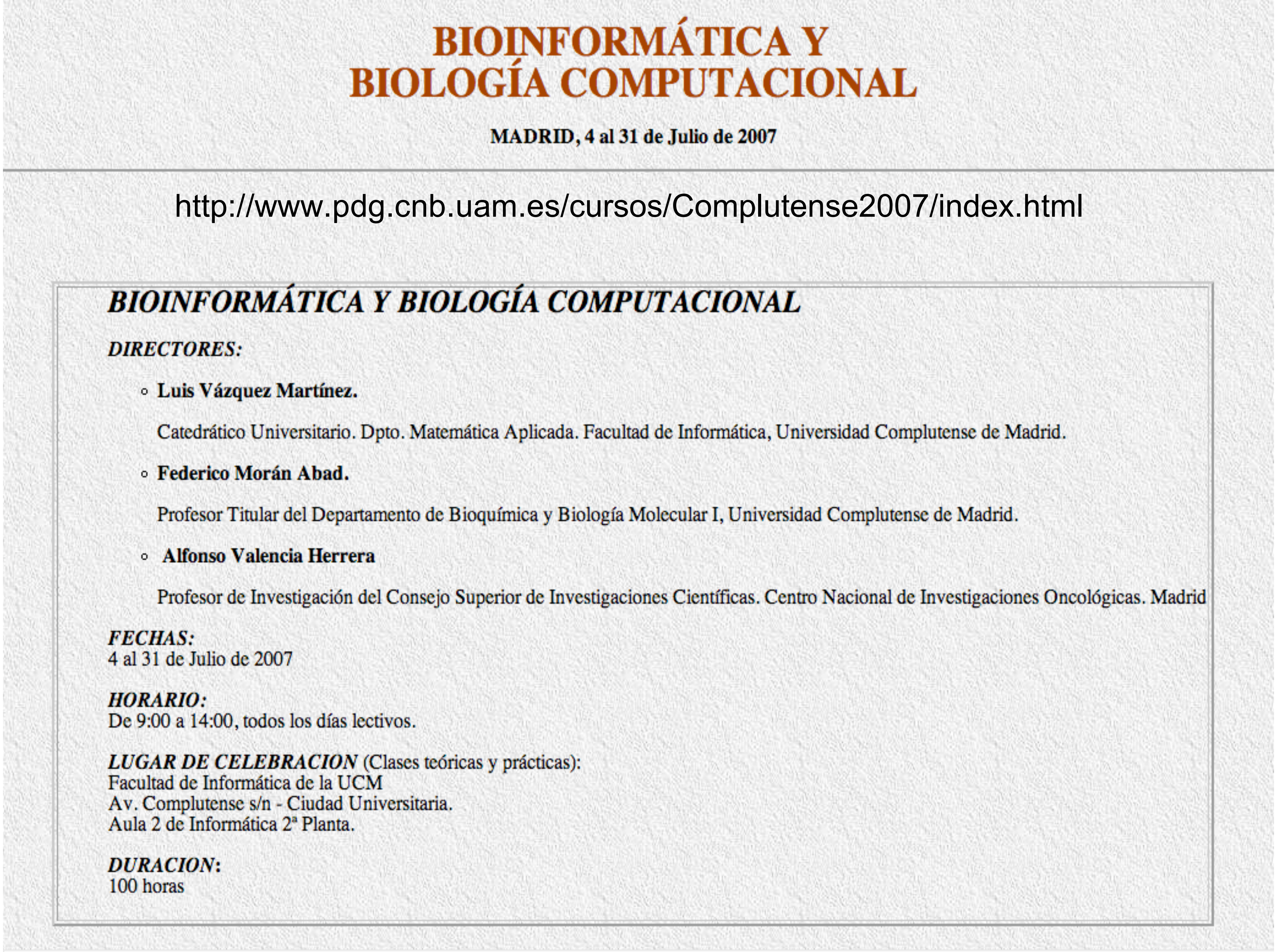


BIOINFORMÁTICA Y BIOLOGÍA COMPUTACIONAL

MADRID, 4 al 31 de Julio de 2007

http://www.pdg.cnb.uam.es/cursos/Complutense2007/index.html

BIOINFORMÁTICA Y BIOLOGÍA COMPUTACIONAL

DIRECTORES:

- **Luis Vázquez Martínez.**

  Catedrático Universitario. Dpto. Matemática Aplicada. Facultad de Informática, Universidad Complutense de Madrid.

- **Federico Morán Abad.**

  Profesor Titular del Departamento de Bioquímica y Biología Molecular I, Universidad Complutense de Madrid.

- **Alfonso Valencia Herrera**

  Profesor de Investigación del Consejo Superior de Investigaciones Científicas. Centro Nacional de Investigaciones Oncológicas. Madrid

FECHAS:
4 al 31 de Julio de 2007

HORARIO:
De 9:00 a 14:00, todos los días lectivos.

LUGAR DE CELEBRACION (Clases teóricas y prácticas):
Facultad de Informática de la UCM
Av. Complutense s/n - Ciudad Universitaria.
Aula 2 de Informática 2ª Planta.

***DURACION*:**
100 horas

Introduccion a Bioinformática Y Biología Computacional

U. Complutense de Verano

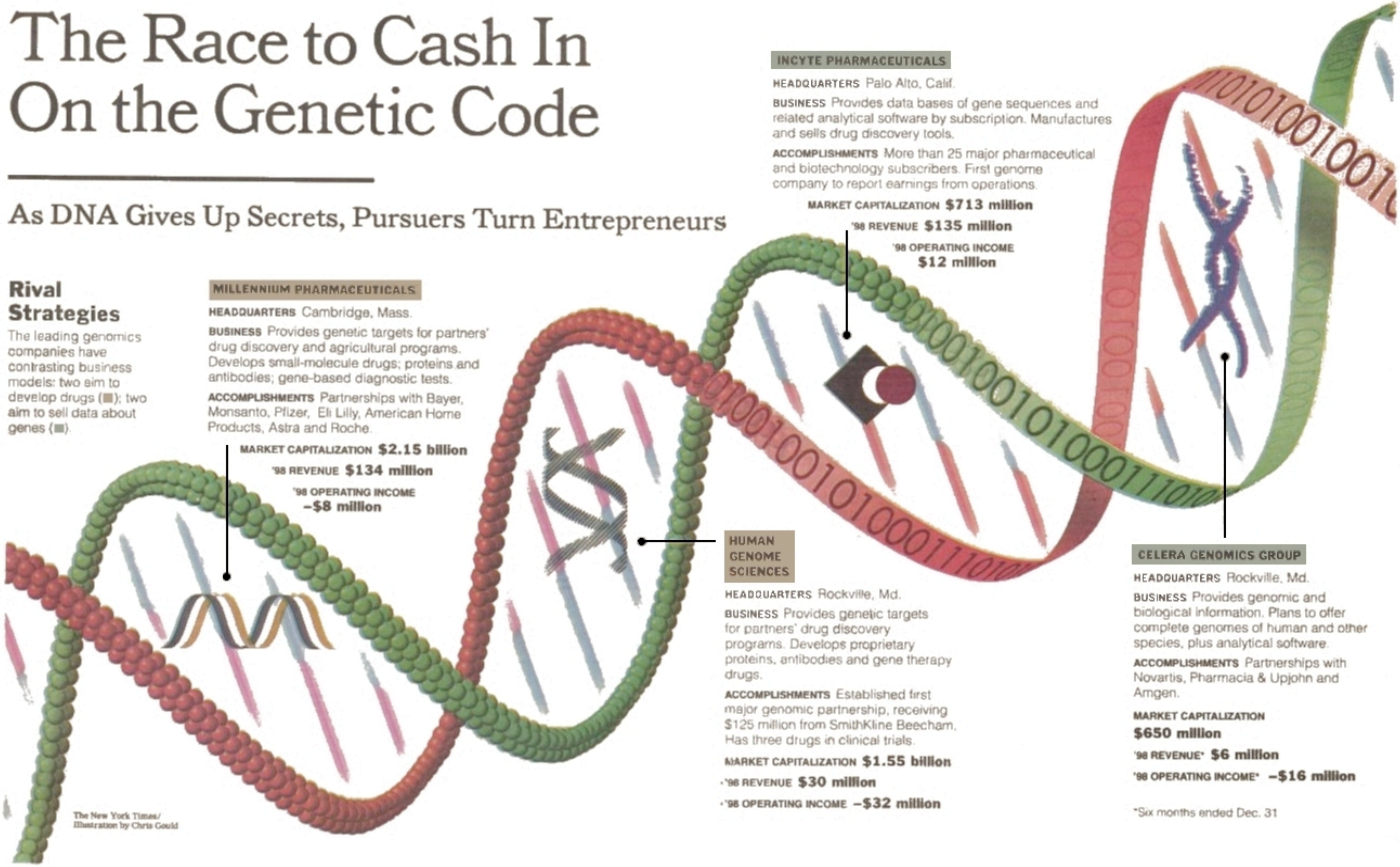
Alfonso Valencia
S-CompBio
CNIO, Madrid

Madrid, 2007

The New York Times
Money&Business
Sunday, August 29, 1999
Section 3

The Race to Cash In On the Genetic Code

As DNA Gives Up Secrets, Pursuers Turn Entrepreneurs



Rival Strategies

The leading genomics companies have contrasting business models: two aim to develop drugs (■); two aim to sell data about genes (■).

MILLENNIUM PHARMACEUTICALS

HEADQUARTERS Cambridge, Mass.

BUSINESS Provides genetic targets for partners' drug discovery and agricultural programs. Develops small-molecule drugs; proteins and antibodies; gene-based diagnostic tests.

ACCOMPLISHMENTS Partnerships with Bayer, Monsanto, Pfizer, Eli Lilly, American Home Products, Astra and Roche.

MARKET CAPITALIZATION **$2.15 billion**

'98 REVENUE **$134 million**

'98 OPERATING INCOME **–$8 million**

INCYTE PHARMACEUTICALS

HEADQUARTERS Palo Alto, Calif.

BUSINESS Provides data bases of gene sequences and related analytical software by subscription. Manufactures and sells drug discovery tools.

ACCOMPLISHMENTS More than 25 major pharmaceutical and biotechnology subscribers. First genome company to report earnings from operations.

MARKET CAPITALIZATION **$713 million**

'98 REVENUE **$135 million**

'98 OPERATING INCOME **$12 million**

HUMAN GENOME SCIENCES

HEADQUARTERS Rockville, Md.

BUSINESS Provides genetic targets for partners' drug discovery programs. Develops proprietary proteins, antibodies and gene therapy drugs.

ACCOMPLISHMENTS Established first major genomic partnership, receiving $125 million from SmithKline Beecham. Has three drugs in clinical trials.

MARKET CAPITALIZATION **$1.55 billion**

'98 REVENUE **$30 million**

'98 OPERATING INCOME **–$32 million**

CELERA GENOMICS GROUP

HEADQUARTERS Rockville, Md.

BUSINESS Provides genomic and biological information. Plans to offer complete genomes of human and other species, plus analytical software.

ACCOMPLISHMENTS Partnerships with Novartis, Pharmacia & Upjohn and Amgen.

MARKET CAPITALIZATION
$650 million

'98 REVENUE* **$6 million**

'98 OPERATING INCOME* **–$16 million**

*Six months ended Dec. 31

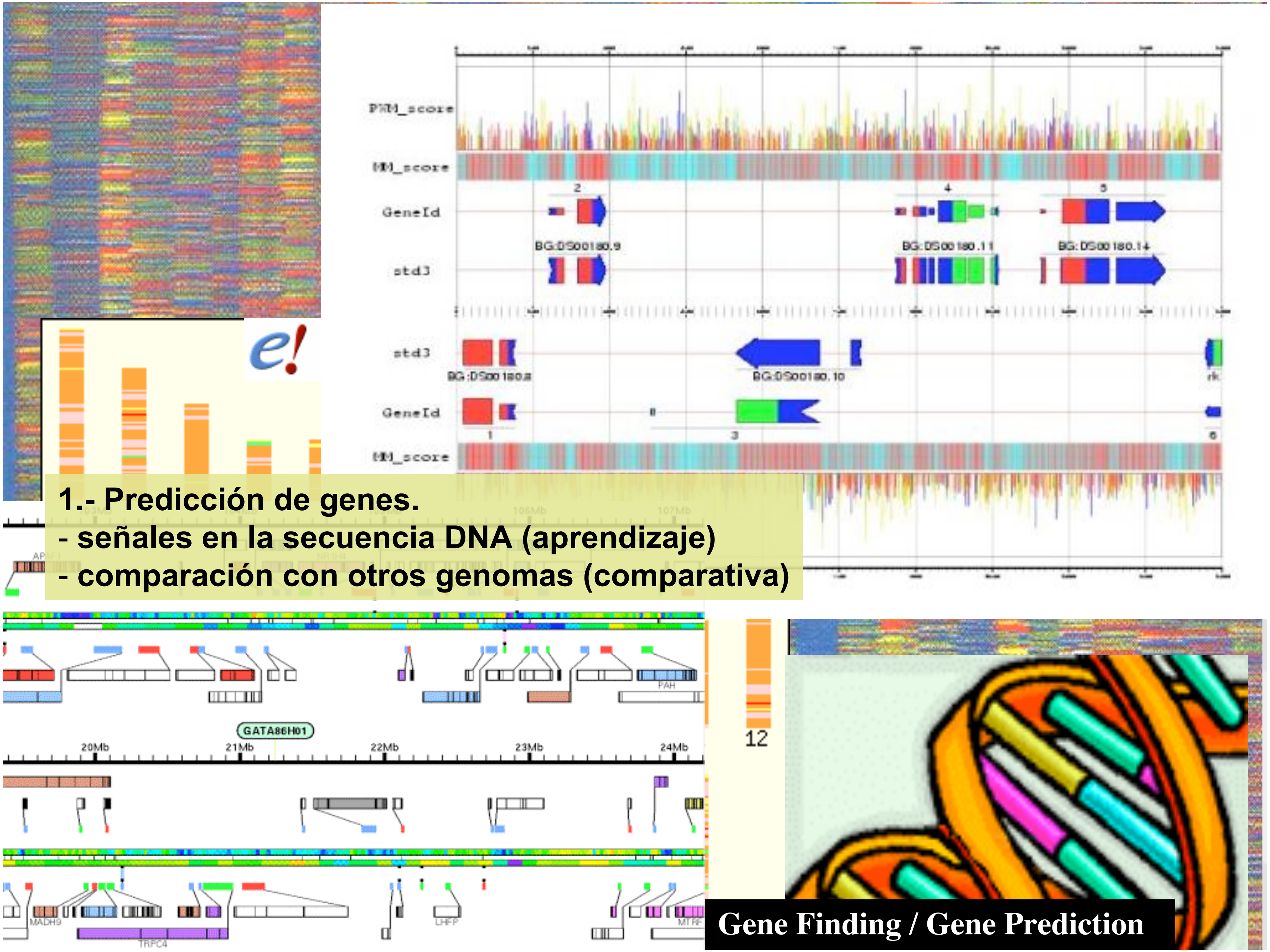
The New York Times/
Illustration by Chris Gould

The Future?

Here's my sequence …

Personalized Medicine

Source: *New Yorker*



1.- Predicción de genes.

- señales en la secuencia DNA (aprendizaje)

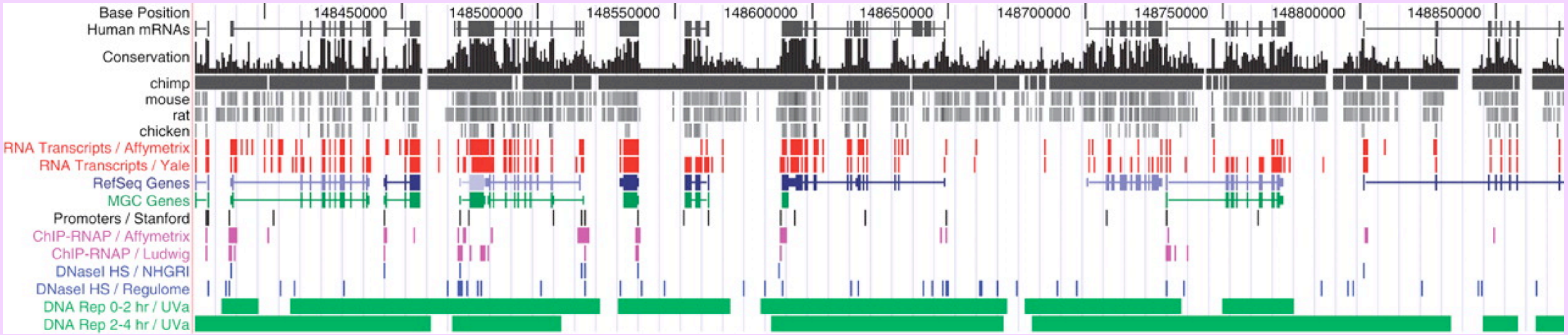
- comparación con otros genomas (comparativa)

Gene Finding / Gene Prediction

The ENCODE Project

The Encyclopaedia of DNA Elements (ENCODE) is an NIH-backed multi-million dollar project, that brings together an international consortium of scientists in industry and academia with the aim of identifying all the functional elements in the human genome.

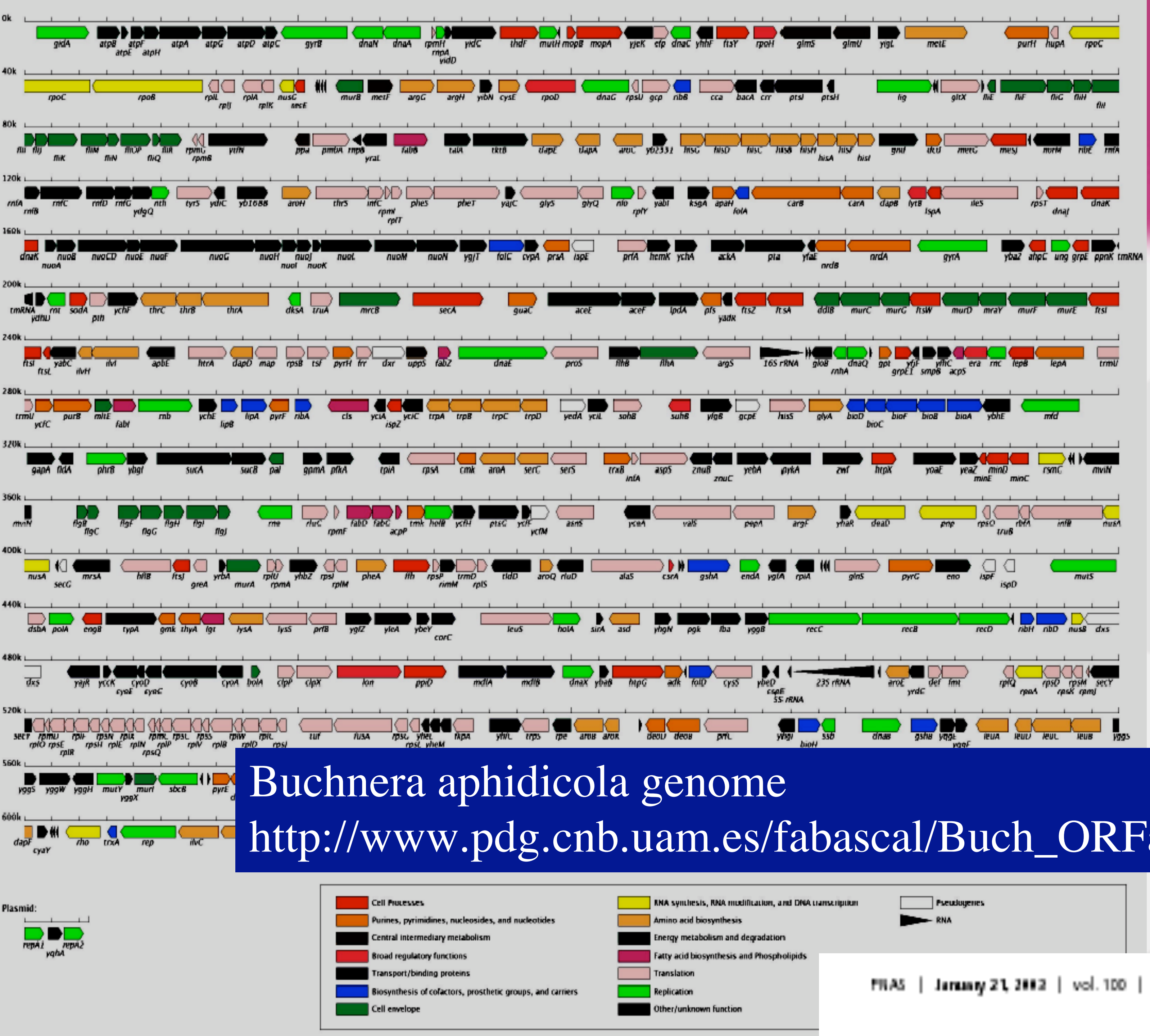
The efforts of the ENCODE consortium are focused entirely on DNA aspects of the project, such as transcription sites, TRANSFRAGS, non–protein-coding genes and sequences that mediate chromosome structure and dynamics.



The pilot project is studying 30MB from 44 regions comprising 1% of the genome.

15MB come from regions chosen for their scientific interest, the rest are chosen via a stratified random sampling method.

by Michael Tress, CNIO


0k
40k
80k
120k
160k
200k
240k
280k
320k
360k
400k
440k
480k
520k
560k
600k
Plasmid:
repA1
repA2
yqhA
Cell Processes
Purines, pyrimidines, nucleosides, and nucleotides
Central intermediary metabolism
Broad regulatory functions
Transport/binding proteins
Biosynthesis of cofactors, prosthetic groups, and carriers
Cell envelope
RNA synthesis, RNA modification, and DNA transcription
Amino acid biosynthesis
Energy metabolism and degradation
Fatty acid biosynthesis and Phospholipids
Translation
Replication
Other/unknown function
Pseudogenes
RNA

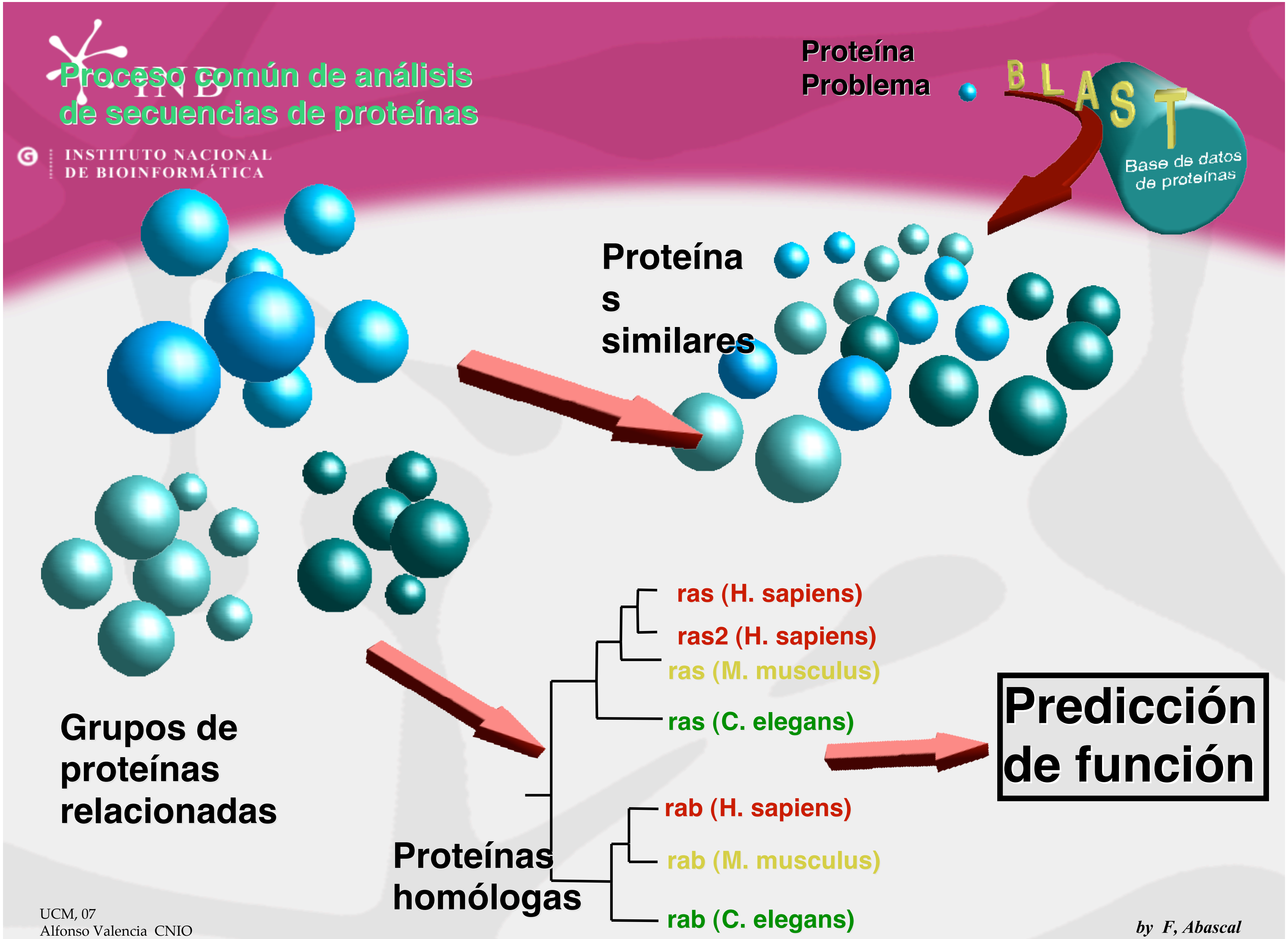
Buchnera aphidicola genome

http://www.pdg.cnb.uam.es/fabascal/Buch_ORFand_ww

PNAS | January 21, 2003 | vol. 100 | no. 2 | 581–586

Proceso común de análisis de secuencias de proteínas

INSTITUTO NACIONAL
DE BIOINFORMÁTICA



Proteína
Problema

BLAST

Base de datos
de proteínas

Proteína
s
similares

Grupos de
proteínas
relacionadas

Proteínas
homólogas

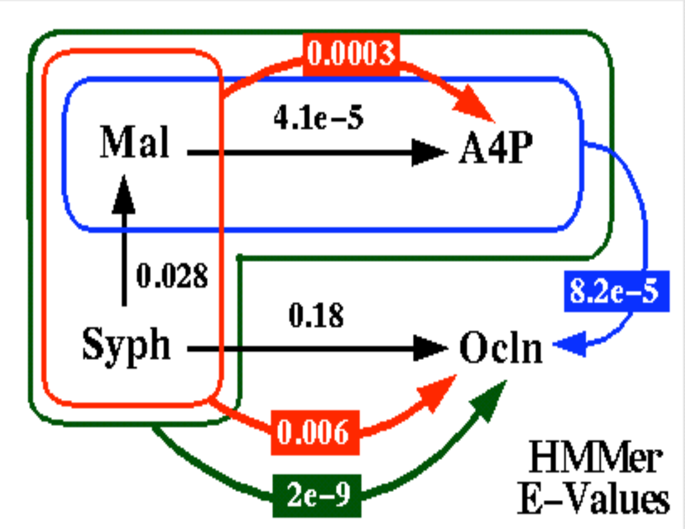
ras (H. sapiens)
ras2 (H. sapiens)
ras (M. musculus)
ras (C. elegans)
rab (H. sapiens)
rab (M. musculus)
rab (C. elegans)

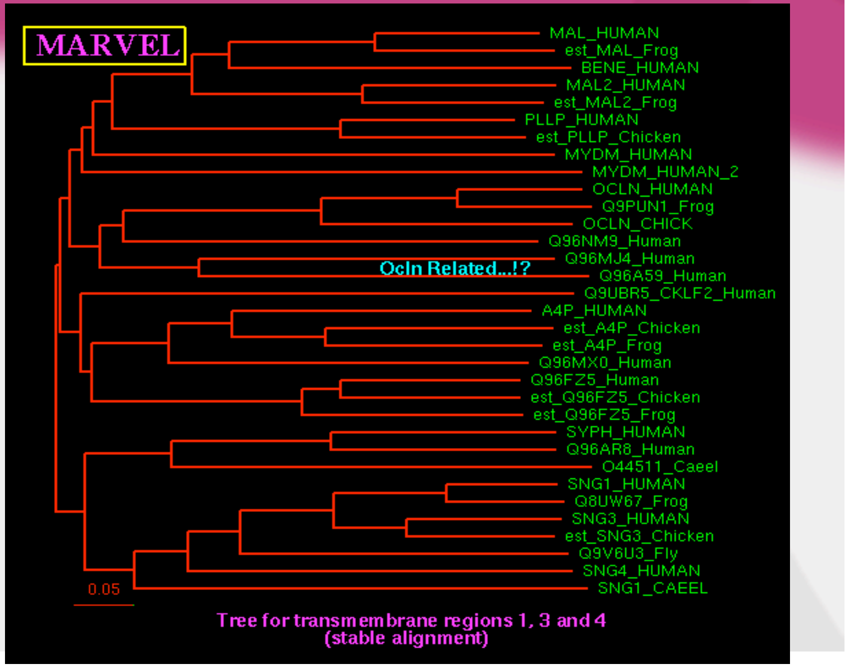
Predicción de función

UCM, 07
Alfonso Valencia CNIO

by F, Abascal

MARVEL: a conserved domain involved in membrane apposition events.
Sanchez-Pulido et al., TIBS 2002





Tree for transmembrane regions 1, 3 and 4
(stable alignment)

MARVEL novel domain

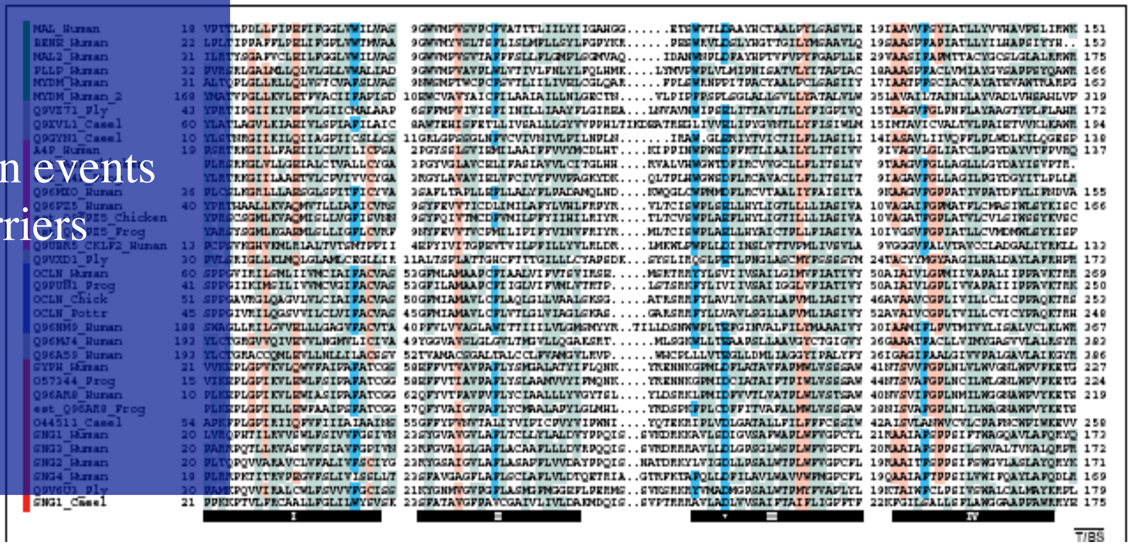
four transmembrane-helix architecture

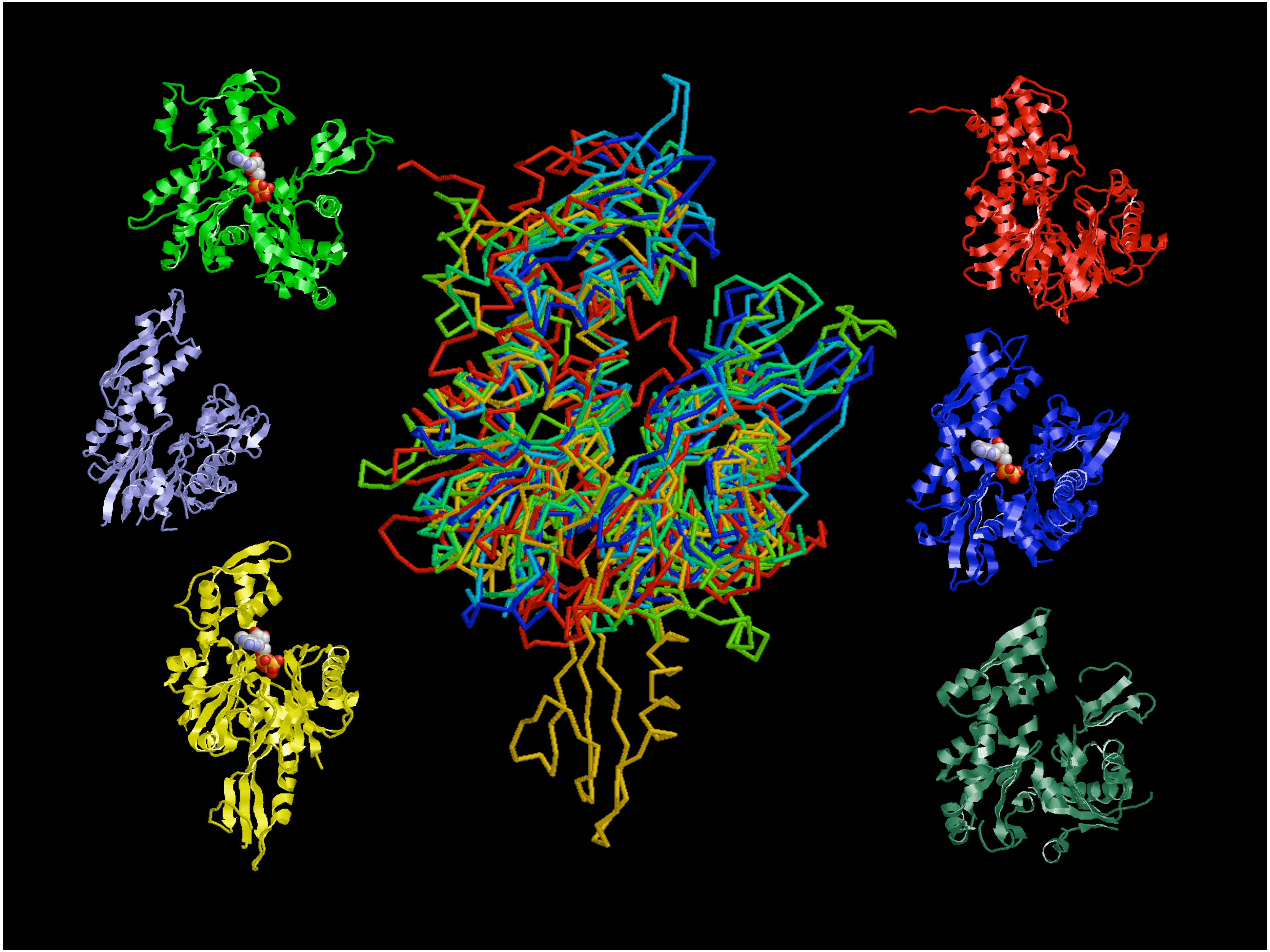
- myelin and lymphocyte (MAL),
- physins,
- gyrins and
- occludin families.

cholesterol-rich membrane apposition events

- biogenesis of vesicular transport carriers
- tight junction regulation.

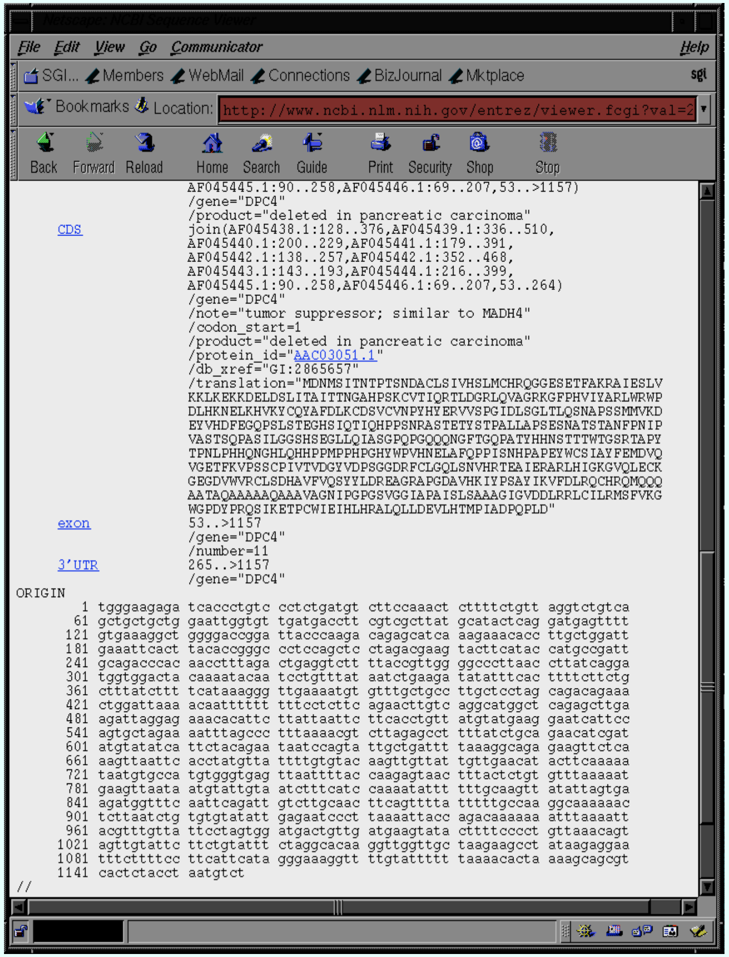
Human diseases: schizophrenia and inflammation.



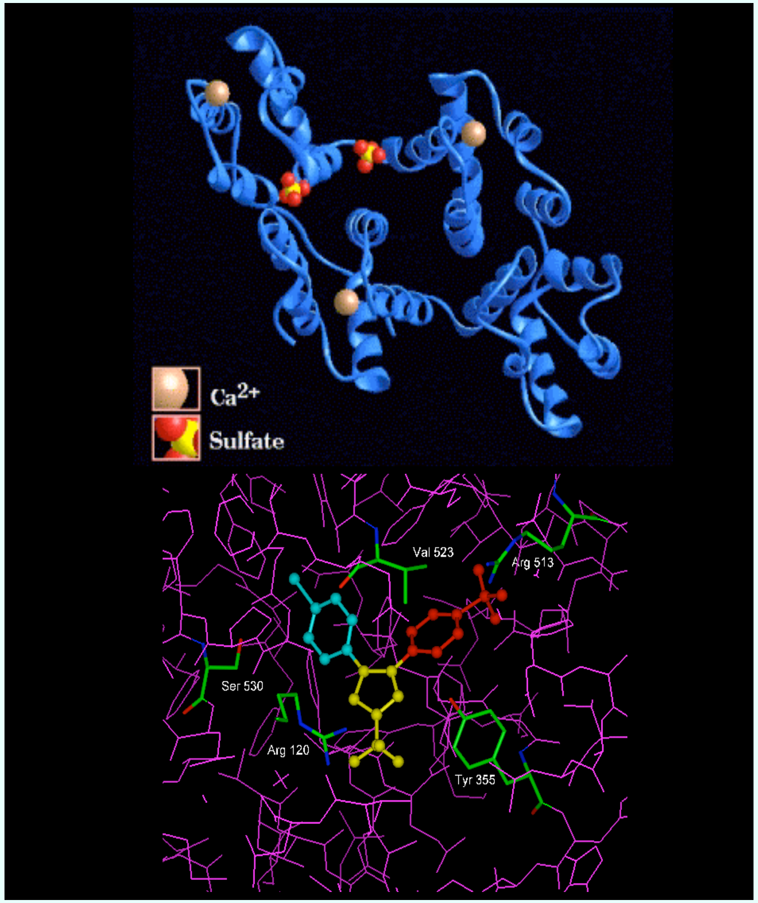


PREDICCION ESTRUCTURA PROTEINAS

2.- Predicción de estructura de proteínas.
- como problema físico: simulaciones, campos de fuerza
- Informático: aprendizaje (secuencia > estructura conocidas), estadísticas, librerías


Netscape: NCBI Sequence Viewer
File Edit View Go Communicator Help
SGI... Members WebMail Connections BizJournal Mktplace
Bookmarks Location: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=2
Back Forward Reload Home Search Guide Print Security Shop Stop

```
                     AF045445.1:90..258,AF045446.1:69..207,53..>1157)
                     /gene="DPC4"
                     /product="deleted in pancreatic carcinoma"
     CDS             join(AF045438.1:128..376,AF045439.1:336..510,
                     AF045440.1:200..229,AF045441.1:179..391,
                     AF045442.1:138..257,AF045442.1:352..468,
                     AF045443.1:143..193,AF045444.1:216..399,
                     AF045445.1:90..258,AF045446.1:69..207,53..264)
                     /gene="DPC4"
                     /note="tumor suppressor; similar to MADH4"
                     /codon_start=1
                     /product="deleted in pancreatic carcinoma"
                     /protein_id="AAC03051.1"
                     /db_xref="GI:2865657"
                     /translation="MDNMSITNTPTSNDACLSIVHSLMCHRQGGESETFAKRAIESLV
                     KKLKEKKDELDSLITAITTNGAHPSKCVTIQRTLDGRLQVAGRKGFPHVIYARLWRWP
                     DLHKNELKHVKYCQYAFDLKCDSVCVNPYHYERVVSPGIDLSGLTLQSNAPSSMMVKD
                     EYVHDFEGQPSLSTEGHSIQTIQHPPSNRASTETYSTPALLAPSESNATSTANFPNIP
                     VASTSQPASILGGSHSEGLLQIASGPQPGQQQNGFTGQPATYHHNSTTTWTGSRTAPY
                     TPNLPHHQNGHLQHHPPMPPHPGHYWPVHNELAFQPPISNHPAPEYWCSIAYFEMDVQ
                     VGETFKVPSSCPIVTVDGYVDPSGGDRFCLGQLSNVHRTEAIERARLHIGKGVQLECK
                     GEGDVWVRCLSDHAVFVQSYYLDREAGRAPGDAVHKIYPSAYIKVFDLRQCHRQMQQQ
                     AATAQAAAAAQAAAVAGNIPGPGSVGGIAPAISLSAAAGIGVDDLRRLCILRMSFVKG
                     WGPDYPRQSIKETPCWIEIHLHRALQLLDEVLHTMPIADPQPLD"
     exon            53..>1157
                     /gene="DPC4"
                     /number=11
     3'UTR           265..>1157
                     /gene="DPC4"
ORIGIN
        1 tgggaagaga tcaccctgtc cctctgatgt cttccaaact cttttctgtt aggtctgtca
       61 gctgctgctg gaattggtgt tgatgacctt cgtcgcttat gcatactcag gatgagtttt
      121 gtgaaaggct ggggaccgga ttacccaaga cagagcatca aagaaacacc ttgctggatt
      181 gaaattcact tacaccgggc cctccagctc ctagacgaag tacttcatac catgccgatt
      241 gcagacccac aacctttaga ctgaggtctt ttaccgttgg ggcccttaac cttatcagga
      301 tggtggacta caaaatacaa tcctgtttat aatctgaaga tatatttcac ttttcttctg
      361 ctttatcttt tcataaaggg ttgaaaatgt gtttgctgcc ttgctcctag cagacagaaa
      421 ctggattaaa acaatttttt tttcctcttc agaacttgtc aggcatggct cagagcttga
      481 agattaggag aaacacattc ttattaattc ttcacctgtt atgtatgaag gaatcattcc
      541 agtgctagaa aatttagccc tttaaaacgt cttagagcct tttatctgca gaacatcgat
      601 atgtatatca ttctacagaa taatccagta ttgctgattt taaaggcaga gaagttctca
      661 aagttaattc acctatgtta ttttgtgtac aagttgttat tgttgaacat acttcaaaaa
      721 taatgtgcca tgtgggtgag ttaattttac caagagtaac tttactctgt gtttaaaaat
      781 gaagttaata atgtattgta atctttcatc caaaatattt tttgcaagtt atattagtga
      841 agatggtttc aattcagatt gtcttgcaac ttcagtttta tttttgccaa ggcaaaaaac
      901 tcttaatctg tgtgtatatt gagaatccct taaaattacc agacaaaaaa atttaaaatt
      961 acgtttgtta ttcctagtgg atgactgttg atgaagtata cttttcccct gttaaacagt
     1021 agttgtattc ttctgtattt ctaggcacaa ggttggttgc taagaagcct ataagaggaa
     1081 tttcttttcc ttcattcata gggaaaggtt ttgtattttt taaaacacta aaagcagcgt
     1141 cactctacct aatgtct
//
```



Estructura de Proteínas

MAKEFGIPAA
VAGTVLNVVE
AGGWVTTIV
S ILTAVGSGGL
SLLAAAGRES
IKAYLKKEIK
KKGKRAVIAW

Principios Físicos: no en los próximos años!!

Informática (copiar de los casos conoc

- Modelado por homología

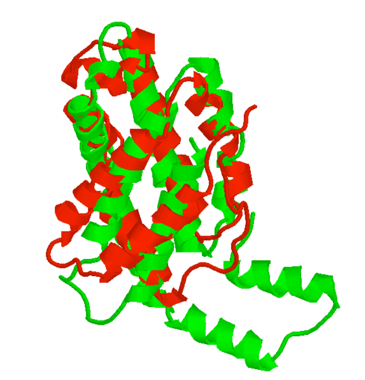
```
 Score = 85.1 bits (208), Expect = 6e-19
  Identities = 27/56 (48%), Positives = 42/56 (74%), Gaps = 1/56

Query: 2  FIAIYDYKAETEEDLTIKKGEKLEIIEKEGD-WWKAKAIGSGEIGYIPANYIA
           F+A+YDY+A TE+DL+  KGEK +I+       WW+A+++ +GE GYIP+NY+A
Sbjct: 8  FVALYDYEARTEDDLSFHKGEKFQILNSSEGDWWEARSLTTGETGYIPSNYVA
```

- Threading

MAKEFGIPAA
VAGTVLNVVE
AGGWVTTIV
S ILTAVGSGGL
SLLAAAGRES
IKAYLKKEIK
KKGKRAVIAW

PDB



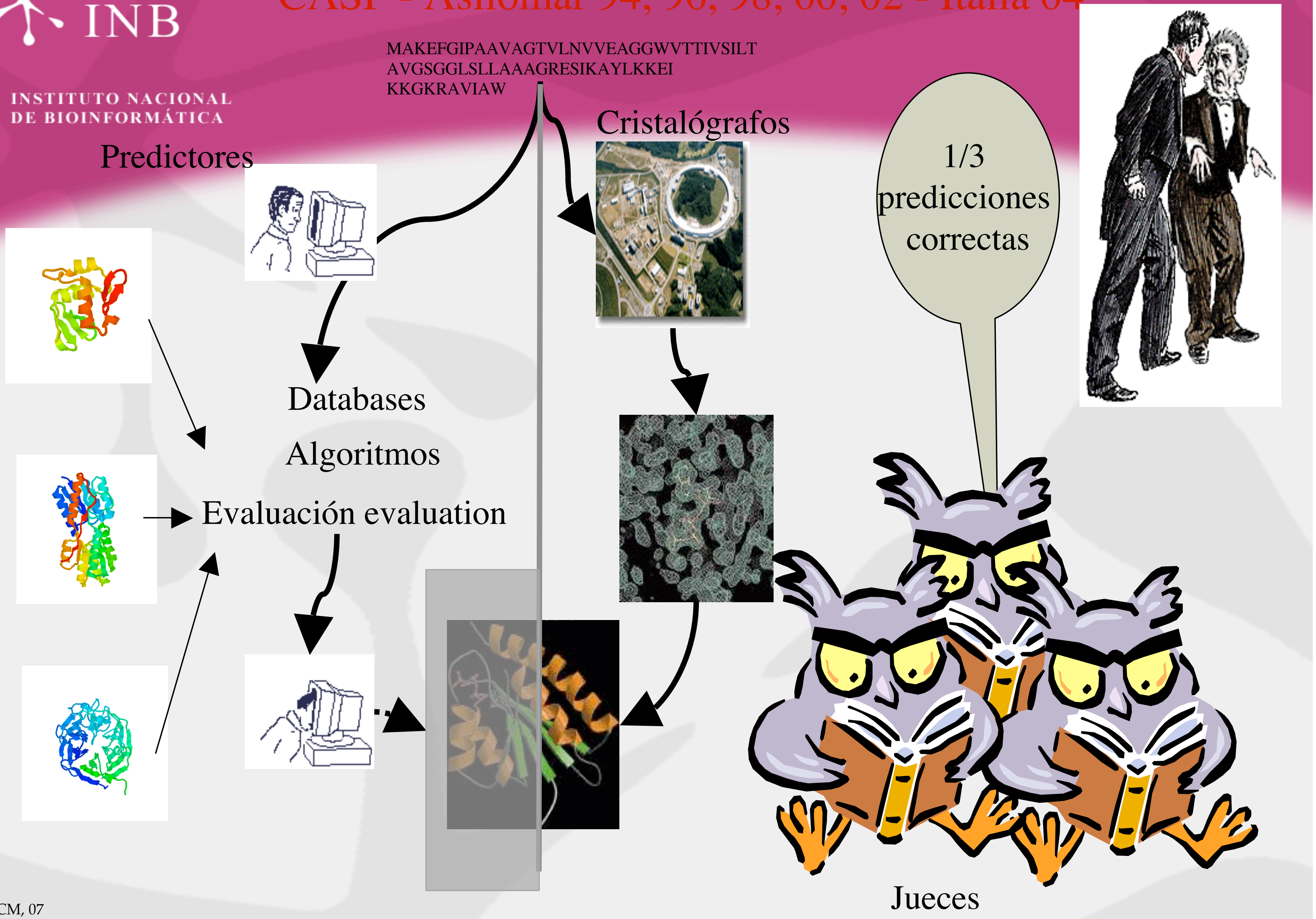
Orden de predicciones

Niveles de PseudoEnergía

By D. Devos

UCM, 07
Alfonso Valencia CNIO

CASP - Asilomar 94, 96, 98, 00, 02 - Italia 04



MAKEFGIPAAVAGTVLNVVEAGGWVTTIVSILT
AVGSGGLSLLAAAGRESIKAYLKKEI
KKGKRAVIAW

Predictores

Cristalógrafos

1/3 predicciones correctas

Databases

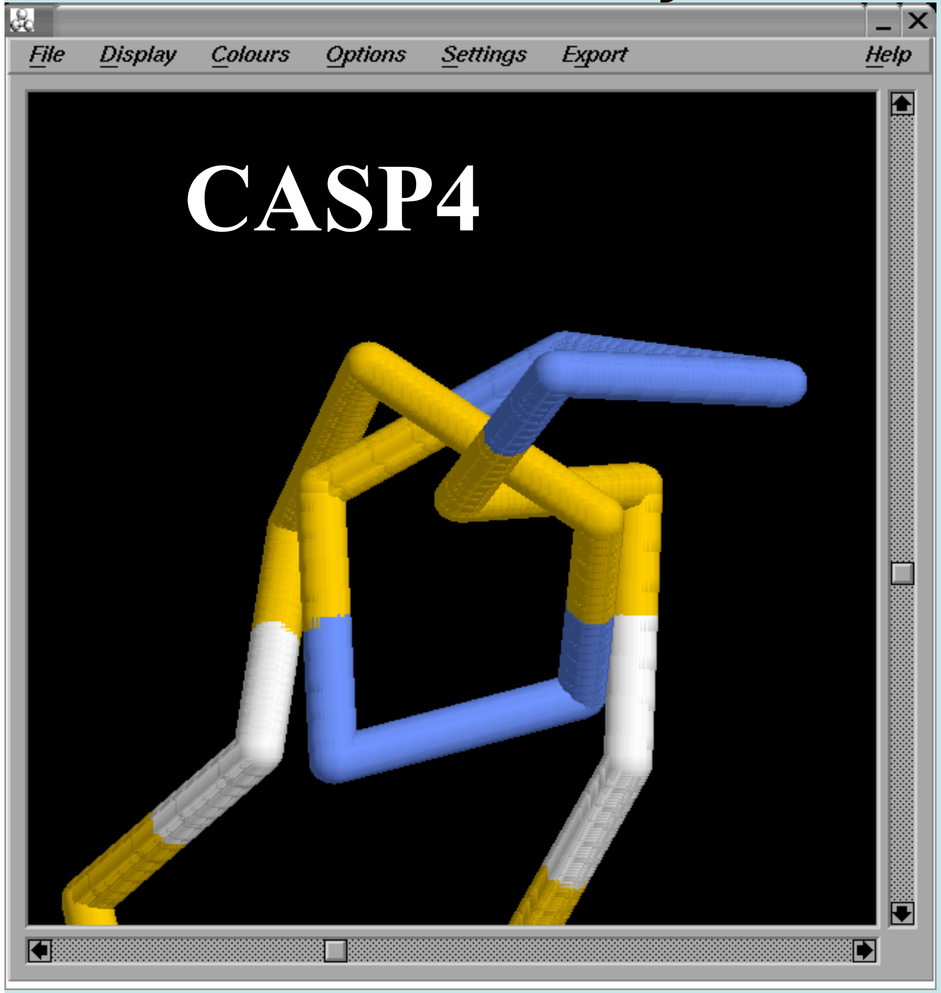
Algoritmos

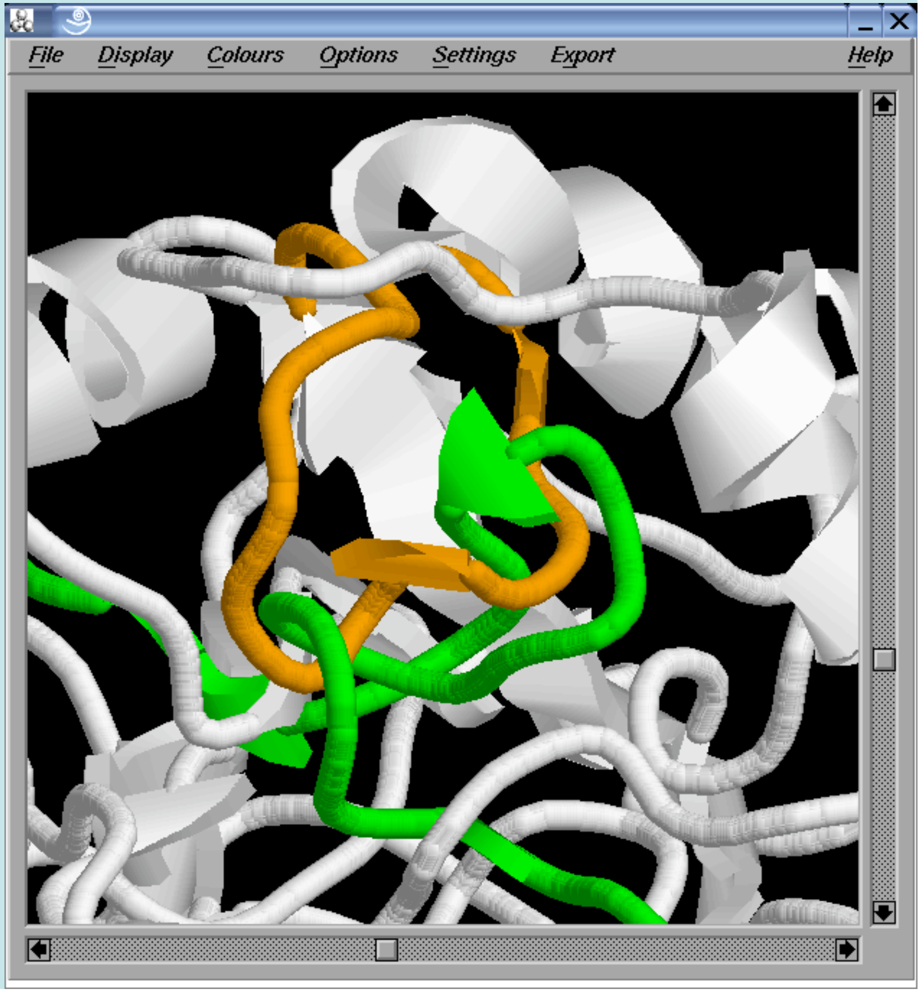
Evaluación evaluation

Jueces

By D. Devos

You think you have seen knots ...

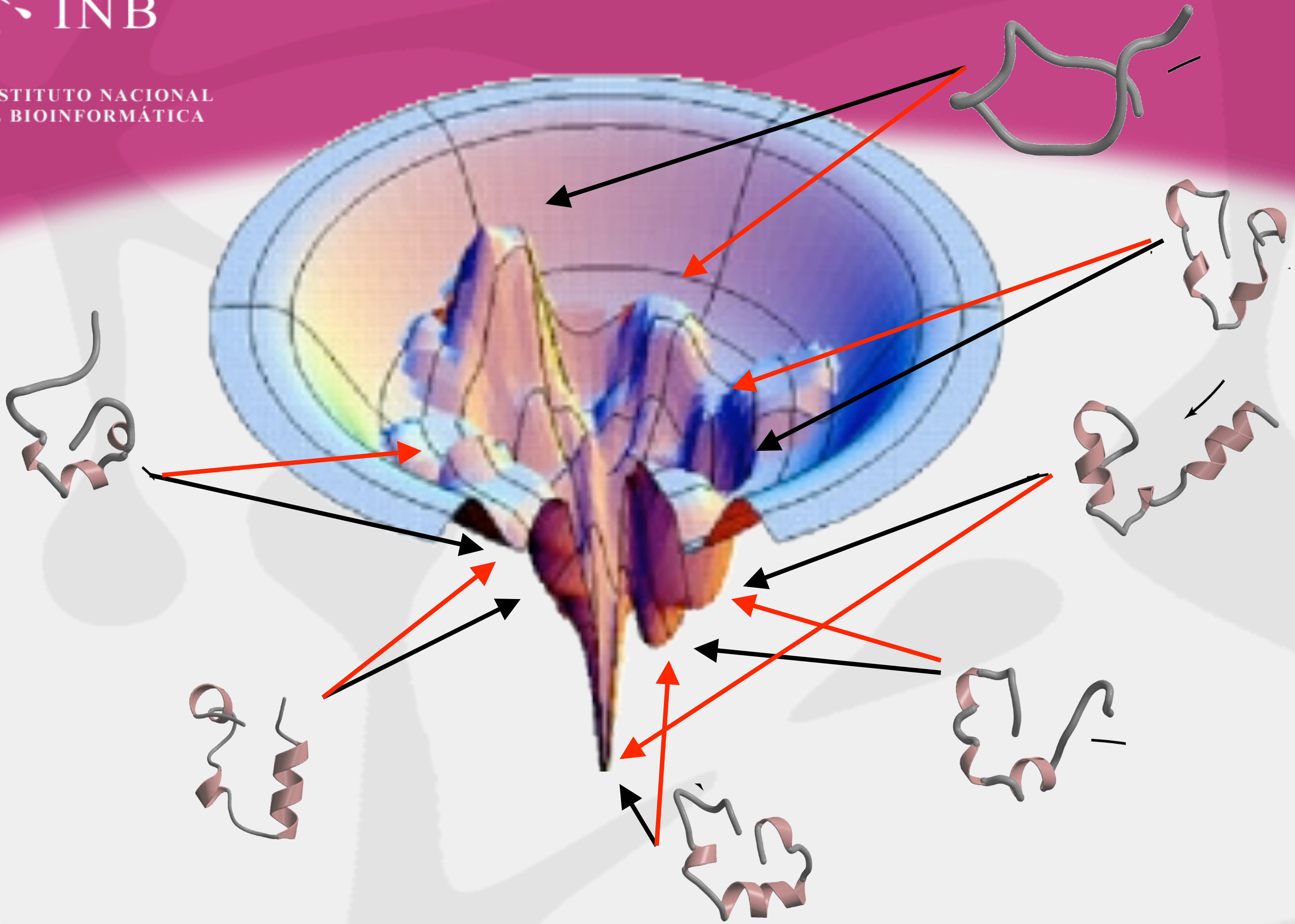




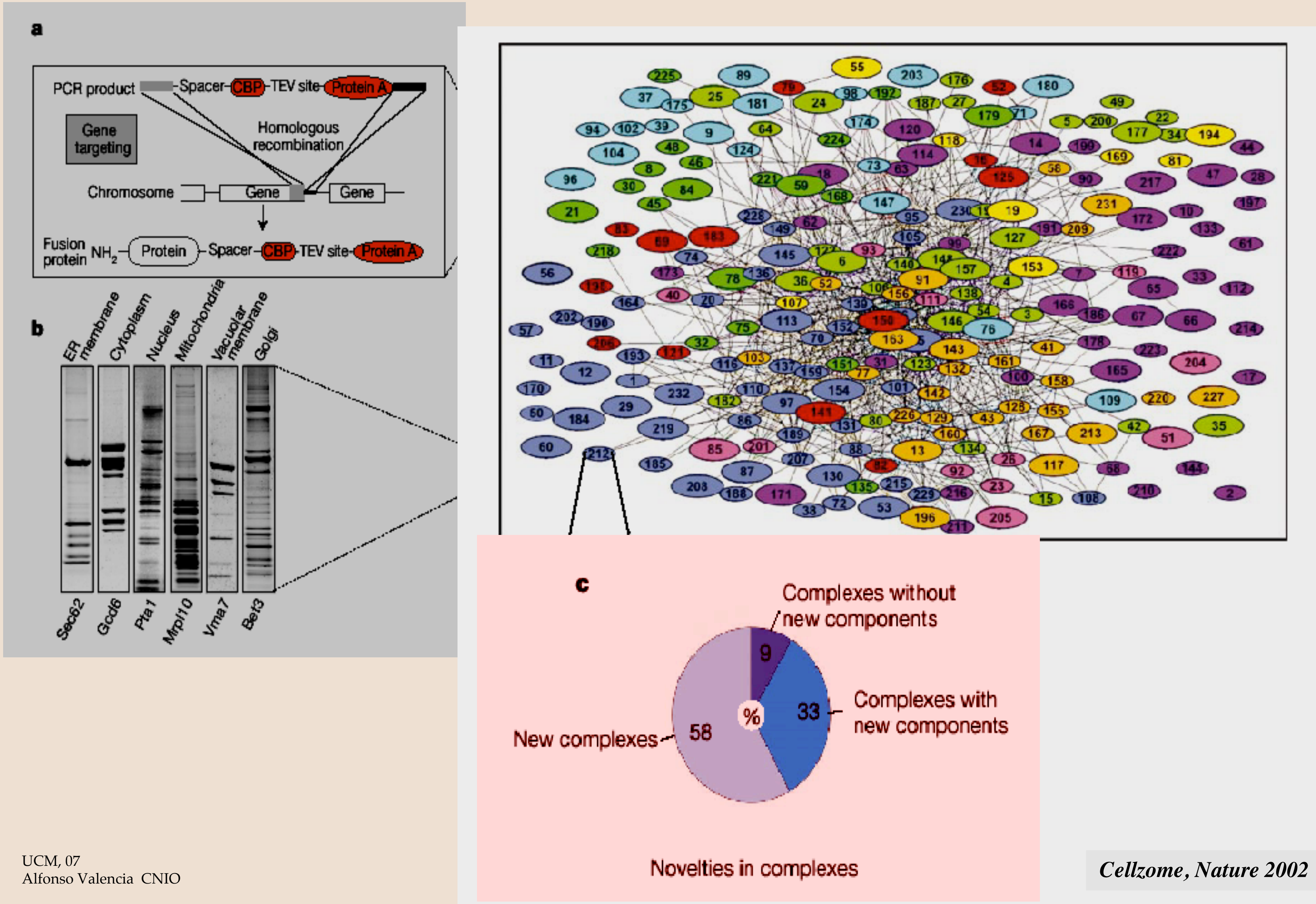
UCM, 07
Alfonso Valencia CNIO

INB

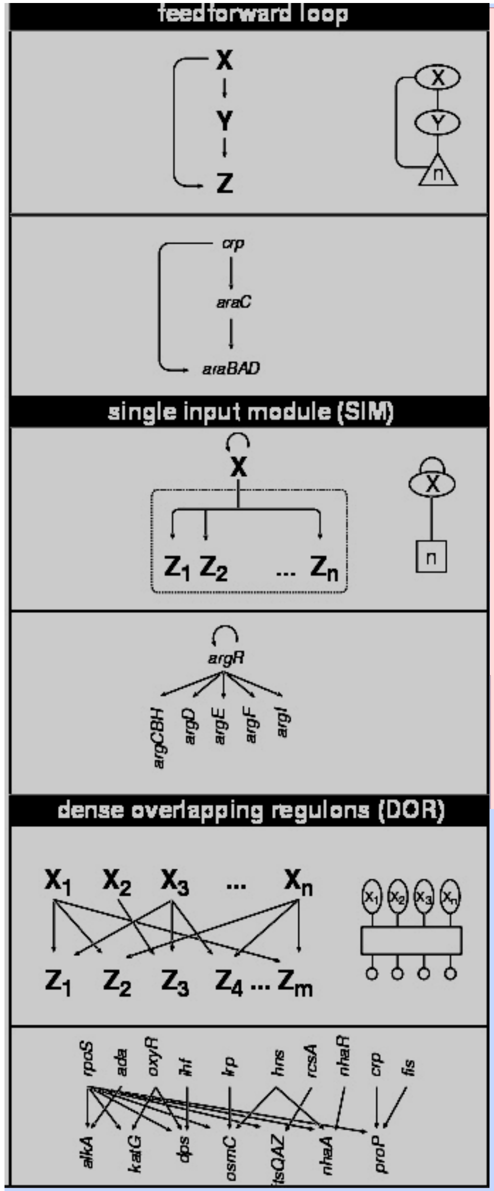
INSTITUTO NACIONAL DE BIOINFORMÁTICA

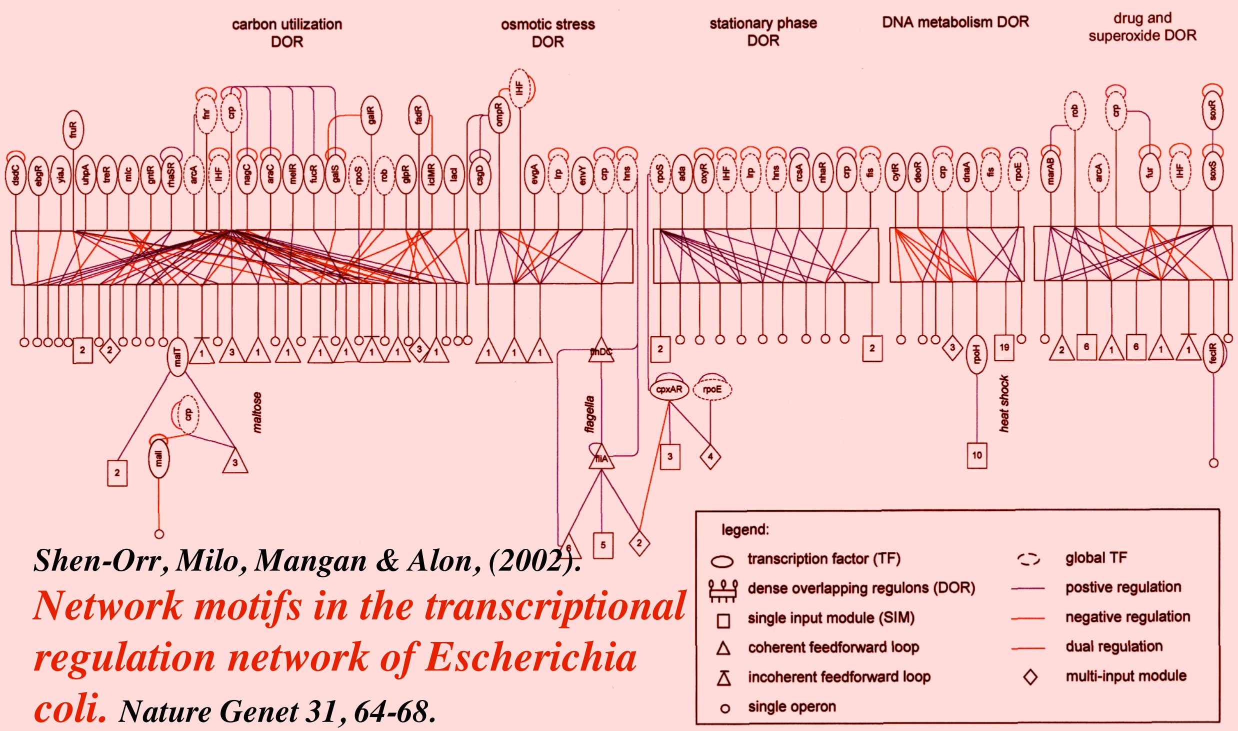


UCM, 07
Alfonso Valencia CNIO

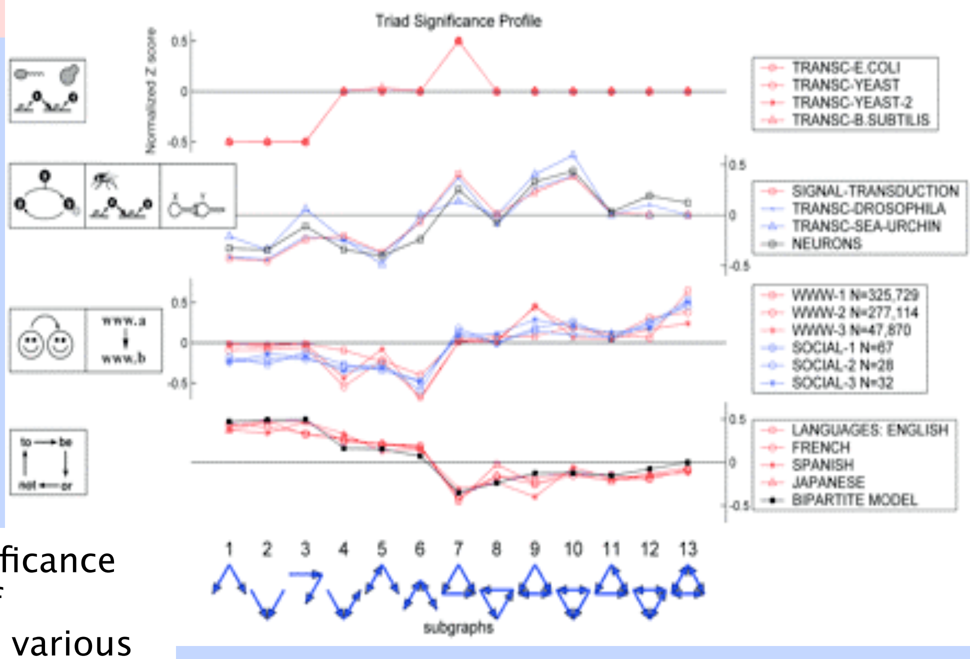


Cellzome, Nature 2002



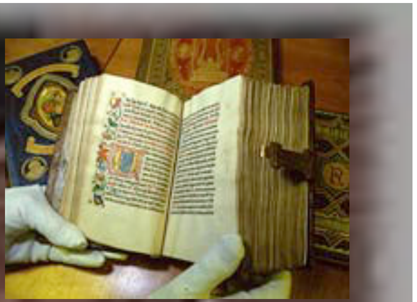


Shen-Orr, Milo, Mangan & Alon, (2002). ***Network motifs in the transcriptional regulation network of Escherichia coli.*** *Nature Genet 31, 64-68.*



The triad significance profile (TSP) of networks from various

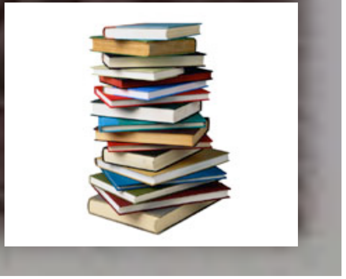
Milo, Itzkovitz, Kashtan, Levitt, Shen-Orr, Ayzenshtat, Sheffer, Alon, (2004) ***Superfamilies of Evolved and Designed Networks.*** ***Science, 303, 1538-1542***

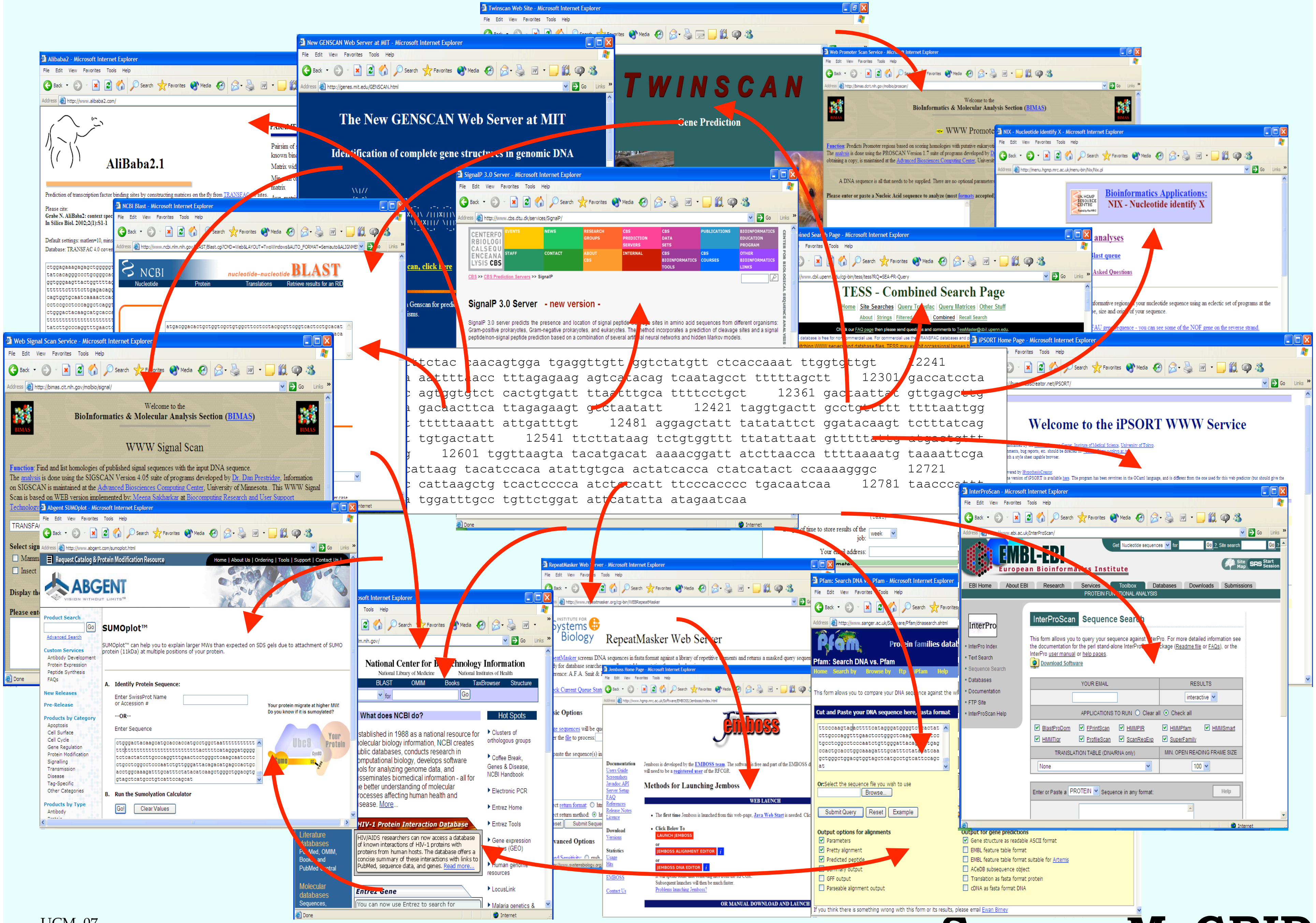


- 14 M publications
- 10M accessible in Medline
- 8M abstracts in Medline

- Increasing number of accessible journals (full text)

- Information organized in web repositories




Twinscan Web Site - Microsoft Internet Explorer
TWINSCAN
Gene Prediction
New GENSCAN Web Server at MIT - Microsoft Internet Explorer
The New GENSCAN Web Server at MIT
Identification of complete gene structures in genomic DNA
Alibaba2 - Microsoft Internet Explorer
AliBaba2.1
Web Promoter Scan Service - Microsoft Internet Explorer
BioInformatics & Molecular Analysis Section (BIMAS)
WWW Promoter
NIX - Nucleotide identify X - Microsoft Internet Explorer
Bioinformatics Applications:
NIX - Nucleotide identify X
NCBI Blast - Microsoft Internet Explorer
NCBI
nucleotide-nucleotide BLAST
SignalP 3.0 Server - Microsoft Internet Explorer
CBS
SignalP 3.0 Server - new version -
TESS - Combined Search Page
Web Signal Scan Service - Microsoft Internet Explorer
BioInformatics & Molecular Analysis Section (BIMAS)
WWW Signal Scan
iPSORT Home Page - Microsoft Internet Explorer
Welcome to the iPSORT WWW Service
InterProScan - Microsoft Internet Explorer
EMBL-EBI
European Bioinformatics Institute
InterProScan Sequence Search
Abgent SUMOplot - Microsoft Internet Explorer
ABGENT
SUMOplot™
National Center for Biotechnology Information
What does NCBI do?
HIV-1 Protein Interaction Database
Entrez Gene
RepeatMasker Web Server - Microsoft Internet Explorer
Institute for Systems Biology
RepeatMasker Web Server
Jemboss Home Page - Microsoft Internet Explorer
emboss
Methods for Launching Jemboss
Pfam: Search DNA vs. Pfam - Microsoft Internet Explorer
Pfam
Protein families database
Pfam: Search DNA vs. Pfam
tctcac caacagtgga tgaggttgtt ggtctatgtt ctcaccaaat ttggtgttgt 12241
aattttacc tttagagaag agtcatacag tcaatagcct tttttagctt 12301 gaccatccta
agtggtgtct cactgtgatt ttaatttgca ttttcctgct 12361 gactaattat gttgagcttg
gacaacttca ttagagaagt gcctaatatt 12421 taggtgactt gcctgttttt ttttaattgg
tttttaaatt attgatttgt 12481 aggagctatt tatatattct ggatacaagt tctttatcag
tgtgactatt 12541 ttcttataag tctgtggttt ttatattaat gtttttattg atgactgttt
12601 tggttaagta tacatgacat aaaacggatt atcttaacca ttttaaaatg taaaattcga
cattaag tacatcacca atattgtgca actatcacca ctatcatact ccaaaagggc 12721
cattaagctg tcactcccca atctatccatt ttcccacccc tgacaatcaa 12781 taacccattt
tggatttgcc tgttctggat attcatatta atagaatcaa

UCM, 07
Alfonso Valencia CNIO

Source: MyGRID

INB
INSTITUTO NACIONAL DE BIOINFO…

Analisis sistemático de familias de proteinas utilizando tecnología web (biomoby, contexto INB)

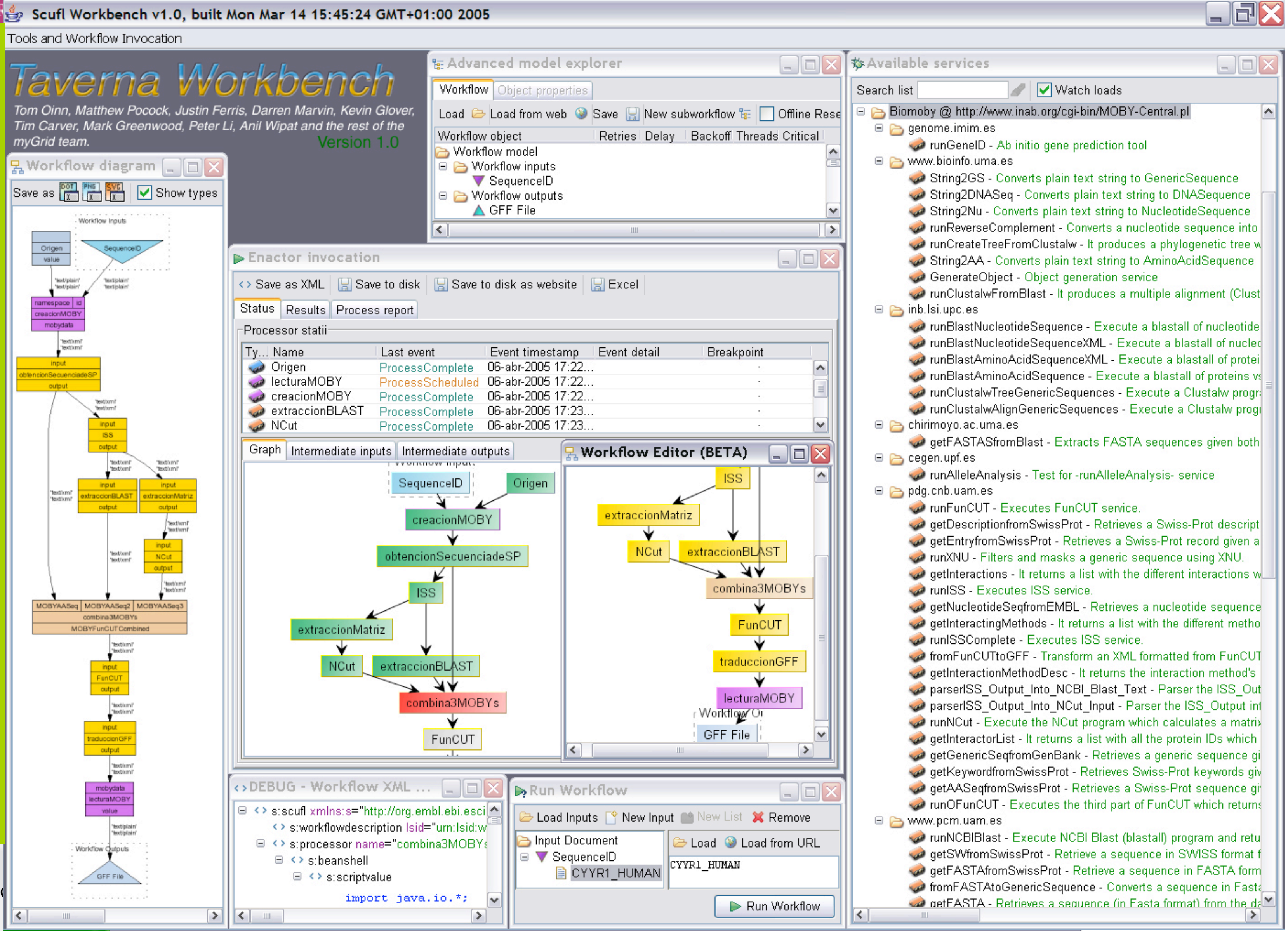
12 CPUs = 1.5 min

Human genome (30.000 sequences)

Weekly updates

Other genomes

More complex workflows



Scufl Workbench v1.0, built Mon Mar 14 15:45:24 GMT+01:00 2005

Tools and Workflow Invocation

Taverna Workbench
Tom Oinn, Matthew Pocock, Justin Ferris, Darren Marvin, Kevin Glover, Tim Carver, Mark Greenwood, Peter Li, Anil Wipat and the rest of the myGrid team.
Version 1.0

Advanced model explorer

Workflow | Object properties

Load | Load from web | Save | New subworkflow | □ Offline Rese

Workflow object	Retries	Delay	Backoff	Threads	Critical
Workflow model					
Workflow inputs					
SequenceID					
Workflow outputs					
GFF File					

Enactor invocation

Save as XML | Save to disk | Save to disk as website | Excel

Status | Results | Process report

Processor statii

Ty…	Name	Last event	Event timestamp	Event detail	Breakpoint
	Origen	ProcessComplete	06-abr-2005 17:22…		·
	lecturaMOBY	ProcessScheduled	06-abr-2005 17:22…		·
	creacionMOBY	ProcessComplete	06-abr-2005 17:22…		·
	extraccionBLAST	ProcessComplete	06-abr-2005 17:23…		·
	NCut	ProcessComplete	06-abr-2005 17:23…		·

Graph | Intermediate inputs | Intermediate outputs

Workflow Editor (BETA)

Available services

Search list | ☑ Watch loads

- Biomoby @ http://www.inab.org/cgi-bin/MOBY-Central.pl
  - genome.imim.es
    - runGeneID - Ab initio gene prediction tool
  - www.bioinfo.uma.es
    - String2GS - Converts plain text string to GenericSequence
    - String2DNASeq - Converts plain text string to DNASequence
    - String2Nu - Converts plain text string to NucleotideSequence
    - runReverseComplement - Converts a nucleotide sequence into
    - runCreateTreeFromClustalw - It produces a phylogenetic tree w
    - String2AA - Converts plain text string to AminoAcidSequence
    - GenerateObject - Object generation service
    - runClustalwFromBlast - It produces a multiple alignment (Clust
  - inb.lsi.upc.es
    - runBlastNucleotideSequence - Execute a blastall of nucleotide
    - runBlastNucleotideSequenceXML - Execute a blastall of nuclec
    - runBlastAminoAcidSequenceXML - Execute a blastall of protei
    - runBlastAminoAcidSequence - Execute a blastall of proteins vs
    - runClustalwTreeGenericSequences - Execute a Clustalw progra
    - runClustalwAlignGenericSequences - Execute a Clustalw progr
  - chirimoyo.ac.uma.es
    - getFASTASfromBlast - Extracts FASTA sequences given both
  - cegen.upf.es
    - runAlleleAnalysis - Test for -runAlleleAnalysis- service
  - pdg.cnb.uam.es
    - runFunCUT - Executes FunCUT service.
    - getDescriptionfromSwissProt - Retrieves a Swiss-Prot descript
    - getEntryfromSwissProt - Retrieves a Swiss-Prot record given a
    - runXNU - Filters and masks a generic sequence using XNU.
    - getInteractions - It returns a list with the different interactions w
    - runISS - Executes ISS service.
    - getNucleotideSeqfromEMBL - Retrieves a nucleotide sequence
    - getInteractingMethods - It returns a list with the different metho
    - runISSComplete - Executes ISS service.
    - fromFunCUTtoGFF - Transform an XML formatted from FunCUT
    - getInteractionMethodDesc - It returns the interaction method's
    - parserISS_Output_Into_NCBI_Blast_Text - Parser the ISS_Out
    - parserISS_Output_Into_NCut_Input - Parser the ISS_Output int
    - runNCut - Execute the NCut program which calculates a matrix
    - getInteractorList - It returns a list with all the protein IDs which
    - getGenericSeqfromGenBank - Retrieves a generic sequence gi
    - getKeywordfromSwissProt - Retrieves Swiss-Prot keywords giv
    - getAASeqfromSwissProt - Retrieves a Swiss-Prot sequence gi
    - runOFunCUT - Executes the third part of FunCUT which returns
  - www.pcm.uam.es
    - runNCBIBlast - Execute NCBI Blast (blastall) program and retu
    - getSWfromSwissProt - Retrieve a sequence in SWISS format f
    - getFASTAfromSwissProt - Retrieve a sequence in FASTA form
    - fromFASTAtoGenericSequence - Converts a sequence in Fasta
    - getFASTA - Retrieves a sequence (in Fasta format) from the da

DEBUG - Workflow XML …

```
<> s:scufl xmlns:s="http://org.embl.ebi.esci
   <> s:workflowdescription lsid="urn:lsid:w
   <> s:processor name="combina3MOBYs
      <> s:beanshell
         <> s:scriptvalue
               import java.io.*;
```

Run Workflow

Load Inputs | New Input | New List | Remove

Input Document
SequenceID
CYYR1_HUMAN

Load | Load from URL

```
CYYR1_HUMAN
```

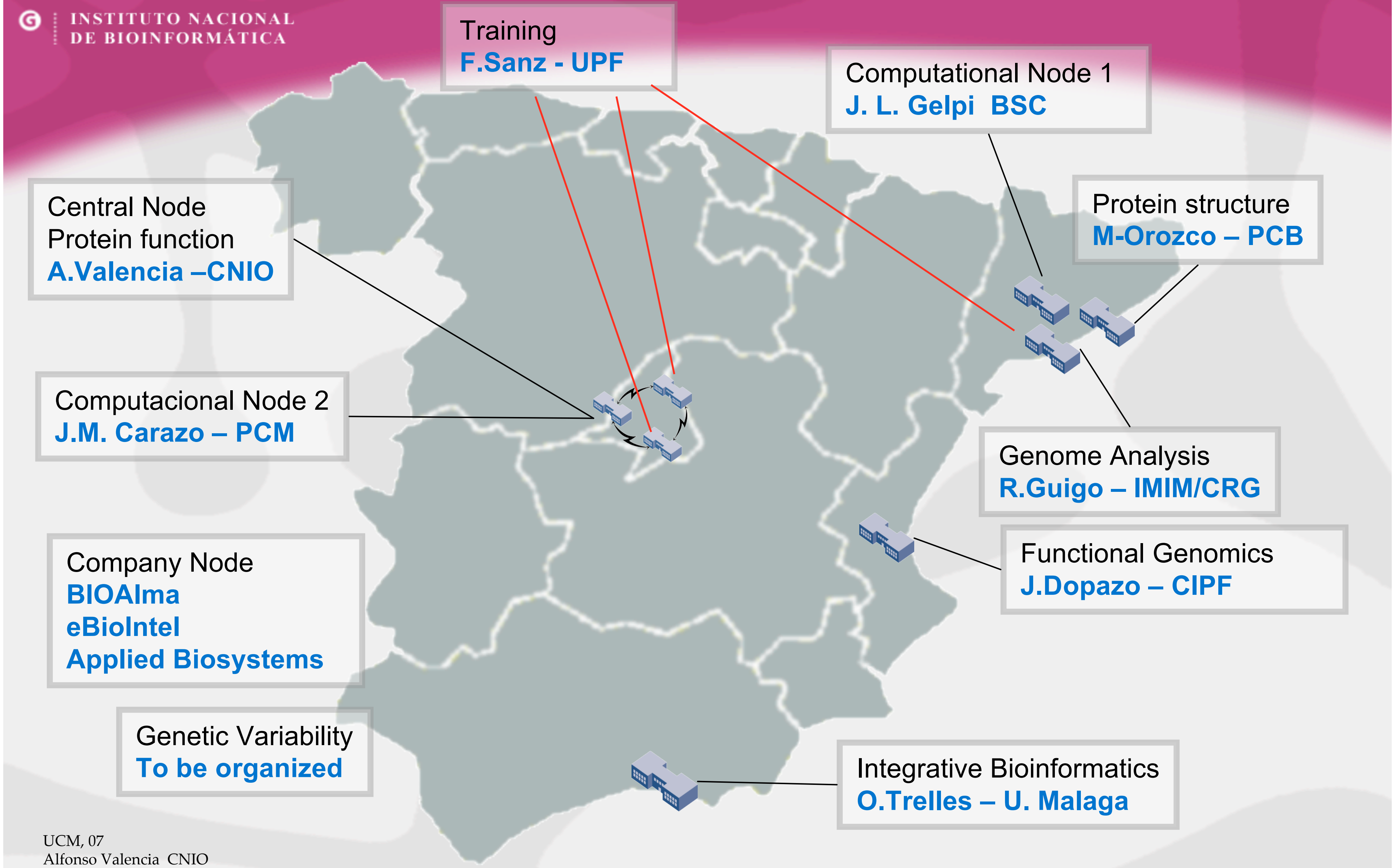
Run Workflow

UCM, 07
Alfonso Valencia CNI

INB

INSTITUTO NACIONAL DE BIOINFORMÁTICA

A Virtual Institute



Training
F.Sanz - UPF

Computational Node 1
J. L. Gelpi BSC

Central Node
Protein function
A.Valencia –CNIO

Protein structure
M-Orozco – PCB

Computacional Node 2
J.M. Carazo – PCM

Genome Analysis
R.Guigo – IMIM/CRG

Company Node
BIOAlma
eBioIntel
Applied Biosystems

Functional Genomics
J.Dopazo – CIPF

Genetic Variability
To be organized

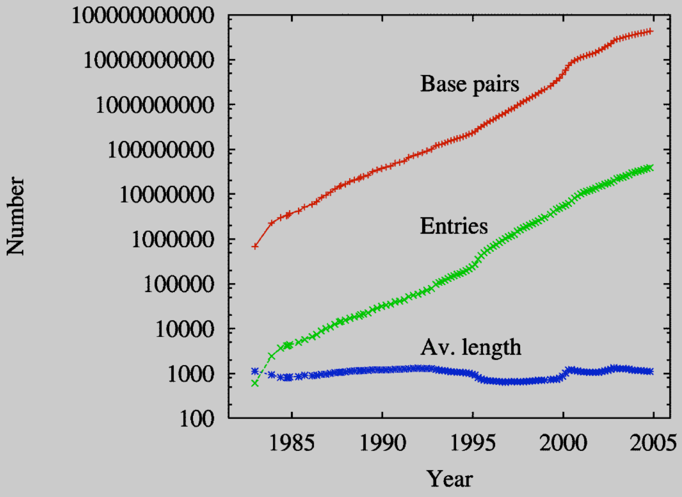
Integrative Bioinformatics
O.Trelles – U. Malaga

UCM, 07
Alfonso Valencia CNIO

```
TCCAAACCCAGGCTCTCTCCCAAA
TGAAGTAGACAATGCCCAGTGTTT
GCTTGTAATCCCAGCACTGTGGGA
CAGGCGTGGTAGCACATACCTGTA
AGGGCAACAAGAGGCAAAATGGCG
AATGGGCCACGGAATCACTCATTC
GGCAGGGAATGGGGAAAAACAGGG
TGCCTGCGTCCAGGGCTGGGTTGG
CTCCCCGCCACAACATACATCCCA
AATGAAGACATGAGATTCCTCTGC
GCTTTTGTAACCTCTCCATGTGAG
GTGTCACAGACACAGAGTAAACTT
TCAGTATGGCGCTGTCCTGGGTTC
GGGTGAGTGCCTGGGCTAGCCCTG
TTCTGGGTCCCCAGGGTGAAATTC
GAGTAAACTTTTGCTGGGCTCCAA
TCCTGGGTTCTTACAGTCCTGAGC
GCTAGCCCTGTCCTGAGCACATGG
GGTGAAATTCTCACCAGCCCAGGG
GGGGGGCAATGACTGATCCTCAGG
GTCAAAGATTCAGCATCAAGCCCC
GAGATCCAAGCAACCTTCTTTTAC
CTGGCCAGAAGACCTGAGCAAGTC
CGATTGGCCACTTCTCCTCGATAA
CCGTCTCCAGATACGGTGAGGGCC
AATAACCACTAACATTTTTGAGCT
TCACACAGCTTATAATTAGAACTA
GAGGTGGTACCTTCAACTATGTCC
CTCCTCTGTAAAACGGGGAGAAAG
CGGTGGCCCAAAGGAGACCCGGGC
AGGACTCCTCACCTGTAAGACAGG
TTCTATGGTAGGCATGCTTAGCAG
CCCACCATGTCCCCAGTCAGTCTC
GATGTCATGTACACCGACTGGAAA
GCTGCACAGCTAGGCAGATCTTCT
AACCAGGGAGGACCTGAAAGCTAA
CCAGGCTTTCACCCCAACCTCCAC
ATTTTTGCTACAAATGAAAATTAC
CACCGATGAGCTGGTTCCAATTTT
CTCCTGCCCCATCCCCAGCCCCAG
TGGGTCCCAAGGCCACCCTGCTCC
AGTGGTTCTTAAGGGTCTGAGCTC
TGGTTGAGCTGGTTCCACAGGCCA
AGTGTGAGCCACTGGAGAAGCAGC
CCGACATGTGTACCTCAGCTTTTT
```

GenBank Release 144.0 — October 15, 2004

Species	Genome size	Bases	Entries
Homo sapiens	3,400,000,000	10,965,381,932	8,338,229
Mus musculus	3,454,200,000	6,774,229,516	6,071,679
Rattus norvegicus	2,900,000,000	5,644,398,664	985,869
Danio rerio	1,900,000,000	1,957,414,191	771,409
Zea mays	5,000,000,000	1,455,760,045	2,292,596
Oryza sativa	5,000,000,000	779,829,843	336,051
Drosophila melanogaster	180,000,000	754,291,835	482,102
Bos taurus	3,651,500,000	650,653,065	884,892
Gallus gallus	1,200,000,000	605,802,046	697,037
Arabidopsis thaliana	100,000,000	584,114,192	845,876
Canis familiaris	3,355,500,000	582,919,466	1,015,724
Xenopus tropicalis	3,355,500,000	464,627,420	560,218
Pan troglodytes	3,577,500,000	439,544,237	193,505
Ciona intestinalis	200,000,000	418,098,823	693,084
Brassica oleracea	759,500,000	403,897,848	595,915
Macaca mulatta	3,543,000,000	372,152,352	55,192
Medicago truncatula	400,000,000	327,501,222	348,369
Triticum aestivum	16,978,500,000	311,942,146	570,595
Xenopus laevis	3,100,000,000	298,427,470	455,955
Caenorhabditis elegans	100,000,000	283,634,604	309,719
Total		43,194,602,655	38,941,263



```
CTAGTACAGGGACTGTG
CAGGCGCGGTGGCTCAT
AAAAATACAAAATTAGC
GCCATTGCACTCCAGCC
CATAAGGTAGGAACGCA
AGCCTGCCTGCTCAGAT
GTCCCTTTTTGGAAATG
CATCCTTCACCCCAACT
CAGAGGGATAGGAATTG
GGCAGCCATACTCTAGG
CAAGTGCTGGCAGGATT
CCACGTGCCTCCTGGTC
CAACGCCACCCTGGACC
AGCTCTGCCTTTTTCTC
ATTGTGTCACAGACACA
GTCTCAGTATGGCGCTG
ACCGGGTGAGTGCCTGG
CTCTTCTGGGTCCCCAG
GCCTGGAGGGGCTGCCT
TCTCCTCCCAGCTGACA
ACAATAAGTCGGTTCAG
CCCCCGATCTCAGCTGT
CCCGGGACTGTGATGGG
CCAGCGGGAGAATGGGA
GGCTTTGGGCACAGAGA
AAATAGTTTCCCCAAGG
GAAGCCTGTATGTTGGA
CTCTGGGCCTCATTTTC
AGCAGGGGCTGGGCACA
TCCTTAGGGACACCCCT
AGAACTGGGGGCTGTCT
CCTGGCCTCCGCCGGGC
TGTGCATTCCCAGGTCA
GAAAGCTGCCAGGCAAG
GGGAGAAGTCCCTGTGA
CAGACGCGTCCCAAACC
CCACAGTCACCATCCCG
TACCTGTGAAATCACAG
CCAGGGATGTGAACTAC
GGCTAGGGAGCCTCGAA
TGTGCTCAGTAAATGCC
TGCCAGGGTTGGAGCCC
CGCTGCCTTCTAGGATA
CATCCTGCCCCTCCAAC
GGCAGGAGGGTTCACCG
```

MareNostrum

4.812 IBM PowerPC 970FX processors 2,2 GHz

(2.406 dual 64-bit processor blade nodes).

9,6 TB Main Memory

(4GB ECC 333 DDR memory per node).

42,35 Tflops (peak).

140 + 96 TB disk.

3 networks:

Myrinet, Gigabit, 10/100 Ethernet.

Linux 2.6 cluster (SuSe).

Diskless network support.

Collaboration with the INB node at the BSC

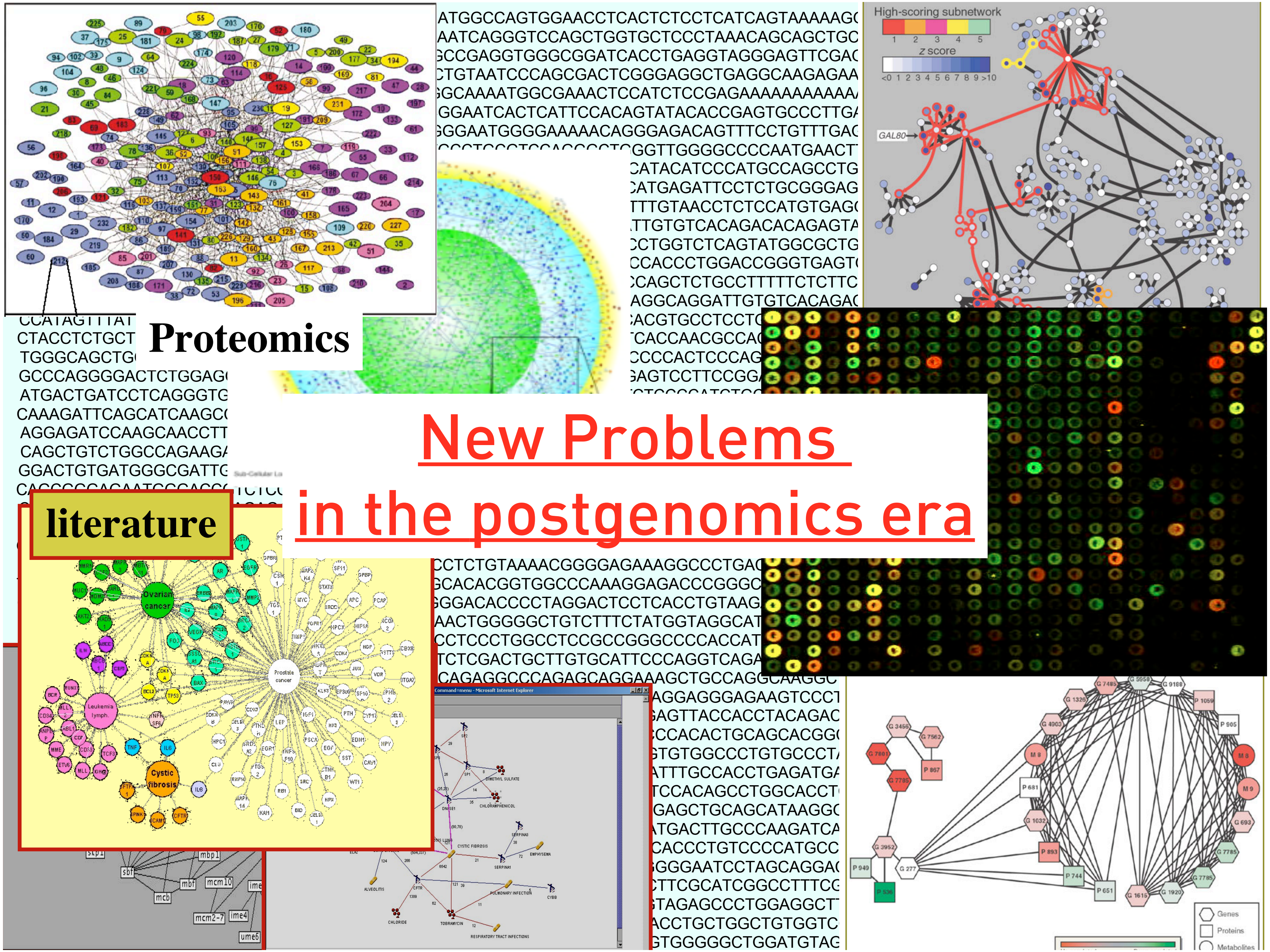
INB: Max. 40% capacity of the external use.

Department of Comp. Biol. (M. Orozco director)

Access / parallelization help service

eman ta zabal zazu

Barcelona Supercomputing Center



New Problems in the postgenomics era

Proteomics

literature

INB

INSTITUTO NACIONAL DE BIOINFORMÁTICA

CEGEN

Centro Nacional de Genotipado

Colaboración con proyectos de genómica. Diseño e implementación de workflos para el análisis de datos de SNPs (NV3)

Experimental design (linkage, pathway, etc)

Problem 1: feed the monster. E.g. Illumina: 150.000 genotipes at a time

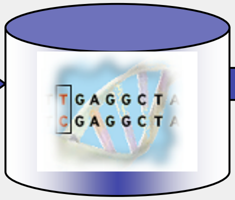
Computer-aided selection. PupaSNP and PupasView

Conde et al. 2004, 2005, NAR

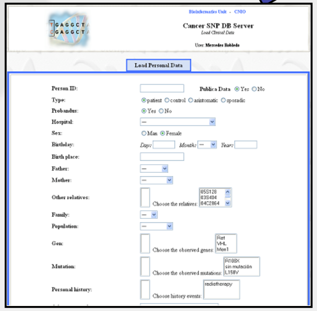
PupaSNP PupasView

moby

Problem 2: store results...



Cancer SNPs DB server



Bioinformatics Unit - CNIO

Cancer SNP DB Server

Load Clinical Data

User: **Mercedes Robledo**

Load Personal Data

Person ID: Publica Data ● Yes ○ No

Type: ● patient ○ control ○ asintomatic ○ sporadic

Probandus: ● Yes ○ No

Hospital: —

Sex: ○ Man ● Female

Birthday: *Day:* *Month:* — *Year:*

Birth place:

Father: —

Mother: —

Other relatives: Choose the relatives: 05S128 03S404 04C2864

Family: —

Population: —

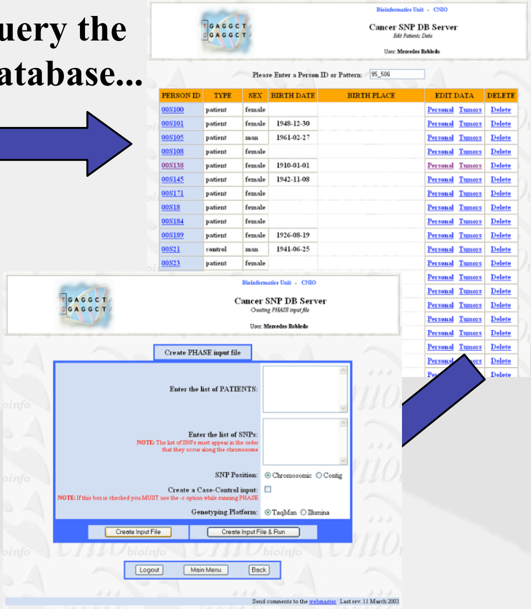
Gen: Choose the observed genes: Ret VHL Men1

Mutation: Choose the observed mutations: R108X sin mutación L158V

Personal history: Choose history events: radiotherapy

...along with clinical data

Problem 3: query the database...



Bioinformatics Unit - CNIO

Cancer SNP DB Server

Edit Patients Data

User: **Mercedes Robledo**

Please Enter a Person ID or Pattern: 95_506

PERSON ID	TYPE	SEX	BIRTH DATE	BIRTH PLACE	EDIT DATA	DELETE
00S100	patient	female			Personal Tumors	Delete
00S101	patient	female	1948-12-30		Personal Tumors	Delete
00S105	patient	man	1961-02-27		Personal Tumors	Delete
00S108	patient	female			Personal Tumors	Delete
00S138	patient	female	1910-01-01		Personal Tumors	Delete
00S145	patient	female	1942-11-08		Personal Tumors	Delete
00S171	patient	female			Personal Tumors	Delete
00S18	patient	female			Personal Tumors	Delete
00S184	patient	female			Personal Tumors	Delete
00S189	patient	female	1926-08-19		Personal Tumors	Delete
00S21	control	man	1941-06-25		Personal Tumors	Delete
00S23	patient	female			Personal Tumors	Delete

Bioinformatics Unit - CNIO

Cancer SNP DB Server

Creating PHASE input file

User: **Mercedes Robledo**

Create PHASE input file

Enter the list of PATIENTS:

Enter the list of SNPs:
NOTE: The list of SNPs must appear in the order that they occur along the chromosome

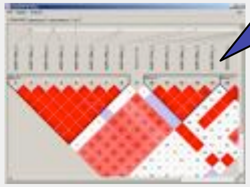
SNP Position: ● Chromosomic ○ Contig

Create a Case-Control input: □
NOTE: If this box is checked you MUST use the -c option while running PHASE

Genotyping Platform: ● TaqMan ○ Illumina

Send comments to the webmaster Last rev. 11 March 2003

...and submit to analysis programs



LD, Case-control, haplotypes, ODD ratios, etc.

October 2004: 45.000 SNPs designed

PDG Protein Design Group

www.cnio.es

INB

BioSapiens Network

ECCB

BioCreAtIvE

COMBIO

ENFIN Enabling Systems Biology

EMERGENCE COST
CA on Synthetic Biology

www.pdg.cnb.uam.es

www.inba.org

Curso UCM07 Personasl CNIO / CNB.

- Secuencias: L.Sanchez, J.C.Sanchez, F. Abascal, J.M.G-Izarzugaza, A. Rojas
- Prediccion genes: J.J.Wesselink
- Estructuras: D. de Juan, G. Lopez, M. Tress
- Text Mining: M.Krallinger
- Arrays: J.C.Oliveros, G.Gomez
- Redes: F.Pazos, I. Cases
- Web services: J.M. Fernandez, J.M.Rodriguez
- Introduccion y Resumen: A. Valencia

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

cnio

Alfonso Valencia