&

***n* genes are significantly over-expressed**

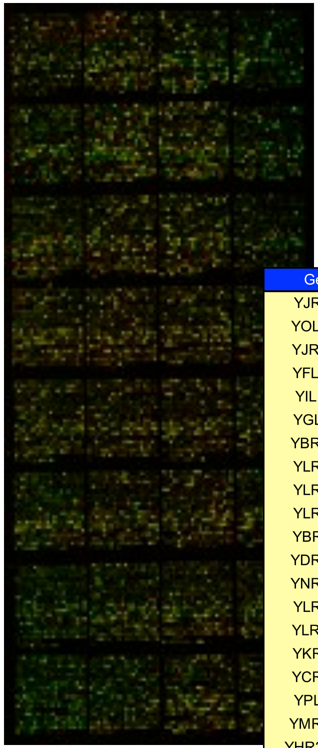


Gene Annotation Co-ocurrence discovery

10000 (*N*) genes analyzed

&

***n* genes are significantly over-expressed**



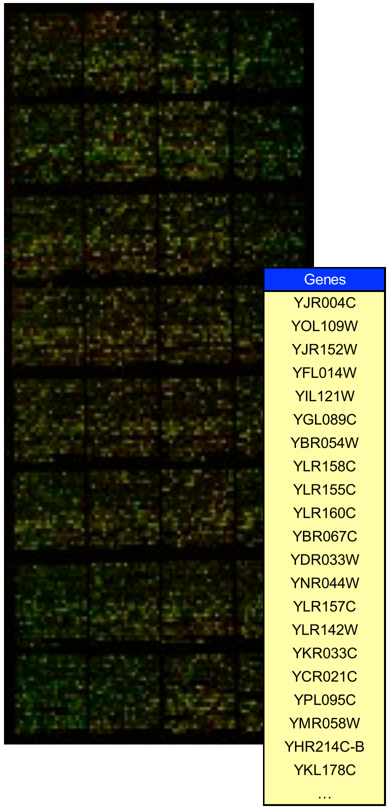
Genes
YJR004C
YOL109W
YJR152W
YFL014W
YIL121W
YGL089C
YBR054W
YLR158C
YLR155C
YLR160C
YBR067C
YDR033W
YNR044W
YLR157C
YLR142W
YKR033C
YCR021C
YPL095C
YMR058W
YHR214C-B
YKL178C
…

Gene Annotation Co-ocurrence discovery

10000 (*N*) genes analyzed

&

***n* genes are significantly over-expressed**



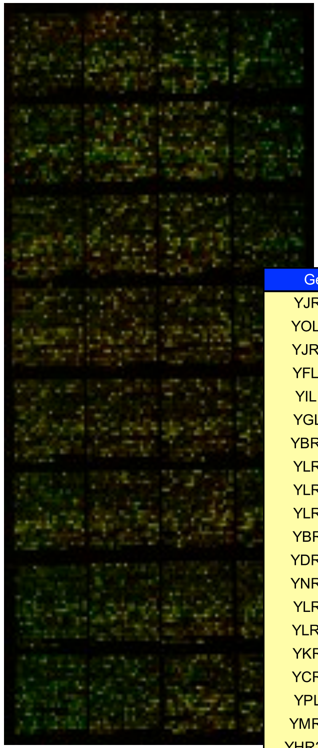
Genes
YJR004C
YOL109W
YJR152W
YFL014W
YIL121W
YGL089C
YBR054W
YLR158C
YLR155C
YLR160C
YBR067C
YDR033W
YNR044W
YLR157C
YLR142W
YKR033C
YCR021C
YPL095C
YMR058W
YHR214C-B
YKL178C
…

Gene Annotation Co-ocurrence discovery

10000 (*N*) genes analyzed

&

n genes are significantly over-expressed



Genes
YJR004C
YOL109W
YJR152W
YFL014W
YIL121W
YGL089C
YBR054W
YLR158C
YLR155C
YLR160C
YBR067C
YDR033W
YNR044W
YLR157C
YLR142W
YKR033C
YCR021C
YPL095C
YMR058W
YHR214C-B
YKL178C
…

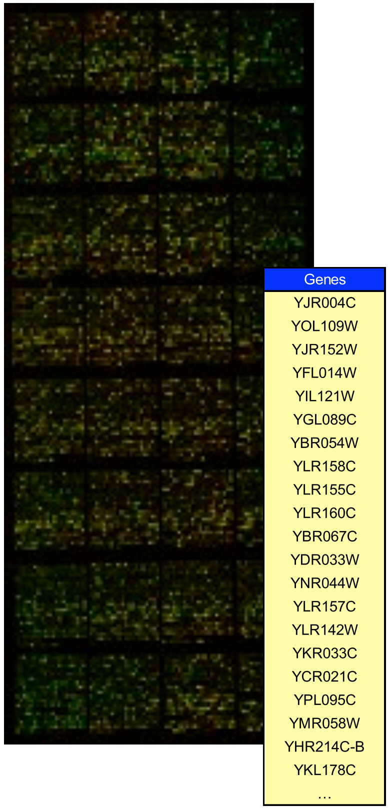
the Gene Ontology

Gene Annotation Co-ocurrence discovery

10000 (*N*) genes analyzed

&

n genes are significantly over-expressed



Genes
YJR004C
YOL109W
YJR152W
YFL014W
YIL121W
YGL089C
YBR054W
YLR158C
YLR155C
YLR160C
YBR067C
YDR033W
YNR044W
YLR157C
YLR142W
YKR033C
YCR021C
YPL095C
YMR058W
YHR214C-B
YKL178C
…

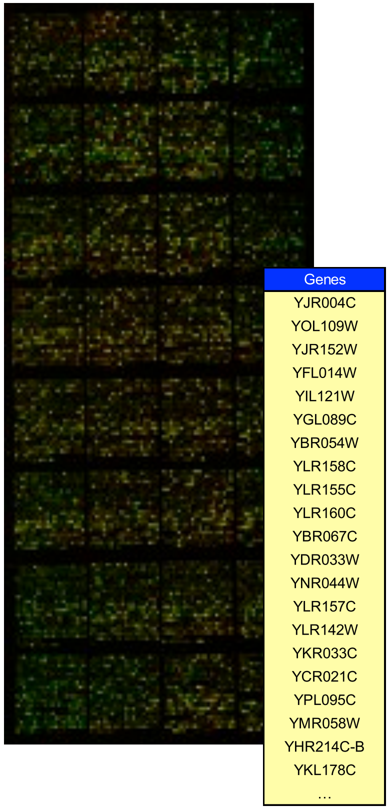
Find combinations of terms that appear in at least *x* genes

Gene Annotation Co-ocurrence discovery

10000 (*N*) genes analyzed

&

***n* genes are significantly over-expressed**



Genes
YJR004C
YOL109W
YJR152W
YFL014W
YIL121W
YGL089C
YBR054W
YLR158C
YLR155C
YLR160C
YBR067C
YDR033W
YNR044W
YLR157C
YLR142W
YKR033C
YCR021C
YPL095C
YMR058W
YHR214C-B
YKL178C
…

Find combinations of terms that appear in at least *x* genes

x* genes with a term/s combination in *n

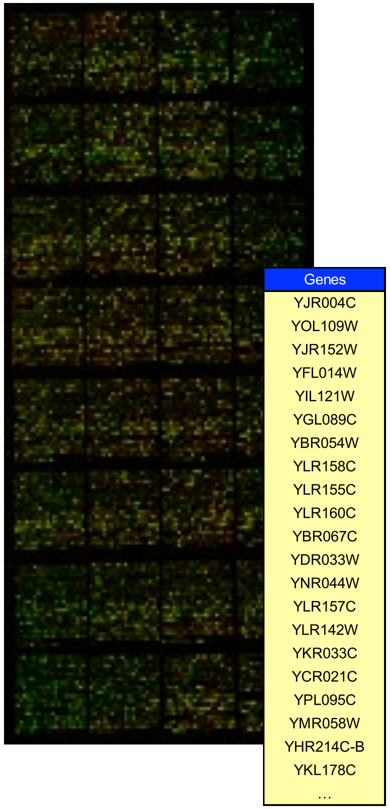
M* genes with a term/s combination in *N

Gene Annotation Co-ocurrence discovery

10000 (*N*) genes analyzed

&

n genes are significantly over-expressed



Genes
YJR004C
YOL109W
YJR152W
YFL014W
YIL121W
YGL089C
YBR054W
YLR158C
YLR155C
YLR160C
YBR067C
YDR033W
YNR044W
YLR157C
YLR142W
YKR033C
YCR021C
YPL095C
YMR058W
YHR214C-B
YKL178C
…

the Gene Ontology

Find combinations of terms that appear in at least *x* genes

x* genes with a term/s combination in *n
M* genes with a term/s combination in *N

Probability of having *x* of *n* genes having an annotation to a GO term, given that in the reference list *M* of *N* genes have that annotation

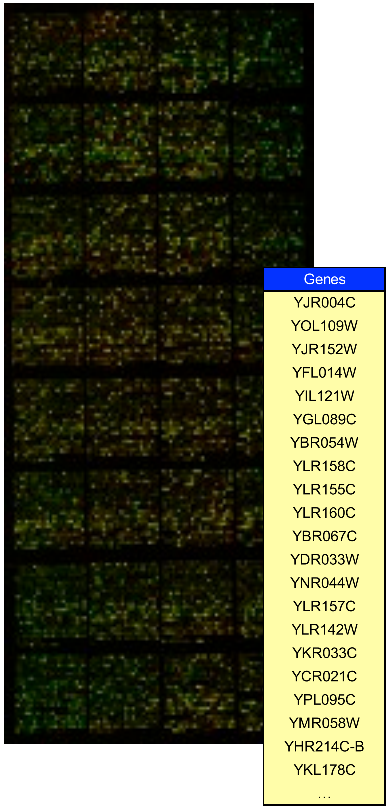
$$P = \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}}$$

Gene Annotation Co-ocurrence discovery

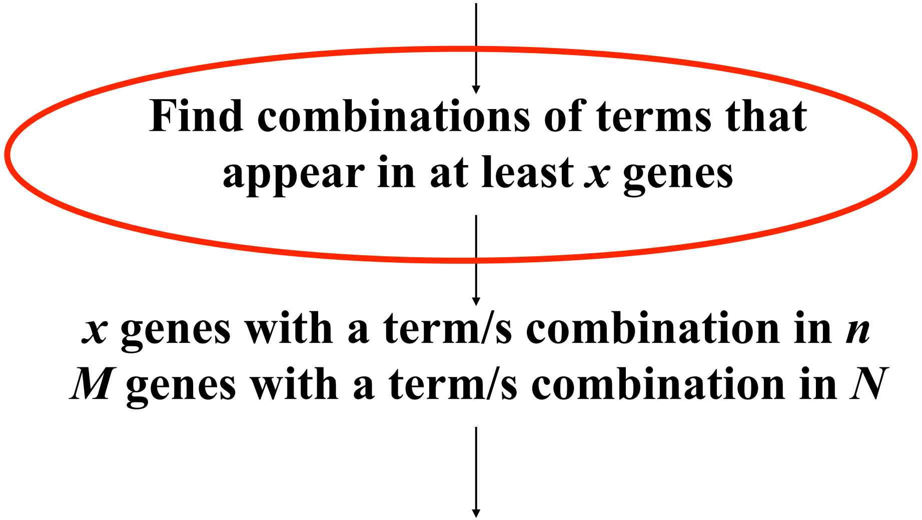
10000 (*N*) genes analyzed

&

***n* genes are significantly over-expressed**



Genes
YJR004C
YOL109W
YJR152W
YFL014W
YIL121W
YGL089C
YBR054W
YLR158C
YLR155C
YLR160C
YBR067C
YDR033W
YNR044W
YLR157C
YLR142W
YKR033C
YCR021C
YPL095C
YMR058W
YHR214C-B
YKL178C
…



Find combinations of terms that appear in at least *x* genes

x* genes with a term/s combination in *n

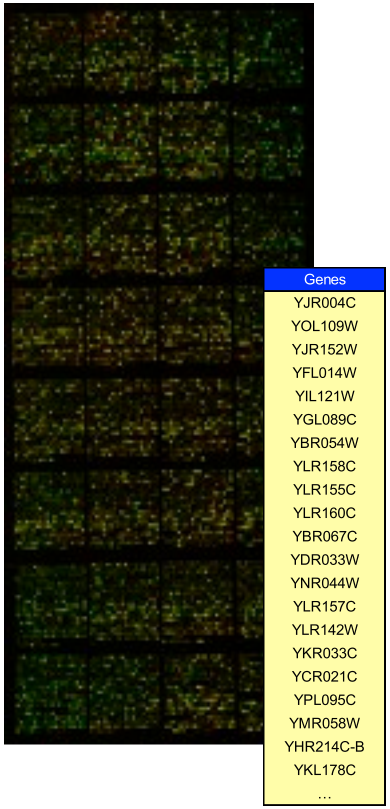
M* genes with a term/s combination in *N

Probability of having *x* of *n* genes having an annotation to a GO term, given that in the reference list *M* of *N* genes have that annotation

$$P = \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}}$$

Gene Annotation Co-ocurrence discovery

10000 (*N*) genes analyzed
&
***n* genes are significantly over-expressed**



Genes
YJR004C
YOL109W
YJR152W
YFL014W
YIL121W
YGL089C
YBR054W
YLR158C
YLR155C
YLR160C
YBR067C
YDR033W
YNR044W
YLR157C
YLR142W
YKR033C
YCR021C
YPL095C
YMR058W
YHR214C-B
YKL178C
...

Find combinations of terms that appear in at least *x* genes

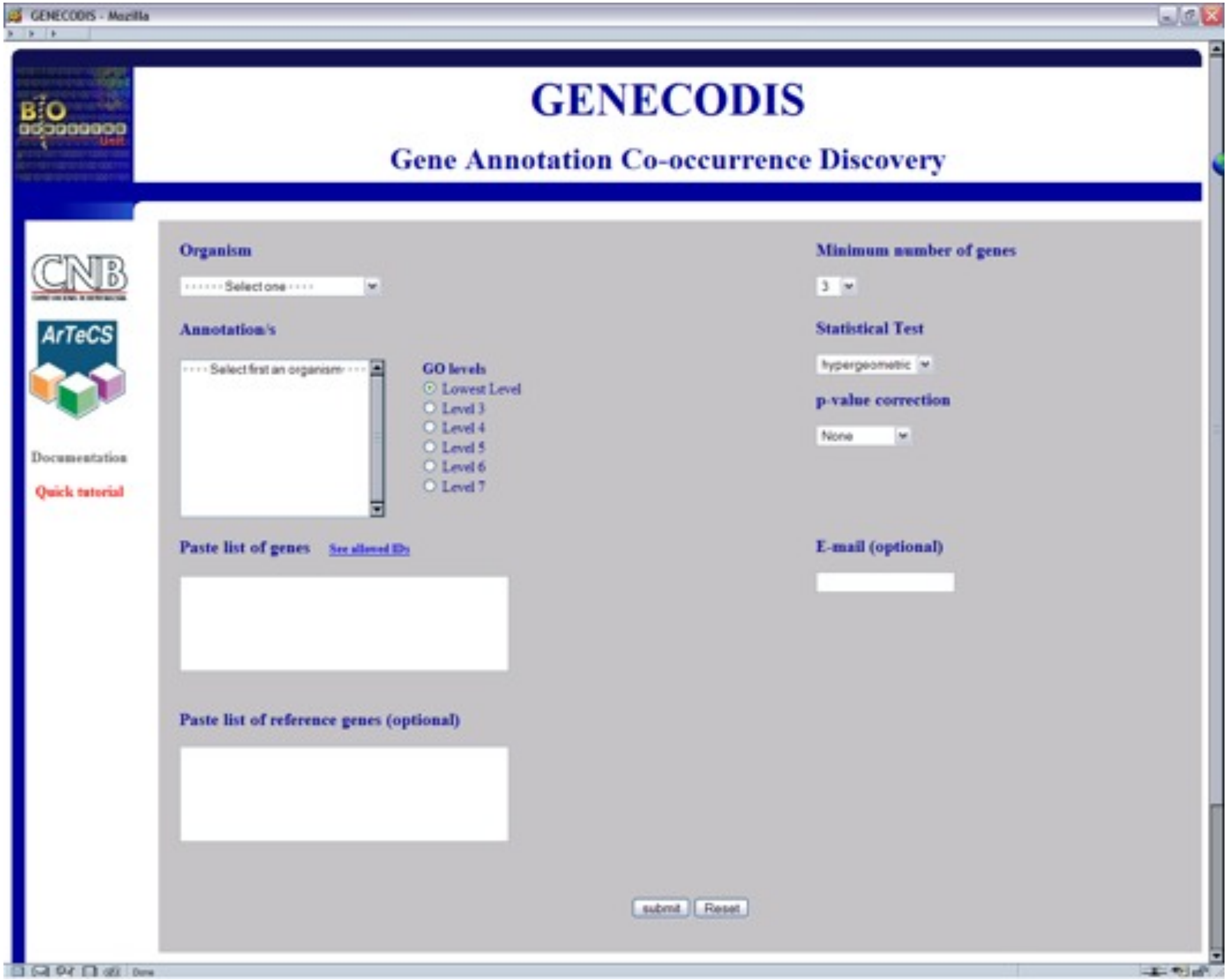
x* genes with a term/s combination in *n
M* genes with a term/s combination in *N

Probability of having *x* of *n* genes having an annotation to a GO term, given that in the reference list *M* of *N* genes have that annotation

$$P = \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}}$$

Pedro Carmona-Saez, Monica Chagoyen, Francisco Tirado, Jose M Carazo and Alberto Pascual-Montano. GENECODIS: A web-based tool for finding significant concurrent annotations in gene lists. Genome Biology. 2007 Jan 4;8(1):R3 Highly accessed

GENECODIS: http://genecodis.dacya.ucm.es/



GENECODIS - Mozilla

GENECODIS

Gene Annotation Co-occurrence Discovery

Organism

······· Select one ·····

Minimum number of genes

3

Annotation/s

···· Select first an organism ····

GO levels

- ● Lowest Level
- ○ Level 3
- ○ Level 4
- ○ Level 5
- ○ Level 6
- ○ Level 7

Statistical Test

hypergeometric

p-value correction

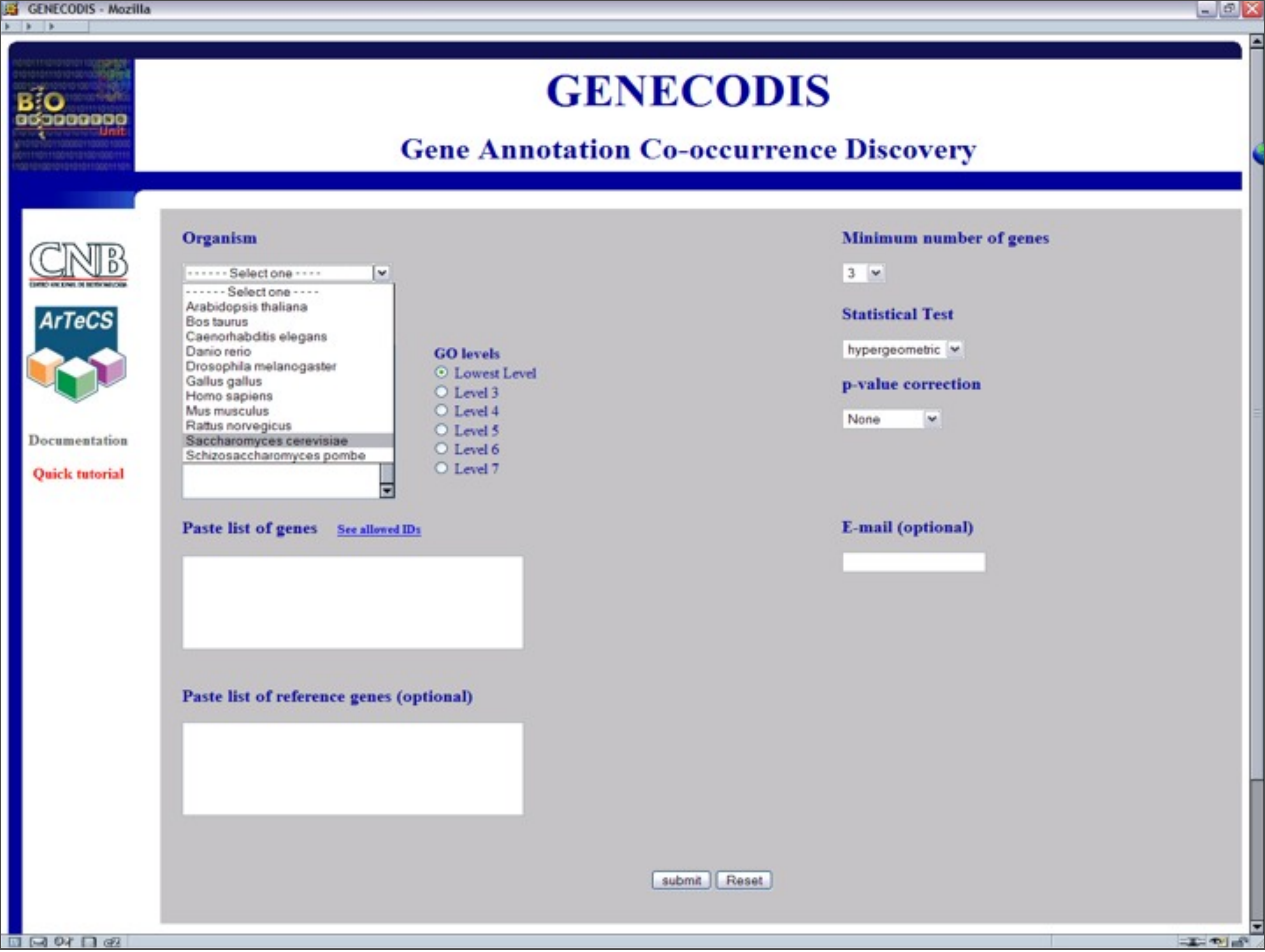
None

Paste list of genes

See allowed IDs

E-mail (optional)

Paste list of reference genes (optional)



GENECODIS - Mozilla

GENECODIS

Gene Annotation Co-occurrence Discovery

Documentation

Quick tutorial

Organism

------ Select one ----

- ------ Select one ----
- Arabidopsis thaliana
- Bos taurus
- Caenorhabditis elegans
- Danio rerio
- Drosophila melanogaster
- Gallus gallus
- Homo sapiens
- Mus musculus
- Rattus norvegicus
- Saccharomyces cerevisiae
- Schizosaccharomyces pombe

GO levels

- ● Lowest Level
- ○ Level 3
- ○ Level 4
- ○ Level 5
- ○ Level 6
- ○ Level 7

Minimum number of genes

3

Statistical Test

hypergeometric

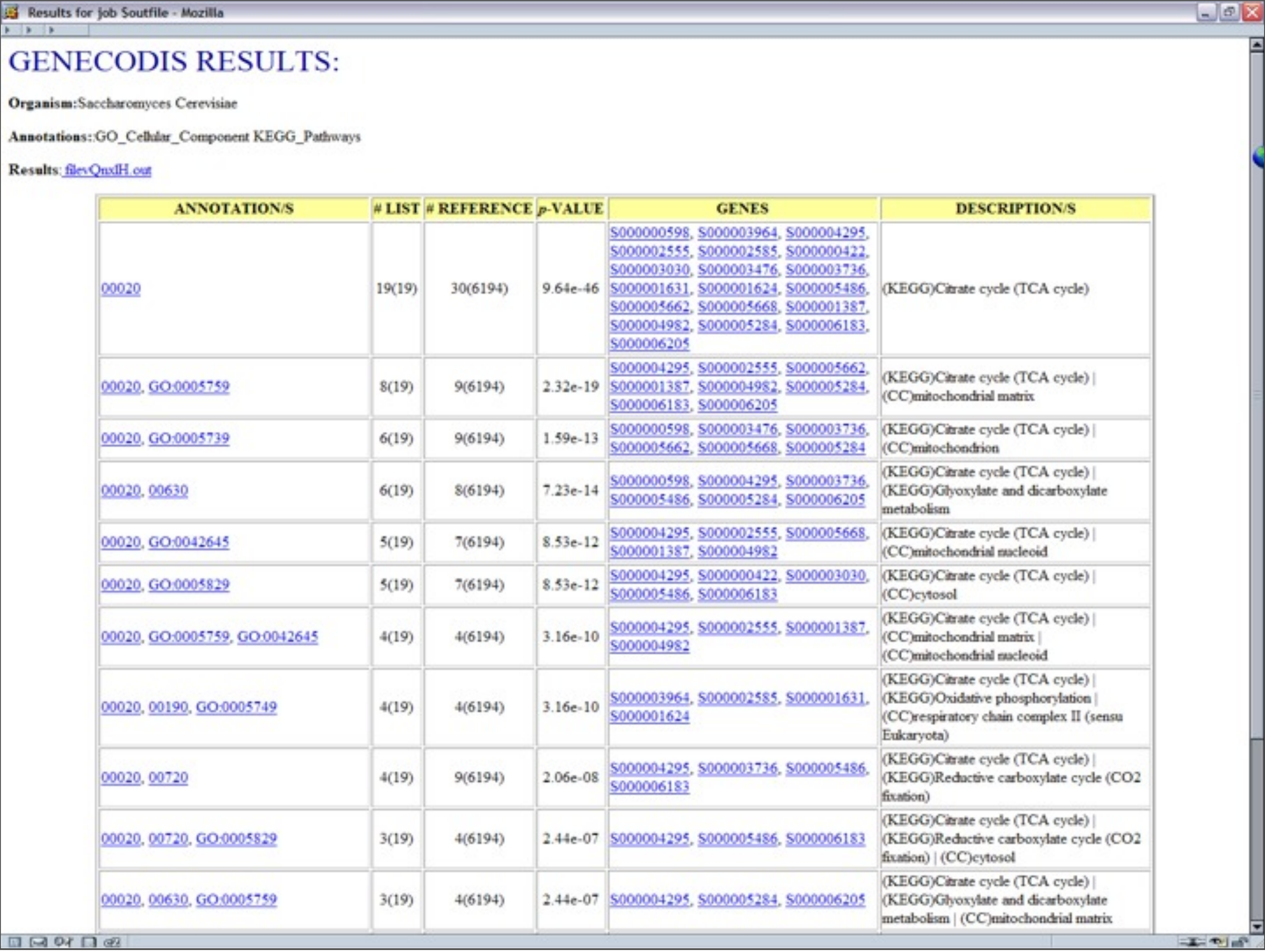
p-value correction

None

Paste list of genes See allowed IDs

E-mail (optional)

Paste list of reference genes (optional)



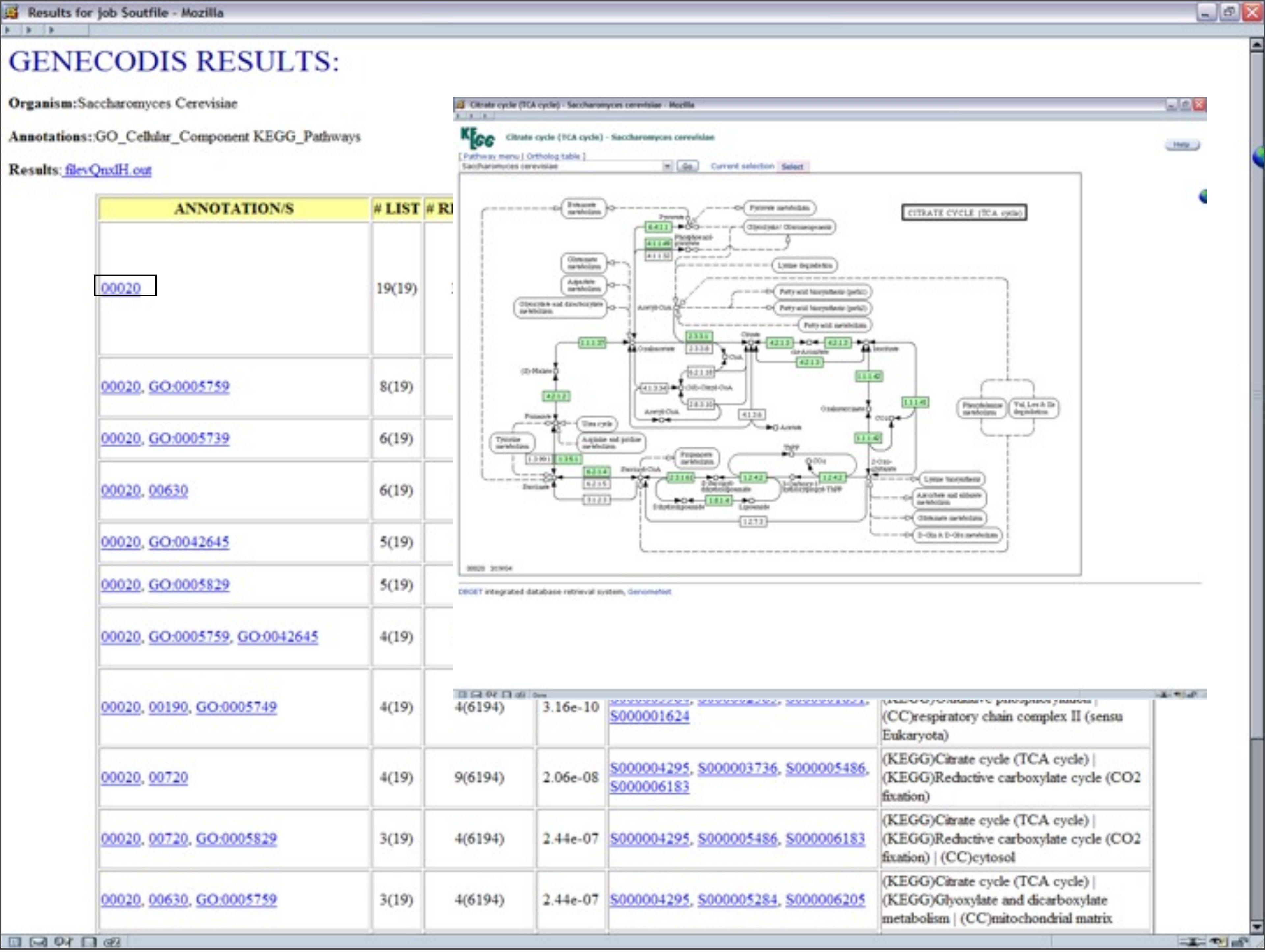
GENECODIS RESULTS:

Organism: Saccharomyces Cerevisiae

Annotations::GO_Cellular_Component KEGG_Pathways

Results: filevQnxlH.out

ANNOTATION/S	# LIST	# REFERENCE	*p*-VALUE	GENES	DESCRIPTION/S
00020	19(19)	30(6194)	9.64e-46	S000000598, S000003964, S000004295, S000002555, S000002585, S000000422, S000003030, S000003476, S000003736, S000001631, S000001624, S000005486, S000005662, S000005668, S000001387, S000004982, S000005284, S000006183, S000006205	(KEGG)Citrate cycle (TCA cycle)
00020, GO:0005759	8(19)	9(6194)	2.32e-19	S000004295, S000002555, S000005662, S000001387, S000004982, S000005284, S000006183, S000006205	(KEGG)Citrate cycle (TCA cycle) | (CC)mitochondrial matrix
00020, GO:0005739	6(19)	9(6194)	1.59e-13	S000000598, S000003476, S000003736, S000005662, S000005668, S000005284	(KEGG)Citrate cycle (TCA cycle) | (CC)mitochondrion
00020, 00630	6(19)	8(6194)	7.23e-14	S000000598, S000004295, S000003736, S000005486, S000005284, S000006205	(KEGG)Citrate cycle (TCA cycle) | (KEGG)Glyoxylate and dicarboxylate metabolism
00020, GO:0042645	5(19)	7(6194)	8.53e-12	S000004295, S000002555, S000005668, S000001387, S000004982	(KEGG)Citrate cycle (TCA cycle) | (CC)mitochondrial nucleoid
00020, GO:0005829	5(19)	7(6194)	8.53e-12	S000004295, S000000422, S000003030, S000005486, S000006183	(KEGG)Citrate cycle (TCA cycle) | (CC)cytosol
00020, GO:0005759, GO:0042645	4(19)	4(6194)	3.16e-10	S000004295, S000002555, S000001387, S000004982	(KEGG)Citrate cycle (TCA cycle) | (CC)mitochondrial matrix | (CC)mitochondrial nucleoid
00020, 00190, GO:0005749	4(19)	4(6194)	3.16e-10	S000003964, S000002585, S000001631, S000001624	(KEGG)Citrate cycle (TCA cycle) | (KEGG)Oxidative phosphorylation | (CC)respiratory chain complex II (sensu Eukaryota)
00020, 00720	4(19)	9(6194)	2.06e-08	S000004295, S000003736, S000005486, S000006183	(KEGG)Citrate cycle (TCA cycle) | (KEGG)Reductive carboxylate cycle (CO2 fixation)
00020, 00720, GO:0005829	3(19)	4(6194)	2.44e-07	S000004295, S000005486, S000006183	(KEGG)Citrate cycle (TCA cycle) | (KEGG)Reductive carboxylate cycle (CO2 fixation) | (CC)cytosol
00020, 00630, GO:0005759	3(19)	4(6194)	2.44e-07	S000004295, S000005284, S000006205	(KEGG)Citrate cycle (TCA cycle) | (KEGG)Glyoxylate and dicarboxylate metabolism | (CC)mitochondrial matrix



GENECODIS RESULTS:

Organism: Saccharomyces Cerevisiae

Annotations::GO_Cellular_Component KEGG_Pathways

Results: filevQnxlH.out

ANNOTATION/S	# LIST	# RI			
00020	19(19)				
00020, GO:0005759	8(19)				
00020, GO:0005739	6(19)				
00020, 00630	6(19)				
00020, GO:0042645	5(19)				
00020, GO:0005829	5(19)				
00020, GO:0005759, GO:0042645	4(19)				
00020, 00190, GO:0005749	4(19)	4(6194)	3.16e-10	S000001624	(CC)respiratory chain complex II (sensu Eukaryota)
00020, 00720	4(19)	9(6194)	2.06e-08	S000004295, S000003736, S000005486, S000006183	(KEGG)Citrate cycle (TCA cycle) | (KEGG)Reductive carboxylate cycle (CO2 fixation)
00020, 00720, GO:0005829	3(19)	4(6194)	2.44e-07	S000004295, S000005486, S000006183	(KEGG)Citrate cycle (TCA cycle) | (KEGG)Reductive carboxylate cycle (CO2 fixation) | (CC)cytosol
00020, 00630, GO:0005759	3(19)	4(6194)	2.44e-07	S000004295, S000005284, S000006205	(KEGG)Citrate cycle (TCA cycle) | (KEGG)Glyoxylate and dicarboxylate metabolism | (CC)mitochondrial matrix

Citrate cycle (TCA cycle) - Saccharomyces cerevisiae

[Pathway menu | Ortholog table]

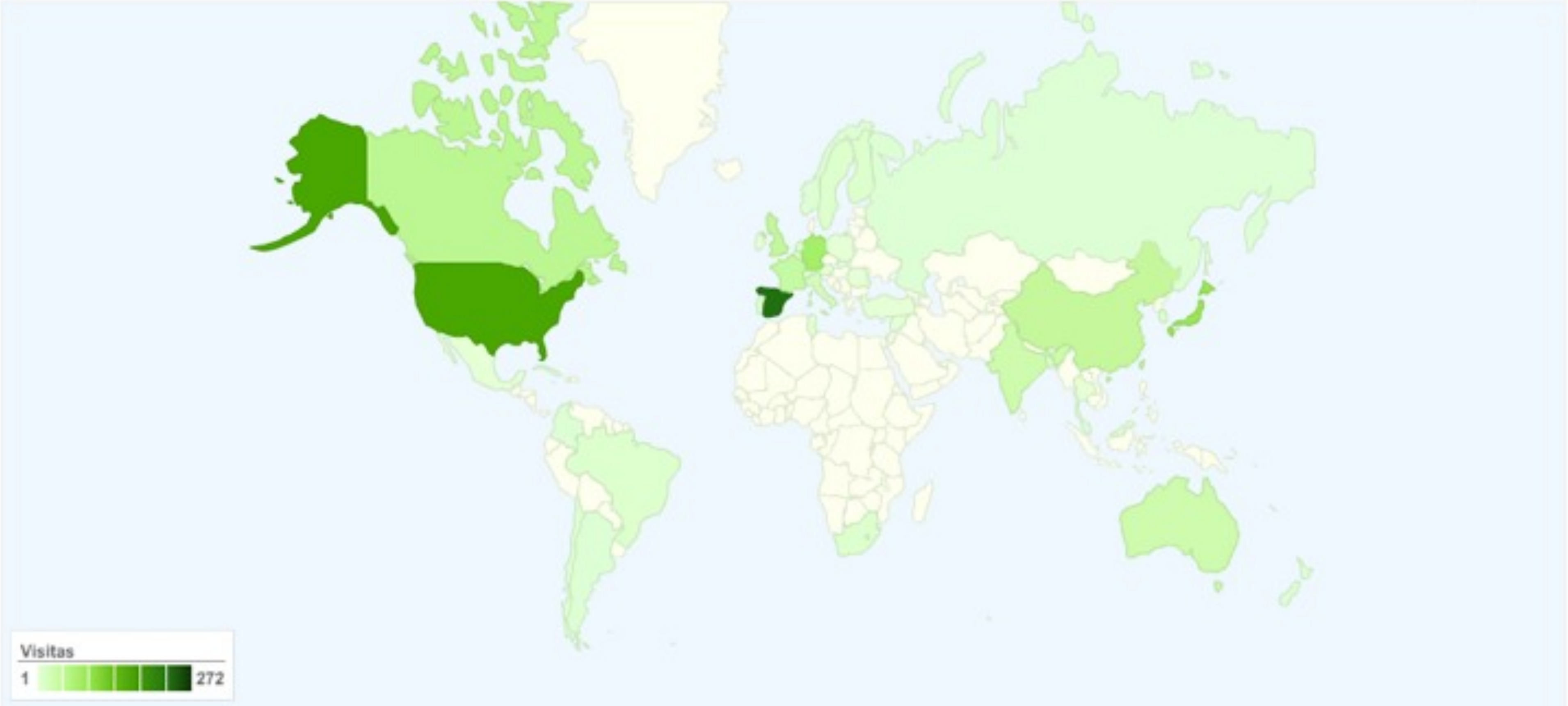
Saccharomyces cerevisiae

Current selection

DBGET integrated database retrieval system, GenomeNet

lunes 25 de julio de 2011

Genecodis statistics (50.000 accesses since Jan 2007!!!!)

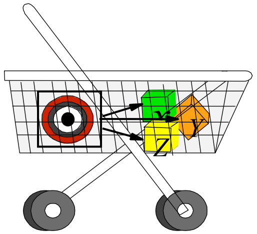


Interpretación de datos de expresión génica:

Anotaciones y análisis de reglas asociativas

DESCUBRIMIENTO DE REGLAS DE ASOCIACION

Detecta conjunto de atributos que co-ocurren frecuentememte, así como Reglas entre ellos



Se ha usado mucho en supermercados para descubrir elementos que se vendían juntos. "Market Basket Analysis"

TID	Items
T1	Bread, Cheese, Apples,Coke
T2	Bread, Apples, Bananas, Peaches
T3	Bread, Milk, Apples, Bananas
T4	Milk, Bananas, Peaches
T5	Apples, Bananas, Sugar, Peaches

Transactions -> Basket

Items-> Products

EJEMPLOS DE REGLAS DE ASOCIACION

$$\overbrace{\text{Apples}}^{\substack{\text{LHS}\\ \text{Antecedente}}} \text{ Þ } \overbrace{\text{Bananas, Peaches}}^{\substack{\text{RHS}\\ \text{Consecuente}}}$$

TID	Items
T1	Bread, Cheese, Apples,Coke
T2	Bread, Apples, Bananas, Peaches
T3	Bread, Milk, Apples, Bananas
T4	Milk, Bananas, Peaches
T5	Apples, Bananas, Sugar, Peaches

Soporte es el porcentaje de registros que contienen una cierta combinación de elementos. Por ejemplo, el 40% de los clientes compra manzanas y melocotones al mismo tiempo.

$$conf = \frac{P(\text{apples} \cup \text{bananas} \cup \text{peaches})}{P(\text{apples})} = 0.5$$

La “confianza” es una medida de la bondad de la regla. Esto es, si el cliente ha comprado un producto, ¿cuál es la probabilidad de que compre otro?

DATOS DE MICRO ARRAYS

					Characteristics	Characteristics	Characteristics	...
Gen	**Function**	**Pathway**	**Promoter**	**Cluster**	**Exp1**	**Exp2**	**Exp3**	...
Gen 1	Cell cycle	ATM Signaling Pathway	Seq1	Cluster 1	-0.47	1.63	0.58	...
Gen 2	Aa Metabolism	Biosynthesis of Lysine	-	Cluster 4	1.01	0.79	0.89	...
Gen 3	Cell cycle	G1/S Chekpoint	Seq1,Seq2	Cluster 1	-0.31	-1.53	-1.29	...
Gen 4	Apoptosis	FAS signaling pathway	Seq3	Cluster 2	0.47	-0.98	-0.19	...
Gen 5	Signal transduction	ATM Signaling Pathway	-	Cluster 1	0.05	0.82	1.82	...
...	...	...	...	...	...	...	...	...

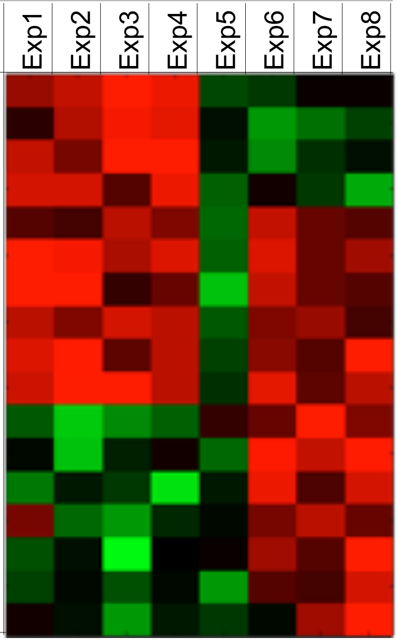
El método puede extraer:

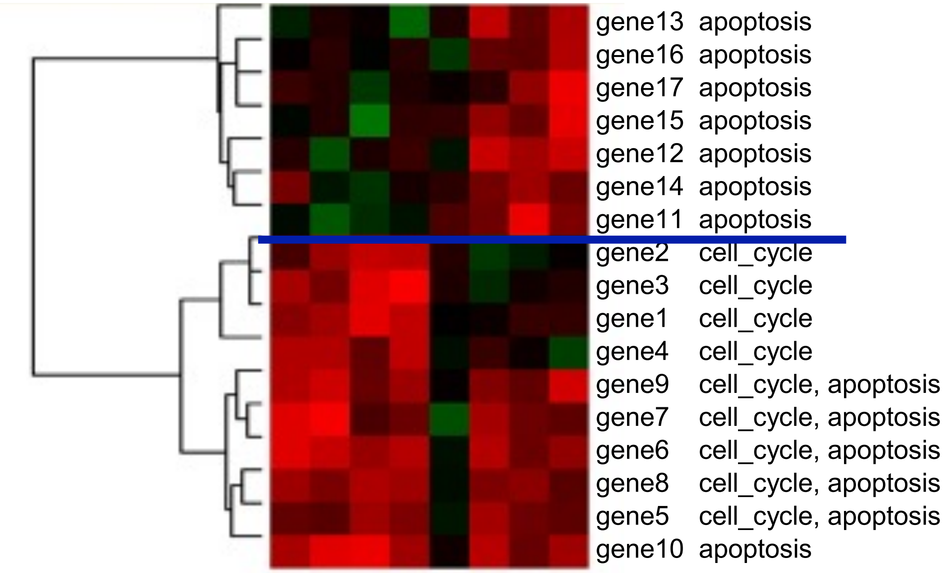
- Reglas entre genes **([+]Gen1->[+]Gen2, [+]Gen3,[-] Gen4)**
- Reglas entre atributos de los genes y condiciones experimentales
- **(Cell Cycle-> [-]Exp1, [+]Exp2)**

- Reglas entre condiciones experimentales **([+]Exp1-> [+]Exp2, [+]Exp3)**
- Reglas entre atributos de los genes **(Cell cycle->Cluster 1)**

ARD AND GENE EXPRESSION DATA ANALYSIS
A NOVEL APPROACH

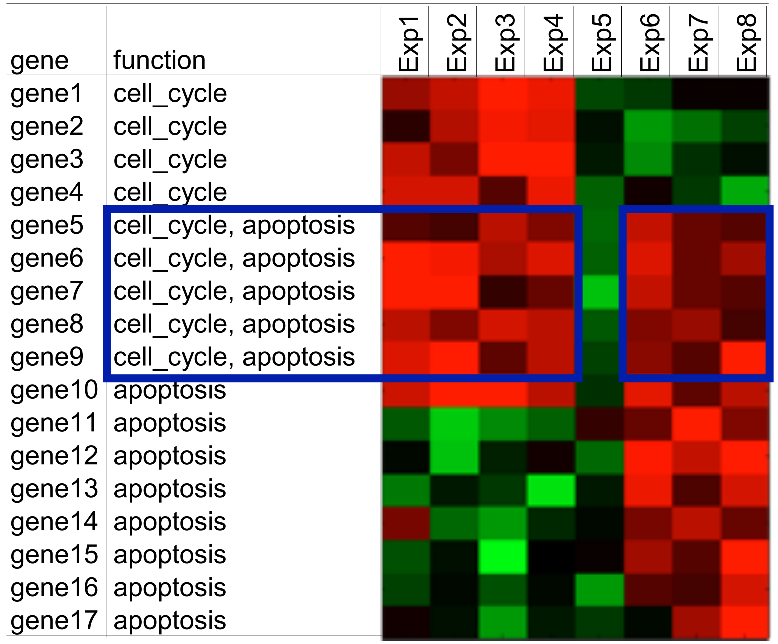
gene	function
gene1	cell_cycle
gene2	cell_cycle
gene3	cell_cycle
gene4	cell_cycle
gene5	cell_cycle, apoptosis
gene6	cell_cycle, apoptosis
gene7	cell_cycle, apoptosis
gene8	cell_cycle, apoptosis
gene9	cell_cycle, apoptosis
gene10	apoptosis
gene11	apoptosis
gene12	apoptosis
gene13	apoptosis
gene14	apoptosis
gene15	apoptosis
gene16	apoptosis
gene17	apoptosis



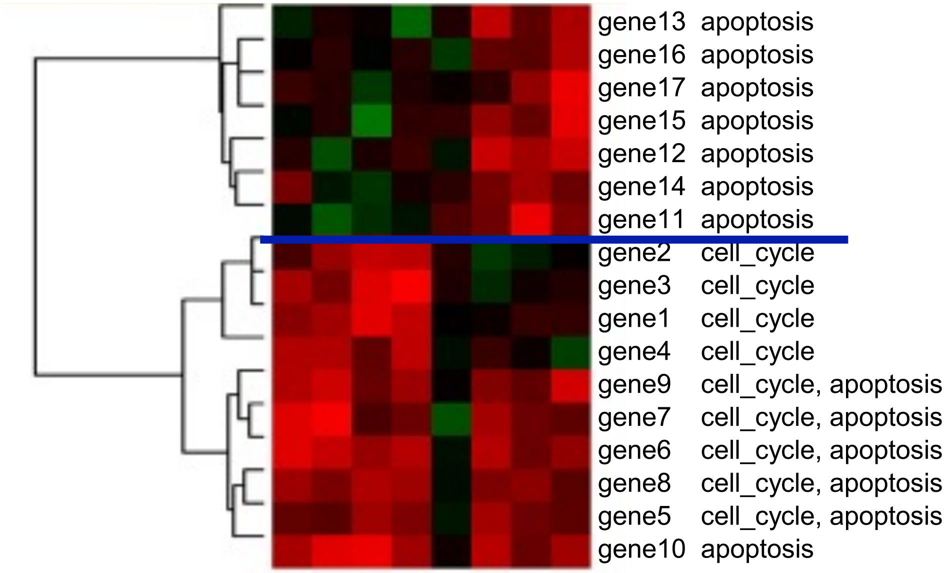


Conf.	supp.	Ante.	Cons.
100	29.412001	apoptosis,cell_cycle	[+]Exp1,[+]Exp2,[+]Exp3,[+]Exp4,[+]Exp6,[+]Exp7,[+]Exp8
100	52.941002	cell_cycle	[+]Exp1,[+]Exp2,[+]Exp3,[+]Exp4
100	76.471001	apoptosis	[+]Exp6,[+]Exp7,[+]Exp8

ARD AND GENE EXPRESSION DATA ANALYSIS
A NOVEL APPROACH



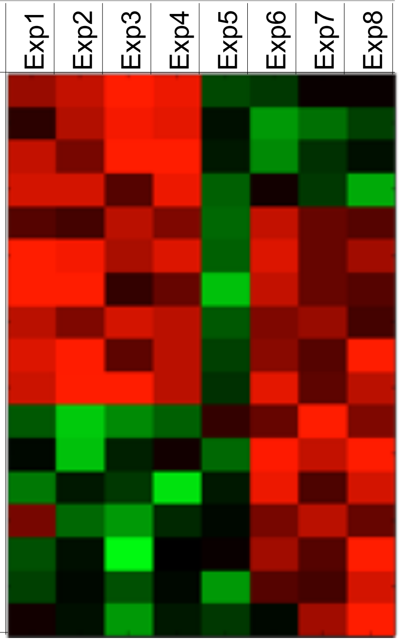
gene	function
gene1	cell_cycle
gene2	cell_cycle
gene3	cell_cycle
gene4	cell_cycle
gene5	cell_cycle, apoptosis
gene6	cell_cycle, apoptosis
gene7	cell_cycle, apoptosis
gene8	cell_cycle, apoptosis
gene9	cell_cycle, apoptosis
gene10	apoptosis
gene11	apoptosis
gene12	apoptosis
gene13	apoptosis
gene14	apoptosis
gene15	apoptosis
gene16	apoptosis
gene17	apoptosis

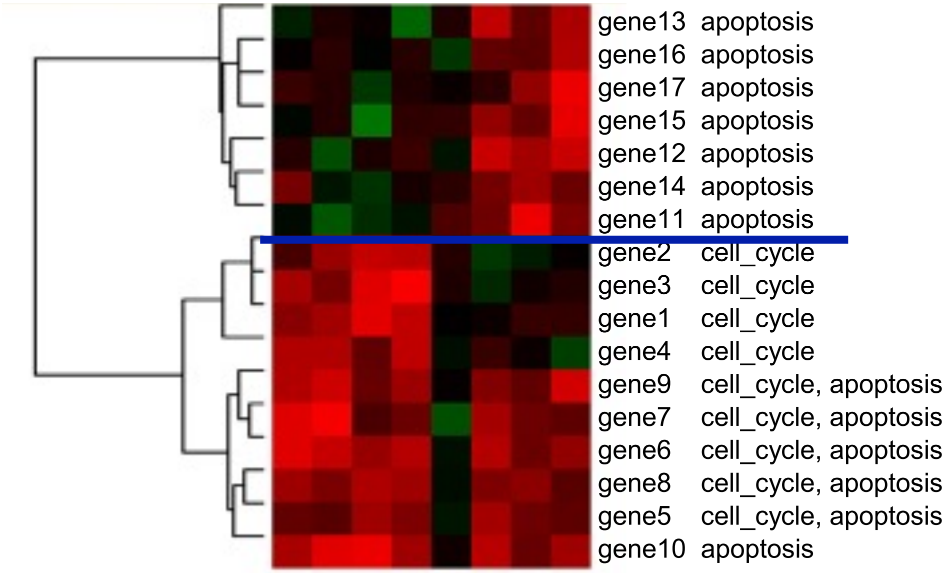


Conf.	supp.	Ante.	Cons.
➡ 100	29.412001	apoptosis,cell_cycle	[+]Exp1,[+]Exp2,[+]Exp3,[+]Exp4,[+]Exp6,[+]Exp7,[+]Exp8
100	52.941002	cell_cycle	[+]Exp1,[+]Exp2,[+]Exp3,[+]Exp4
100	76.471001	apoptosis	[+]Exp6,[+]Exp7,[+]Exp8

ARD AND GENE EXPRESSION DATA ANALYSIS A NOVEL APPROACH

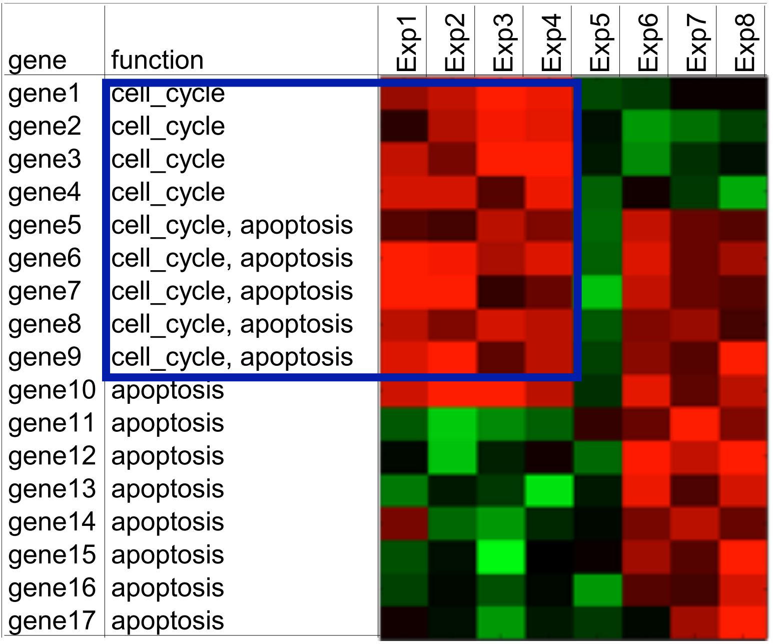
gene	function
gene1	cell_cycle
gene2	cell_cycle
gene3	cell_cycle
gene4	cell_cycle
gene5	cell_cycle, apoptosis
gene6	cell_cycle, apoptosis
gene7	cell_cycle, apoptosis
gene8	cell_cycle, apoptosis
gene9	cell_cycle, apoptosis
gene10	apoptosis
gene11	apoptosis
gene12	apoptosis
gene13	apoptosis
gene14	apoptosis
gene15	apoptosis
gene16	apoptosis
gene17	apoptosis



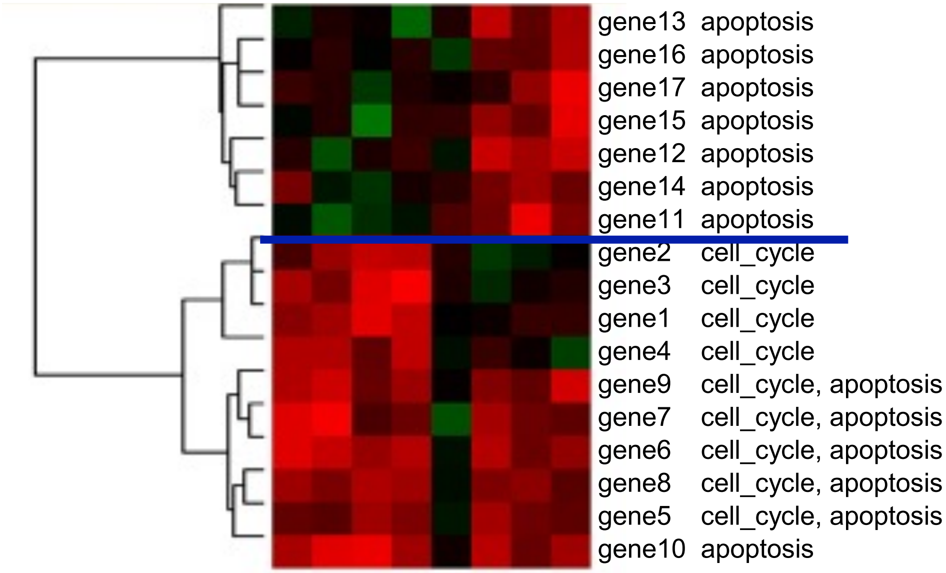


Conf.	supp.	Ante.	Cons.
100	29.412001	apoptosis,cell_cycle	[+]Exp1,[+]Exp2,[+]Exp3,[+]Exp4,[+]Exp6,[+]Exp7,[+]Exp8
100	52.941002	cell_cycle	[+]Exp1,[+]Exp2,[+]Exp3,[+]Exp4
100	76.471001	apoptosis	[+]Exp6,[+]Exp7,[+]Exp8

ARD AND GENE EXPRESSION DATA ANALYSIS A NOVEL APPROACH



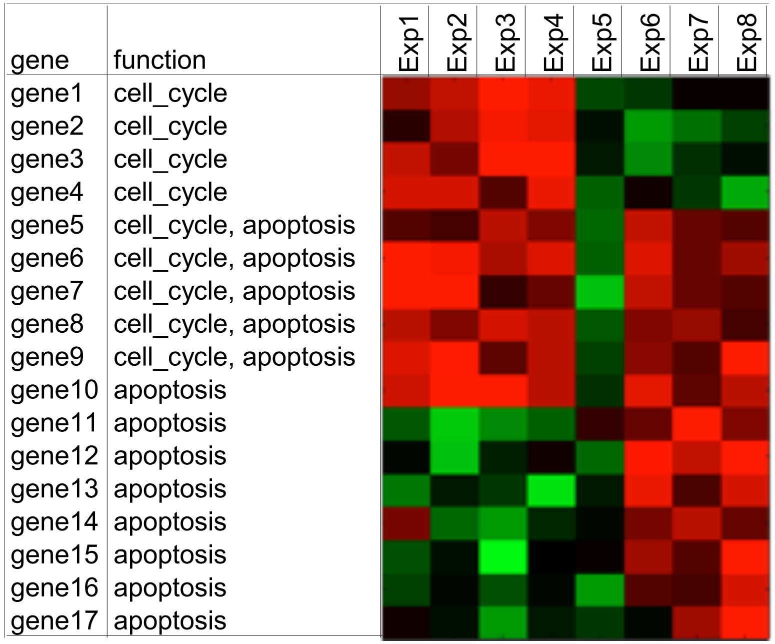
gene	function
gene1	cell_cycle
gene2	cell_cycle
gene3	cell_cycle
gene4	cell_cycle
gene5	cell_cycle, apoptosis
gene6	cell_cycle, apoptosis
gene7	cell_cycle, apoptosis
gene8	cell_cycle, apoptosis
gene9	cell_cycle, apoptosis
gene10	apoptosis
gene11	apoptosis
gene12	apoptosis
gene13	apoptosis
gene14	apoptosis
gene15	apoptosis
gene16	apoptosis
gene17	apoptosis



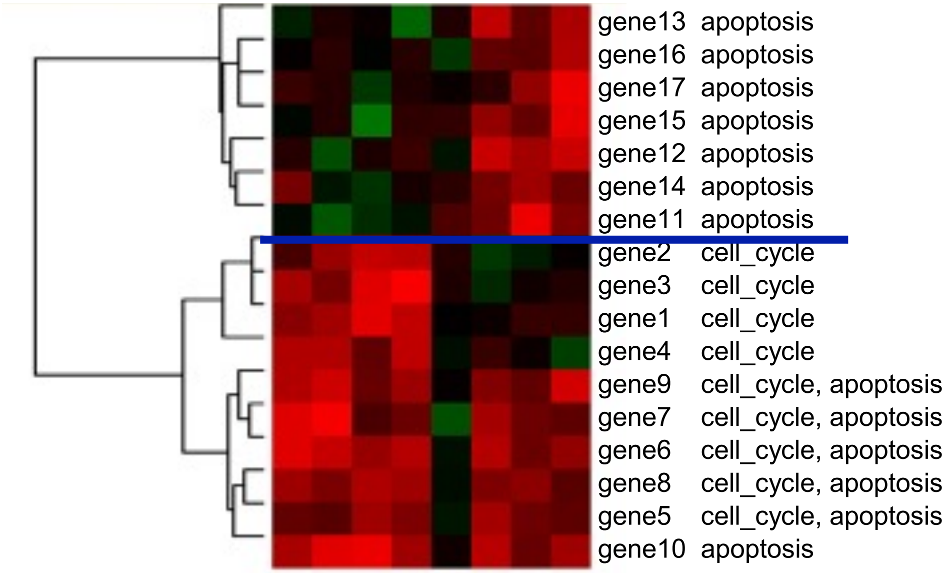
gene13	apoptosis
gene16	apoptosis
gene17	apoptosis
gene15	apoptosis
gene12	apoptosis
gene14	apoptosis
gene11	apoptosis
gene2	cell_cycle
gene3	cell_cycle
gene1	cell_cycle
gene4	cell_cycle
gene9	cell_cycle, apoptosis
gene7	cell_cycle, apoptosis
gene6	cell_cycle, apoptosis
gene8	cell_cycle, apoptosis
gene5	cell_cycle, apoptosis
gene10	apoptosis

Conf.	supp.	Ante.	Cons.
100	29.412001	apoptosis,cell_cycle	[+]Exp1,[+]Exp2,[+]Exp3,[+]Exp4,[+]Exp6,[+]Exp7,[+]Exp8
➡ 100	52.941002	cell_cycle	[+]Exp1,[+]Exp2,[+]Exp3,[+]Exp4
100	76.471001	apoptosis	[+]Exp6,[+]Exp7,[+]Exp8

ARD AND GENE EXPRESSION DATA ANALYSIS
A NOVEL APPROACH

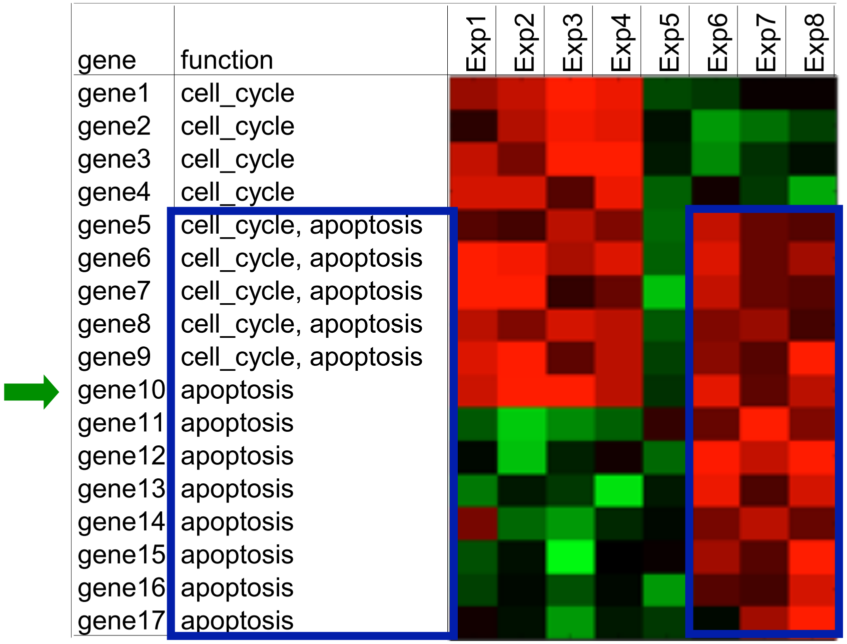


gene	function
gene1	cell_cycle
gene2	cell_cycle
gene3	cell_cycle
gene4	cell_cycle
gene5	cell_cycle, apoptosis
gene6	cell_cycle, apoptosis
gene7	cell_cycle, apoptosis
gene8	cell_cycle, apoptosis
gene9	cell_cycle, apoptosis
gene10	apoptosis
gene11	apoptosis
gene12	apoptosis
gene13	apoptosis
gene14	apoptosis
gene15	apoptosis
gene16	apoptosis
gene17	apoptosis

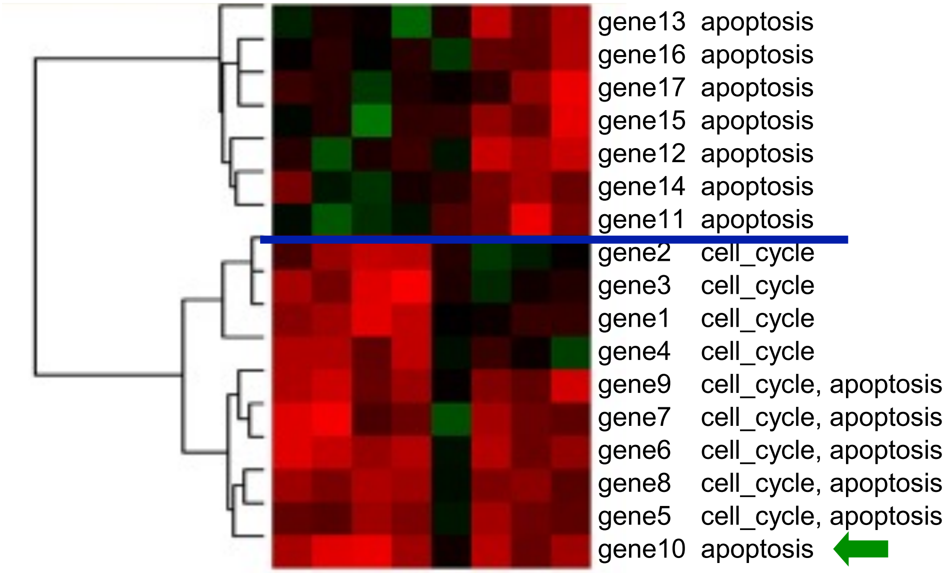


Conf.	supp.	Ante.	Cons.
100	29.412001	apoptosis,cell_cycle	[+]Exp1,[+]Exp2,[+]Exp3,[+]Exp4,[+]Exp6,[+]Exp7,[+]Exp8
100	52.941002	cell_cycle	[+]Exp1,[+]Exp2,[+]Exp3,[+]Exp4
100	76.471001	apoptosis	[+]Exp6,[+]Exp7,[+]Exp8

ARD AND GENE EXPRESSION DATA ANALYSIS
A NOVEL APPROACH



gene	function
gene1	cell_cycle
gene2	cell_cycle
gene3	cell_cycle
gene4	cell_cycle
gene5	cell_cycle, apoptosis
gene6	cell_cycle, apoptosis
gene7	cell_cycle, apoptosis
gene8	cell_cycle, apoptosis
gene9	cell_cycle, apoptosis
gene10	apoptosis
gene11	apoptosis
gene12	apoptosis
gene13	apoptosis
gene14	apoptosis
gene15	apoptosis
gene16	apoptosis
gene17	apoptosis



gene13	apoptosis
gene16	apoptosis
gene17	apoptosis
gene15	apoptosis
gene12	apoptosis
gene14	apoptosis
gene11	apoptosis
gene2	cell_cycle
gene3	cell_cycle
gene1	cell_cycle
gene4	cell_cycle
gene9	cell_cycle, apoptosis
gene7	cell_cycle, apoptosis
gene6	cell_cycle, apoptosis
gene8	cell_cycle, apoptosis
gene5	cell_cycle, apoptosis
gene10	apoptosis

Conf.	supp.	Ante.	Cons.
100	29.412001	apoptosis,cell_cycle	[+]Exp1,[+]Exp2,[+]Exp3,[+]Exp4,[+]Exp6,[+]Exp7,[+]Exp8
100	52.941002	cell_cycle	[+]Exp1,[+]Exp2,[+]Exp3,[+]Exp4
100	76.471001	apoptosis	[+]Exp6,[+]Exp7,[+]Exp8

ASSOCIATION RULES DISCOVERY SOFTWARE

Engene™

Gene-Expression Data Processing and Exploratory Data Analysis

http://www.engene.cnb.uam.es

(García De La Nava, *et al.*, Bioinformatics 2003)

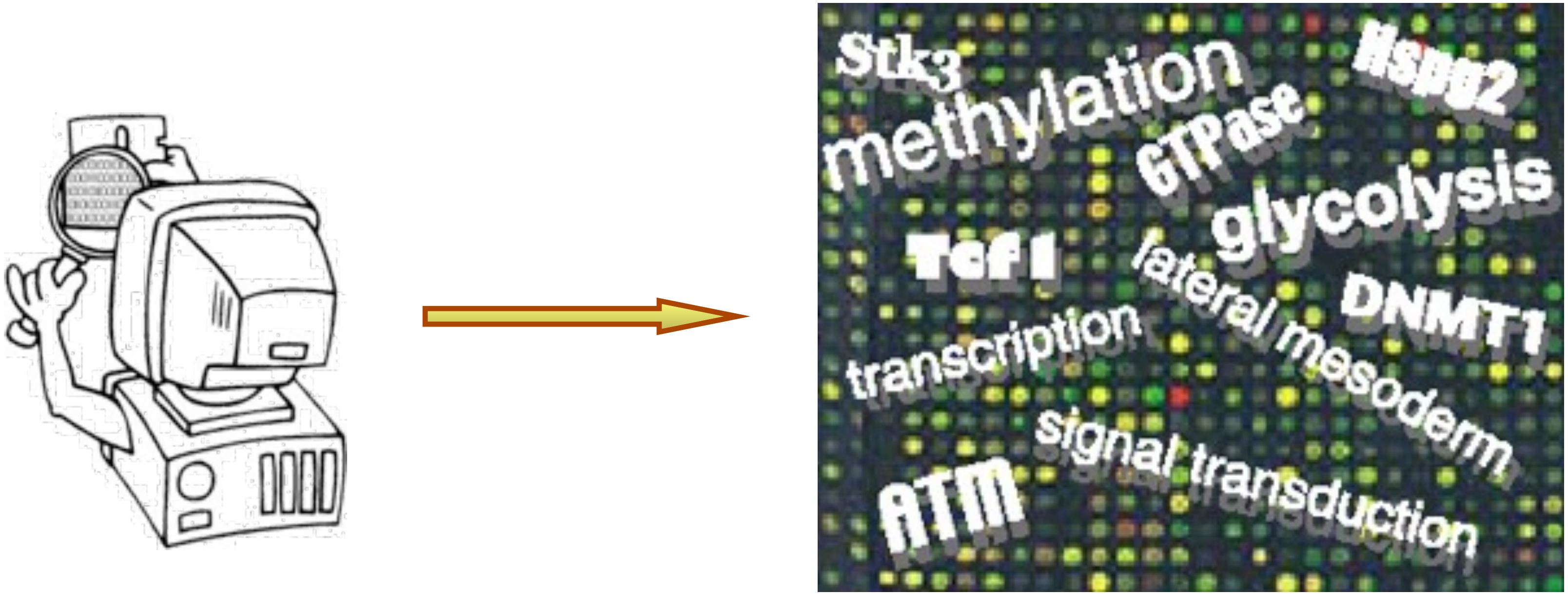
CNB CENTRO NACIONAL DE BIOTECNOLOGIA

Integromics Information Technologies for Life Sciences

Interpretation of gene expression using PubMed:

El caso de “NMF”

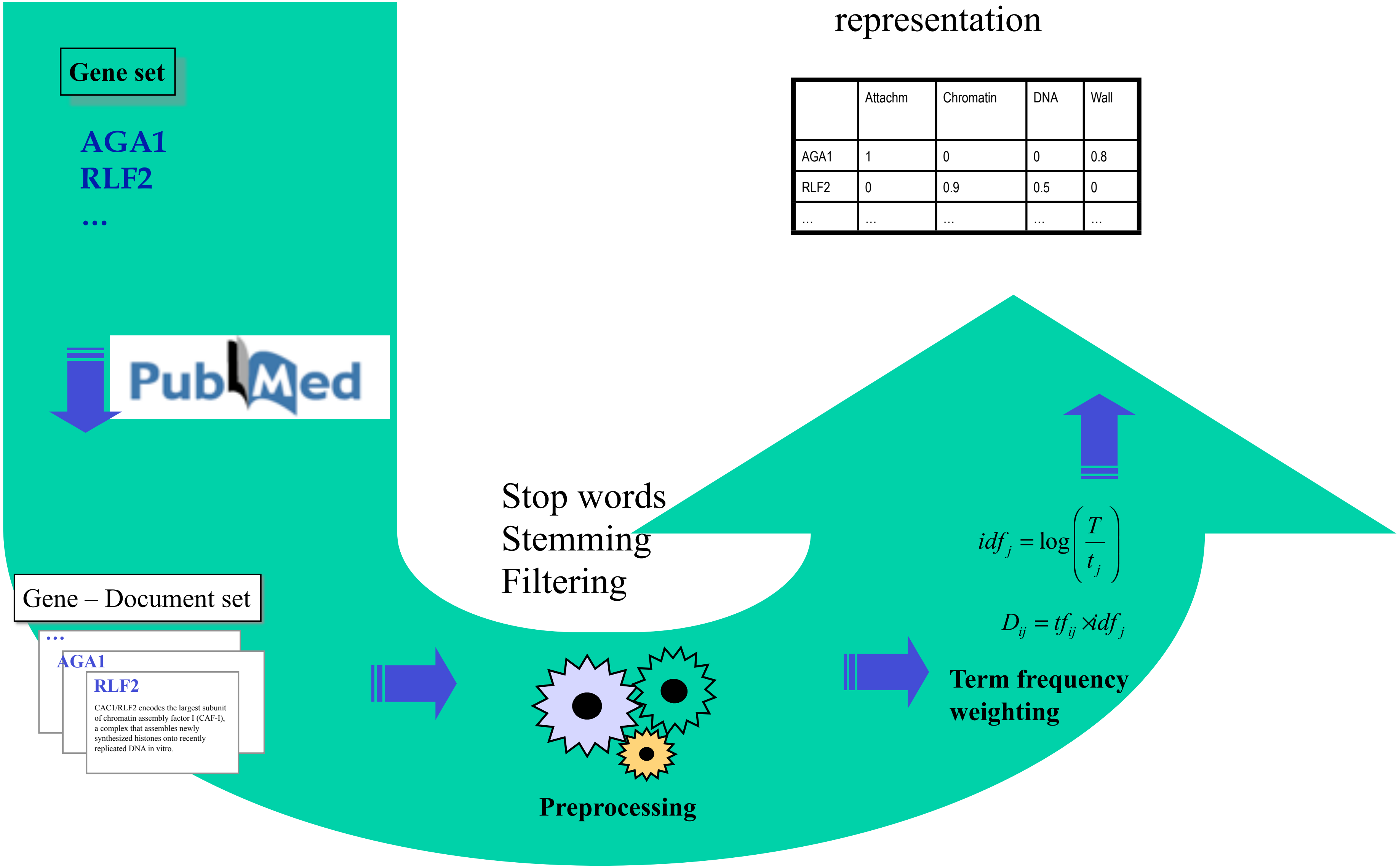
The “maths” Beauty: Text Mining of biomedical data with nsNMF



Discovering semantic features in the literature: a foundation for building functional associations

Chagoyen M, Carmona-Saez P, Shatkay H, Carazo JM, Pascual-Montano A.
Discovering semantic features in the literature: a foundation for building functional associations
BMC Bioinformatics. 2006 Jan 26;7(1):41 Highly accessed

Document processing



Vector space representation

	Attachm	Chromatin	DNA	Wall
AGA1	1	0	0	0.8
RLF2	0	0.9	0.5	0
…	…	…	…	…

Gene set

AGA1
RLF2
…

PubMed

Gene – Document set

…
AGA1
RLF2
CAC1/RLF2 encodes the largest subunit of chromatin assembly factor I (CAF-I), a complex that assembles newly synthesized histones onto recently replicated DNA in vitro.

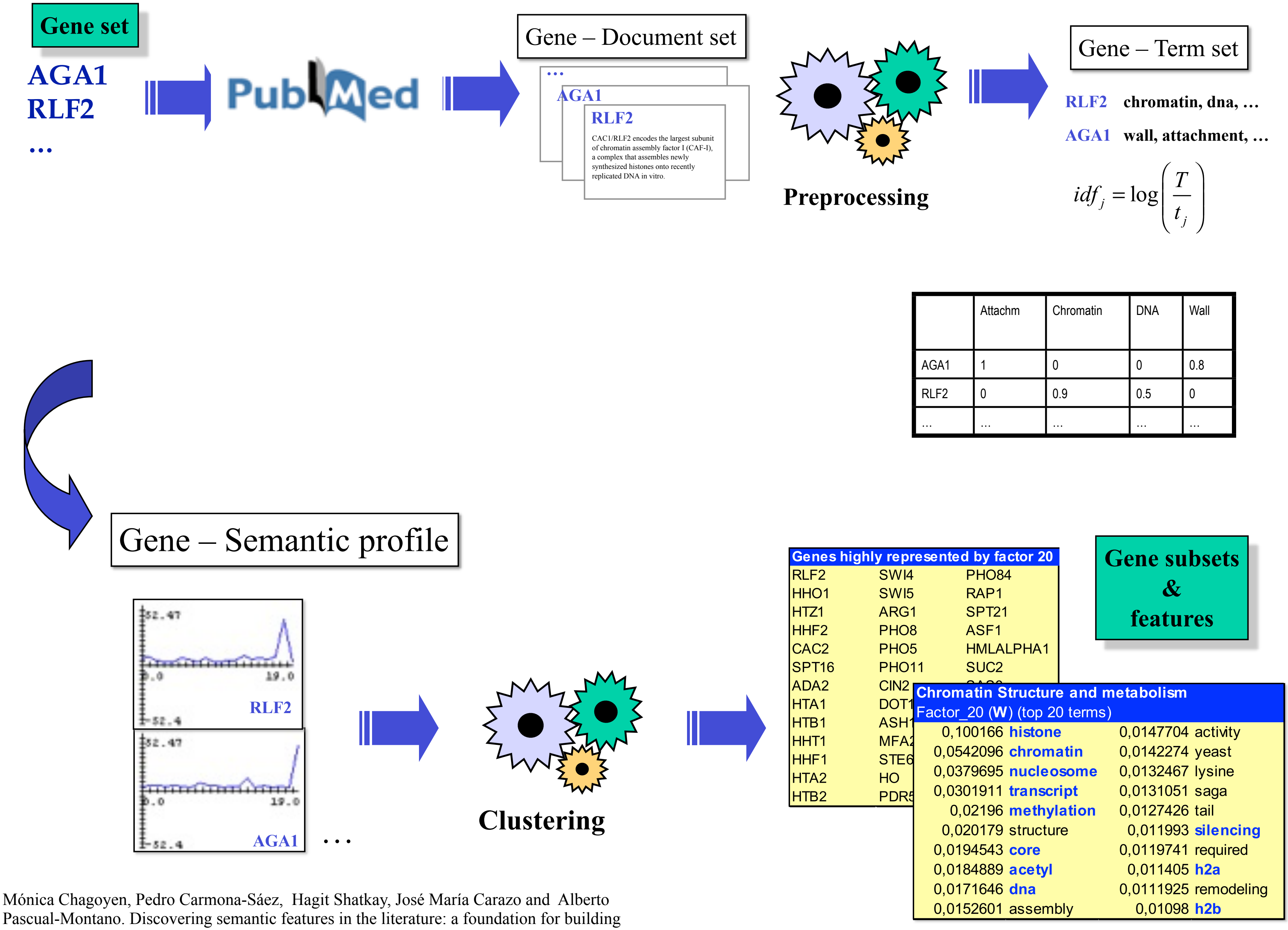
Stop words
Stemming
Filtering

Preprocessing

$$idf_j = \log\left(\frac{T}{t_j}\right)$$

$$D_{ij} = tf_{ij} \times idf_j$$

Term frequency weighting



CAC1/RLF2 encodes the largest subunit of chromatin assembly factor I (CAF-I), a complex that assembles newly synthesized histones onto recently replicated DNA in vitro.

RLF2	chromatin, dna, …
AGA1	wall, attachment, …

$$idf_j = \log\left(\frac{T}{t_j}\right)$$

	Attachm	Chromatin	DNA	Wall
AGA1	1	0	0	0.8
RLF2	0	0.9	0.5	0
…	…	…	…	…

Genes highly represented by factor 20

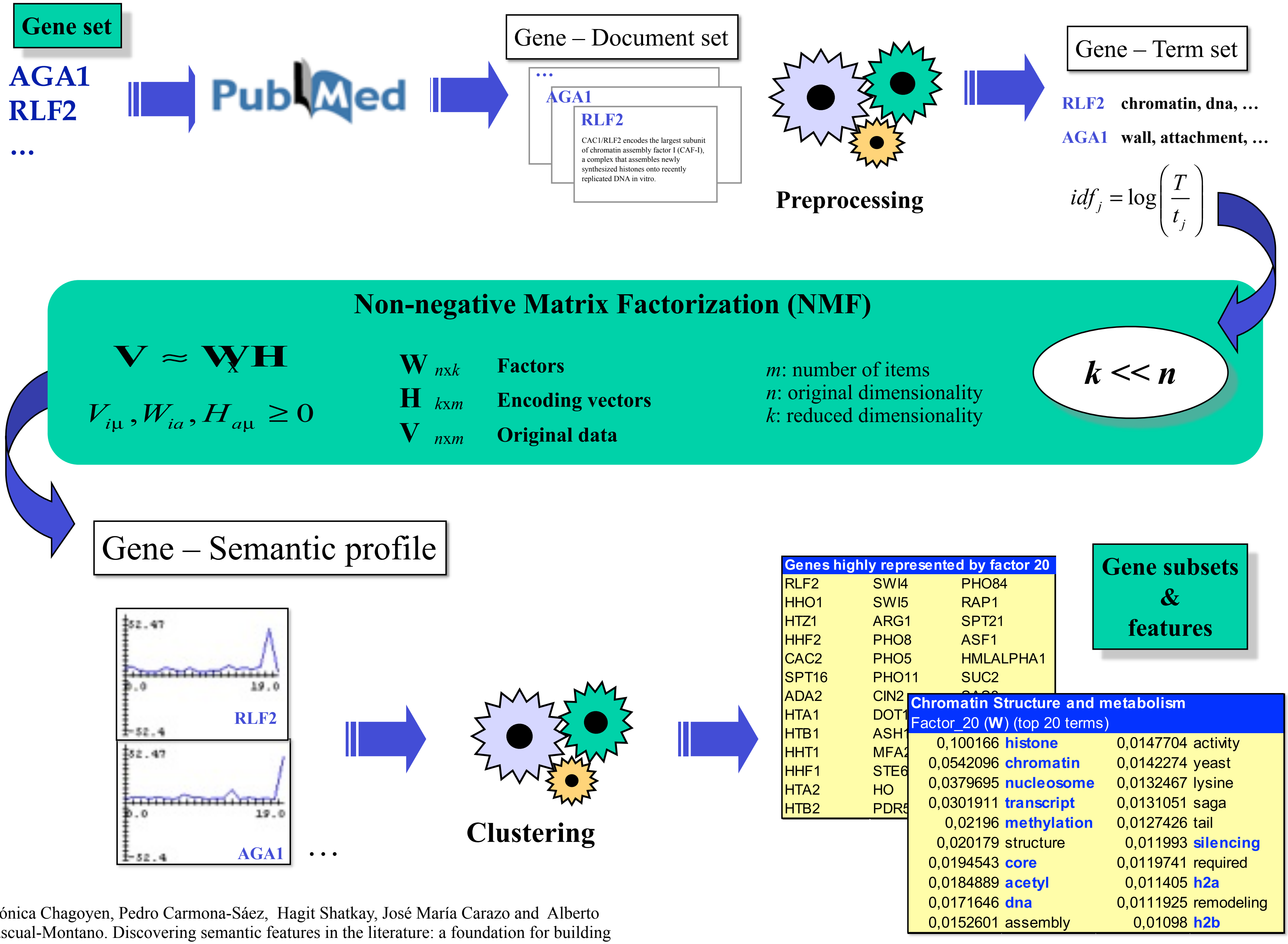
RLF2	SWI4	PHO84
HHO1	SWI5	RAP1
HTZ1	ARG1	SPT21
HHF2	PHO8	ASF1
CAC2	PHO5	HMLALPHA1
SPT16	PHO11	SUC2
ADA2	CIN2	[illegible]
HTA1	DOT1	
HTB1	ASH1	
HHT1	MFA2	
HHF1	STE6	
HTA2	HO	
HTB2	PDR5	

Chromatin Structure and metabolism

Factor_20 (**W**) (top 20 terms)

0,100166	**histone**	0,0147704	activity
0,0542096	**chromatin**	0,0142274	yeast
0,0379695	**nucleosome**	0,0132467	lysine
0,0301911	**transcript**	0,0131051	saga
0,02196	**methylation**	0,0127426	tail
0,020179	structure	0,011993	**silencing**
0,0194543	**core**	0,0119741	required
0,0184889	**acetyl**	0,011405	**h2a**
0,0171646	**dna**	0,0111925	remodeling
0,0152601	assembly	0,01098	**h2b**

Mónica Chagoyen, Pedro Carmona-Sáez, Hagit Shatkay, José María Carazo and Alberto Pascual-Montano. Discovering semantic features in the literature: a foundation for building functional associations. BMC Bioinformatics 2006, 7:41



$$idf_j = \log\left(\frac{T}{t_j}\right)$$

$$\mathbf{V} \approx \mathbf{WH}$$

$$V_{i\mu}, W_{ia}, H_{a\mu} \geq 0$$

$\mathbf{W}$	nxk	**Factors**
$\mathbf{H}$	kxm	**Encoding vectors**
$\mathbf{V}$	nxm	**Original data**

m: number of items
n: original dimensionality
k: reduced dimensionality

Genes highly represented by factor 20		
RLF2	SWI4	PHO84
HHO1	SWI5	RAP1
HTZ1	ARG1	SPT21
HHF2	PHO8	ASF1
CAC2	PHO5	HMLALPHA1
SPT16	PHO11	SUC2
ADA2	CIN2	SAS2
HTA1	DOT1	
HTB1	ASH1	
HHT1	MFA2	
HHF1	STE6	
HTA2	HO	
HTB2	PDR5	

Chromatin Structure and metabolism Factor_20 (**W**) (top 20 terms)			
0,100166	**histone**	0,0147704	activity
0,0542096	**chromatin**	0,0142274	yeast
0,0379695	**nucleosome**	0,0132467	lysine
0,0301911	**transcript**	0,0131051	saga
0,02196	**methylation**	0,0127426	tail
0,020179	structure	0,011993	**silencing**
0,0194543	**core**	0,0119741	required
0,0184889	**acetyl**	0,011405	**h2a**
0,0171646	**dna**	0,0111925	remodeling
0,0152601	assembly	0,01098	**h2b**

Mónica Chagoyen, Pedro Carmona-Sáez, Hagit Shatkay, José María Carazo and Alberto Pascual-Montano. Discovering semantic features in the literature: a foundation for building functional associations. BMC Bioinformatics 2006, 7:41

IEEE Trans Pattern Anal Mach Intell. 2006 Mar;28(3):403-15.

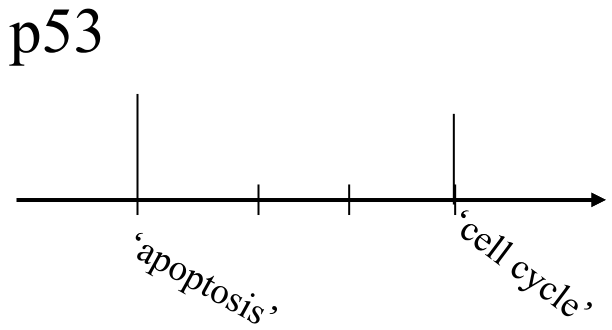
Nonsmooth nonnegative matrix factorization (nsNMF).

Pascual-Montano A, **Carazo JM**, Kochi K, Lehmann D, Pascual-Marqui RD.

.

Advantages

- Low-dimensionality
- Latent semantics
- Non-orthogonality
- Interpretability



Gene representation:

$$\mathbf{V} \approx \mathbf{W} \mathrm{x} \mathbf{H}$$

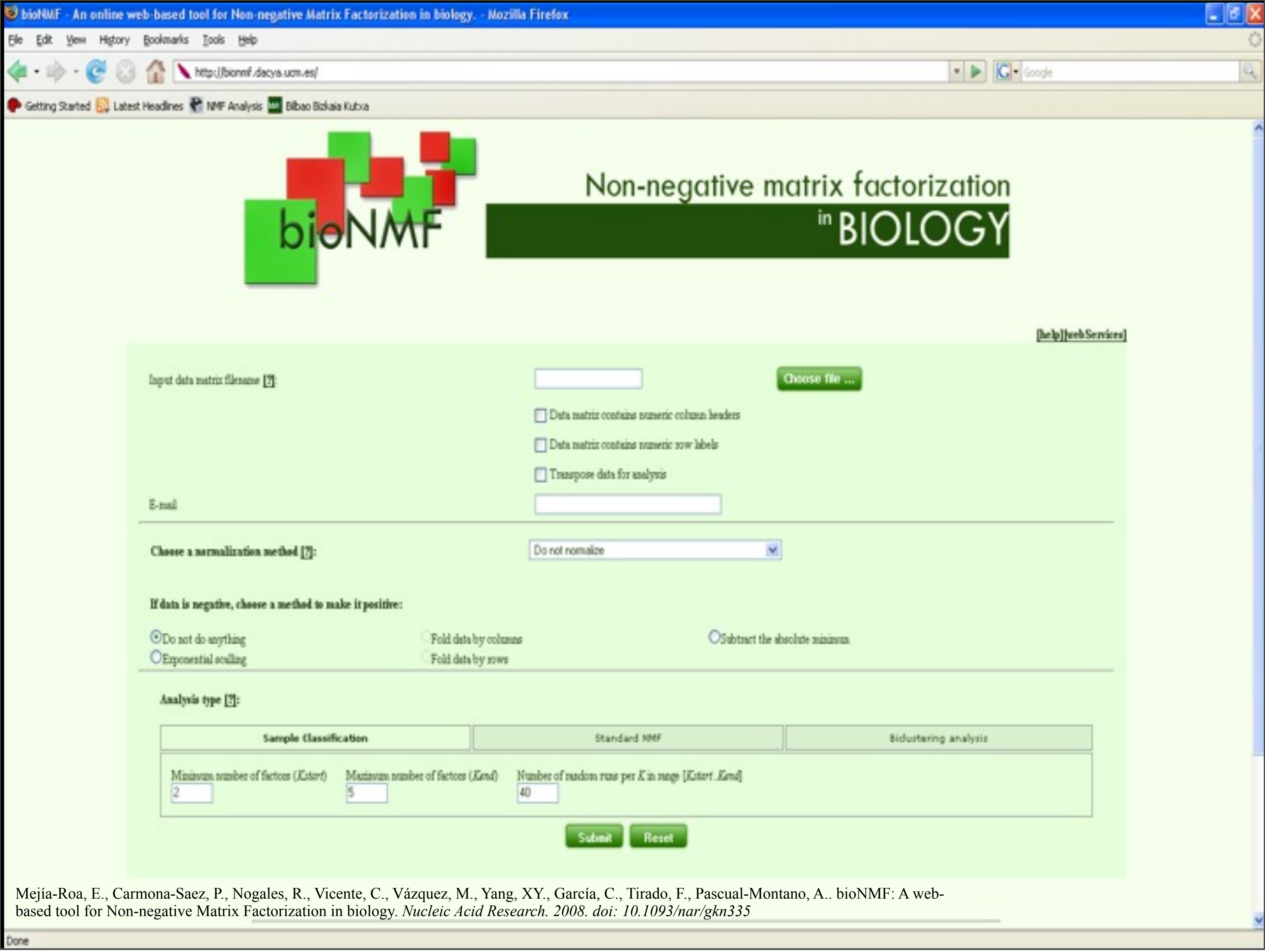
term-frequency vector

feature vector

W → semantic features → biological topics

H → semantic profiles → gene topical profile

Chagoyen M, Carmona-Saez P, Shatkay H, Carazo JM, Pascual-Montano A.
Discovering semantic features in the literature: a foundation for building functional associations
BMC Bioinformatics. 2006 Jan 26;7(1):41 Highly accessed


bioNMF - An online web-based tool for Non-negative Matrix Factorization in biology. - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://bionmf.dacya.ucm.es/
Google
Getting Started Latest Headlines NMF Analysis Bilbao Bizkaia Kutxa
bioNMF
Non-negative matrix factorization in BIOLOGY
[help][webServices]
Choose file ...
Submit
Reset
Done

Input data matrix filename [?]:

☐ Data matrix contains numeric column headers

☐ Data matrix contains numeric row labels

☐ Transpose data for analysis

E-mail

Choose a normalization method [?]: Do not normalize

If data is negative, choose a method to make it positive:

● Do not do anything ○ Fold data by columns ○ Subtract the absolute minimum
○ Exponential scaling ○ Fold data by rows

Analysis type [?]:

Sample Classification	Standard NMF	Biclustering analysis

Minimum number of factors (*Kstart*): 2

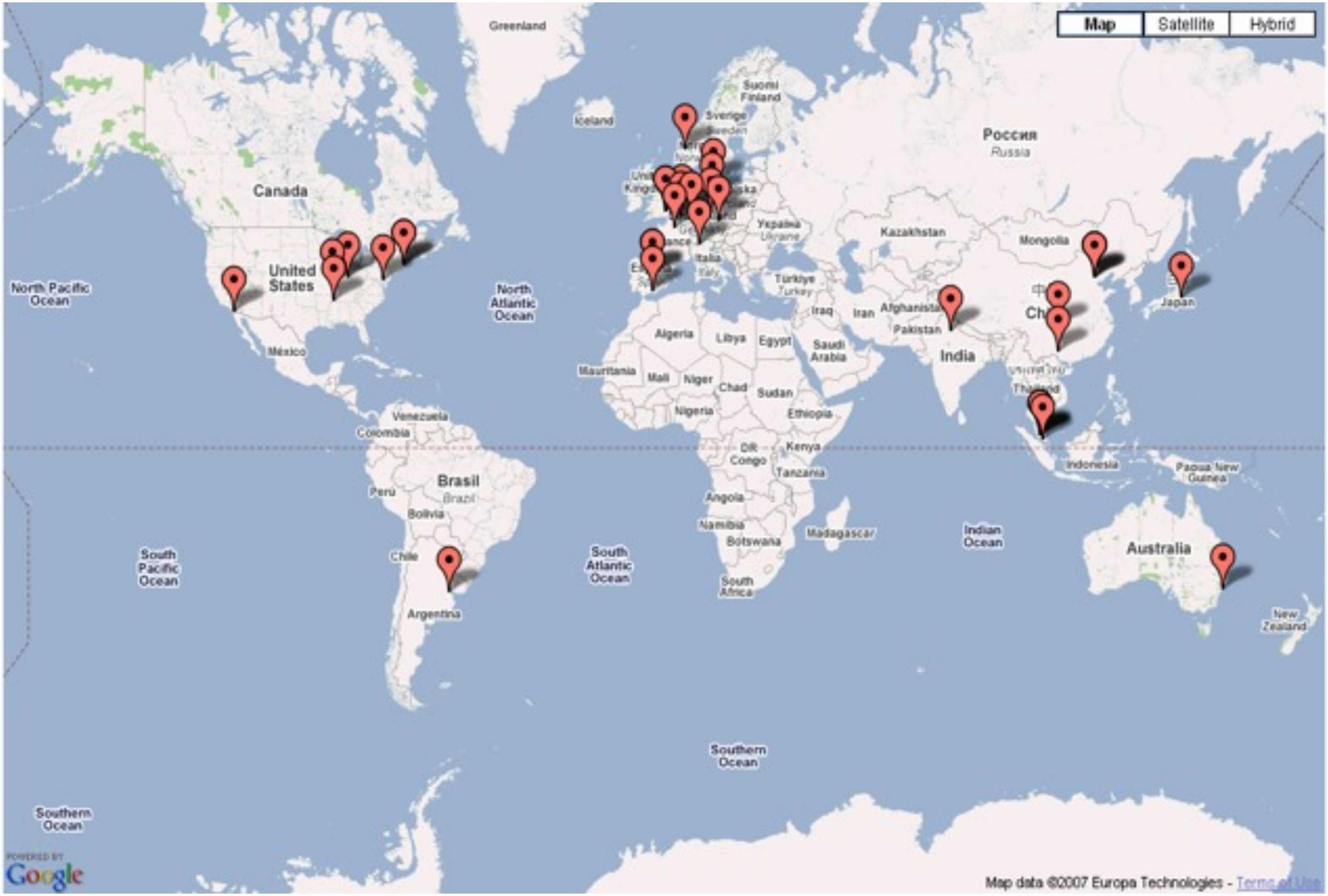
Maximum number of factors (*Kend*): 5

Number of random runs per *K* in range [*Kstart*..*Kend*]: 40

Mejía-Roa, E., Carmona-Saez, P., Nogales, R., Vicente, C., Vázquez, M., Yang, XY., García, C., Tirado, F., Pascual-Montano, A.. bioNMF: A web-based tool for Non-negative Matrix Factorization in biology. *Nucleic Acid Research. 2008. doi: 10.1093/nar/gkn335*

lunes 25 de julio de 2011

bioNMF: statistics (~ 7000 downloads)



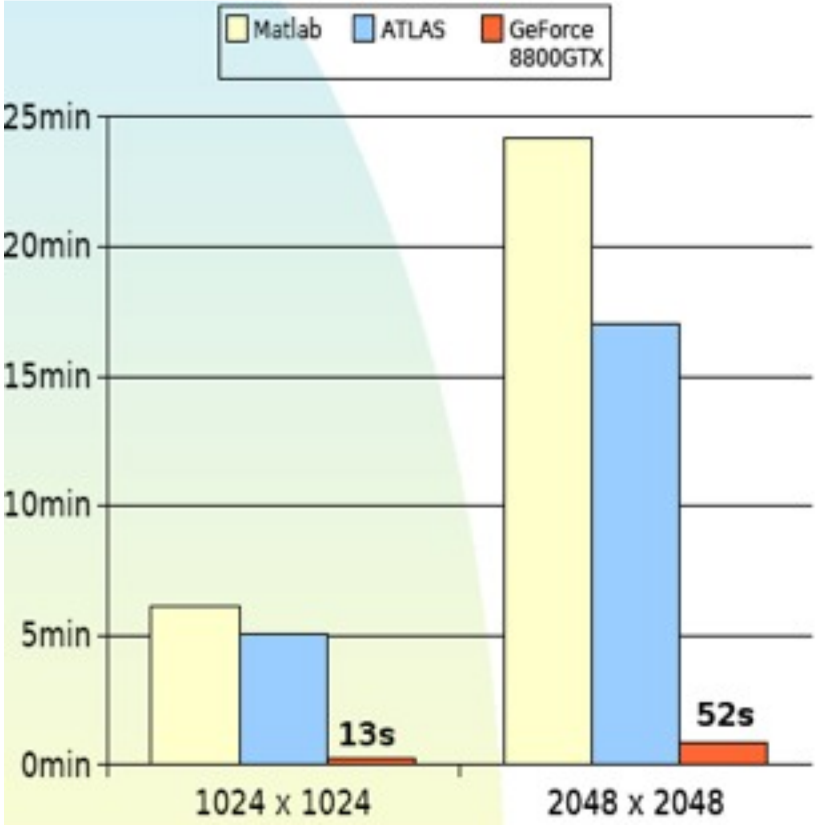
The bioNMF core: NMF

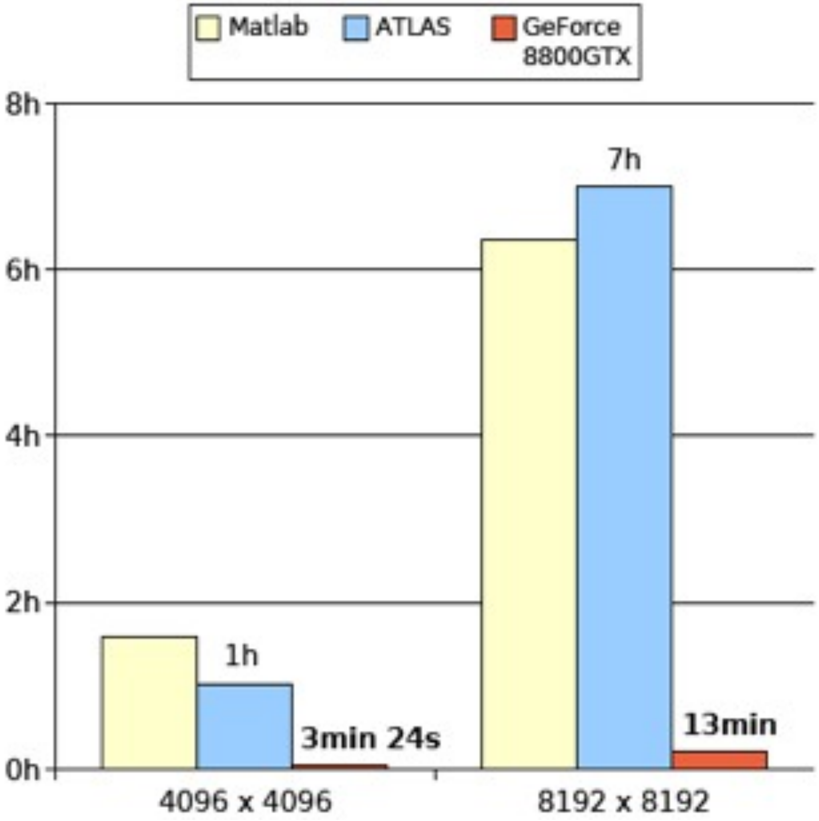
- Multiple possible implementations
  - **C / ATLAS libraries (~ *BLAS*)**
  - *GPGPU*
  - C and MPI

E. Mejía,I. Gómez , M. Prieto, A. Pascual, F. Tirado "Programación bajo un modelo basado en flujos. La factorización NMF como caso de estudio". Procs. XVII Jornadas de Paralelismo, pag. 461-466, Septiembre 2006

NMF in GPU

- Synthetic data matrix.
- Number of factors k = 64.
- 2000 *fixed* loops (no test of convergence).





El Gran Reto: Pasar de la Información al Conocimiento

- Mecanismos de gestión inteligente de grandes volúmenes de datos producida en grandes proyectos colaborativos: LIMS
- Mecanismos para integrar fuentes de datos de datos heterogeneas: Mediadores
- Mecanismos para hacer aflorar “patrones ocultos” en los datos: KDD (Knowledge Discovery and Data Mining)

El Gran Reto: Pasar de la Información al Conocimiento

- Hemos aprendido a “leer” el alfabeto del DNA………. Ahora debemos de entender qué significa!!!

- Es un largo trabajo, pero sabemos en que direcciones proseguir y estamos trabajando!.

The Biocomputing Unit

- *Methods in EM and X-ray Tomo*
  - Dr. Sjors Scheres
  - *Dr. Roberto Marabini (UAM)*
    - Ignacio Arganda and Ana Iriarte (UAM)
  - Dr. Carlos Oscar Sánchez
  - Dr. Roberto Valerio
- *National Institute of Bioinformatics*
  - Dra. Natalia Jiménez-Lozano
  - Joan Segura
  - *Jose Ramón Macias*
  - Juanjo Vega
- *Structural biology of helicases*
  - Dr. Martín Alcorlo
    - Roberto Melero and Marta Rajkiewicz
  - Dra. Sami Kereiche
- *Structural biology of the centrosome*
  - Dra. Rocio González
  - Dr. Johan Busselez

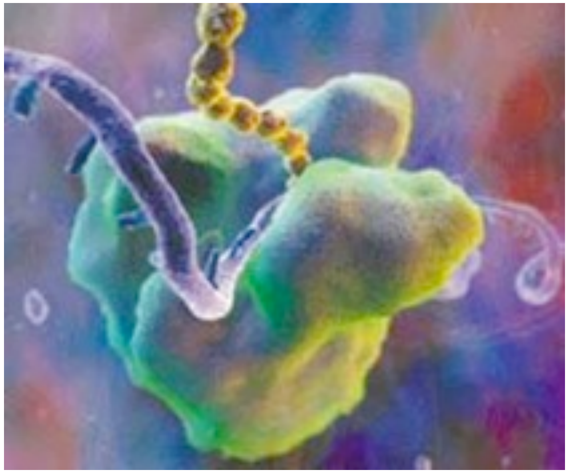
- ***Support:***
  - **Blanca Benítez**
  - **Jesus Cuenca**
- ***Gene Expression Data Analysis-UCM (collaboration with Dr. Alberto Pascual)***
  - **Dr. Federico Abascal**
  - **Mariana Lara**
- ***Main external collaborators***
  - **Prof. Gabor Herman (NYU)**
  - **Prof. Ellen Fanning (Vanderbilt)**
  - **Prof. Xiojiang Cheng (USC)**
  - **Prof. Juan Carlos Alonso (CNB)**
  - **Prof. J. Frank (Columbia)**
  - **Dr.Sergio Marco (Curie)**
  - **Dr. Michel Bornens (Curie)**
  - **Dr.Mikel Valle (Biogune)**
  - **Dra. Carmen San Martín (CNB)**
- **Integromics Inc.**
  - **Philadelphia, Madrid, Granada, Russe and Beijing**

Structural Flexibility, Variability and Function, how can we study them?: The 26S case

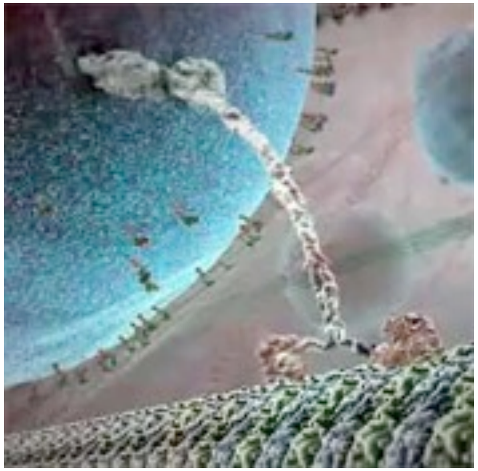
Jose-Maria Carazo, Carlos Sánchez Sorzano, Roberto Marabini

Life based on molecular machines

DNA replication

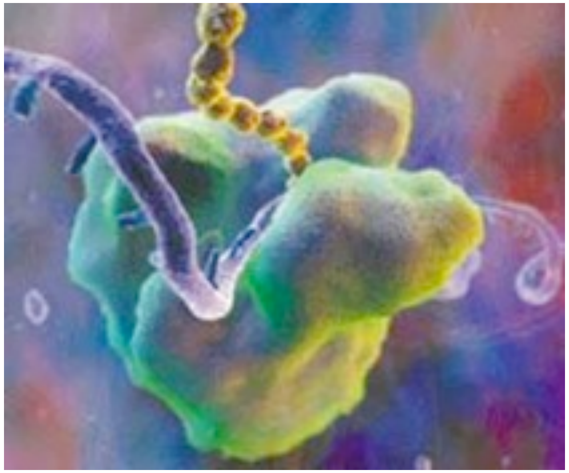


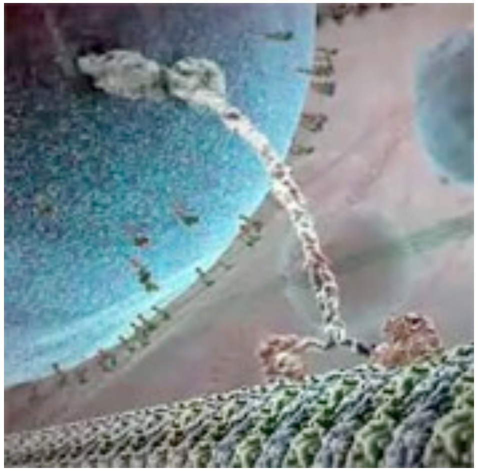
Protein synthesis

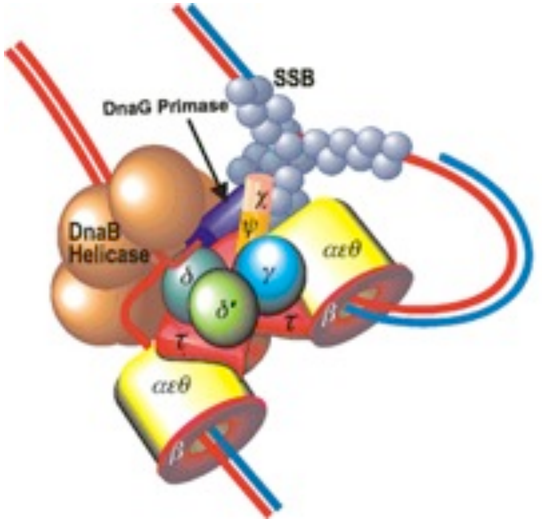


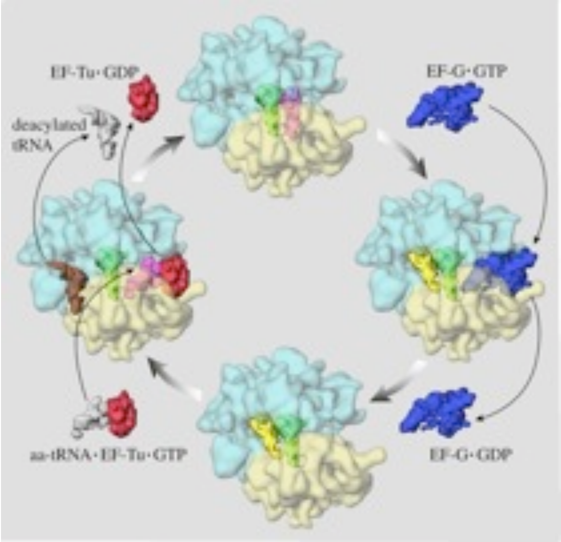
Dynein motion

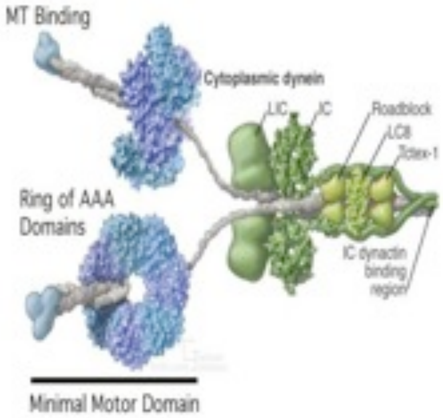
Life based on molecular machines



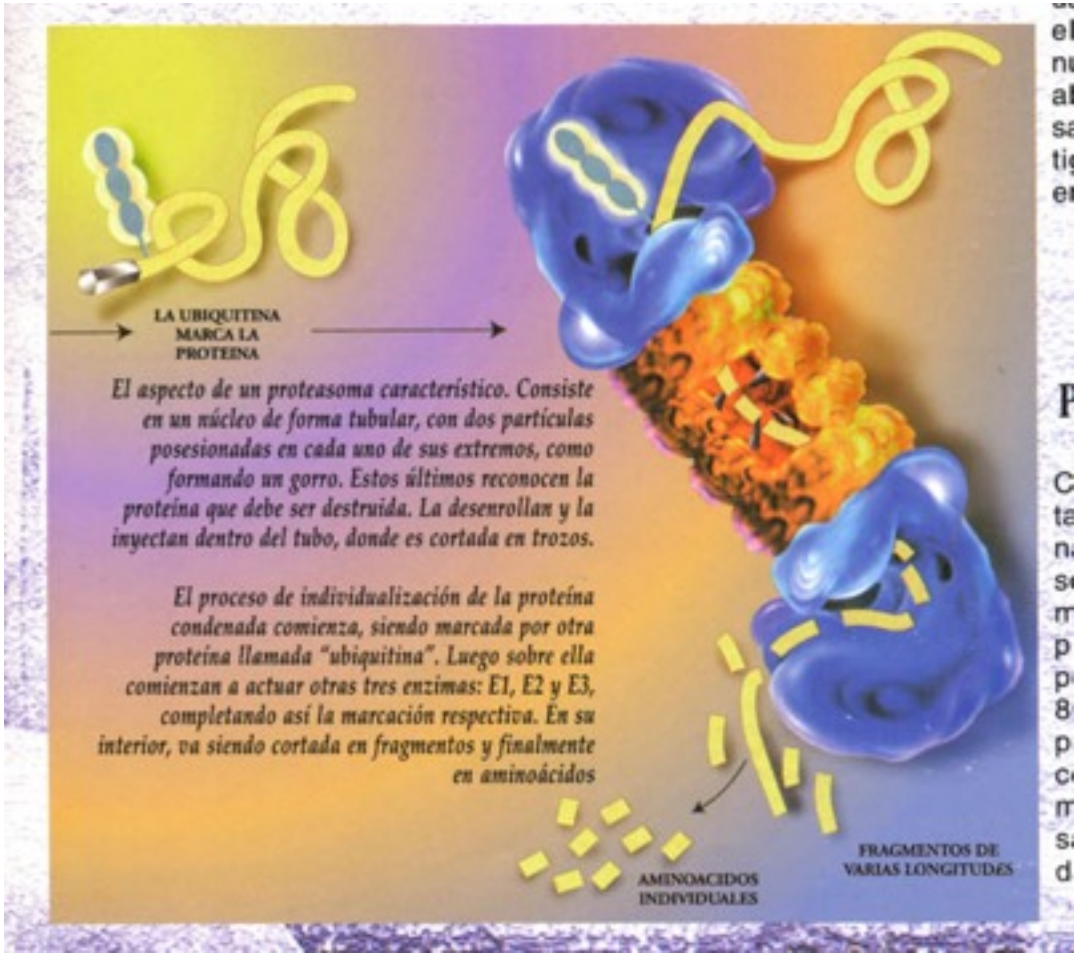








The 26S “Cartoon presentation”

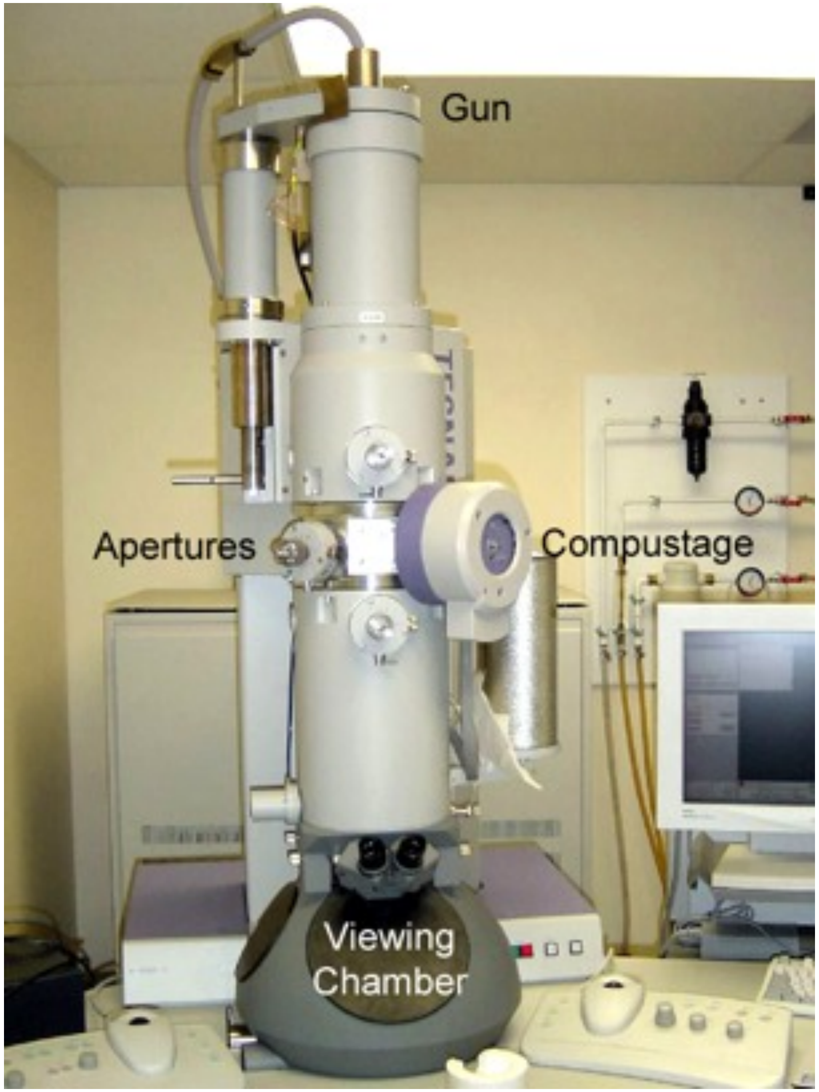


LA UBIQUITINA MARCA LA PROTEINA
FRAGMENTOS DE VARIAS LONGITUDES
AMINOACIDOS INDIVIDUALES

El aspecto de un proteasoma característico. Consiste en un núcleo de forma tubular, con dos partículas posesionadas en cada uno de sus extremos, como formando un gorro. Estos últimos reconocen la proteína que debe ser destruida. La desenrollan y la inyectan dentro del tubo, donde es cortada en trozos.

El proceso de individualización de la proteína condenada comienza, siendo marcada por otra proteína llamada “ubiquitina”. Luego sobre ella comienzan a actuar otras tres enzimas: E1, E2 y E3, completando así la marcación respectiva. En su interior, va siendo cortada en fragmentos y finalmente en aminoácidos

An electron microscope



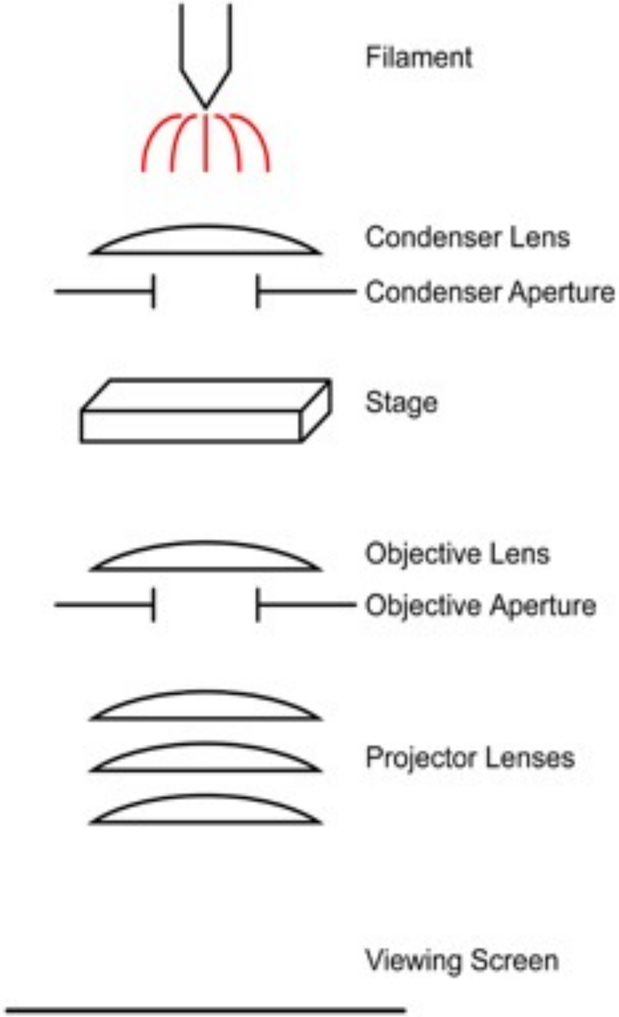
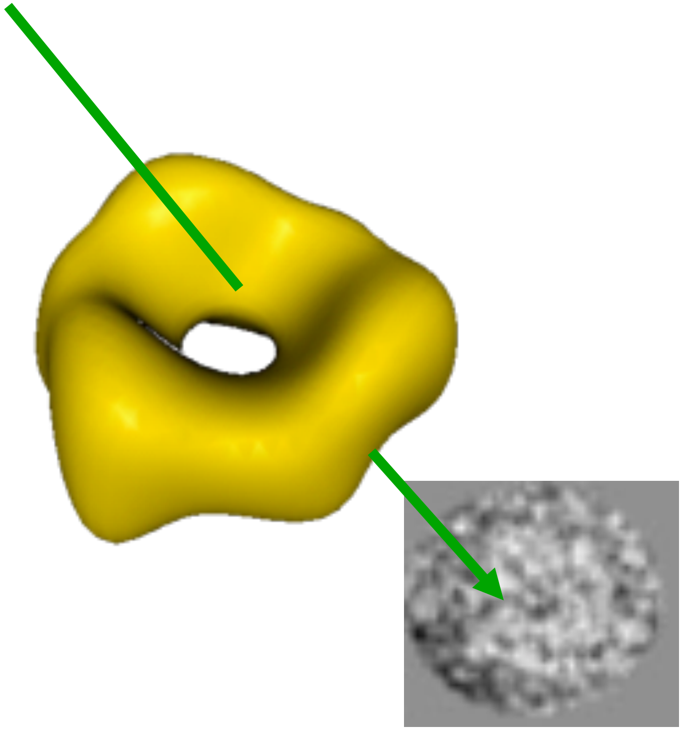


Image formation in 3D-EM

- Under the Weak Phase Object approximation, the Electron Microscopy images are X-ray Transforms of the Coulomb potential of the biological macromolecules



(The inelastic scatering is negligeable versus the elastic scatering, and this latter one can be modelled as a lineal process)

$f : R^n$ ® R

$x \in R^n$

$\omega \in S^{n-1}$

Analogy: Data adquisition for CT

$$f(\mathbf{r}) \approx \sum_{j=1}^{J} x_j b_j(\mathbf{r})$$

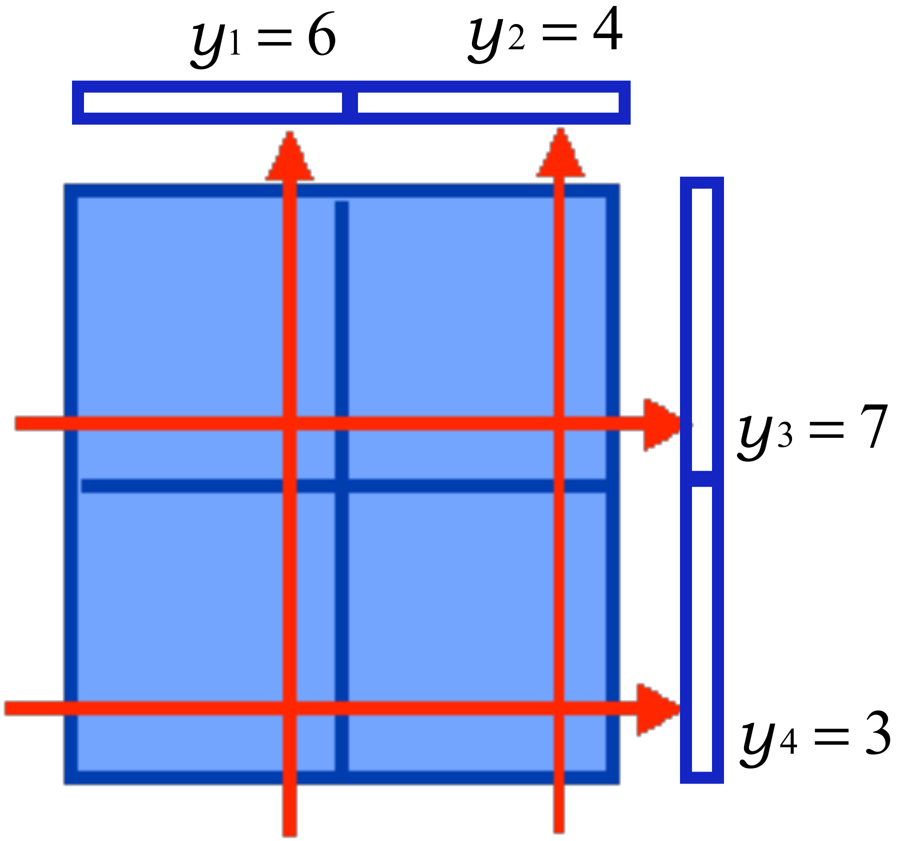
Reconstruction as a linear set of equations

$$y_i \approx \sum_{j=1}^{J} l_{i,j} x_j \qquad \boxed{l_{i,j} = 1,0}$$

Reconstruction as a linear set of equations

$$f(\mathbf{r}) \approx \sum_{j=1}^{J} x_j b_j(\mathbf{r})$$

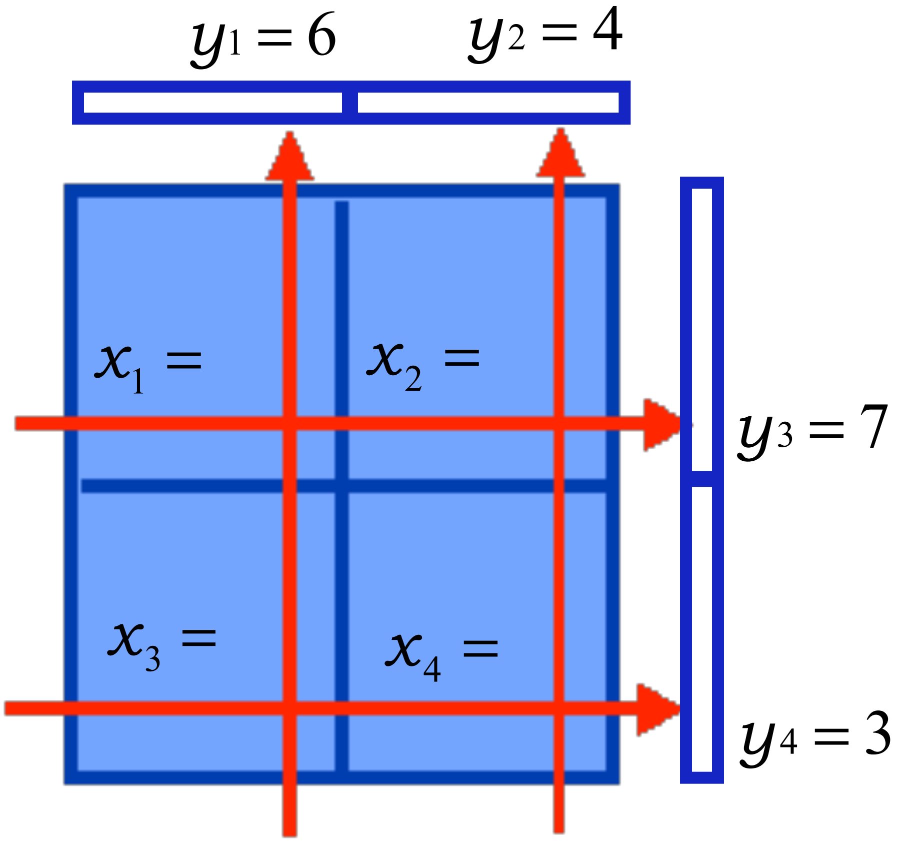
$$y_i \approx \sum_{j=1}^{J} l_{i,j} x_j \qquad \boxed{l_{i,j} = 1,0}$$



Reconstruction as a linear set of equations

$$f(\mathbf{r}) \approx \sum_{j=1}^{J} x_j b_j(\mathbf{r})$$

$$y_i \approx \sum_{j=1}^{J} l_{i,j} x_j \qquad \boxed{l_{i,j} = 1,0}$$

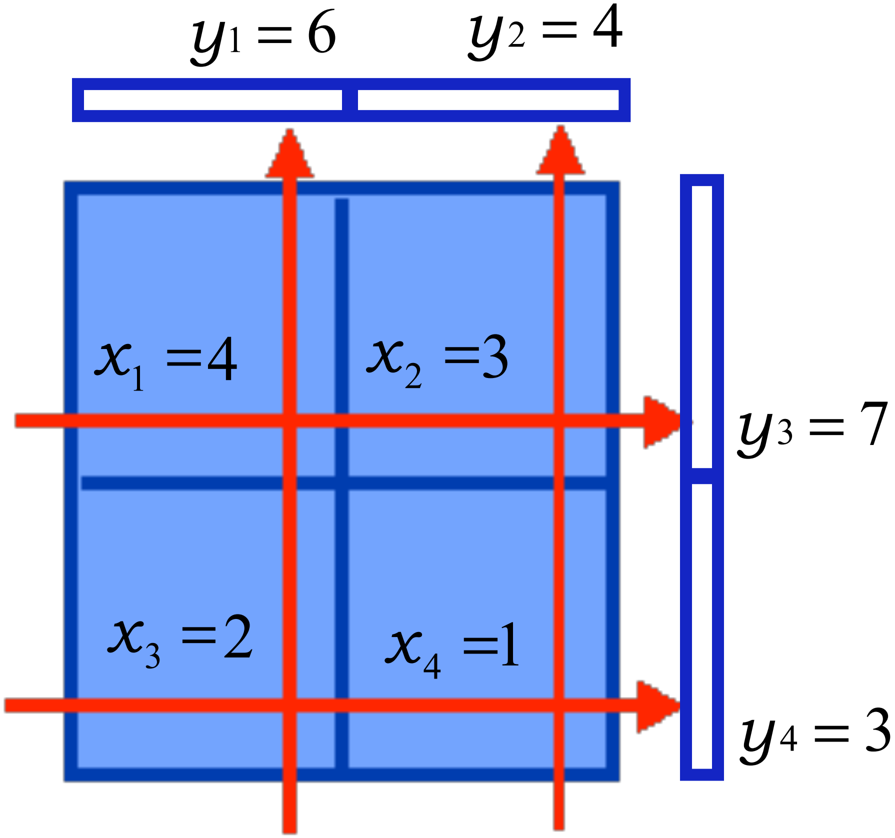


Reconstruction as a linear set of equations

$$f(\mathbf{r}) \approx \sum_{j=1}^{J} x_j b_j(\mathbf{r})$$

$$y_i \approx \sum_{j=1}^{J} l_{i,j} x_j$$

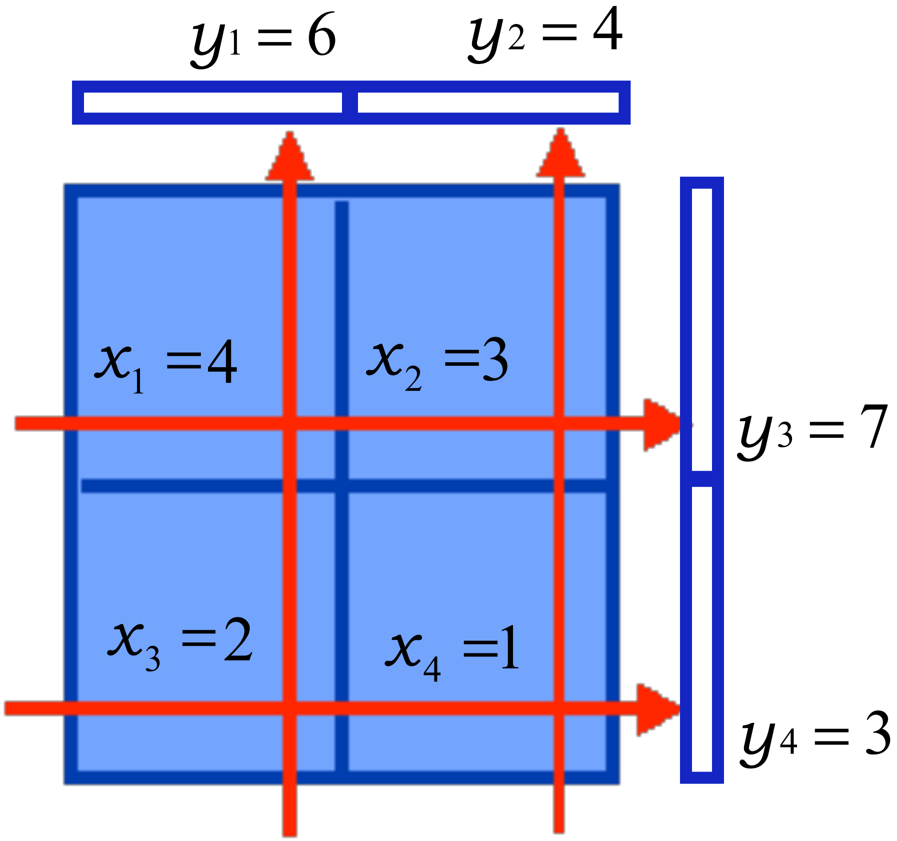
$$\boxed{l_{i,j} = 1,0}$$



Reconstruction as a linear set of equations

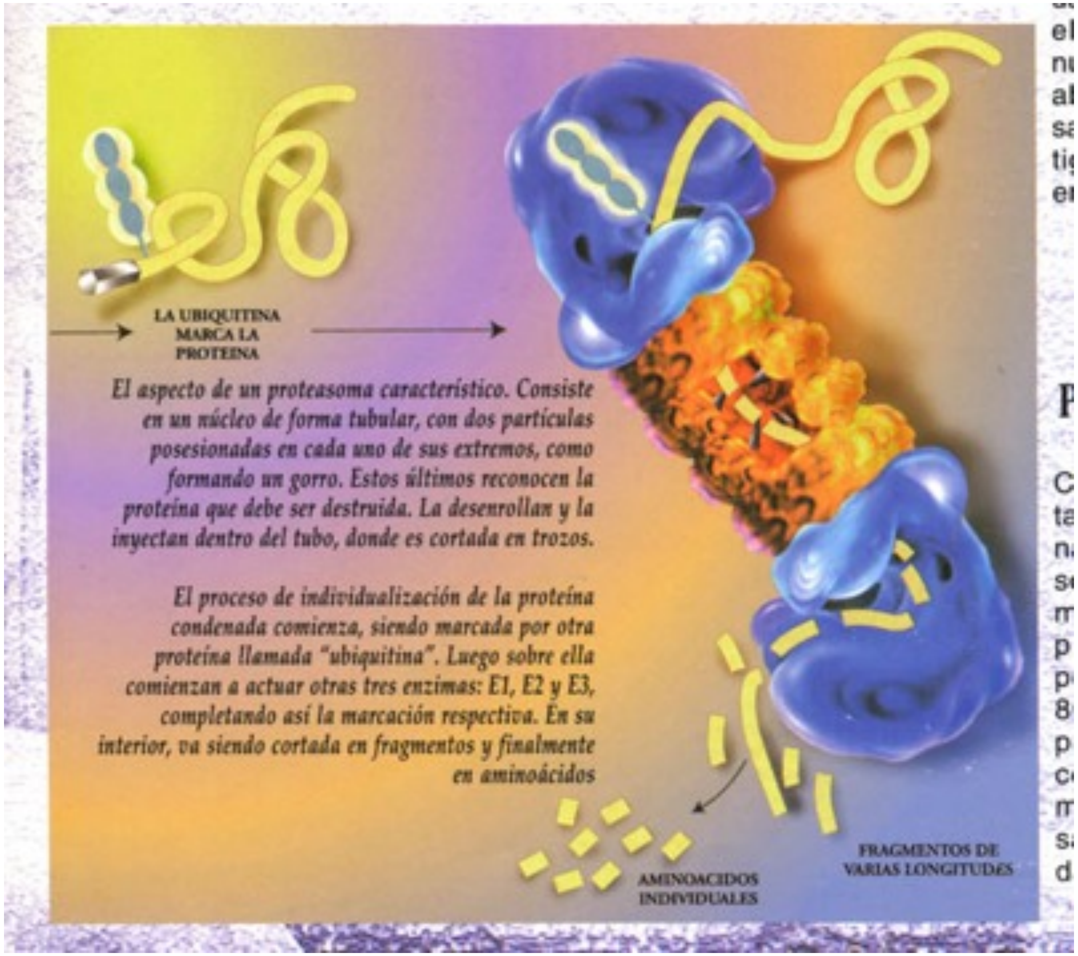
$$f(\mathbf{r}) \approx \sum_{j=1}^{J} x_j b_j(\mathbf{r})$$

$$y_i \approx \sum_{j=1}^{J} l_{i,j} x_j \qquad \boxed{l_{i,j} = 1,0}$$



$$\left\{\begin{array}{l} \left.\begin{array}{l} x_1 + x_3 = 6 \\ x_2 + x_4 = 4 \end{array}\right\} \\ \left.\begin{array}{l} x_1 + x_2 = 7 \\ x_3 + x_4 = 3 \end{array}\right\} \end{array}\right.$$

The 26S “Cartoon presentation”

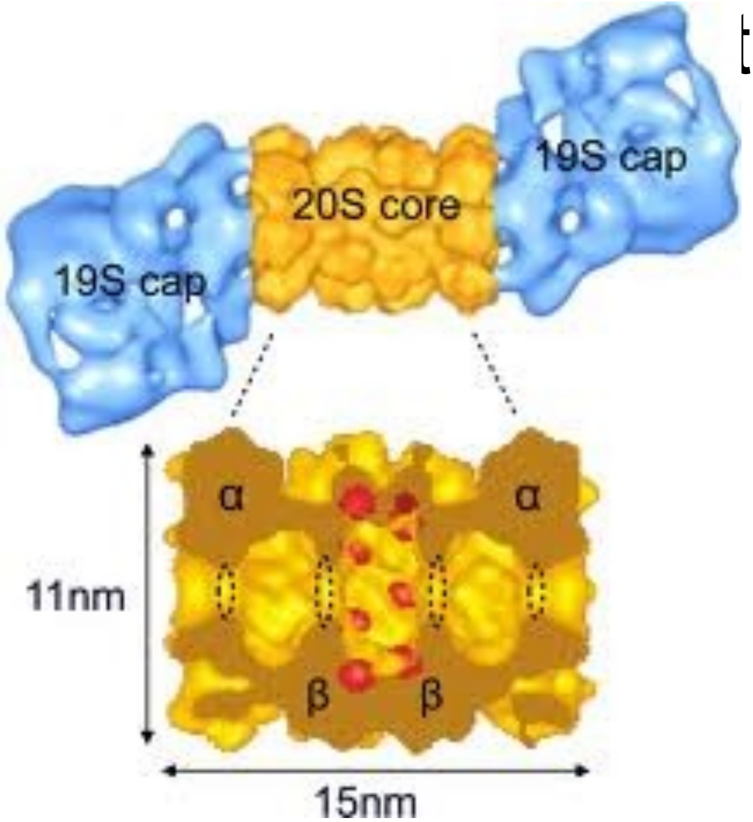


El aspecto de un proteasoma característico. Consiste en un núcleo de forma tubular, con dos partículas posesionadas en cada uno de sus extremos, como formando un gorro. Estos últimos reconocen la proteína que debe ser destruida. La desenrollan y la inyectan dentro del tubo, donde es cortada en trozos.

El proceso de individualización de la proteína condenada comienza, siendo marcada por otra proteína llamada “ubiquitina”. Luego sobre ella comienzan a actuar otras tres enzimas: E1, E2 y E3, completando así la marcación respectiva. En su interior, va siendo cortada en fragmentos y finalmente en aminoácidos

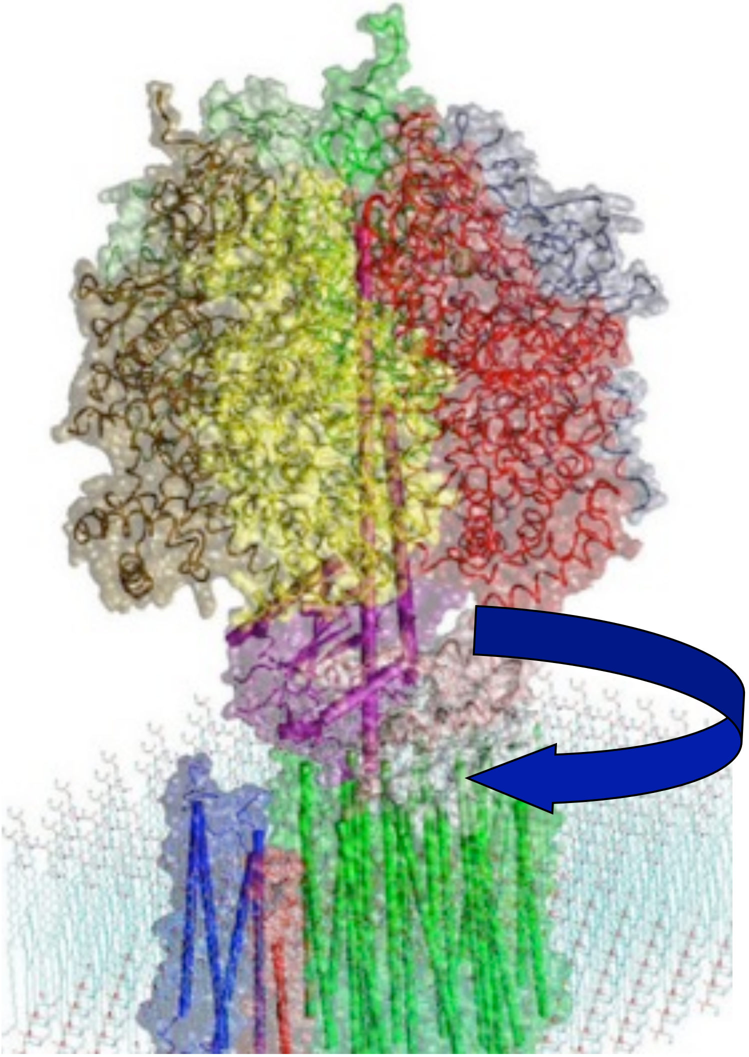
The “26S Case”

- The “20S co ts

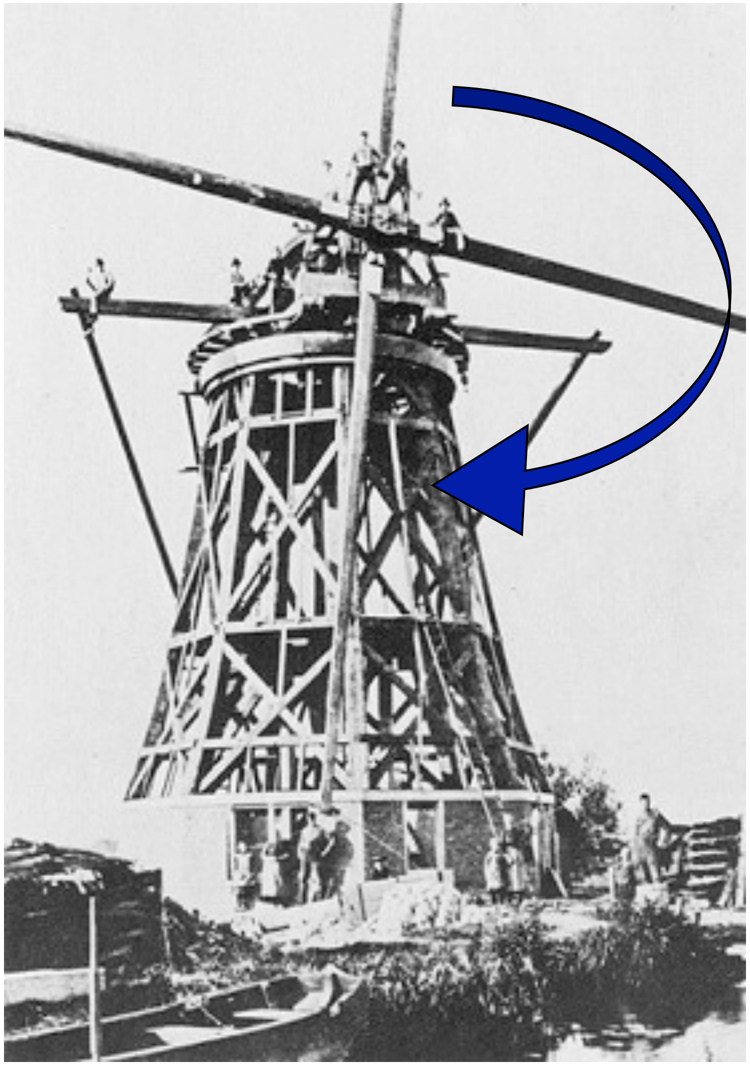


Molecular machines

15 10^{-9} m



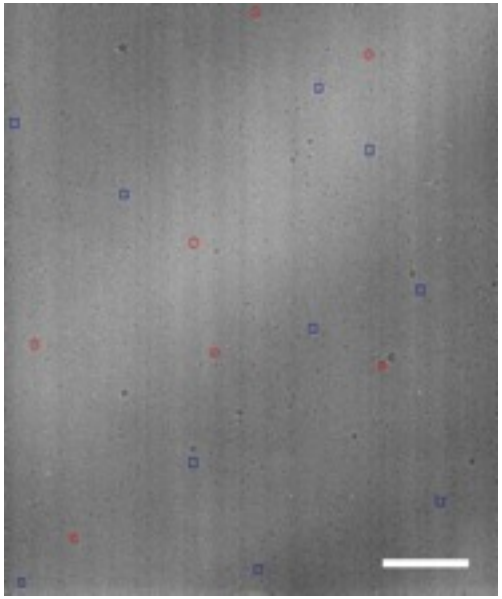
F1-ATPase: Abrahams et al., 1994



15 m

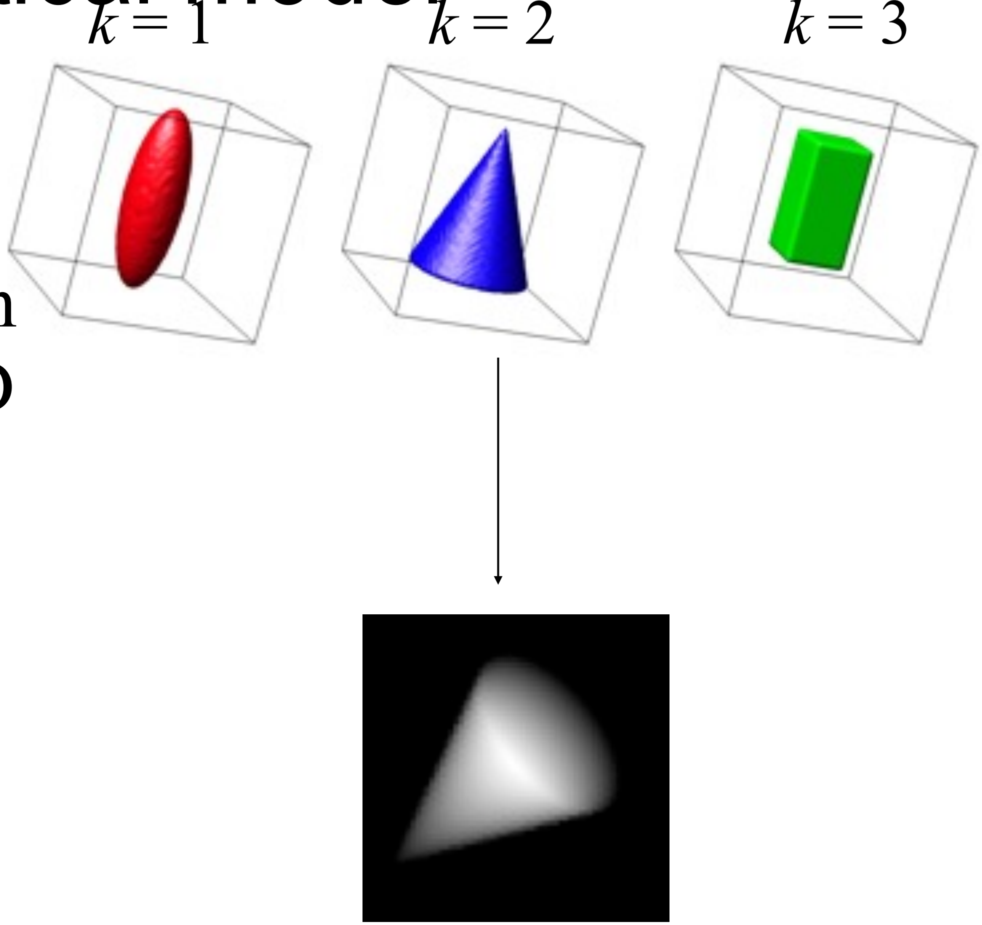
Dutch windmill

An analogy to “conformational changes”

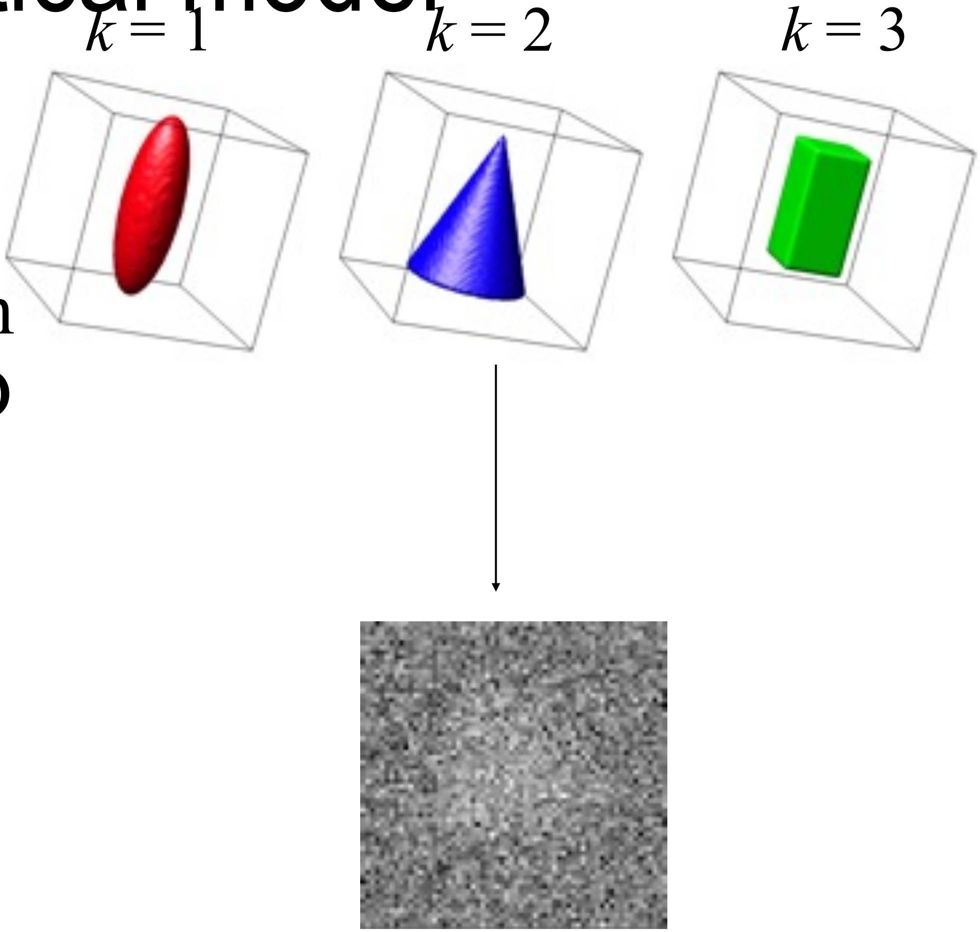


Statistical model

Each image is a projection of one of *K* underlying 3D objects *k*



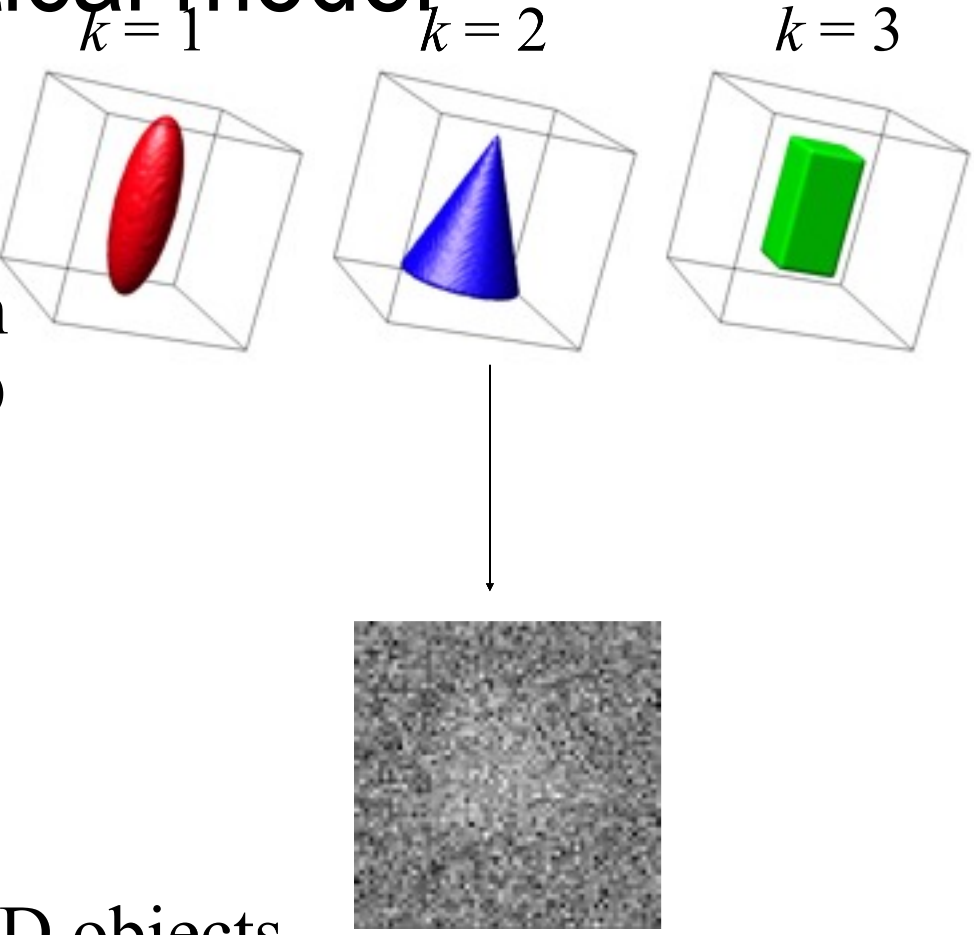
Statistical model



Each image is a projection of one of K underlying 3D objects k

with addition of white Gaussian noise

Statistical model



Each image is a projection of one of K underlying 3D objects k

with addition of
white Gaussian noise

Unknowns: the 3D objects k, orientations

Log-likelihood function

- Adjust model to maximize the log-likelihood of observing the entire dataset:

$$L(\text{model}) = \sum_{i=1}^{N} \ln P(\text{image}_i \mid \text{model})$$

Log-likelihood function

- Adjust model to maximize the log-likelihood of observing the entire dataset:

$$L(\text{model}) = \sum_{i=1}^{N} \ln P(\text{image}_i \mid \text{model})$$

$$= \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \sum_{orient} P(\text{image}_i \mid k, \text{orient.}, \text{model}) P(k, \text{orient.} \mid \text{model})$$

Log-likelihood function

- Adjust model to maximize the log-likelihood of observing the entire dataset:

$$L(\text{model}) = \sum_{i=1}^{N} \ln P(\text{image}_i \mid \text{model})$$

$$= \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \sum_{orient} P(\text{image}_i \mid k, \text{orient.}, \text{model}) P(k, \text{orient.} \mid \text{model})$$

The **model** comprises:

- estimates for the underlying objects
- estimate for the amount of noise (σ)
- statistical distributions of k & orient.

Log-likelihood function

- Adjust model to maximize the log-likelihood of observing the entire dataset:

$$L(\text{model}) = \sum_{i=1}^{N} \ln P(\text{image}_i \mid \text{model})$$

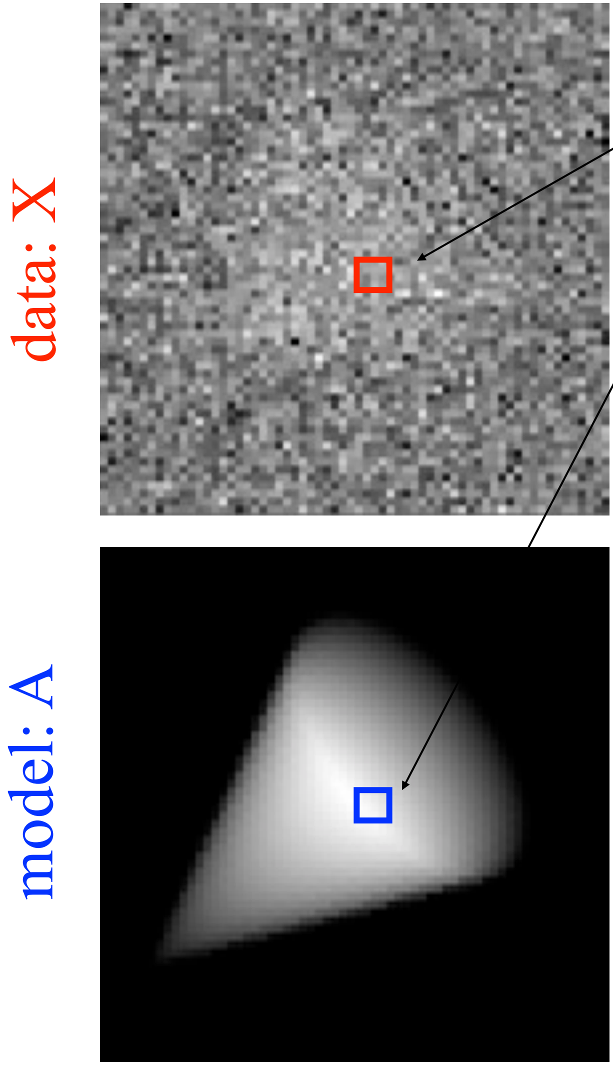
$$= \sum_{i=1}^{N} \ln \sum_{k=1}^{K} \sum_{orient} P(\text{image}_i \mid k, \text{orient.}, \text{model}) P(k, \text{orient.} \mid \text{model})$$

The **model** comprises:

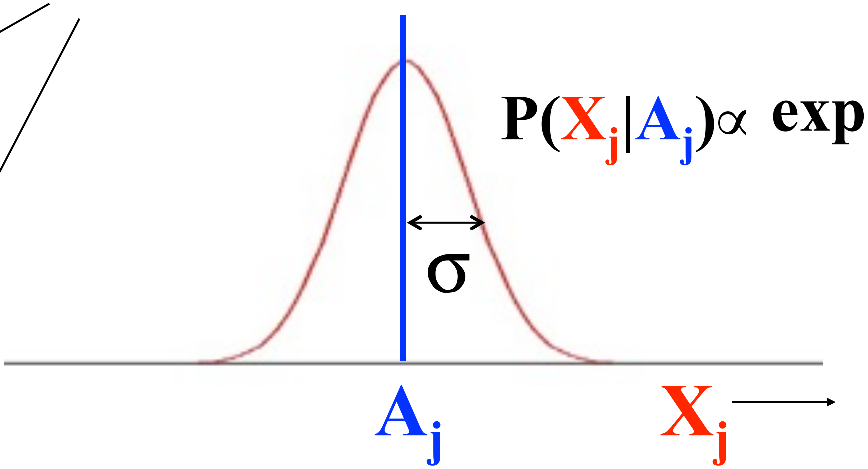
- estimates for the underlying objects
- estimate for the amount of noise (σ)
- statistical distributions of k & orient.

Expectation Maximization

Statistical model



for each pixel j:



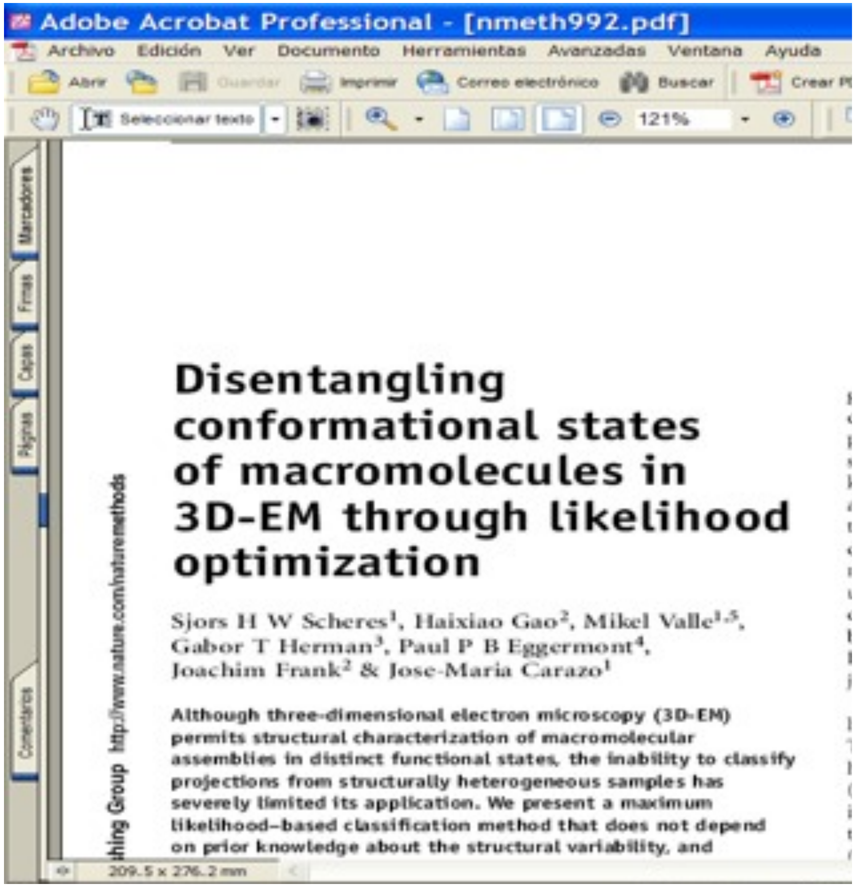
$$P(X_j|A_j) \propto \exp\left(\frac{(X_j - A_j)^2}{-2\sigma^2}\right)$$

White noise =
independence between pixels!

P(data image|model image) ~

$$\prod_j P(X_j|A_j)$$

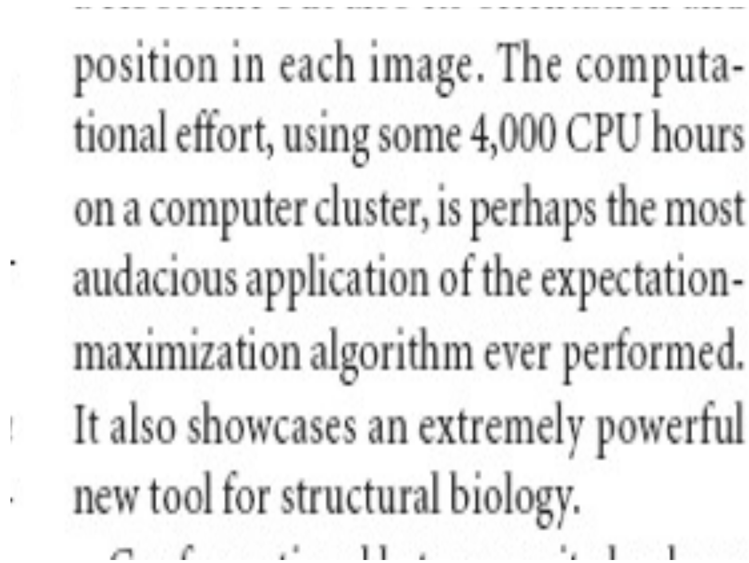
And now, some “maths”: We need to find a (very good) solution to deal with “structurally heterogeneous mixtures”.


Adobe Acrobat Professional - [nmeth992.pdf]
Archivo Edición Ver Documento Herramientas Avanzadas Ventana Ayuda
Abrir
Guardar
Imprimir
Correo electrónico
Buscar
Crear P
Seleccionar texto
121%
Marcadores
Firmas
Capas
Páginas
Comentarios
http://www.nature.com/naturemethods
hing Group
209.5 x 276.2 mm

Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization

Sjors H W Scheres[1], Haixiao Gao[2], Mikel Valle[1,5], Gabor T Herman[3], Paul P B Eggermont[4], Joachim Frank[2] & Jose-Maria Carazo[1]

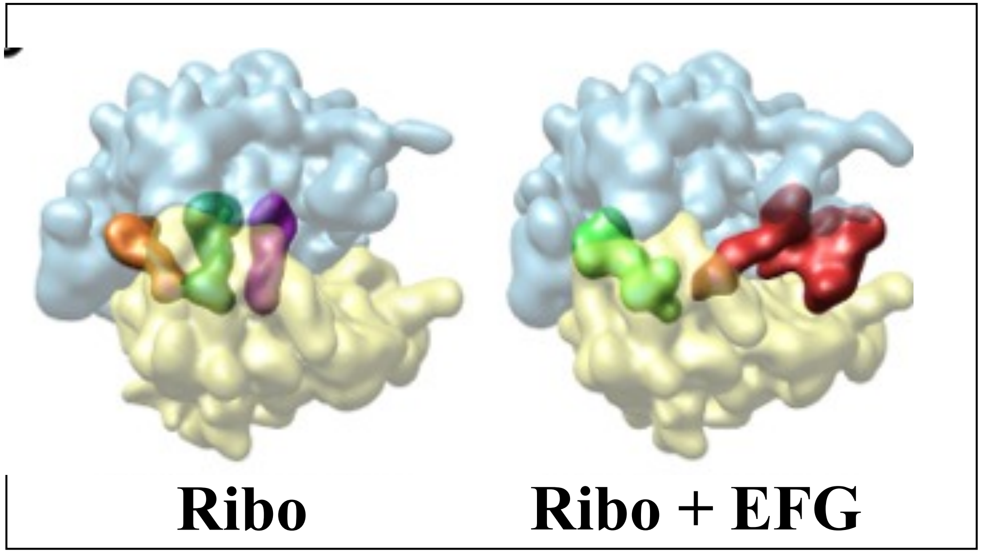
Although three-dimensional electron microscopy (3D-EM) permits structural characterization of macromolecular assemblies in distinct functional states, the inability to classify projections from structurally heterogeneous samples has severely limited its application. We present a maximum likelihood–based classification method that does not depend on prior knowledge about the structural variability, and



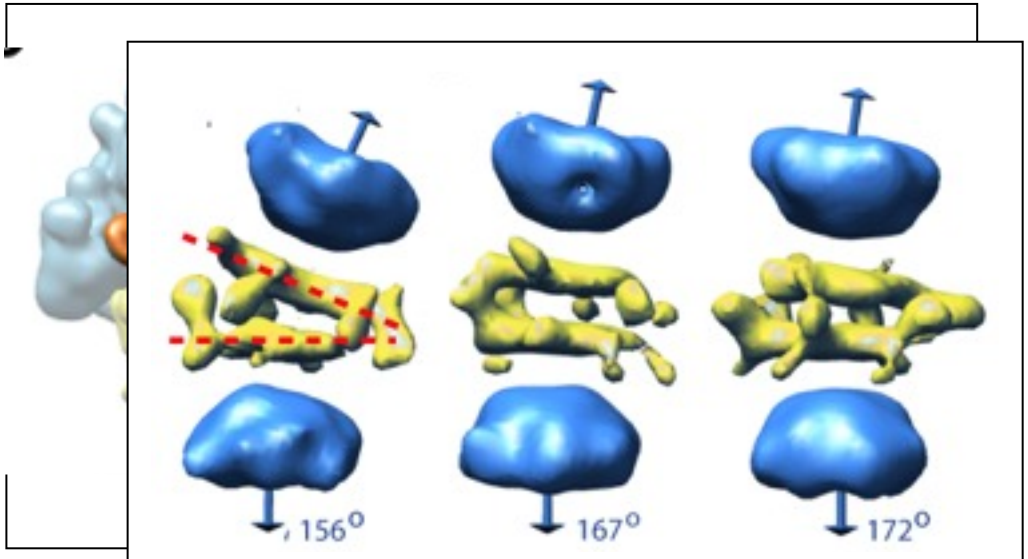
position in each image. The computational effort, using some 4,000 CPU hours on a computer cluster, is perhaps the most audacious application of the expectation-maximization algorithm ever performed. It also showcases an extremely powerful new tool for structural biology.

Nature Methods, 2007; Structure, 2007, 2009; Acta Crys. 2009; JSB 2009,
Structure, JSB, 2010

ML3D: Some applications…

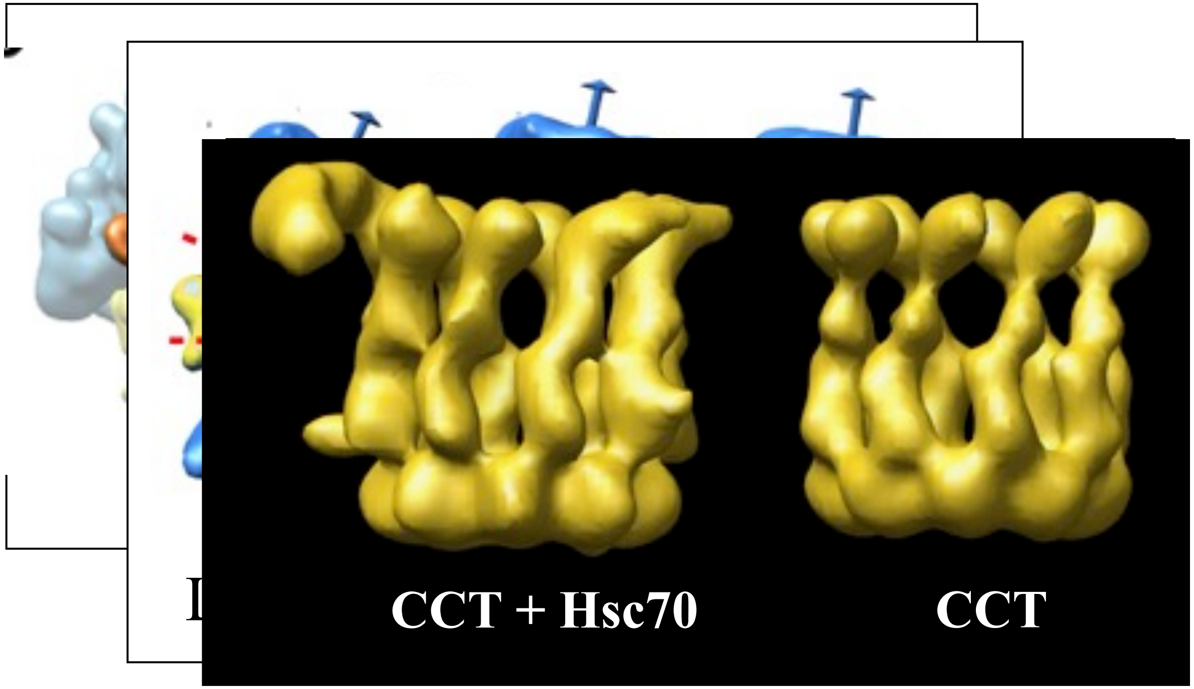


ML3D: Some applications…

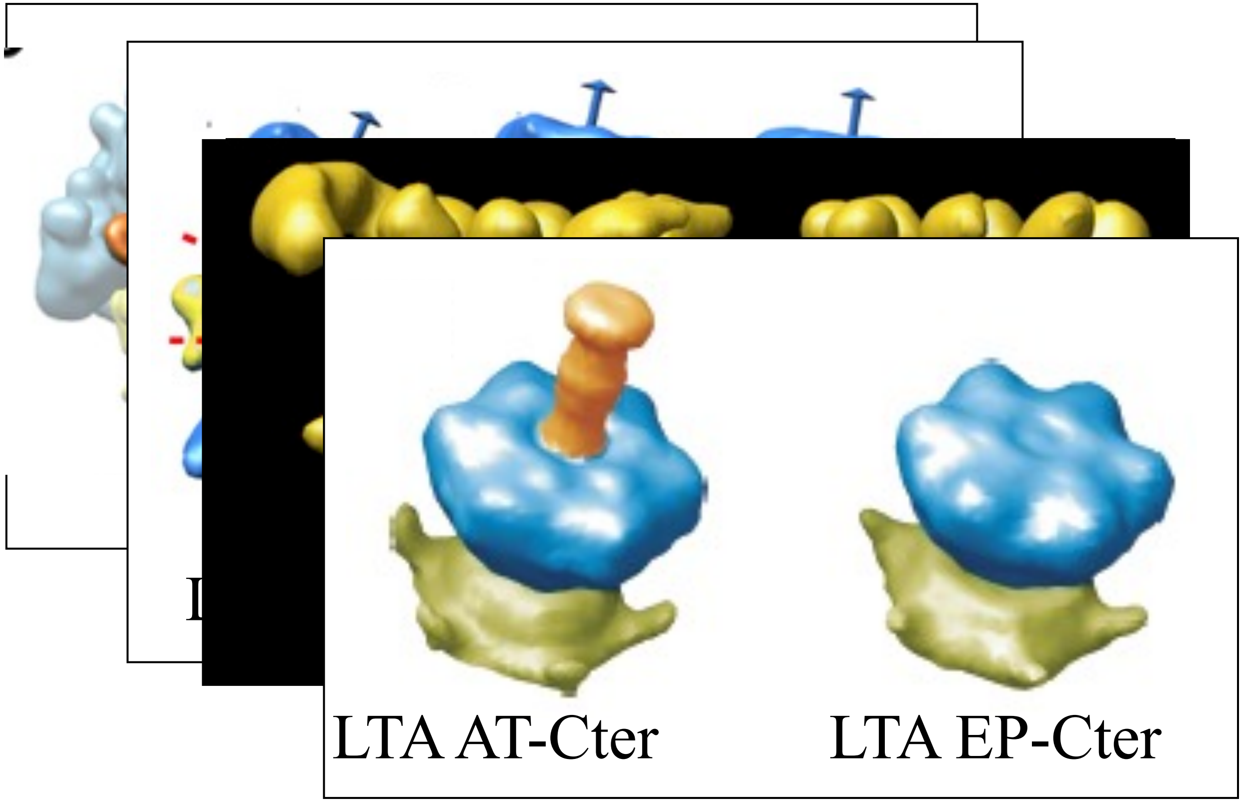


LTA with varying degree of bending

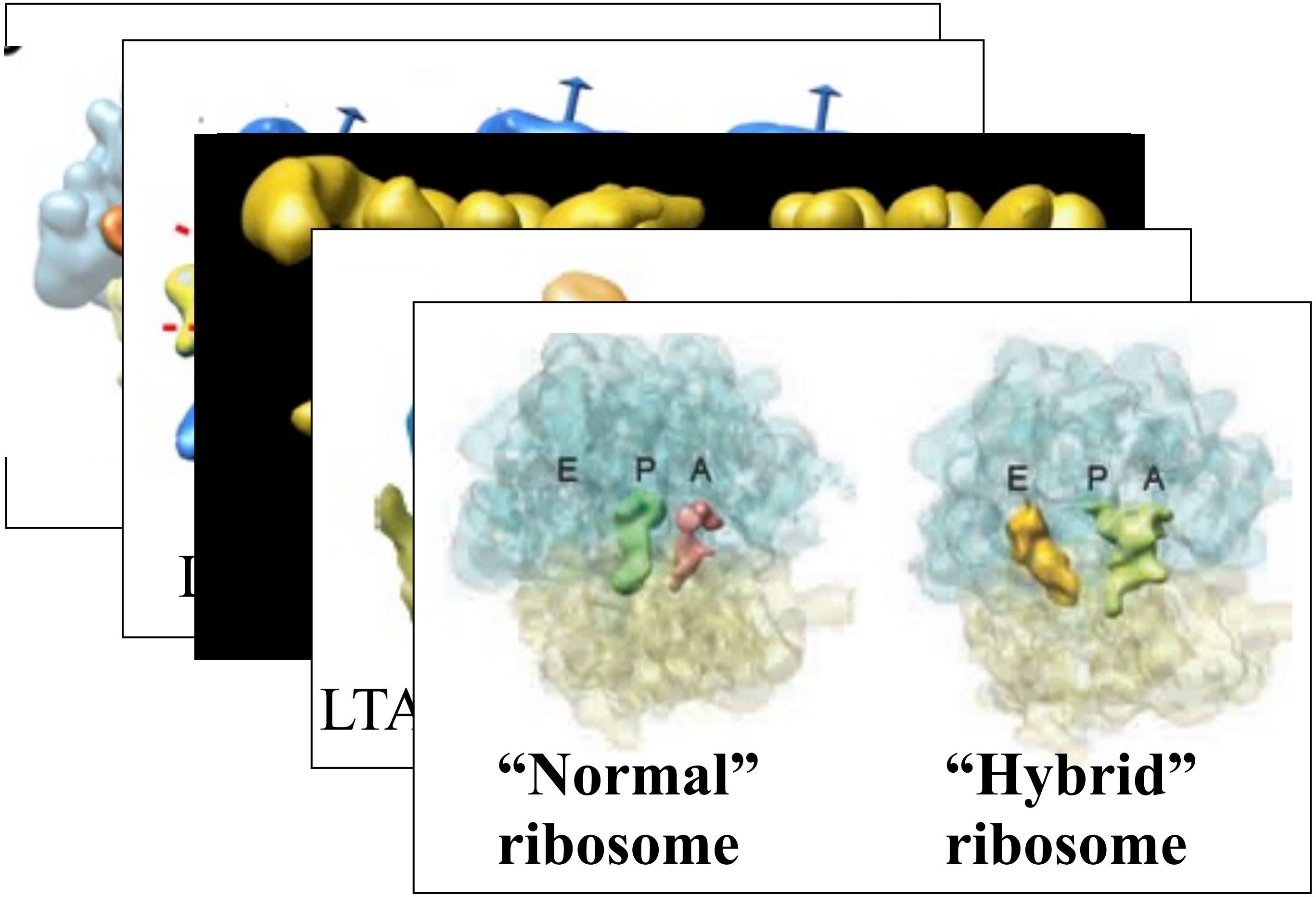
ML3D: Some applications…



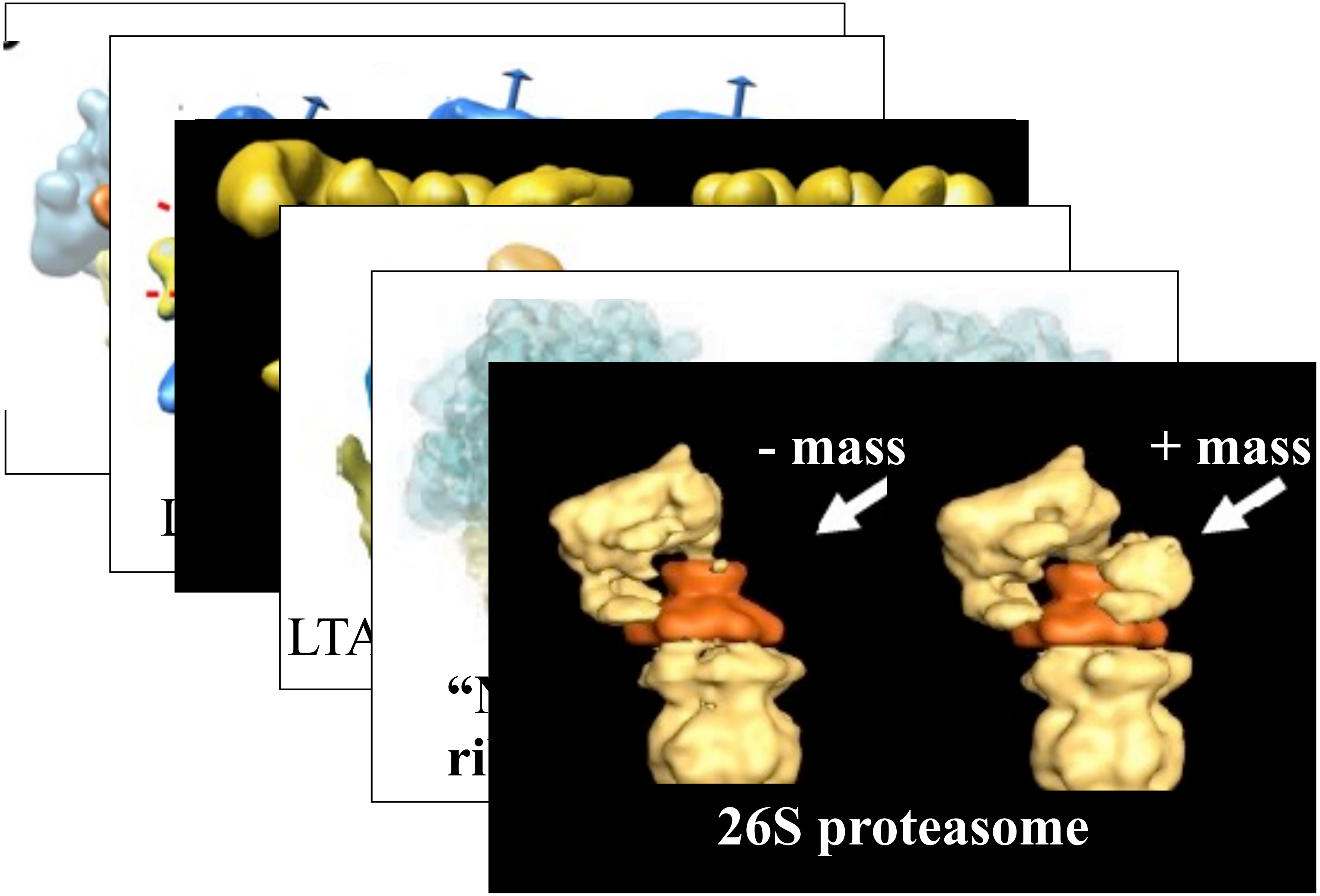
ML3D: Some applications…



ML3D: Some applications…



ML3D: Some applications…

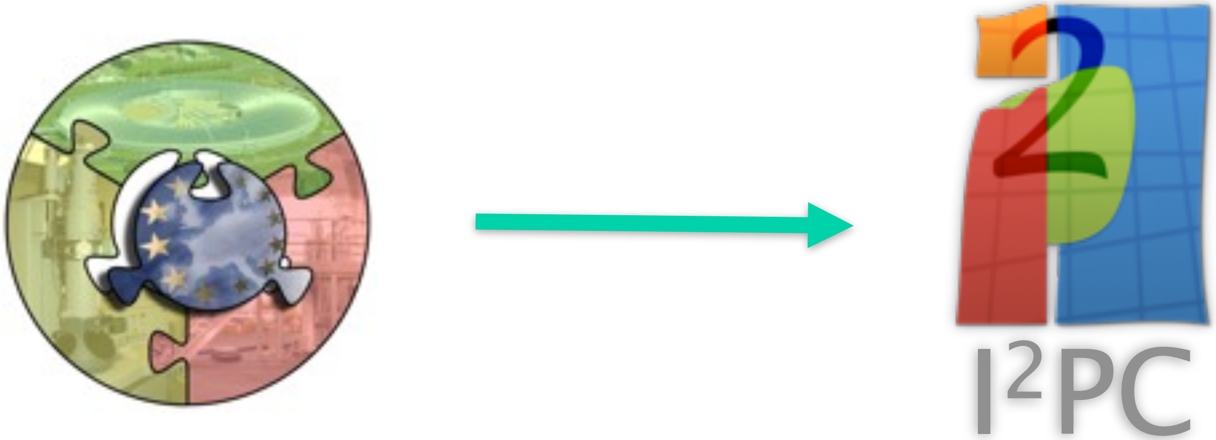


Scipion

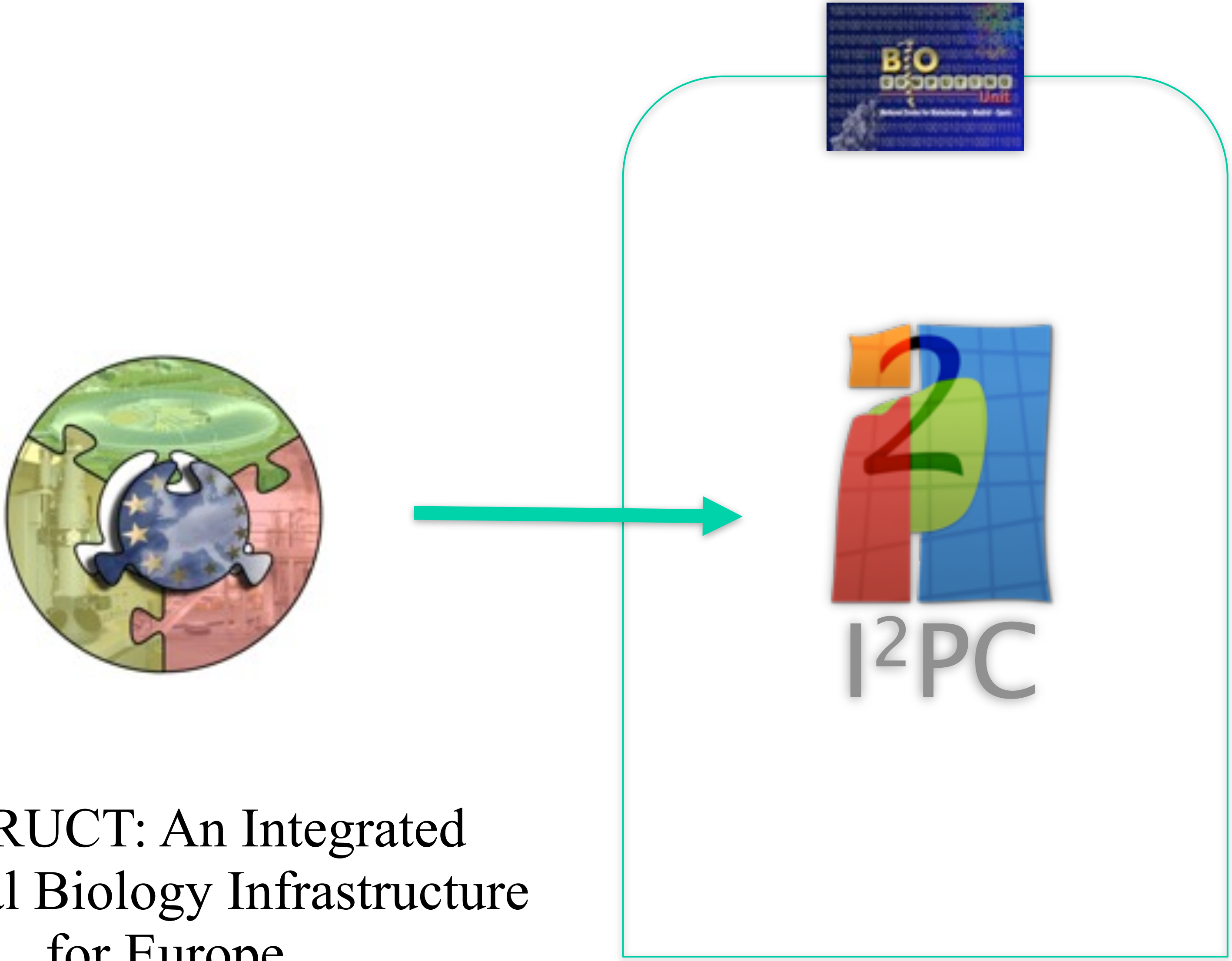
Scipion

an image processing framework
for 3D Electron Microscopy

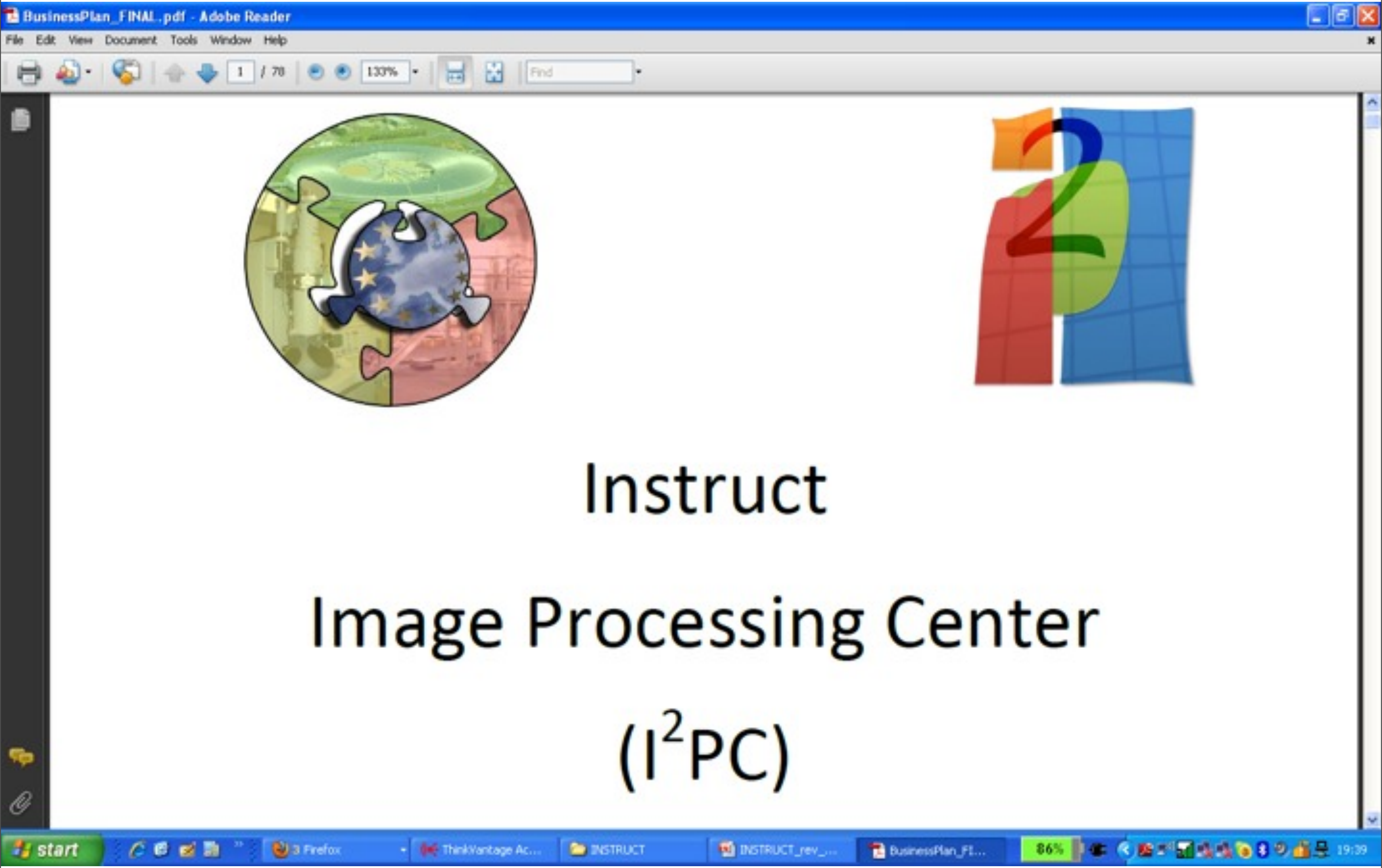
INSTRUCT: An Integrated Structural Biology Infrastructure for Europe



INSTRUCT: An Integrated Structural Biology Infrastructure for Europe



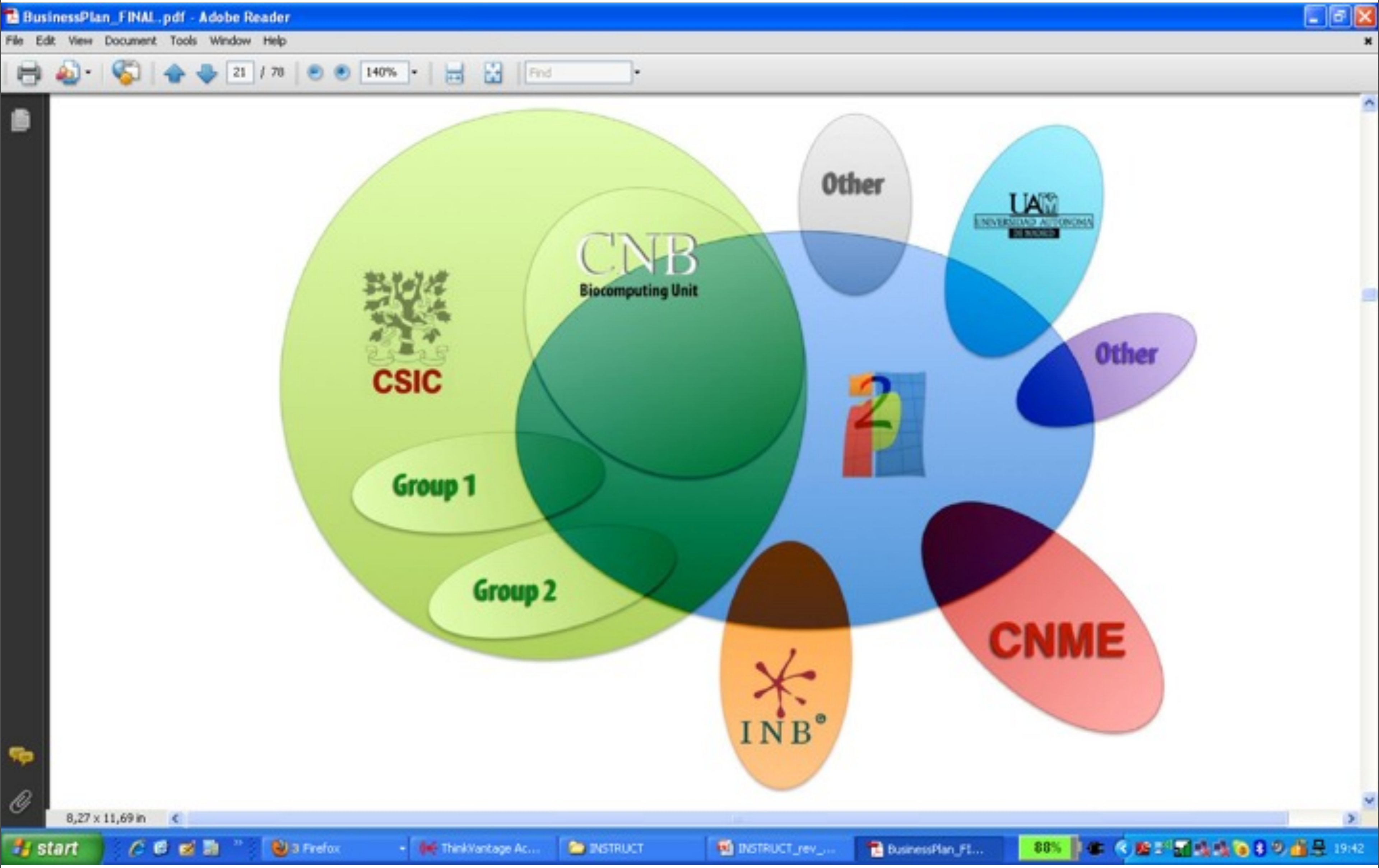
INSTRUCT: An Integrated Structural Biology Infrastructure for Europe

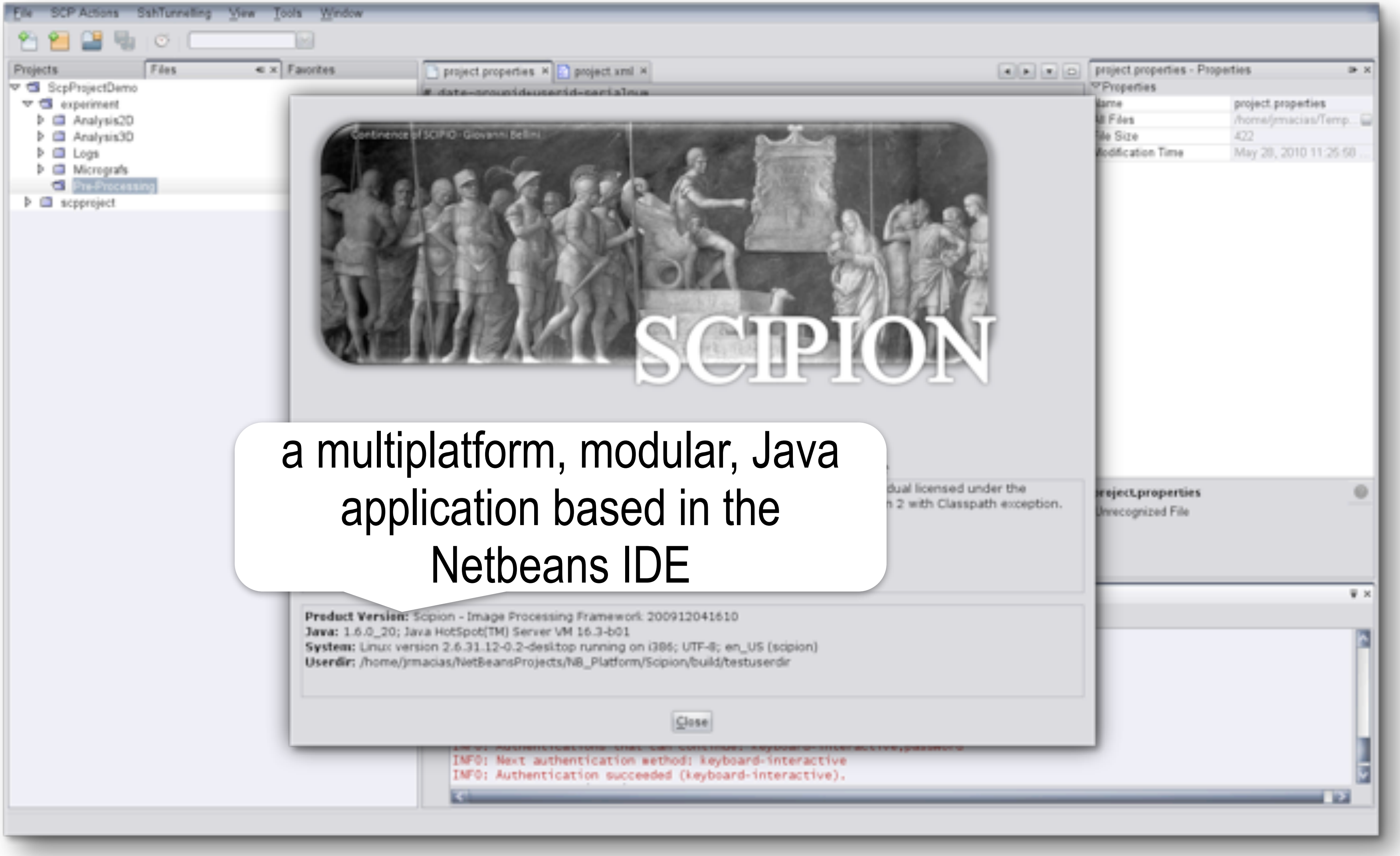


Instruct

Image Processing Center

(I^{2}PC)


BusinessPlan_FINAL.pdf - Adobe Reader
File Edit View Document Tools Window Help
21 / 78
140%
Find
CSIC
CNB
Biocomputing Unit
Group 1
Group 2
Other
UAM
UNIVERSIDAD AUTONOMA
DE MADRID
Other
CNME
INB
8,27 x 11,69 in
start
3 Firefox
ThinkVantage Ac...
INSTRUCT
INSTRUCT_rev_...
BusinessPlan_FI...
88%
19:42



lame	project.properties
ll Files	/home/jrmacias/Temp...
le Size	422
lodification Time	May 28, 2010 11:25:58 ...

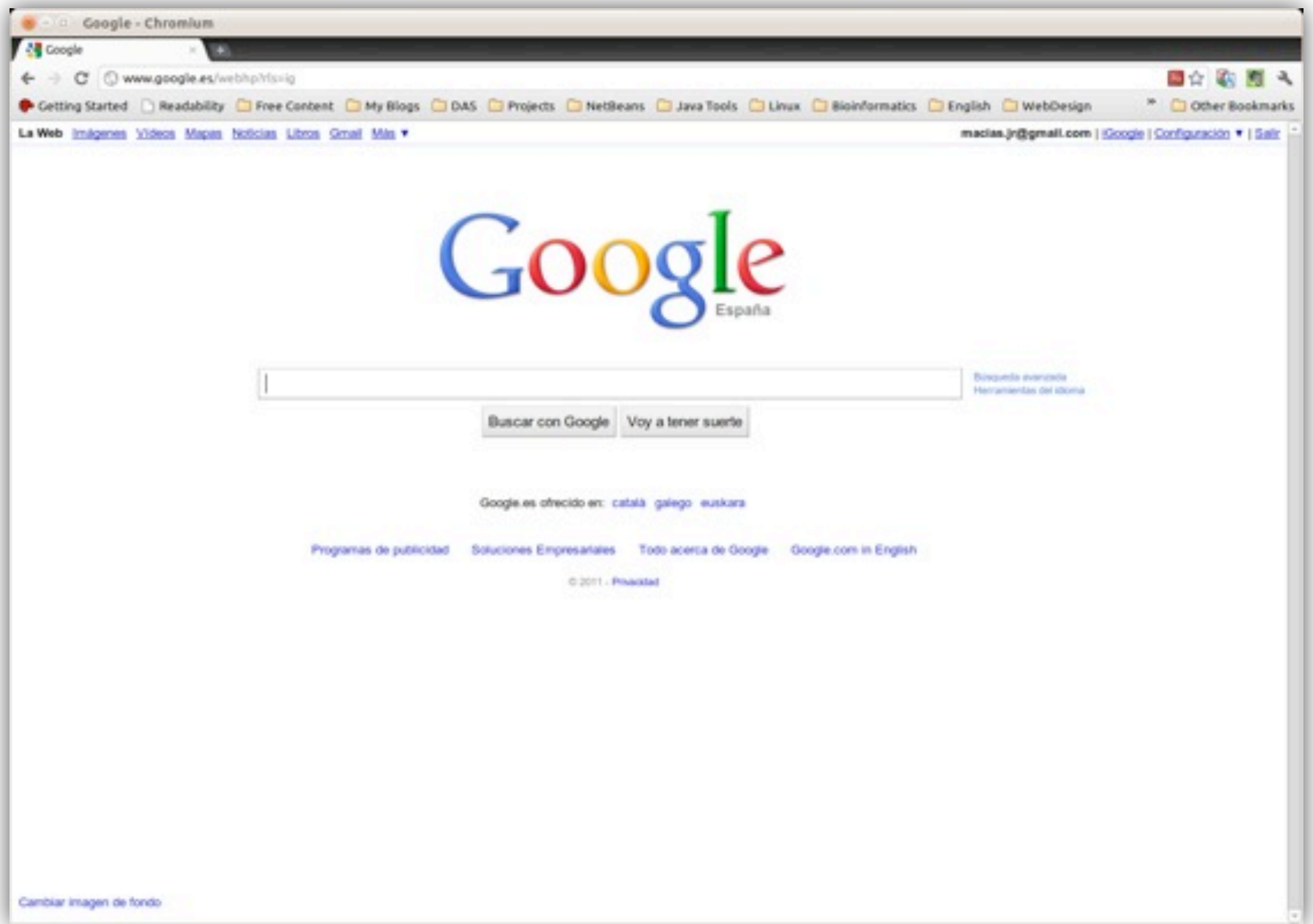
dual licensed under the
n 2 with Classpath exception.

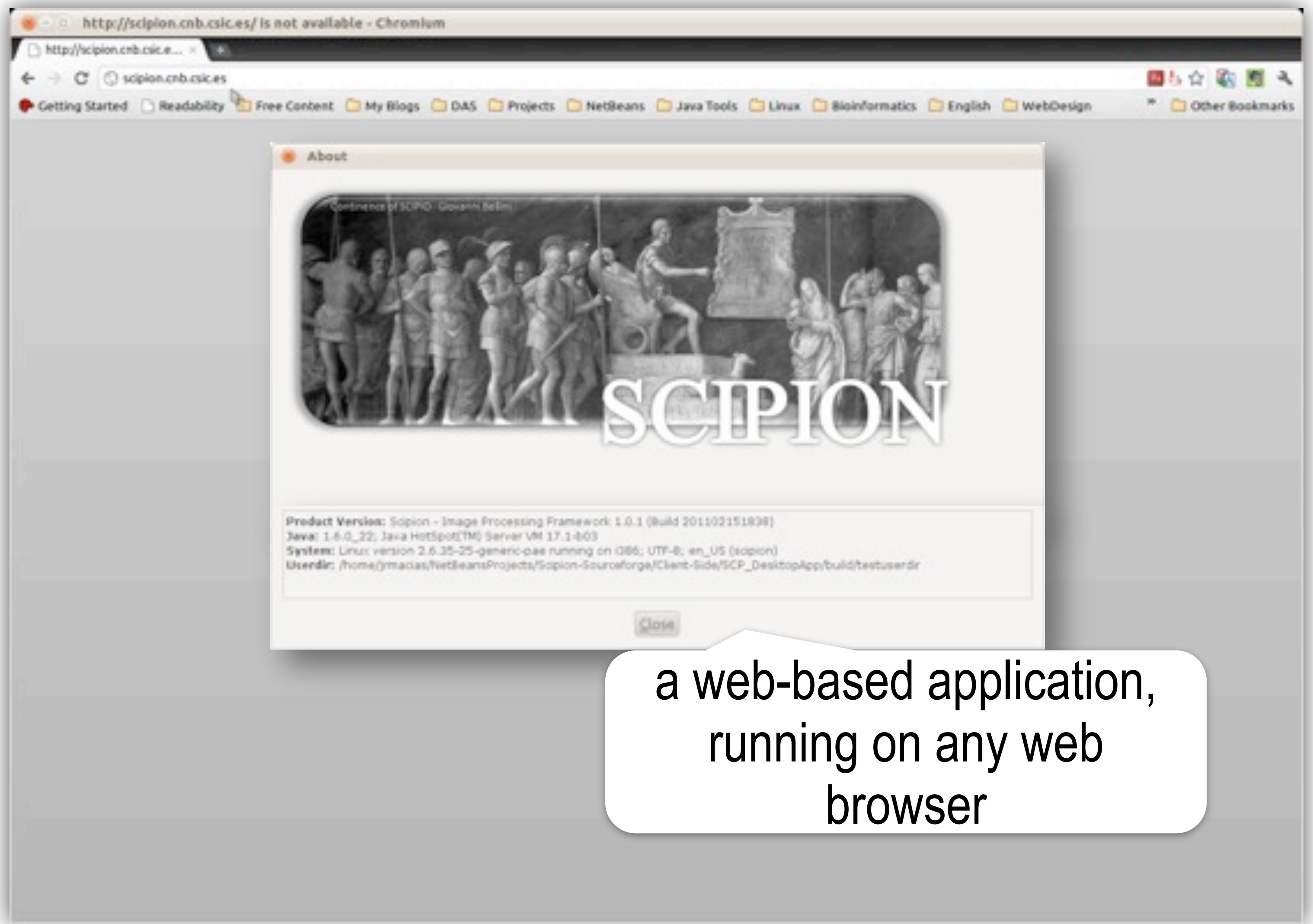
reject.properties
Unrecognized File

Product Version: Scipion - Image Processing Framework 200912041610
Java: 1.6.0_20; Java HotSpot(TM) Server VM 16.3-b01
System: Linux version 2.6.31.12-0.2-desktop running on i386; UTF-8; en_US (scipion)
Userdir: /home/jrmacias/NetBeansProjects/NB_Platform/Scipion/build/testuserdir

```
INFO: Next authentication method: keyboard-interactive
INFO: Authentication succeeded (keyboard-interactive).
```

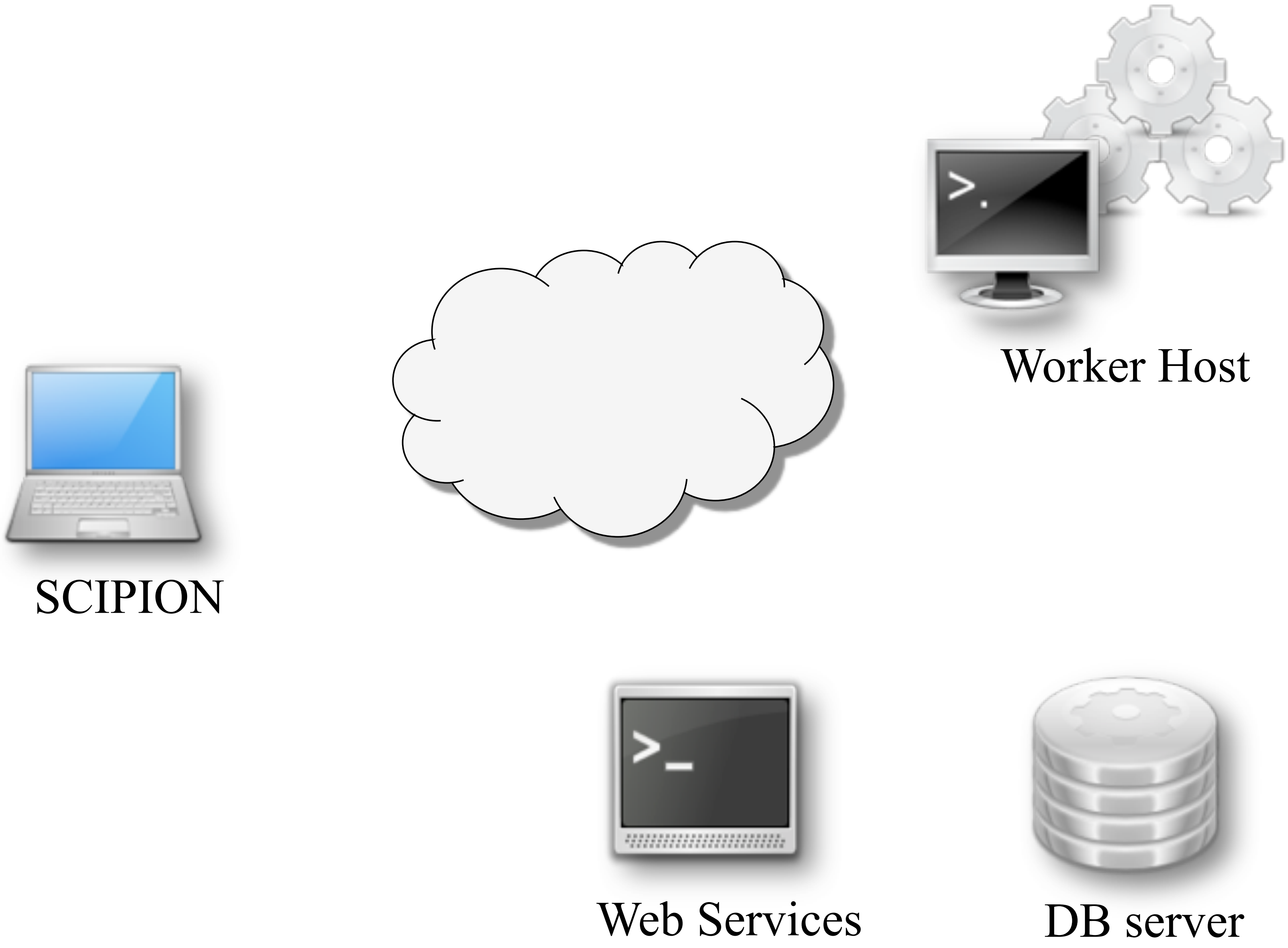
a multiplatform, modular, Java application based in the Netbeans IDE


Google - Chromium
Google
www.google.es/webhp?rls=ig
Getting Started
Readability
Free Content
My Blogs
DAS
Projects
NetBeans
Java Tools
Linux
Bioinformatics
English
WebDesign
Other Bookmarks
La Web
Imágenes
Vídeos
Mapas
Noticias
Libros
Gmail
Más ▾
macias.jr@gmail.com | iGoogle | Configuración ▾ | Salir
Google
España
Búsqueda avanzada
Herramientas del idioma
Buscar con Google
Voy a tener suerte
Google.es ofrecido en: català galego euskara
Programas de publicidad
Soluciones Empresariales
Todo acerca de Google
Google.com in English
© 2011 - Privacidad
Cambiar imagen de fondo



Product Version: Scipion - Image Processing Framework 1.0.1 (Build 201102151838)
Java: 1.6.0_22; Java HotSpot(TM) Server VM 17.1-b03
System: Linux version 2.6.35-25-generic-pae running on i386; UTF-8; en_US (scipion)
Userdir: /home/jmacias/NetBeansProjects/Scipion-Sourceforge/Client-Side/SCP_DesktopApp/build/testuserdir

a web-based application, running on any web browser

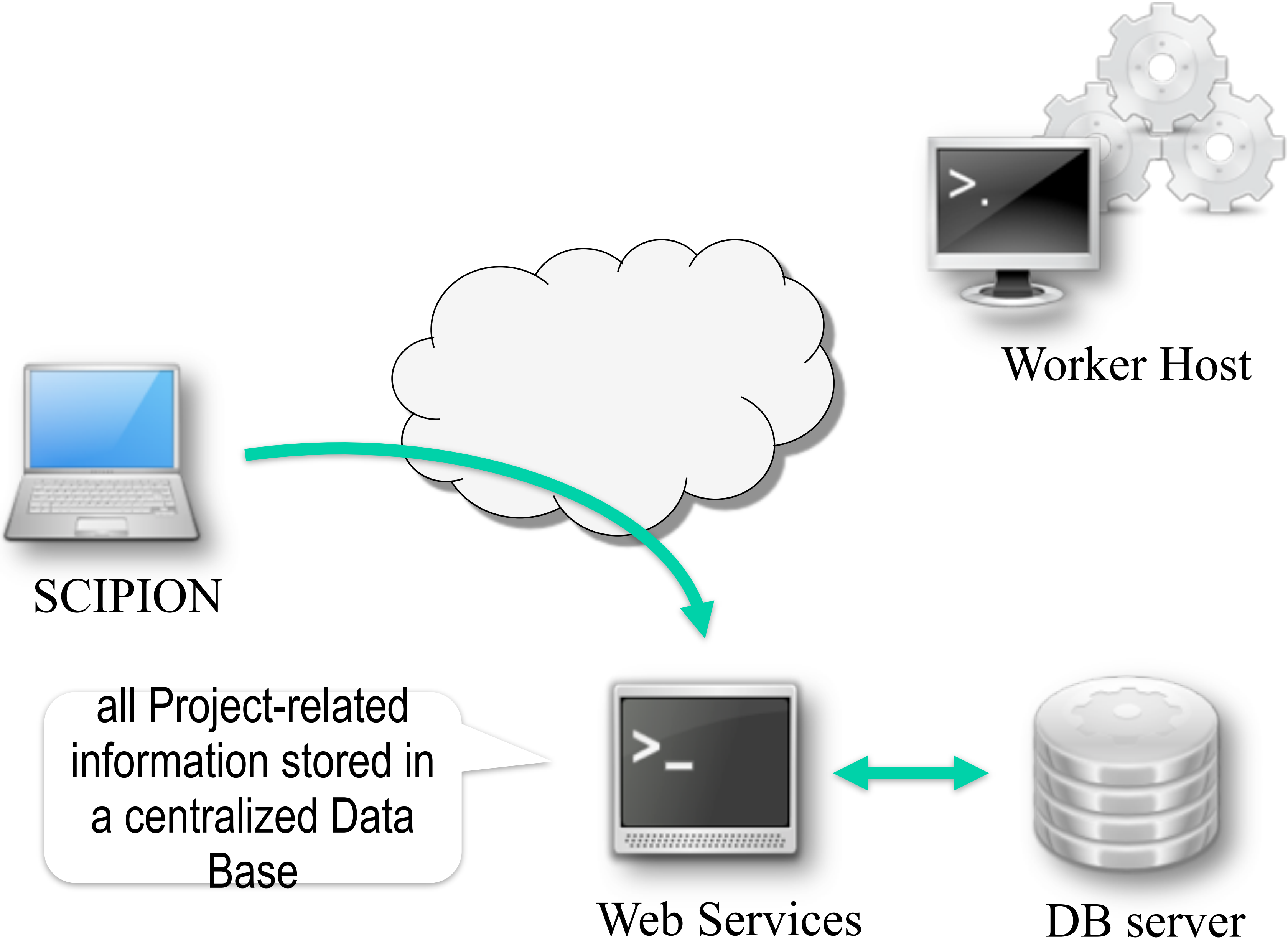


Worker Host

SCIPION

Web Services

DB server



Worker Host

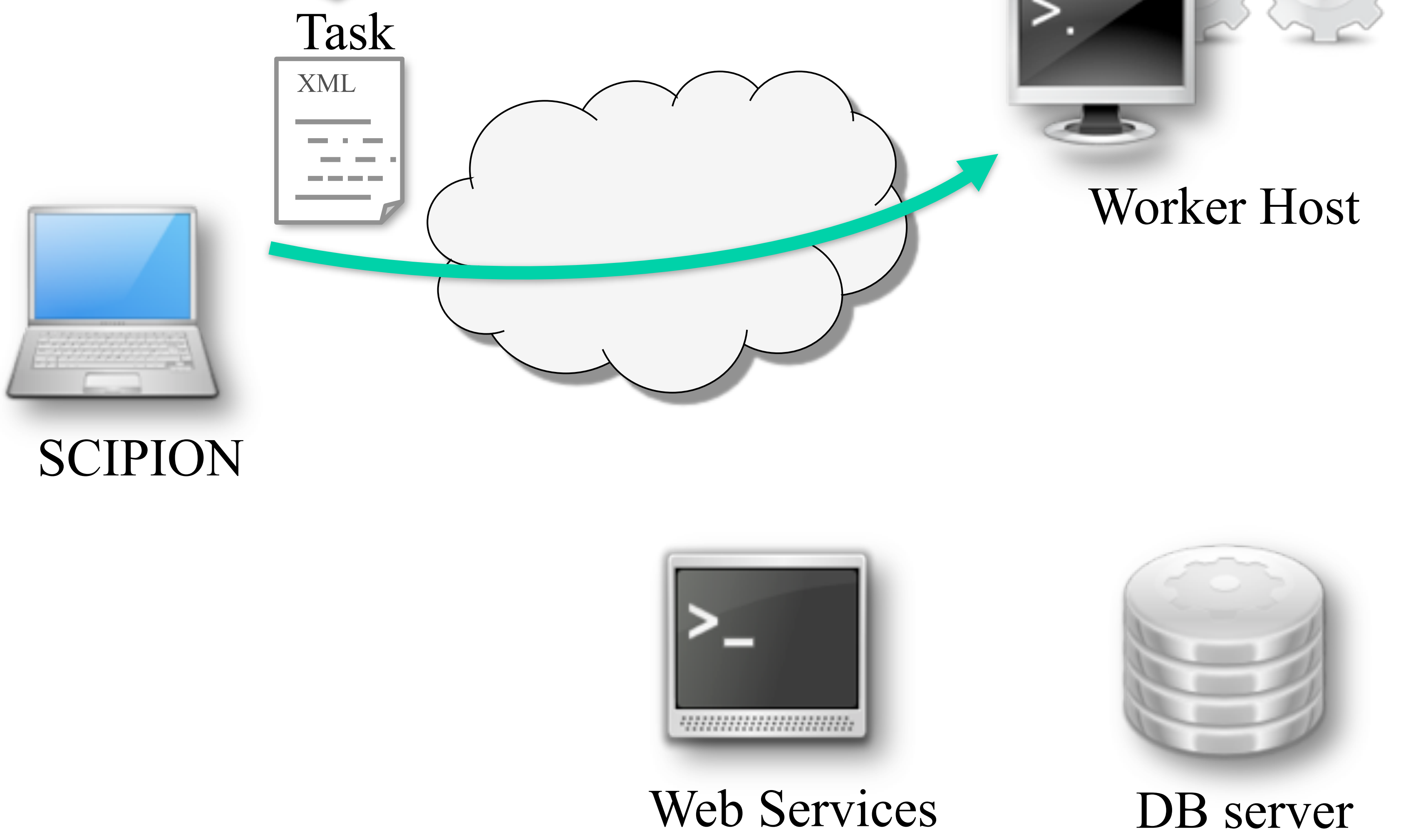
SCIPION

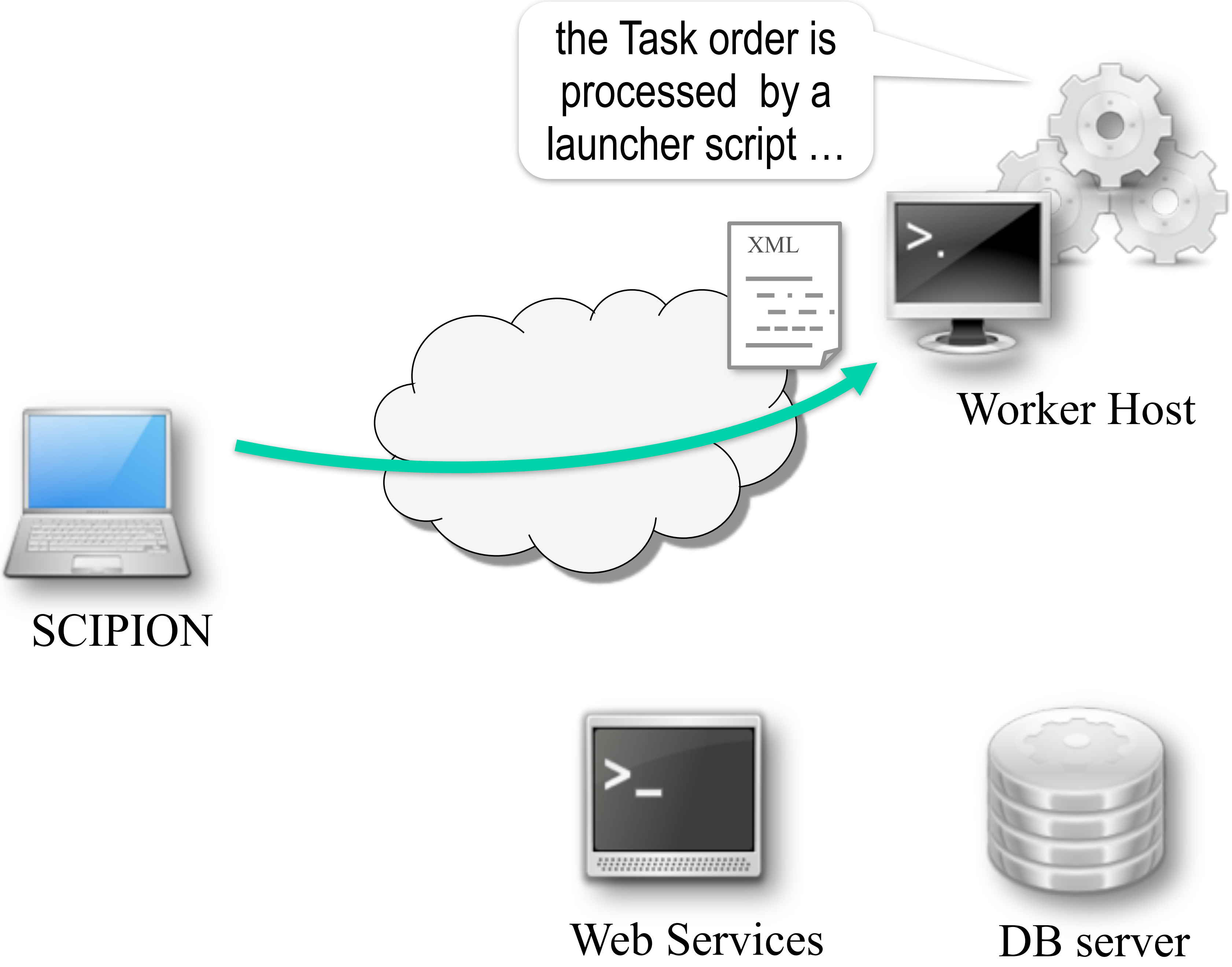
all Project-related information stored in a centralized Data Base

Web Services

DB server

an XML-formatted Task definition is created





the Task order is processed by a launcher script …

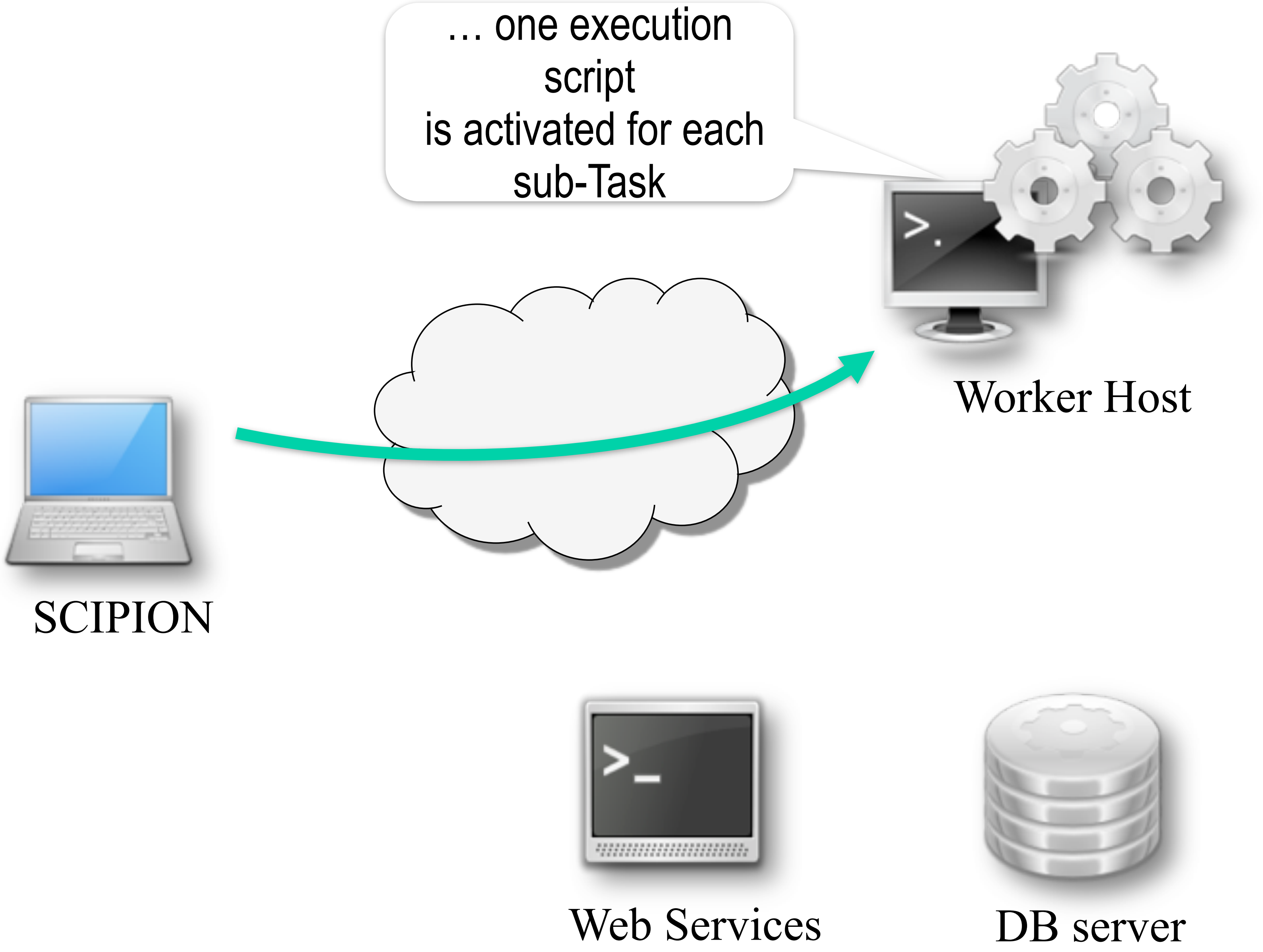
XML

Worker Host

SCIPION

Web Services

DB server



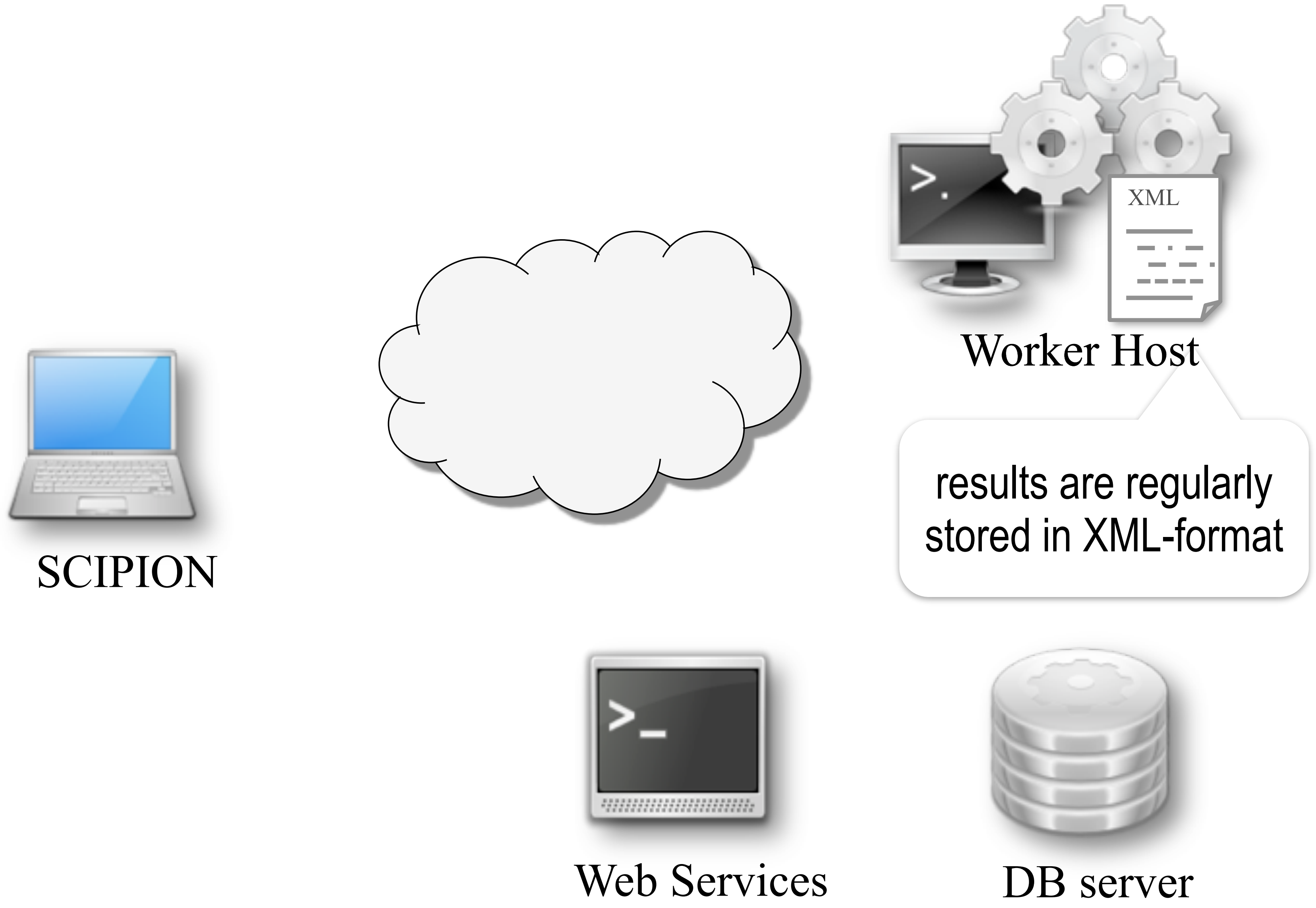
… one execution script
is activated for each sub-Task

Worker Host

SCIPION

Web Services

DB server



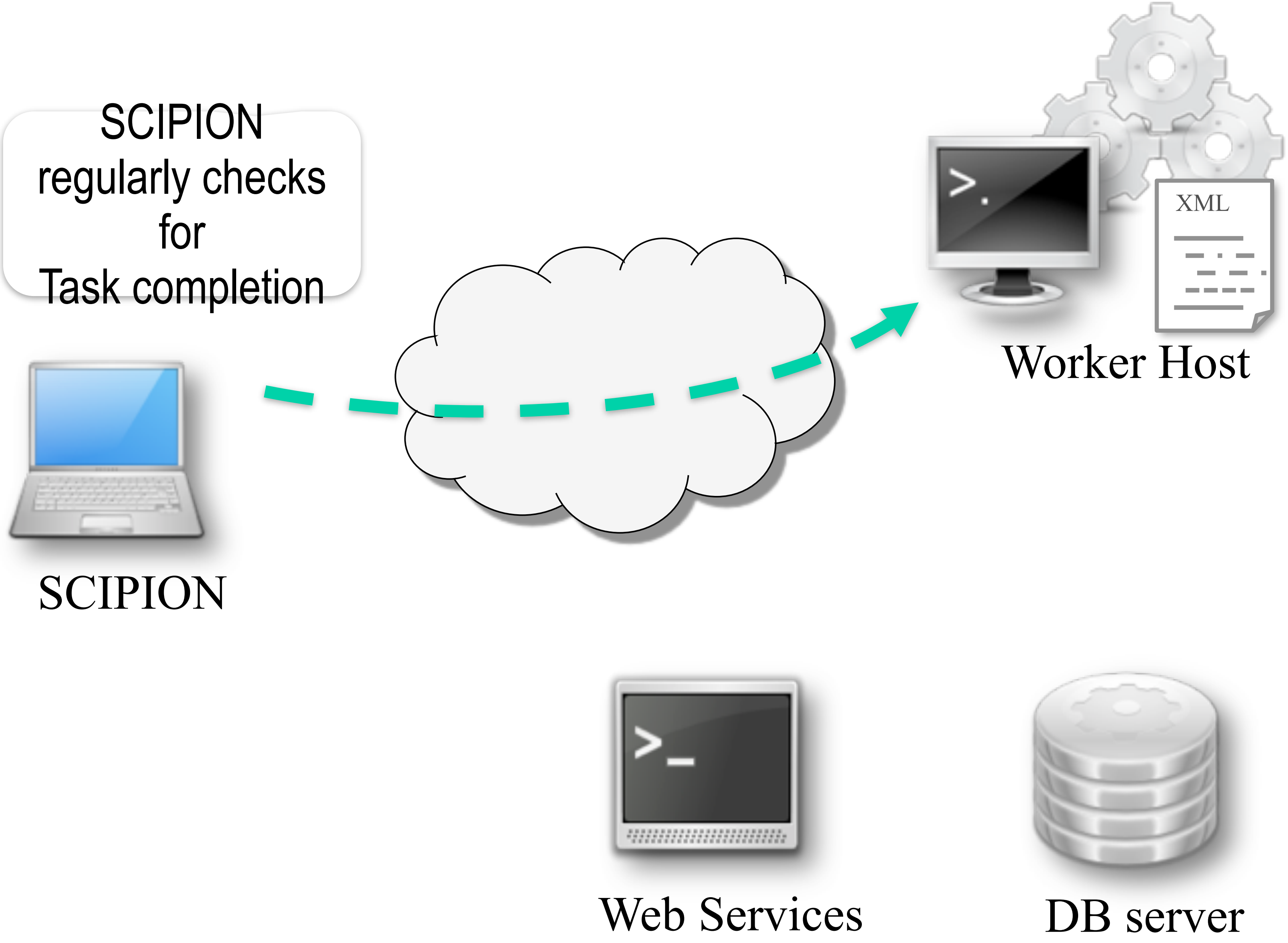
Worker Host

results are regularly stored in XML-format

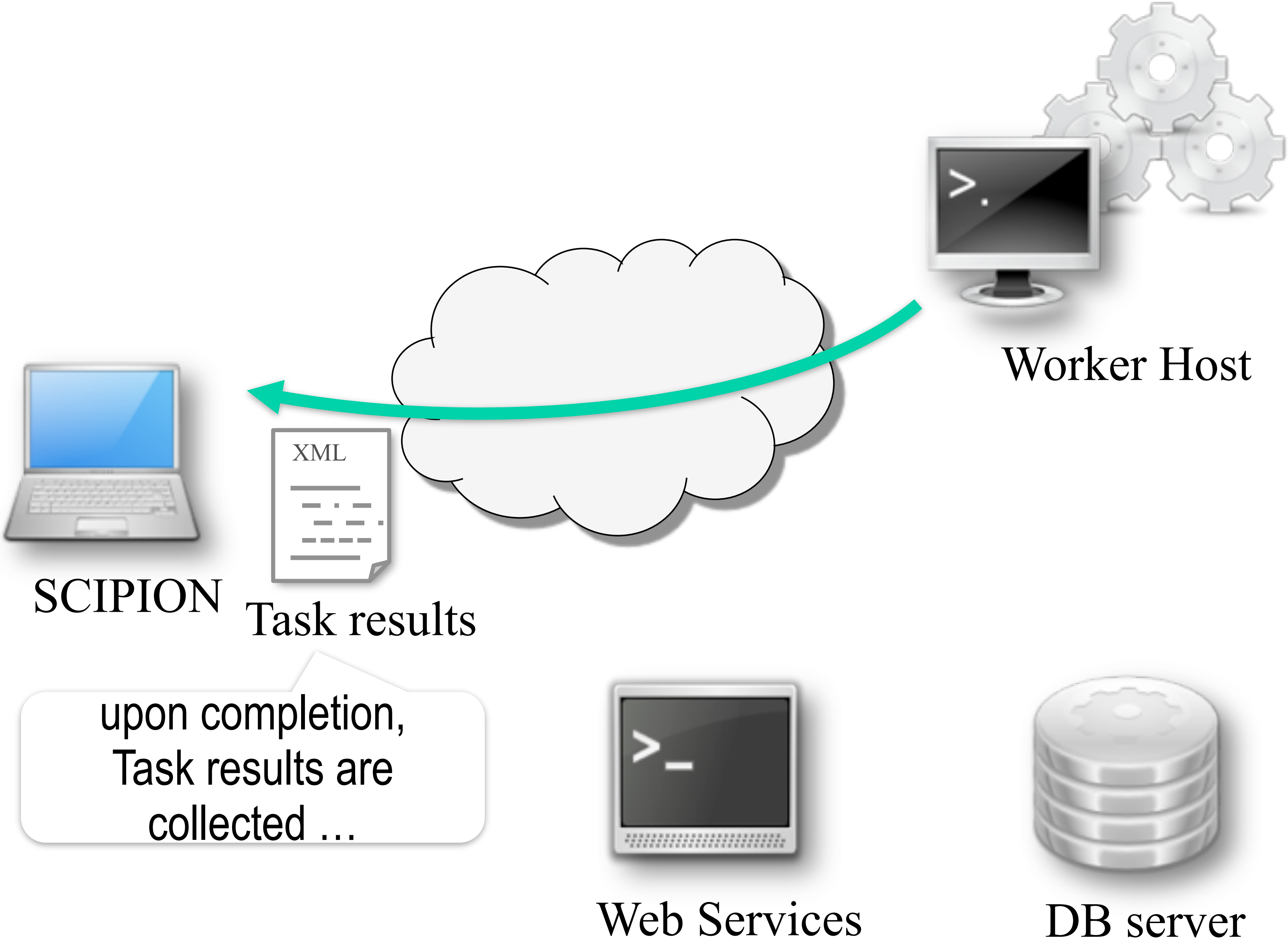
SCIPION

Web Services

DB server



SCIPION regularly checks for Task completion



Worker Host

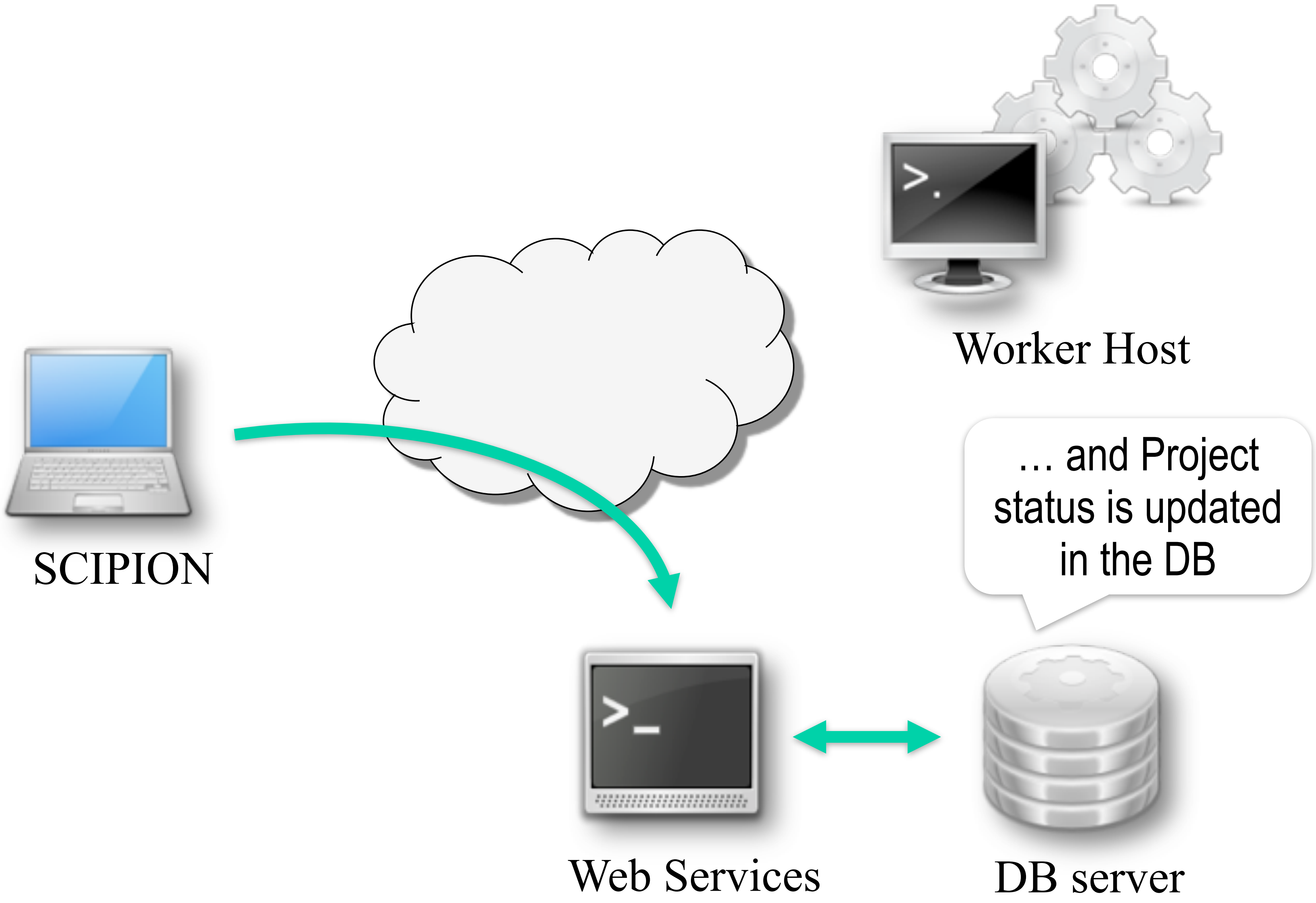
SCIPION

Task results

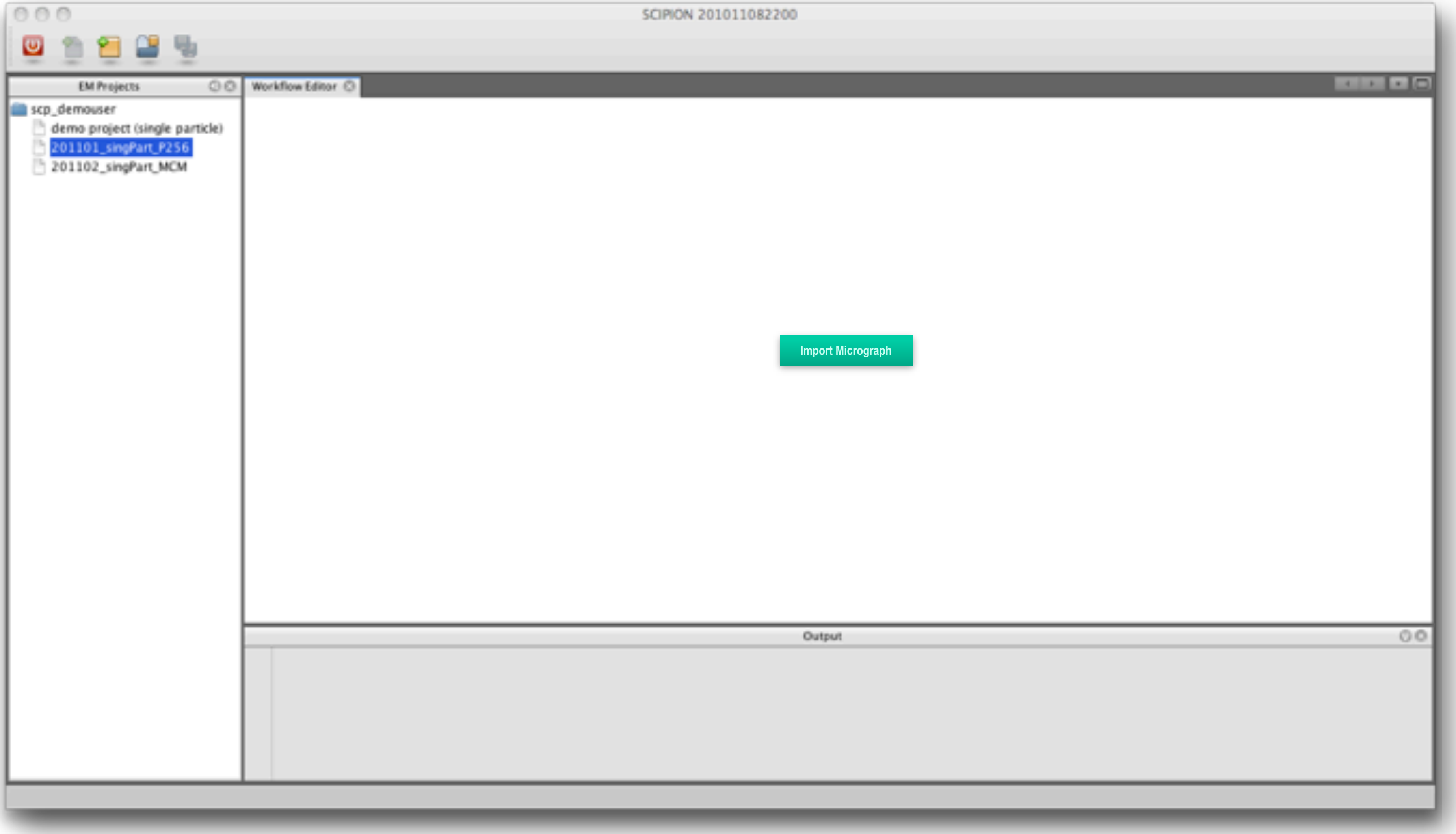
upon completion, Task results are collected …

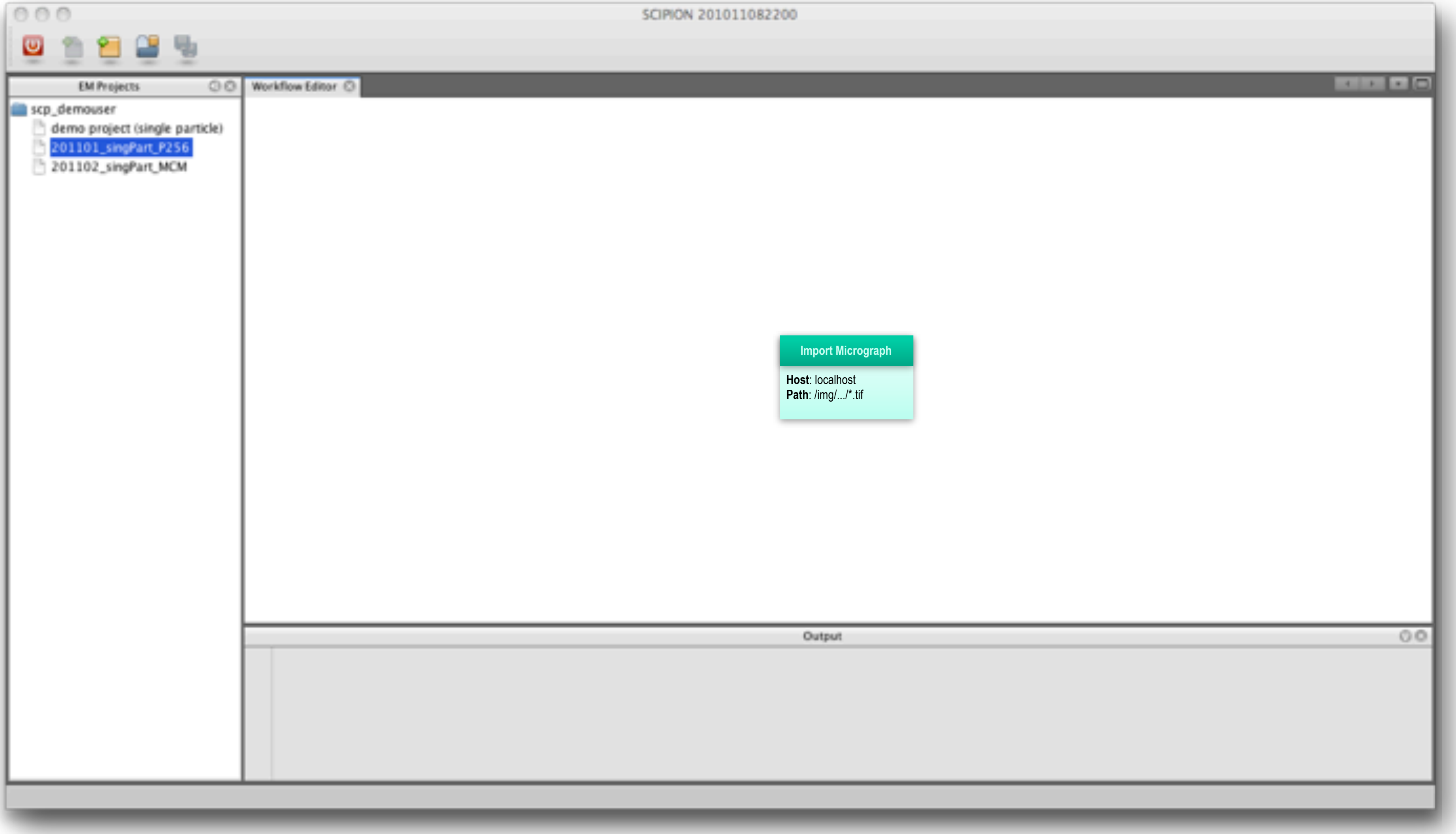
Web Services

DB server

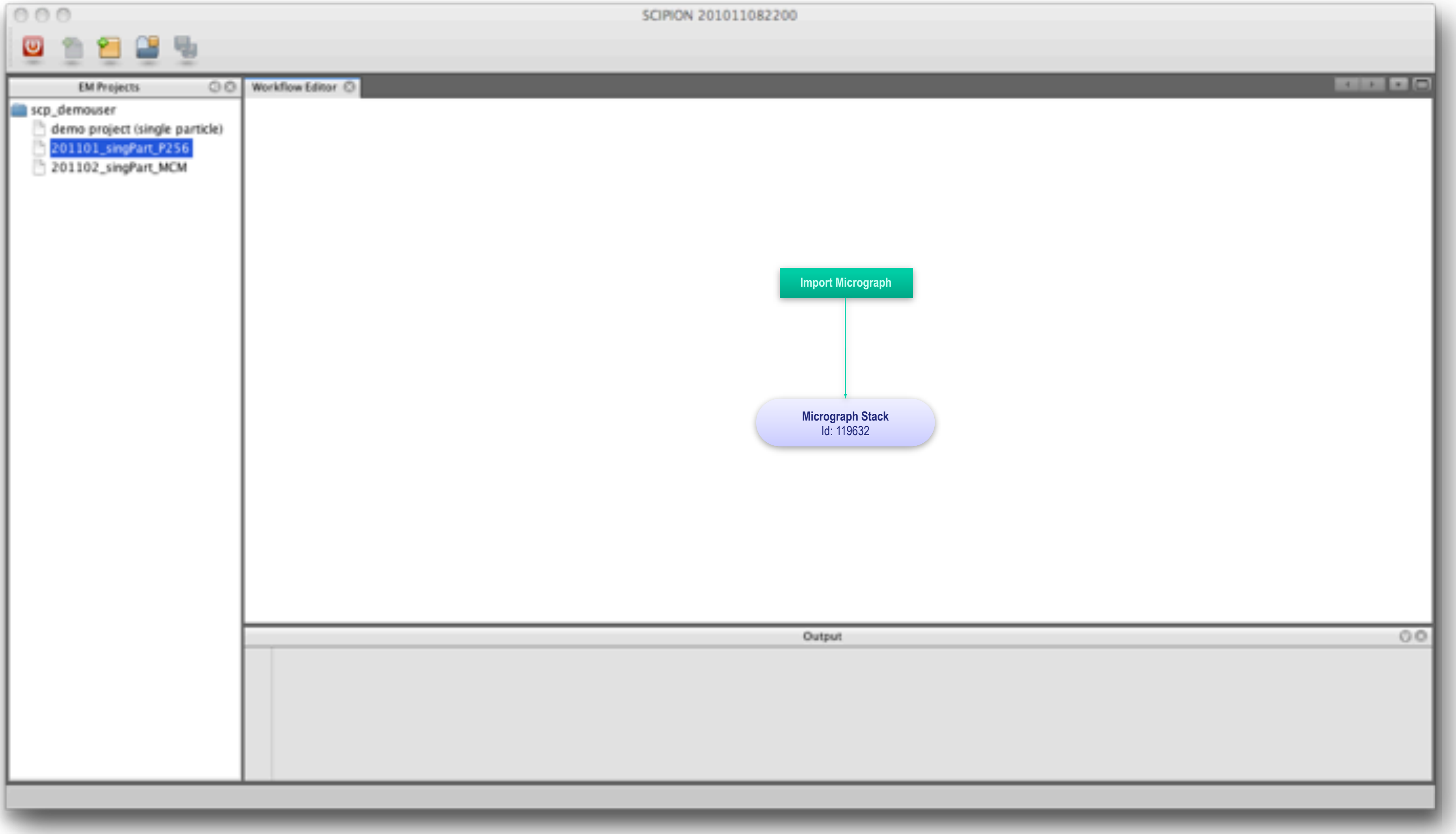


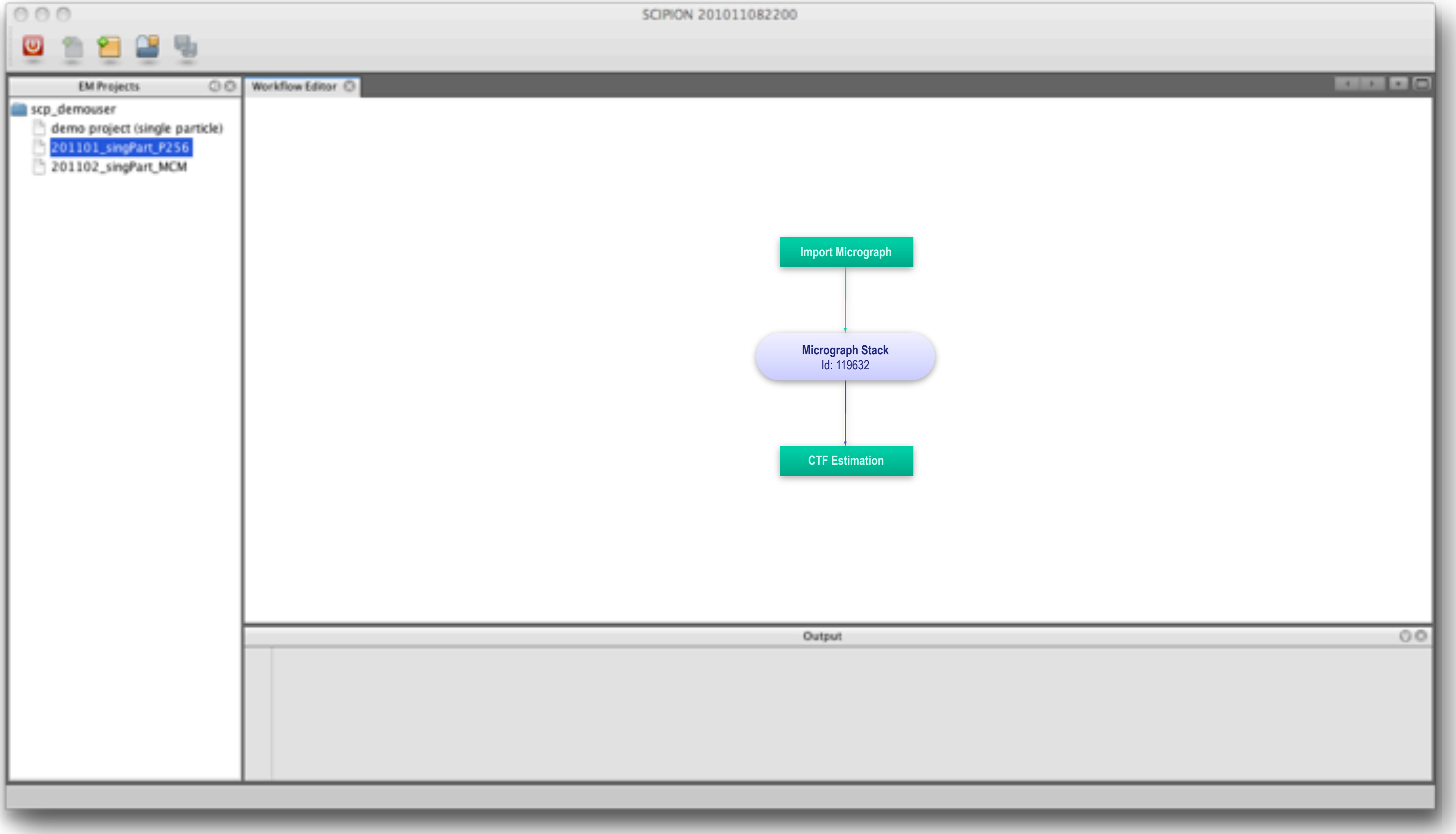
… and Project status is updated in the DB

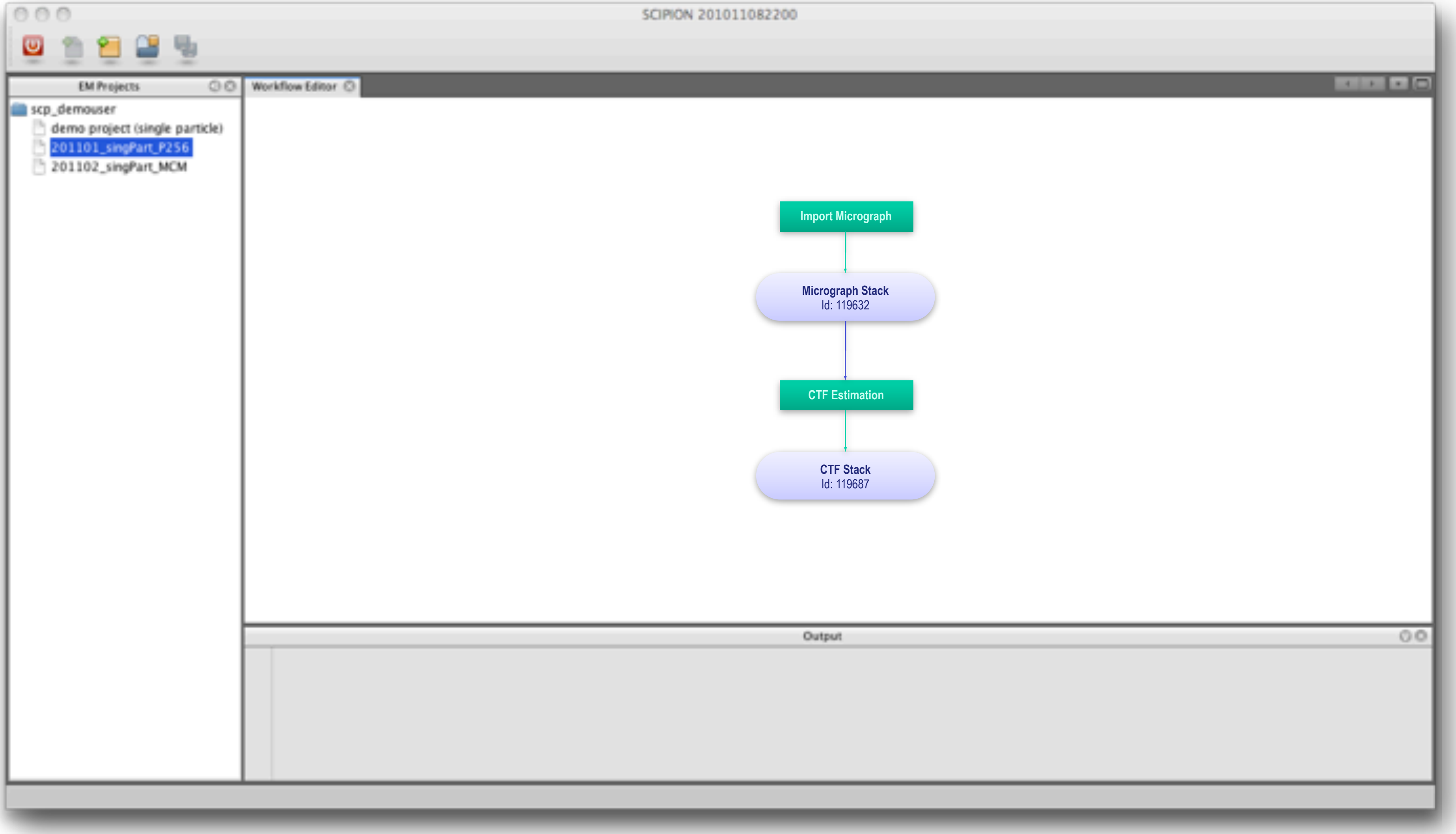

SCIPION 201011082200
EM Projects
scp_demouser
demo project (single particle)
201101_singPart_P256
201102_singPart_MCM
Workflow Editor
Import Micrograph
Output

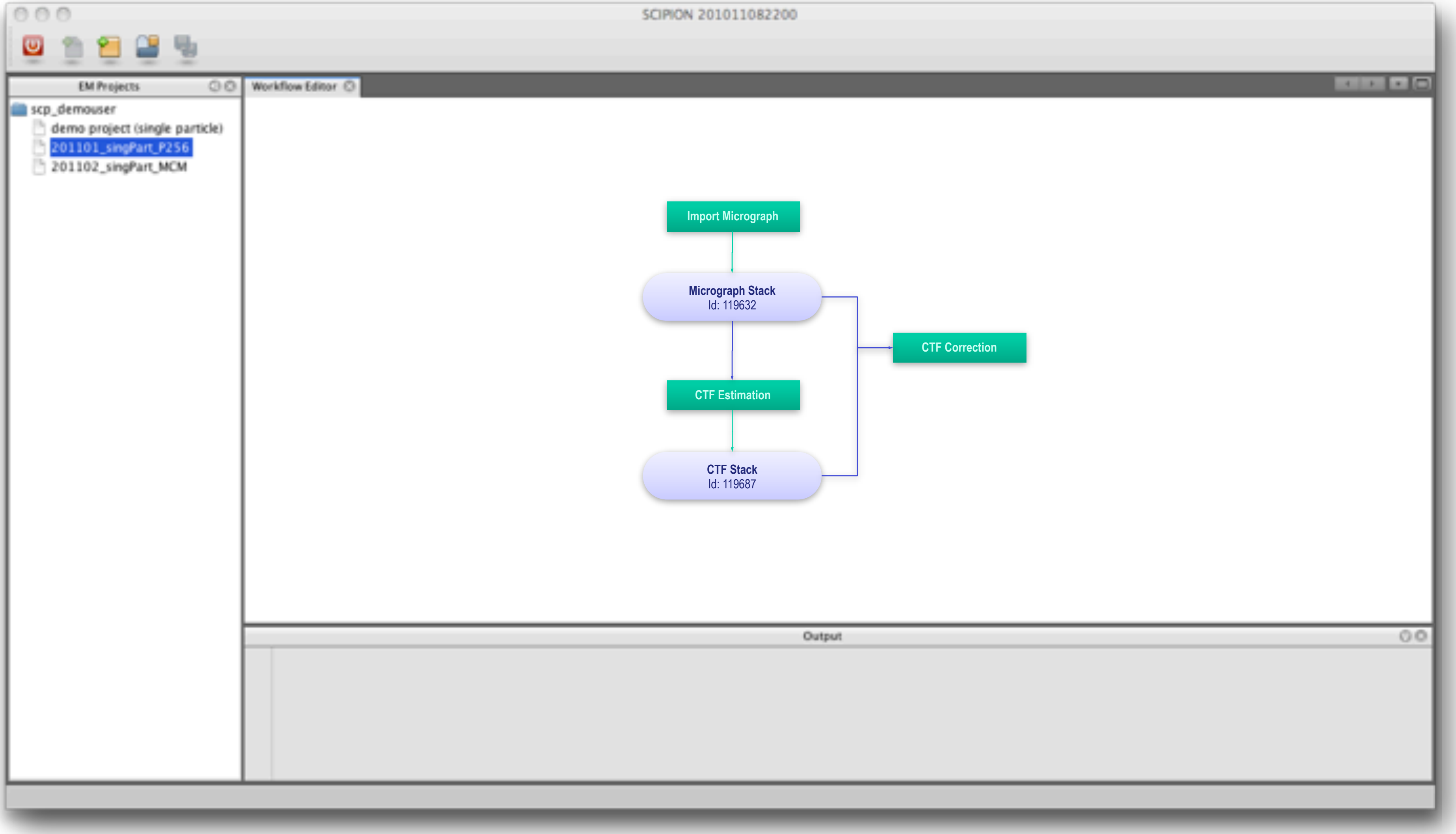


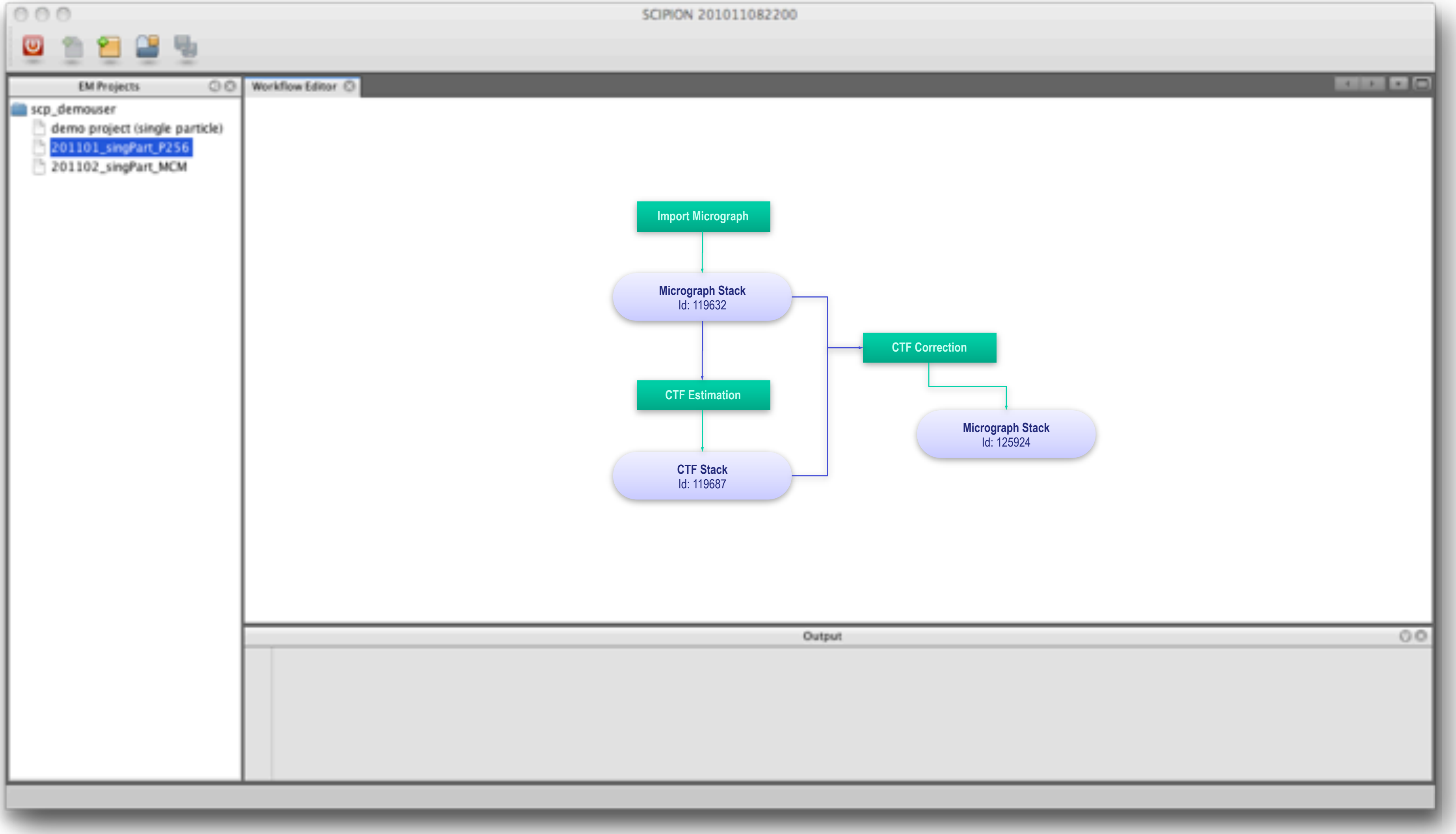
Host: localhost
Path: /img/.../*.tif

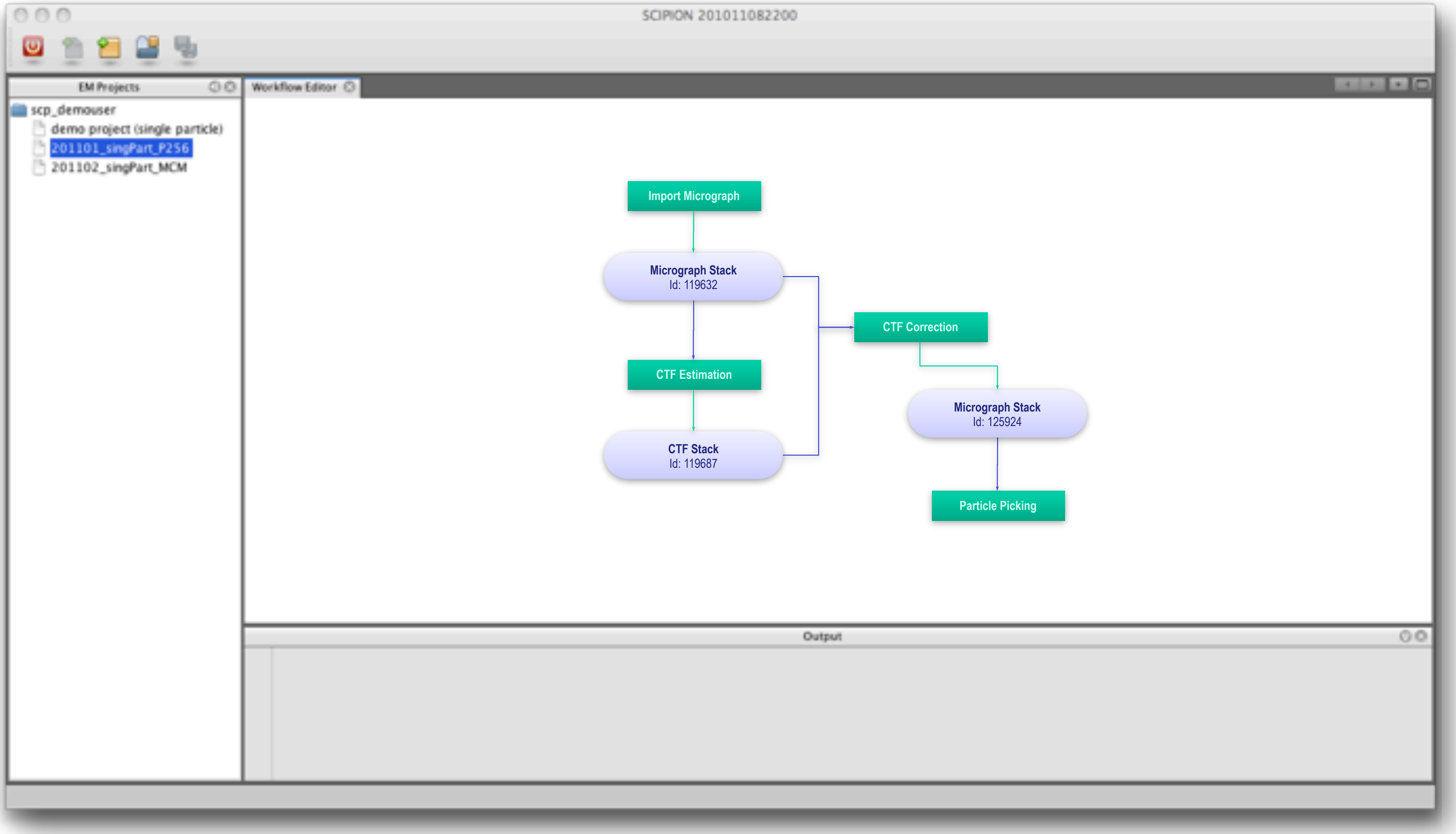

SCIPION 201011082200
EM Projects
scp_demouser
demo project (single particle)
201101_singPart_P256
201102_singPart_MCM
Workflow Editor
Import Micrograph
Micrograph Stack
Id: 119632
Output

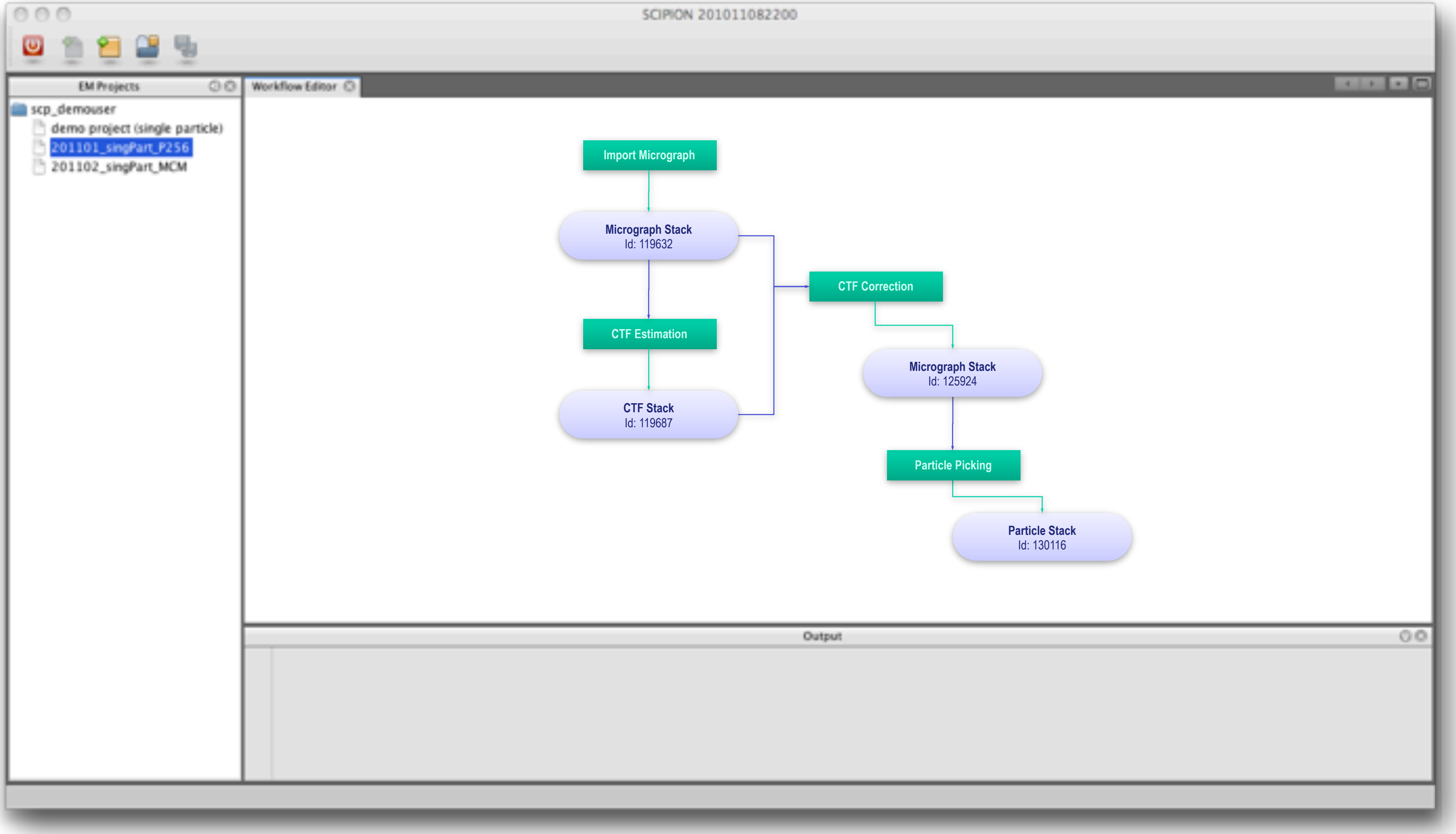

SCIPION 201011082200
EM Projects
scp_demouser
demo project (single particle)
201101_singPart_P256
201102_singPart_MCM
Workflow Editor
Import Micrograph
Micrograph Stack
Id: 119632
CTF Estimation
Output

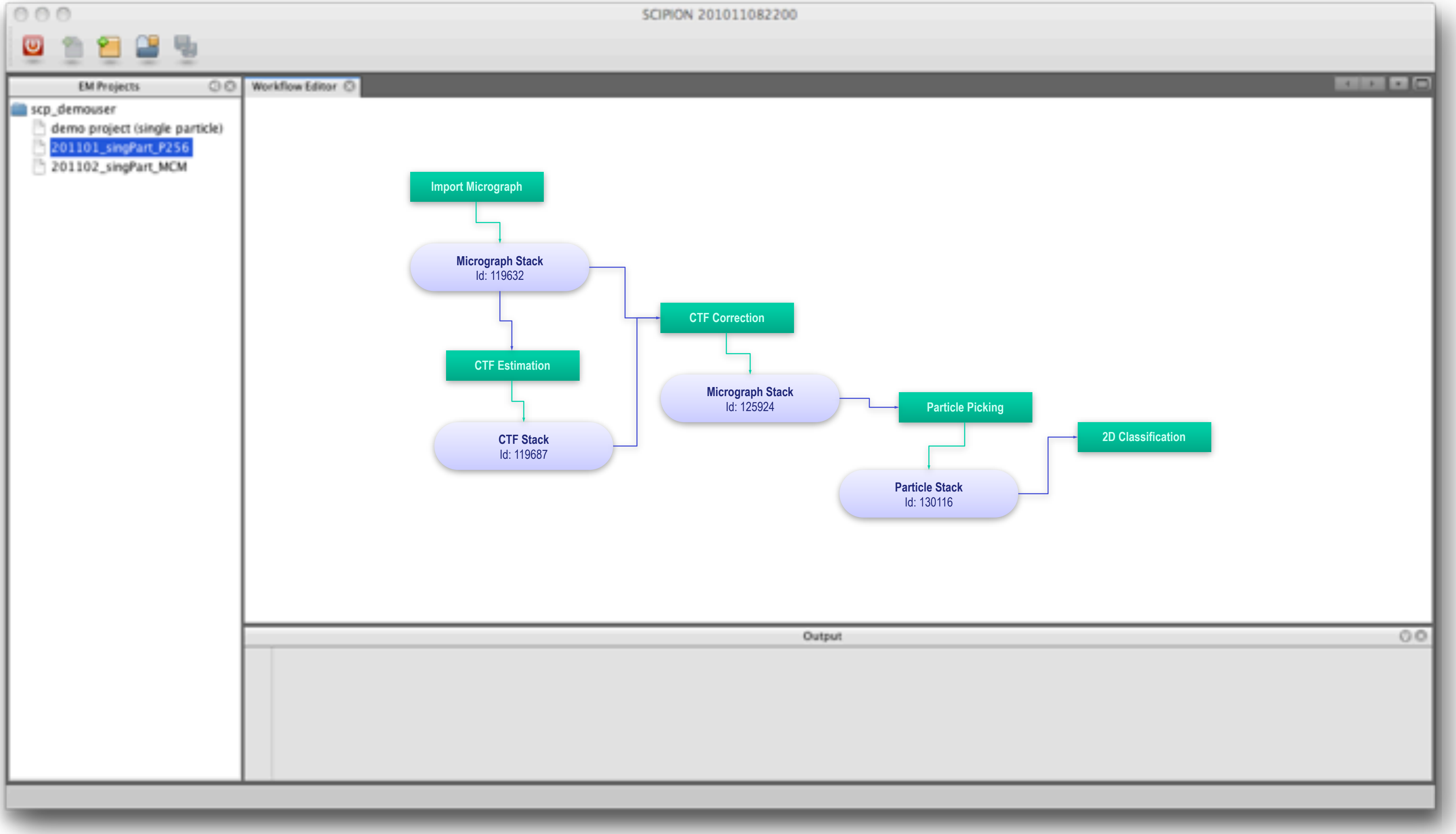

SCIPION 201011082200
EM Projects
scp_demouser
demo project (single particle)
201101_singPart_P256
201102_singPart_MCM
Workflow Editor
Import Micrograph
Micrograph Stack
Id: 119632
CTF Estimation
CTF Stack
Id: 119687
Output

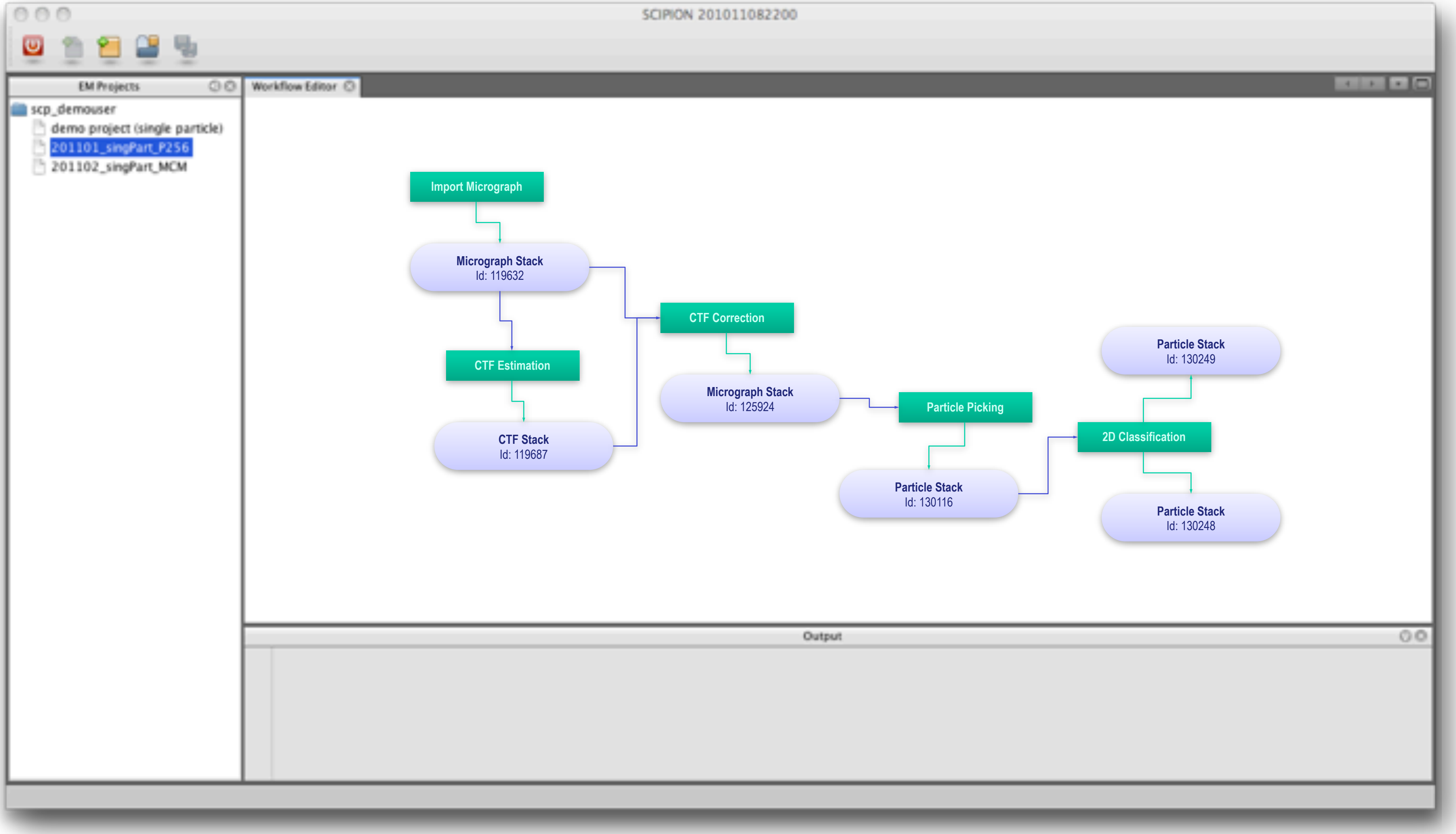

SCIPION 201011082200
EM Projects
scp_demouser
demo project (single particle)
201101_singPart_P256
201102_singPart_MCM
Workflow Editor
Import Micrograph
Micrograph Stack
Id: 119632
CTF Correction
CTF Estimation
CTF Stack
Id: 119687
Output

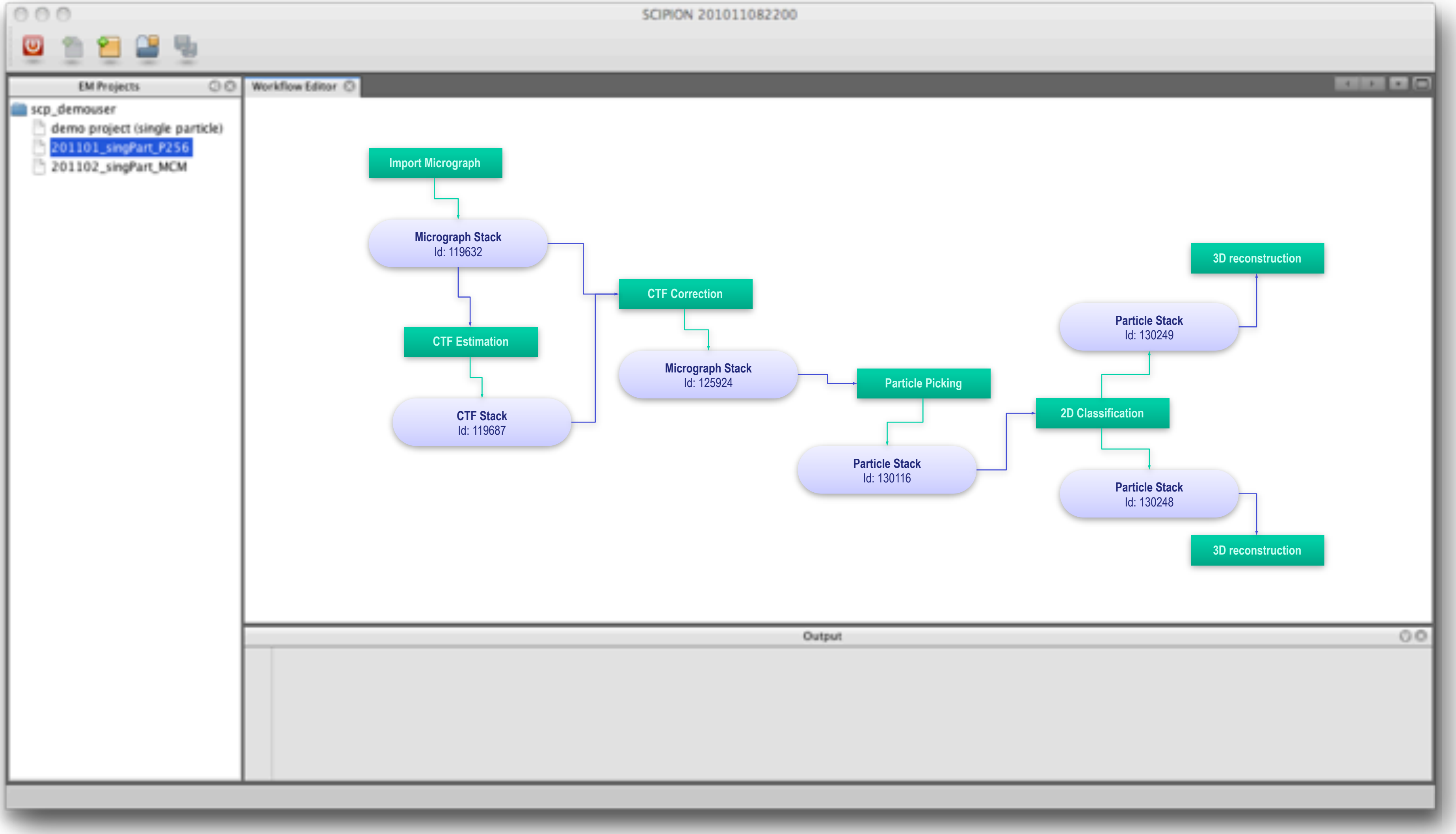

SCIPION 201011082200
EM Projects
scp_demouser
demo project (single particle)
201101_singPart_P256
201102_singPart_MCM
Workflow Editor
Import Micrograph
Micrograph Stack
Id: 119632
CTF Estimation
CTF Stack
Id: 119687
CTF Correction
Micrograph Stack
Id: 125924
Output

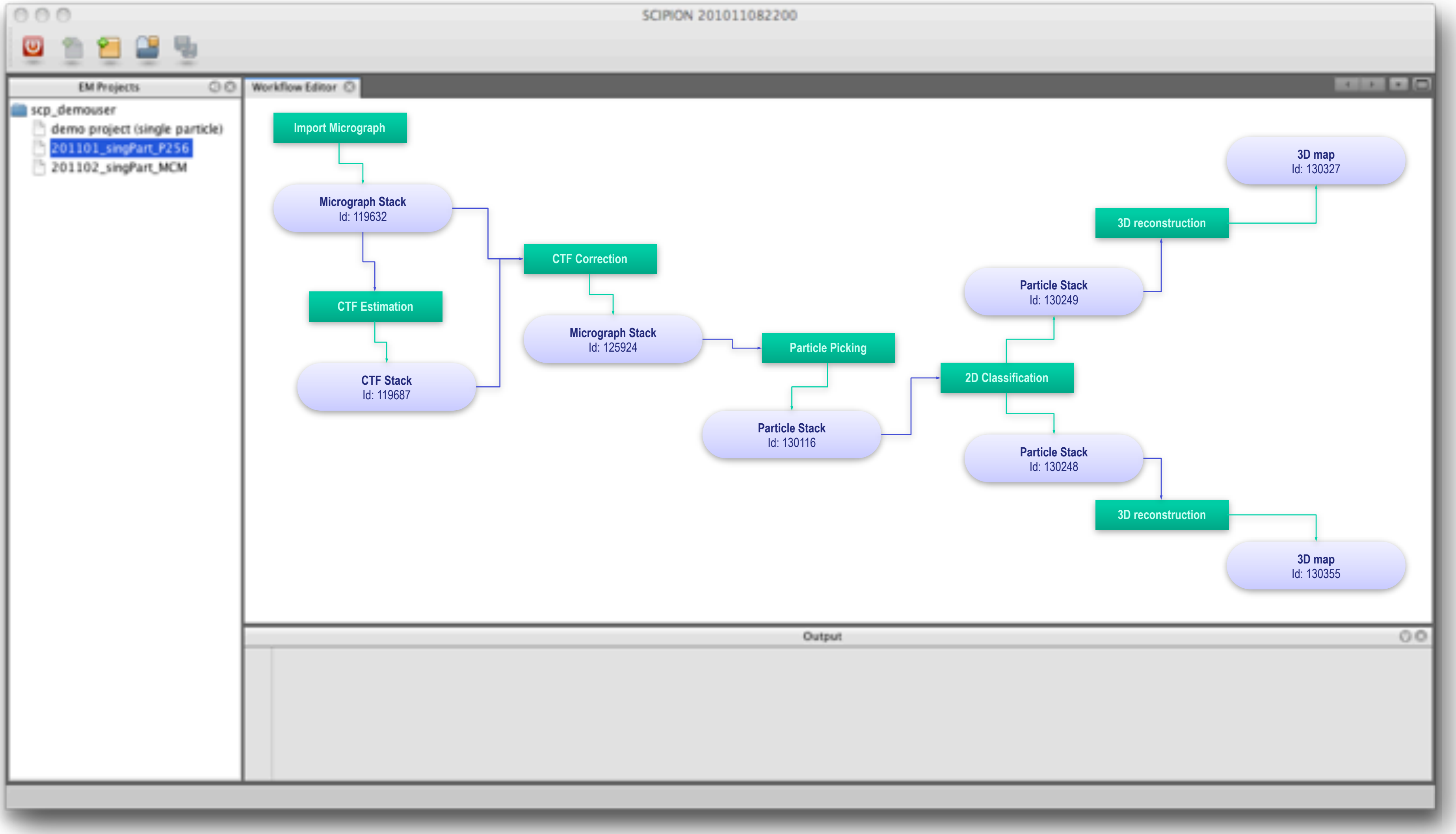

SCIPION 201011082200
EM Projects
scp_demouser
demo project (single particle)
201101_singPart_P256
201102_singPart_MCM
Workflow Editor
Import Micrograph
Micrograph Stack
Id: 119632
CTF Estimation
CTF Stack
Id: 119687
CTF Correction
Micrograph Stack
Id: 125924
Particle Picking
Output

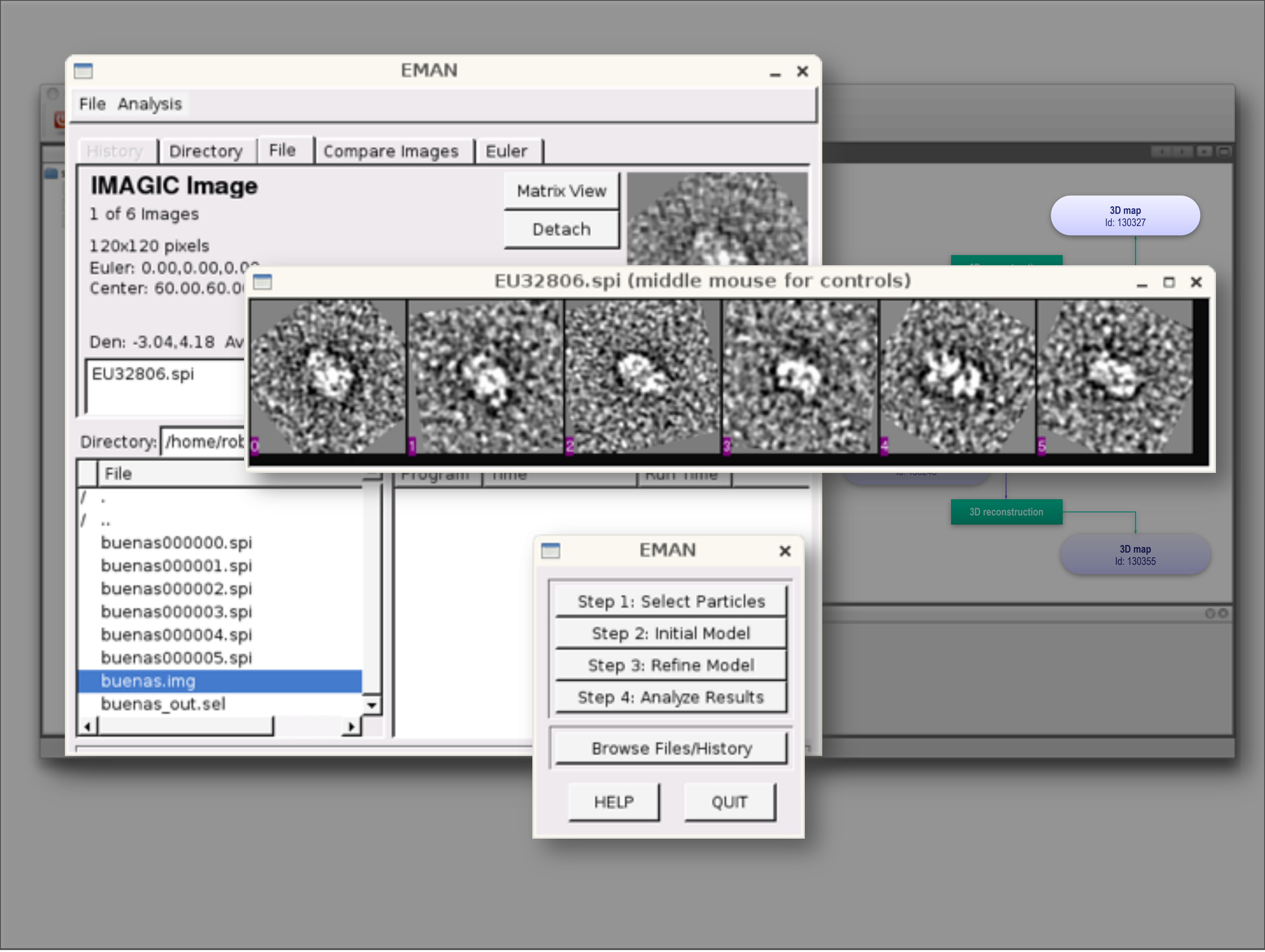

SCIPION 201011082200
EM Projects
scp_demouser
demo project (single particle)
201101_singPart_P256
201102_singPart_MCM
Workflow Editor
Import Micrograph
Micrograph Stack
Id: 119632
CTF Estimation
CTF Stack
Id: 119687
CTF Correction
Micrograph Stack
Id: 125924
Particle Picking
Particle Stack
Id: 130116
Output


SCIPION 201011082200
EM Projects
scp_demouser
demo project (single particle)
201101_singPart_P256
201102_singPart_MCM
Workflow Editor
Import Micrograph
Micrograph Stack
Id: 119632
CTF Estimation
CTF Stack
Id: 119687
CTF Correction
Micrograph Stack
Id: 125924
Particle Picking
Particle Stack
Id: 130116
2D Classification
Output


SCIPION 201011082200
EM Projects
scp_demouser
demo project (single particle)
201101_singPart_P256
201102_singPart_MCM
Workflow Editor
Import Micrograph
Micrograph Stack
Id: 119632
CTF Estimation
CTF Stack
Id: 119687
CTF Correction
Micrograph Stack
Id: 125924
Particle Picking
Particle Stack
Id: 130116
2D Classification
Particle Stack
Id: 130249
Particle Stack
Id: 130248
Output


SCIPION 201011082200
EM Projects
scp_demouser
demo project (single particle)
201101_singPart_P256
201102_singPart_MCM
Workflow Editor
Import Micrograph
Micrograph Stack
Id: 119632
CTF Estimation
CTF Stack
Id: 119687
CTF Correction
Micrograph Stack
Id: 125924
Particle Picking
Particle Stack
Id: 130116
2D Classification
Particle Stack
Id: 130249
3D reconstruction
Particle Stack
Id: 130248
3D reconstruction
Output


SCIPION 201011082200
EM Projects
scp_demouser
demo project (single particle)
201101_singPart_P256
201102_singPart_MCM
Workflow Editor
Import Micrograph
Micrograph Stack
Id: 119632
CTF Estimation
CTF Stack
Id: 119687
CTF Correction
Micrograph Stack
Id: 125924
Particle Picking
Particle Stack
Id: 130116
2D Classification
Particle Stack
Id: 130249
3D reconstruction
3D map
Id: 130327
Particle Stack
Id: 130248
3D reconstruction
3D map
Id: 130355
Output



IMAGIC Image

1 of 6 Images

120x120 pixels

Euler: 0.00,0.00,0.0

Center: 60.00,60.0

Den: -3.04,4.18 Av

EU32806.spi

Directory: /home/rob

File

```
/ .
/ ..
buenas000000.spi
buenas000001.spi
buenas000002.spi
buenas000003.spi
buenas000004.spi
buenas000005.spi
buenas.img
buenas_out.sel
```

Advantages of using SCIPION

Traceability
- covering in detail all the steps involved in a project, registering all the participating parameters, input and output data.

Standardization and Normalization
- of protocols that can then be reviewed and followed by other colleagues, allowing “learning by example”.

Repeatability
- with a new set of parameter values, as a first step towards…

Automation
- reducing the manual intervention in tedious and repetitive duties, so releasing more resources to other tasks.

ACKNOWLEDGEMENTS

- To our colleagues in the 26S team

Nickell S, Beck F, Scheres SH, Korinek A, Förster F, Lasker K, Mihalache O, Sun N, Nagy I, Sali A, Plitzko JM, Mann M, Baumeister W.

- To all the me

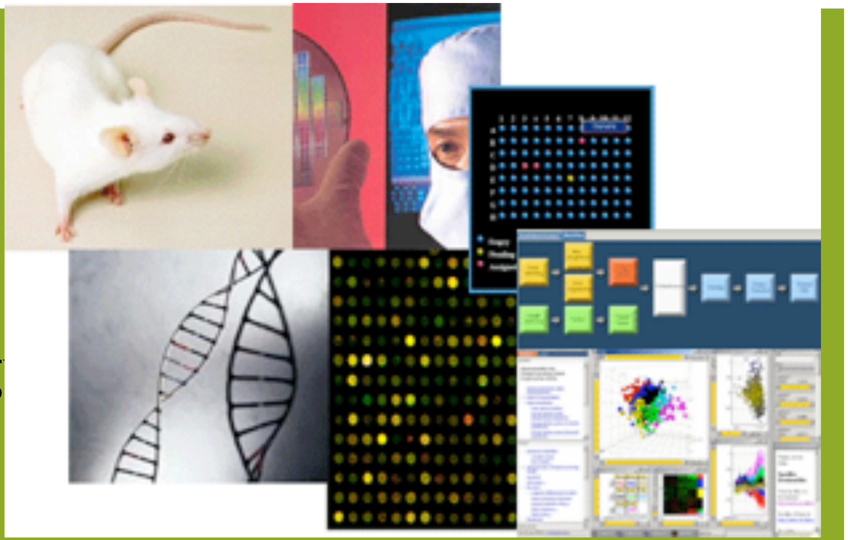
Integromics in the nutshell:

And relationship with other projects

Integromics™ | EXECUTIVE PRESENTATION

J.M.Carazo, Founder

Professor of Molecular Biology
National Center for Biotechnolog



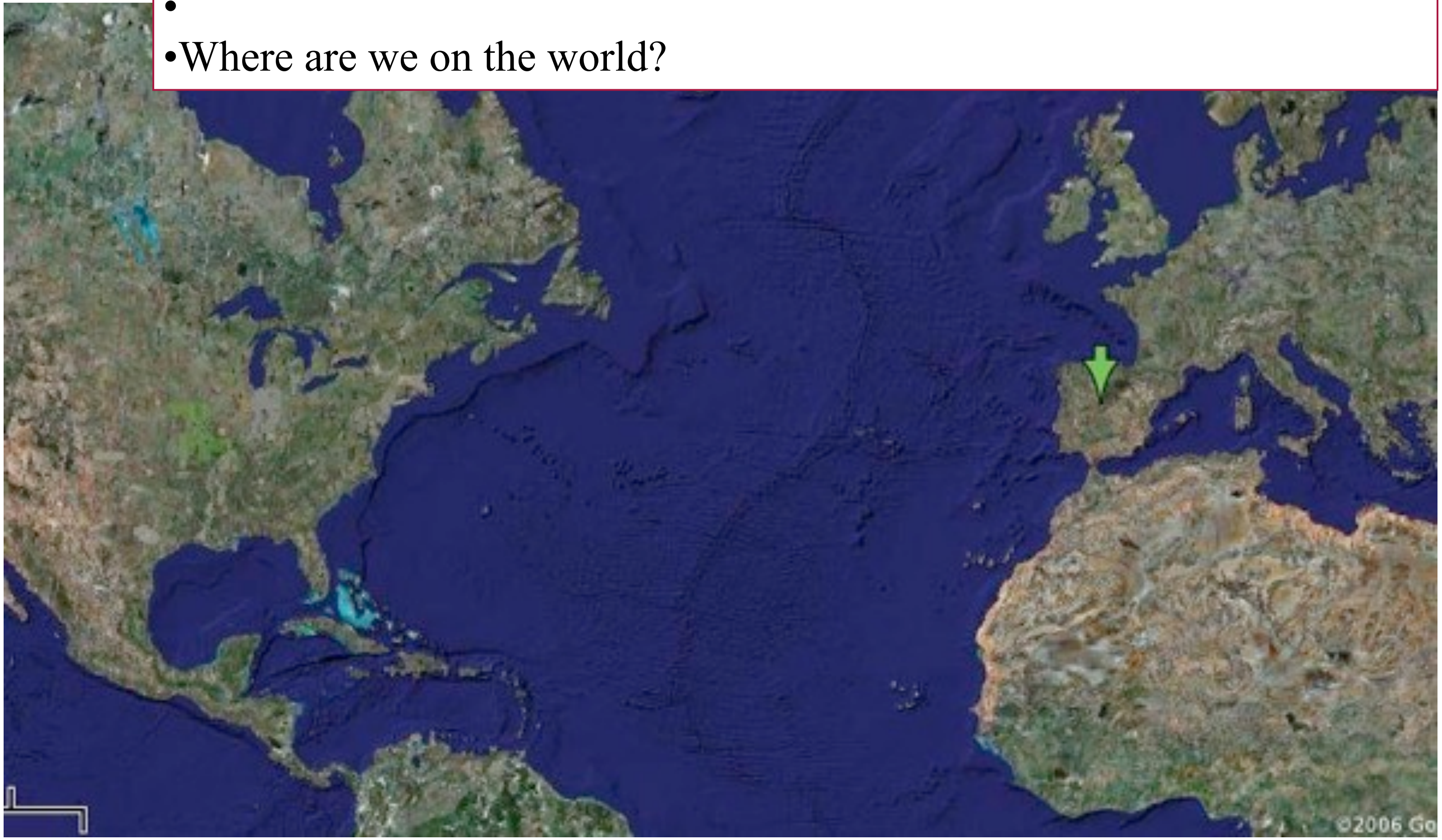
Madrid, Spain

-
- Where are we on the world?

Madrid, Spain

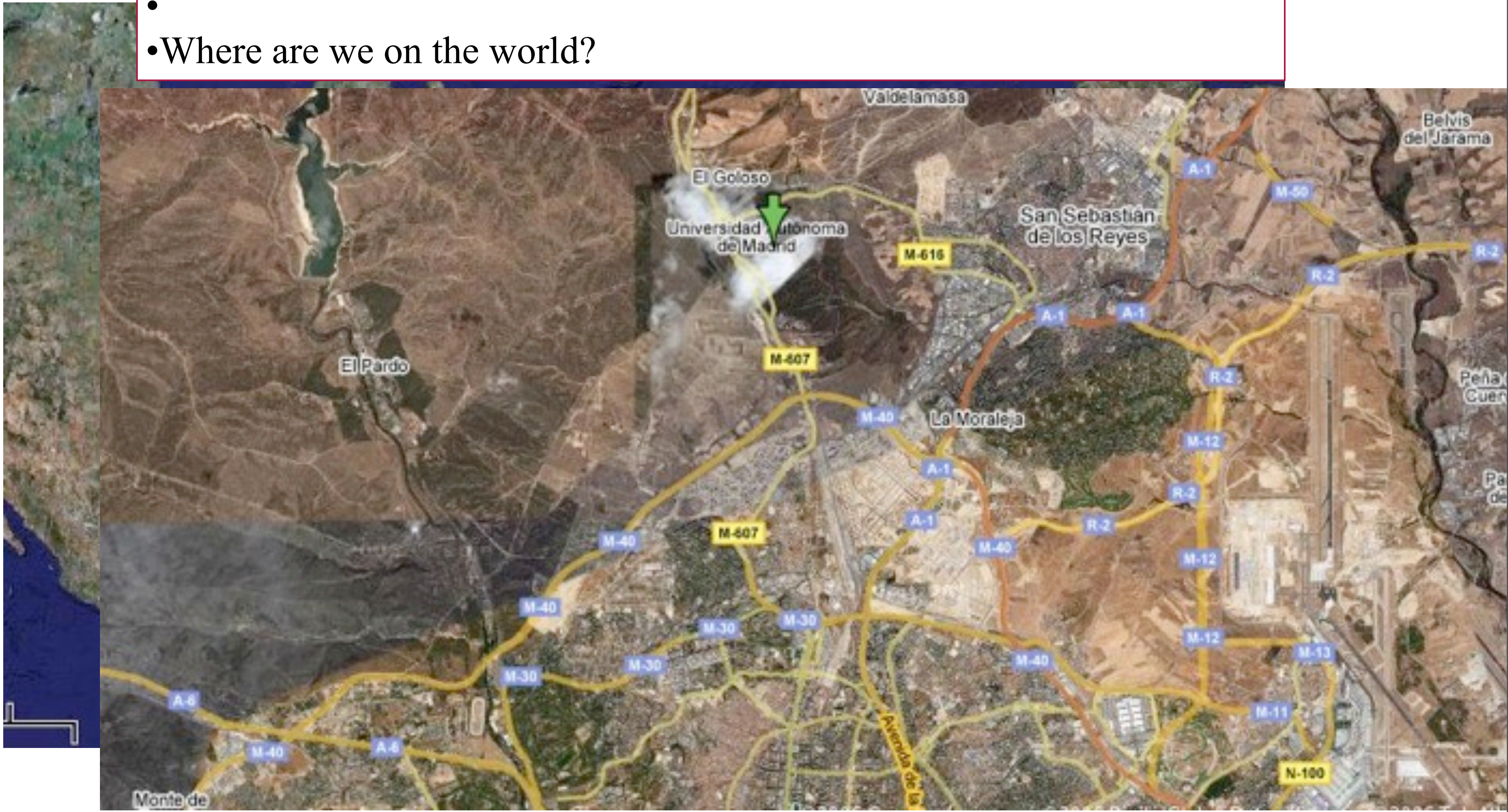
-

- Where are we on the world?



Madrid, Spain

-
- Where are we on the world?



Madrid, Spain

-
- Where are we on the world?

Valdelamasa
Belvis del Jarama
El Goloso
Universidad Autónoma de Madrid
San Sebastián de los Reyes
M-616
A-1
M-50
R-2
A-6
M-40
Monte de
Centro Nacional de Biotecnología

Victor Canivell, PhD (President)

- Doctor en Ciencias Físicas, Universidad de Barcelona
- MBA por ESADE
- Hewlett Packard, Vice President, Europe
- Silicon Graphics, Vice President, Europe
- 3Com, Vice President, Europe
- SSA Global
- Así como diez años en empresas innovadoras tipo start-up (Aspective, ahora Vodafone en Londres, Safelayer y Wisekey ELA en nuestro país)
- Actualmente es miembro del Consejo de dos empresas de biotecnología (Integromics y ERA Biotech), con una marcada vocación internacional

Where we are?

Technological Partners

Offices

Applied Biosystems

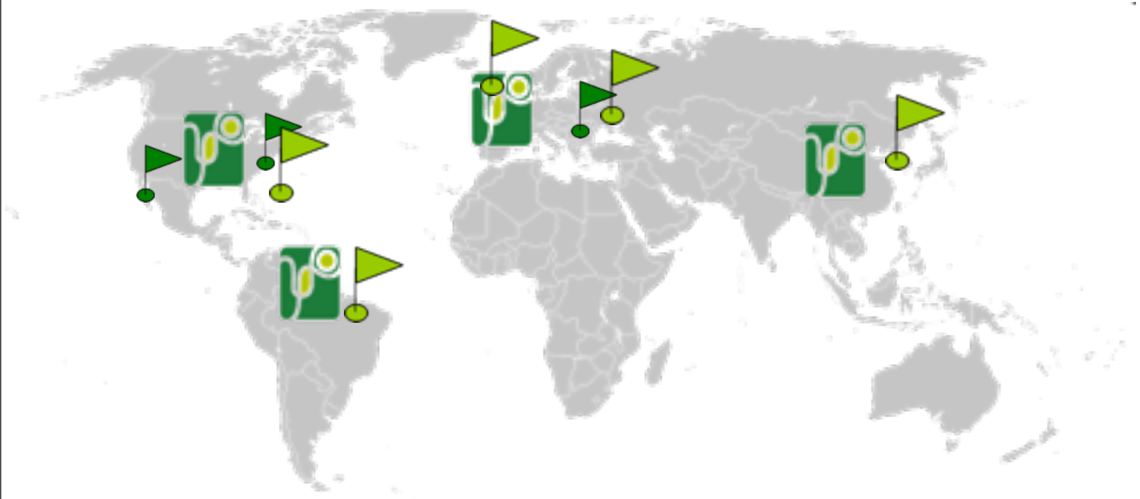
TIBCO The Power of Now™

Spotfire

INGENUITY® SYSTEMS

Integromics

www.integromics.com



Where we are?

Technological Partners

Offices

Applied Biosystems

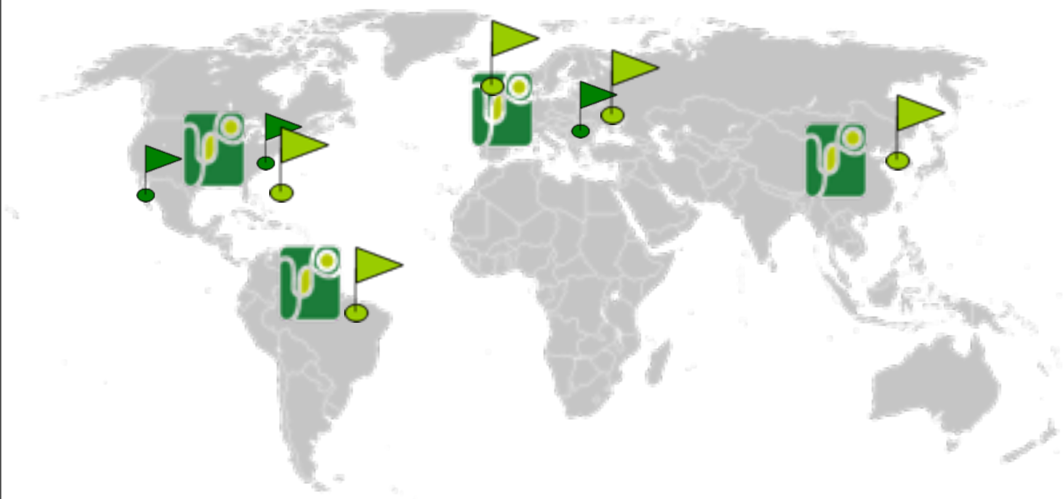
TIBCO The Power of Now™

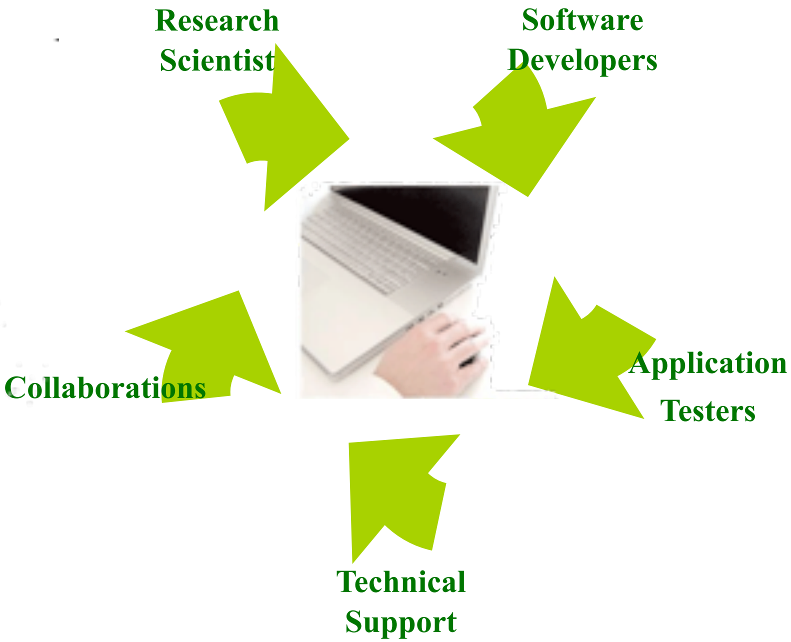
Spotfire

INGENUITY® SYSTEMS

Integromics

www.integromics.com





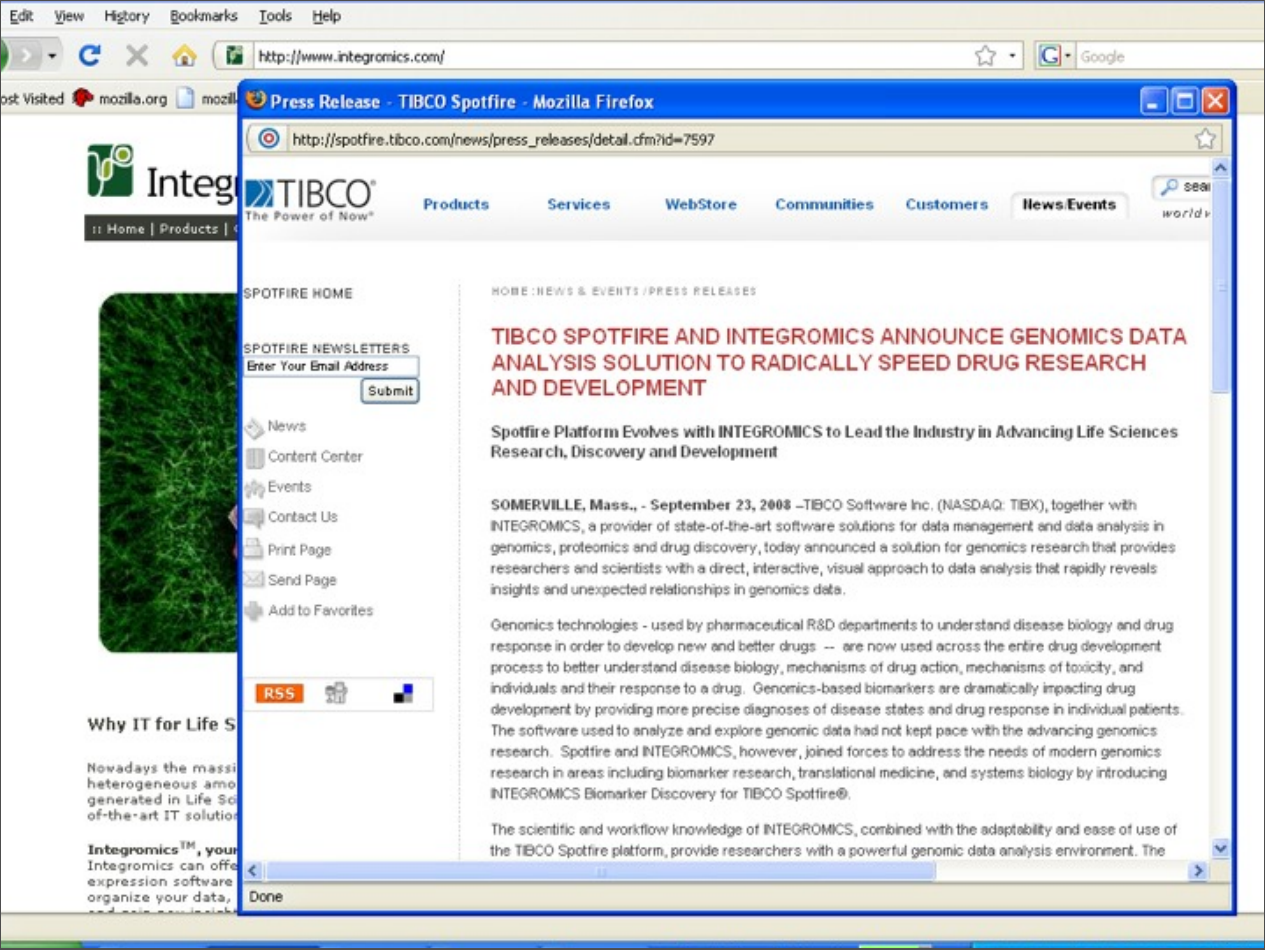
Research Scientist

Software Developers

Collaborations

Application Testers

Technical Support


Edit View History Bookmarks Tools Help
http://www.integromics.com/
Google
Most Visited mozilla.org mozil
Press Release - TIBCO Spotfire - Mozilla Firefox
http://spotfire.tibco.com/news/press_releases/detail.cfm?id=7597
TIBCO The Power of Now
Products Services WebStore Communities Customers News Events
sear
world
Integ
:: Home | Products |
SPOTFIRE HOME
SPOTFIRE NEWSLETTERS
Enter Your Email Address
Submit
News
Content Center
Events
Contact Us
Print Page
Send Page
Add to Favorites
RSS
Done

HOME : NEWS & EVENTS / PRESS RELEASES

TIBCO SPOTFIRE AND INTEGROMICS ANNOUNCE GENOMICS DATA ANALYSIS SOLUTION TO RADICALLY SPEED DRUG RESEARCH AND DEVELOPMENT

Spotfire Platform Evolves with INTEGROMICS to Lead the Industry in Advancing Life Sciences Research, Discovery and Development

SOMERVILLE, Mass., - September 23, 2008 –TIBCO Software Inc. (NASDAQ: TIBX), together with INTEGROMICS, a provider of state-of-the-art software solutions for data management and data analysis in genomics, proteomics and drug discovery, today announced a solution for genomics research that provides researchers and scientists with a direct, interactive, visual approach to data analysis that rapidly reveals insights and unexpected relationships in genomics data.

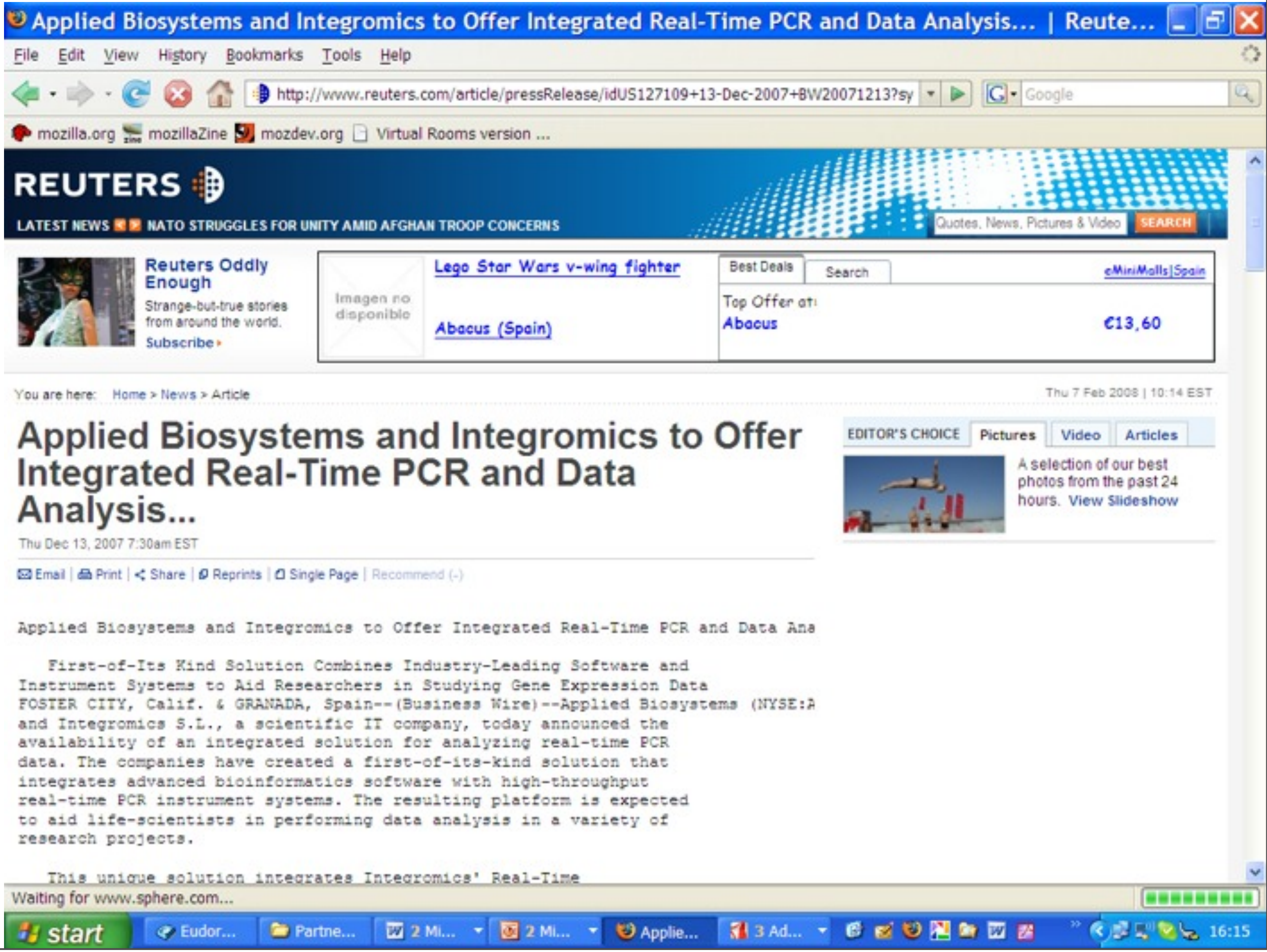
Genomics technologies - used by pharmaceutical R&D departments to understand disease biology and drug response in order to develop new and better drugs -- are now used across the entire drug development process to better understand disease biology, mechanisms of drug action, mechanisms of toxicity, and individuals and their response to a drug. Genomics-based biomarkers are dramatically impacting drug development by providing more precise diagnoses of disease states and drug response in individual patients. The software used to analyze and explore genomic data had not kept pace with the advancing genomics research. Spotfire and INTEGROMICS, however, joined forces to address the needs of modern genomics research in areas including biomarker research, translational medicine, and systems biology by introducing INTEGROMICS Biomarker Discovery for TIBCO Spotfire®.

The scientific and workflow knowledge of INTEGROMICS, combined with the adaptability and ease of use of the TIBCO Spotfire platform, provide researchers with a powerful genomic data analysis environment. The

Why IT for Life S

Nowadays the massi
heterogeneous amo
generated in Life Sc
of-the-art IT solutio

Integromics™, your
Integromics can offe
expression software
organize your data,


Applied Biosystems and Integromics to Offer Integrated Real-Time PCR and Data Analysis... | Reute...
File Edit View History Bookmarks Tools Help
http://www.reuters.com/article/pressRelease/idUS127109+13-Dec-2007+BW20071213?sy
Google
mozilla.org mozillaZine mozdev.org Virtual Rooms version ...
REUTERS
LATEST NEWS NATO STRUGGLES FOR UNITY AMID AFGHAN TROOP CONCERNS
Quotes, News, Pictures & Video
SEARCH
Reuters Oddly Enough
Strange-but-true stories from around the world.
Subscribe
Imagen no disponible
Lego Star Wars v-wing fighter
Abacus (Spain)
Best Deals
Search
eMiniMalls|Spain
Top Offer at:
Abacus
€13,60
EDITOR'S CHOICE
Pictures
Video
Articles
A selection of our best photos from the past 24 hours. View Slideshow
Waiting for www.sphere.com...
start
Eudor...
Partne...
2 Mi...
2 Mi...
Applie...
3 Ad...
16:15

You are here: Home > News > Article

Thu 7 Feb 2008 | 10:14 EST

Applied Biosystems and Integromics to Offer Integrated Real-Time PCR and Data Analysis...

Thu Dec 13, 2007 7:30am EST

Email | Print | Share | Reprints | Single Page | Recommend (-)

Applied Biosystems and Integromics to Offer Integrated Real-Time PCR and Data Ana

First-of-Its Kind Solution Combines Industry-Leading Software and Instrument Systems to Aid Researchers in Studying Gene Expression Data

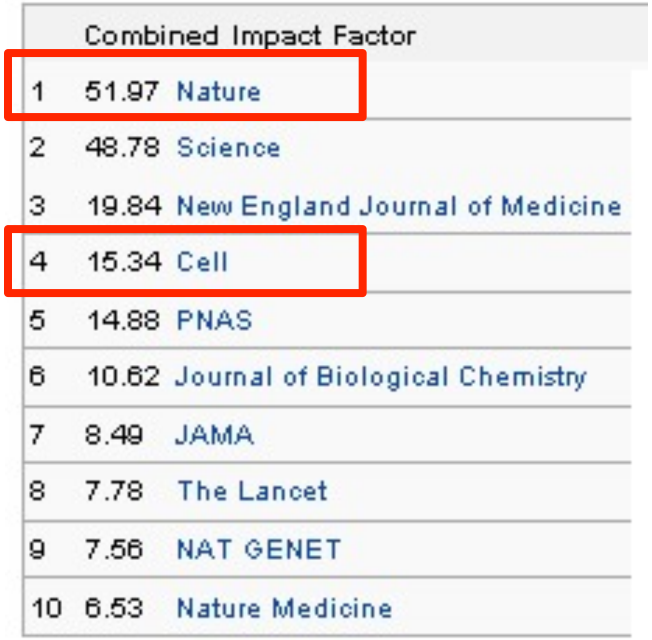
FOSTER CITY, Calif. & GRANADA, Spain--(Business Wire)--Applied Biosystems (NYSE:A and Integromics S.L., a scientific IT company, today announced the availability of an integrated solution for analyzing real-time PCR data. The companies have created a first-of-its-kind solution that integrates advanced bioinformatics software with high-throughput real-time PCR instrument systems. The resulting platform is expected to aid life-scientists in performing data analysis in a variety of research projects.

This unique solution integrates Integromics' Real-Time

The R&D department of Integromics is actively publishing in the most prestigious international scientific journals

- *Published in 2010*
  - Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation.
    Ozsolak F, Kapranov P, **Foissac S**, Kim SW, Fishilevich E, Monaghan AP, John B, Milos PM.
    Cell. 2010 Dec 10;143(6):1018-29.
  - New class of gene-termini-associated human RNAs suggests a novel RNA copying mechanism.
    Kapranov P, Ozsolak F, Kim SW, **Foissac S**, Lipson D, Hart C, Roels S, Borel C, Antonarakis SE, Monaghan AP, John B, Milos PM.
    Nature. 2010 Jul 29;466(7306):642-6.
  - Laboratory information management systems in the "Omics" era.
    González Couto E.
    LifeSciencesLab. 2010 Mar-Apr; 38-40.
  - Data Management, Analysis, Standardization and Reproducibility in a ProteoRed Multicentric Quantitative Proteimics Study with OmicsHub Proteomics Software Tool. **Yankilevich, P**., **J Biomol Tech**. 2010 September; 21(3 Suppl): S35
  - OmicsHub Proteomics Software Tool, **Yankilevich, P**., **, J Biomol Tech**. 2010 September; 21(3 Suppl): S21

- *International Ranking*



	Combined Impact Factor	
1	51.97	Nature
2	48.78	Science
3	19.84	New England Journal of Medicine
4	15.34	Cell
5	14.88	PNAS
6	10.62	Journal of Biological Chemistry
7	8.49	JAMA
8	7.78	The Lancet
9	7.56	NAT GENET
10	6.53	Nature Medicine

(**Integromics** authors highlighted underlined and in **BOLD**)

Integromics™

Integromics in the media

- Spanish 2006:
  - 1st Prize as “Highest Potential Company”

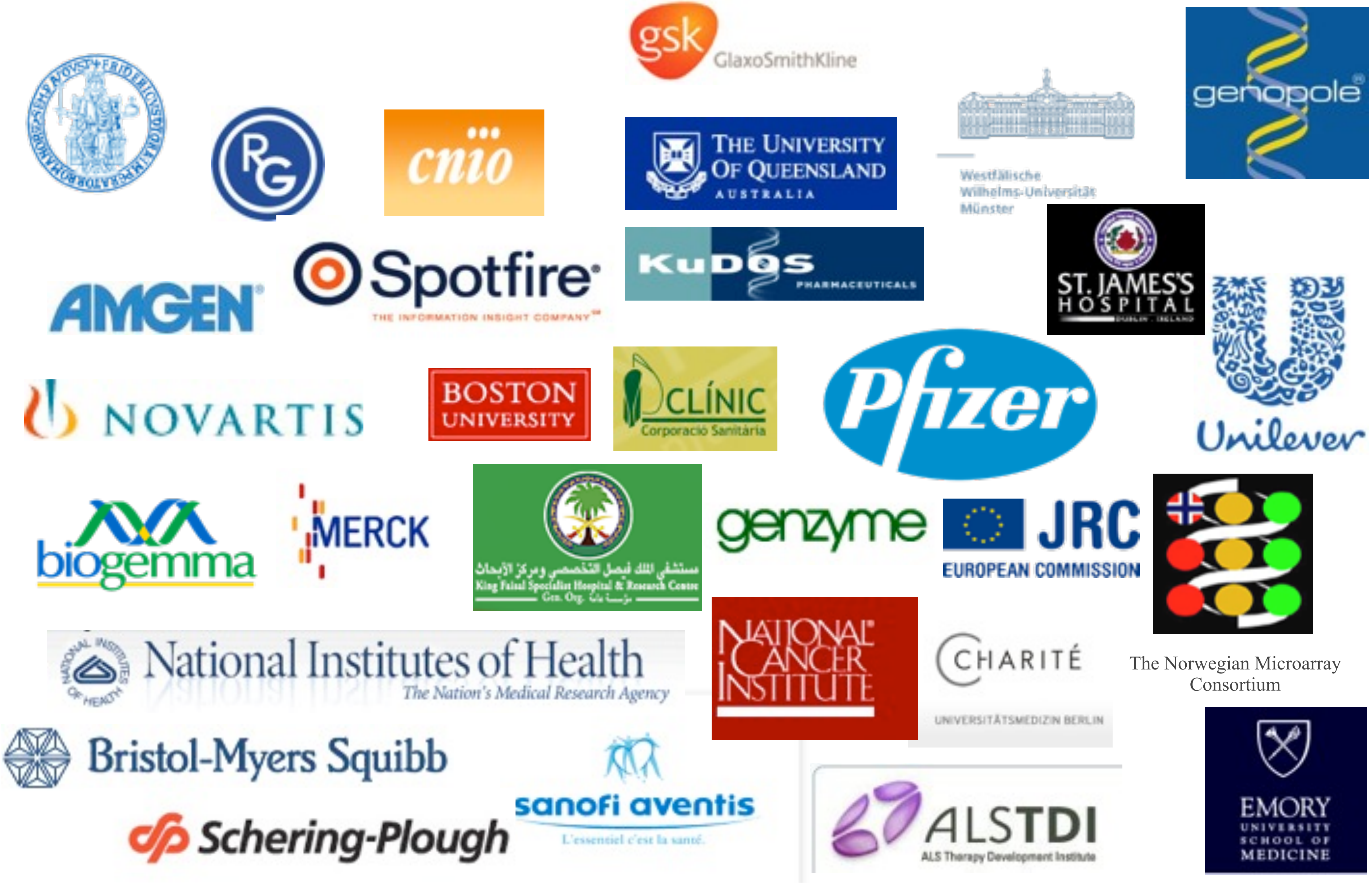
- Europe 2007:
  - 1st Prize “Most Innovative Bioinfo Company”

Integromics™

Integromics in the media

- Spanish 2010:
  - Mejor empresa en I+D+i (Accesit)

Current Customers of Integromics


gsk GlaxoSmithKline
RG
cnio
The University of Queensland Australia
Westfälische Wilhelms-Universität Münster
genopole
AMGEN
Spotfire
THE INFORMATION INSIGHT COMPANY
KuDOS PHARMACEUTICALS
ST. JAMES'S HOSPITAL DUBLIN · IRELAND
Unilever
NOVARTIS
BOSTON UNIVERSITY
CLÍNIC Corporació Sanitària
Pfizer
biogemma
MERCK
مستشفى الملك فيصل التخصصي ومركز الأبحاث
King Faisal Specialist Hospital & Research Centre
Gen. Org. مؤسسة عامة
genzyme
JRC
EUROPEAN COMMISSION
The Norwegian Microarray Consortium
National Institutes of Health
The Nation's Medical Research Agency
NATIONAL CANCER INSTITUTE
CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN
Bristol-Myers Squibb
sanofi aventis
L'essentiel c'est la santé.
Schering-Plough
ALSTDI
ALS Therapy Development Institute
EMORY UNIVERSITY SCHOOL OF MEDICINE