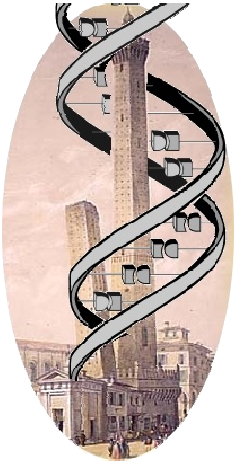
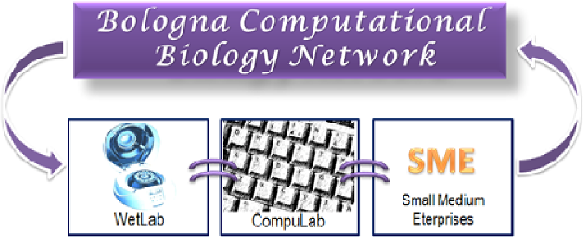
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Bioinformatics and Computational Biology in the post-genomic era



Rita Casadio



Bologna Computational Biology Network

WetLab

CompuLab

SME

Small Medium Eterprises

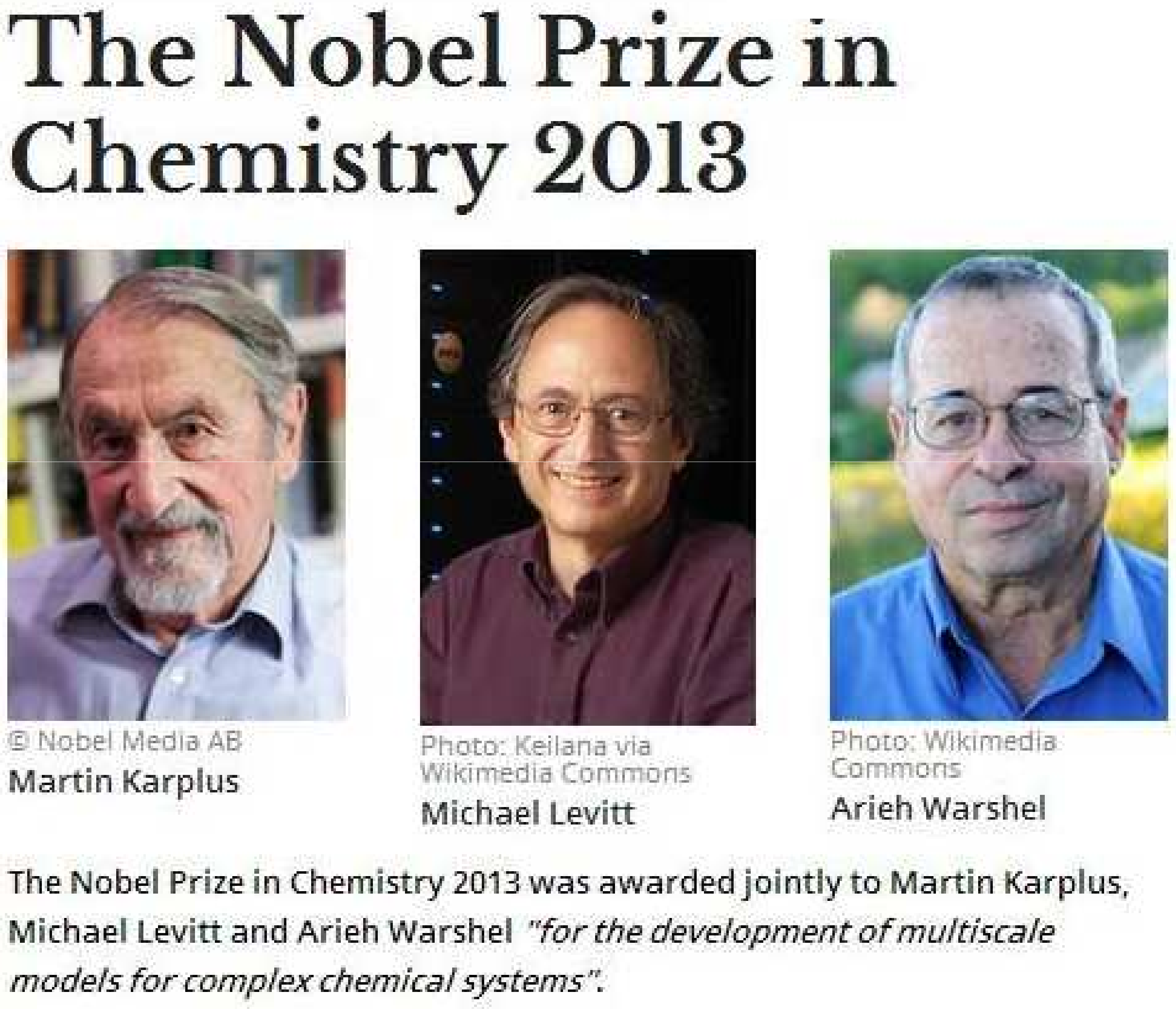
BIOCOMPUTING GROUP
University of Bologna, Italy

AIRBBC

Syllabus:

1) Motivations & definitions
2) The “omic” revolution
3) Next Generation Sequencing Data
4) Data archives & zooming in on biological complexity
5) Open problems in the omic era
6) Annotation pipelines at the Biocomputing group

Motivations



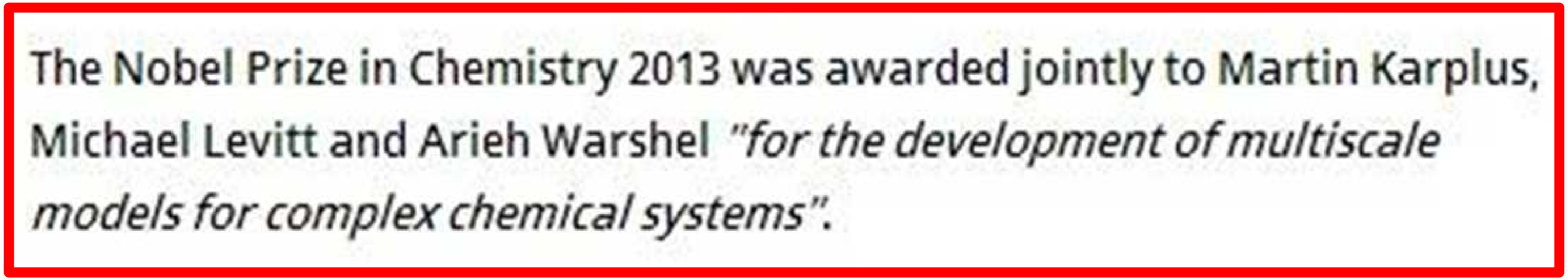
The Nobel Prize in Chemistry 2013

© Nobel Media AB
Martin Karplus

Photo: Keilana via Wikimedia Commons
Michael Levitt

Photo: Wikimedia Commons
Arieh Warshel

The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel *"for the development of multiscale models for complex chemical systems".*

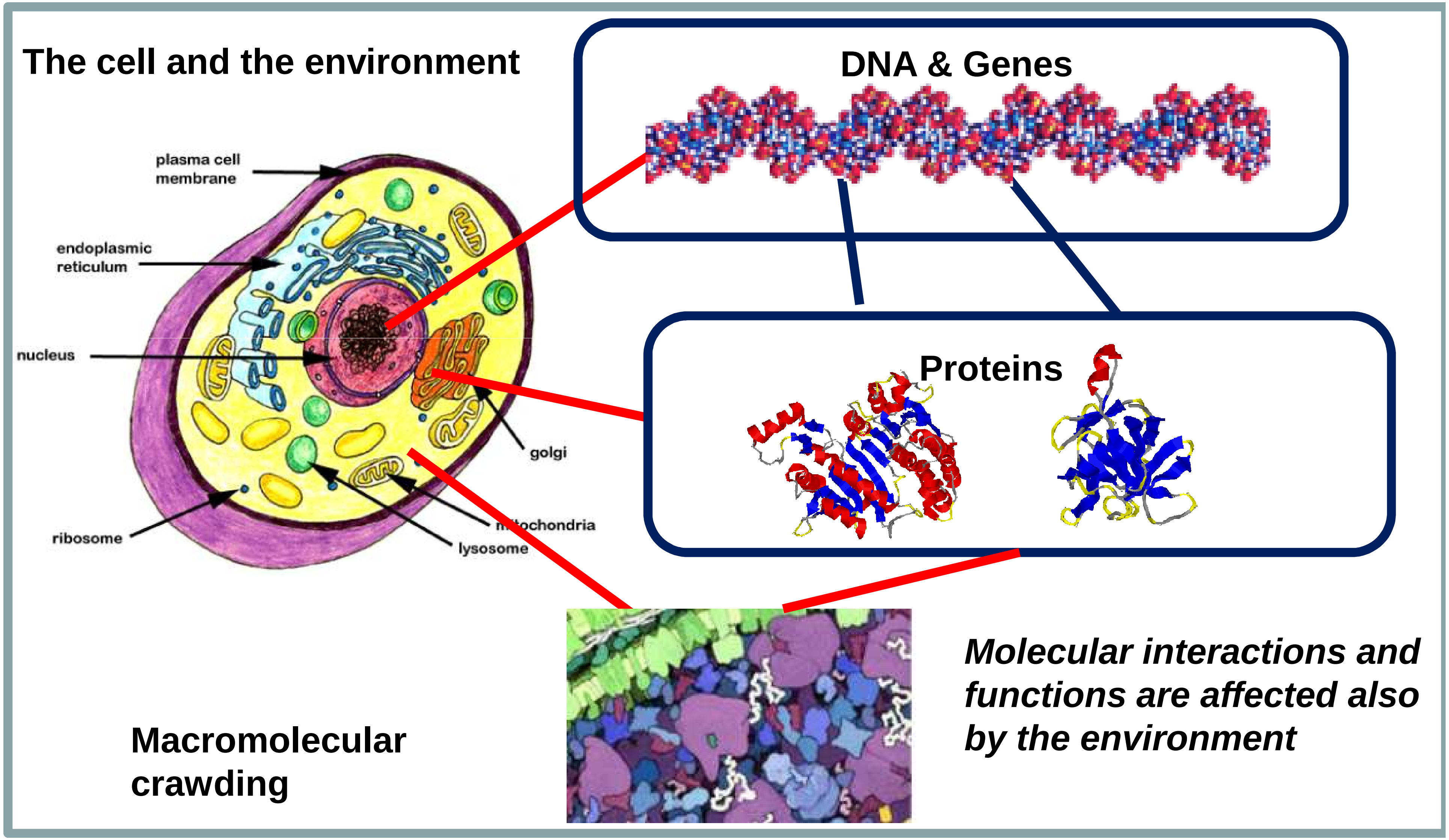


The Nobel Prize in Chemistry 2013 was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel *"for the development of multiscale models for complex chemical systems"*.

1. Complex chemical systems
2. Multiscale models
3. Development of multiscale models

The ingredients of biological complexity at the cell level

From genes to proteins, their interaction and the interplay with the environment



The cell and the environment

DNA & Genes

Proteins

Macromolecular crawding

Molecular interactions and functions are affected also by the environment

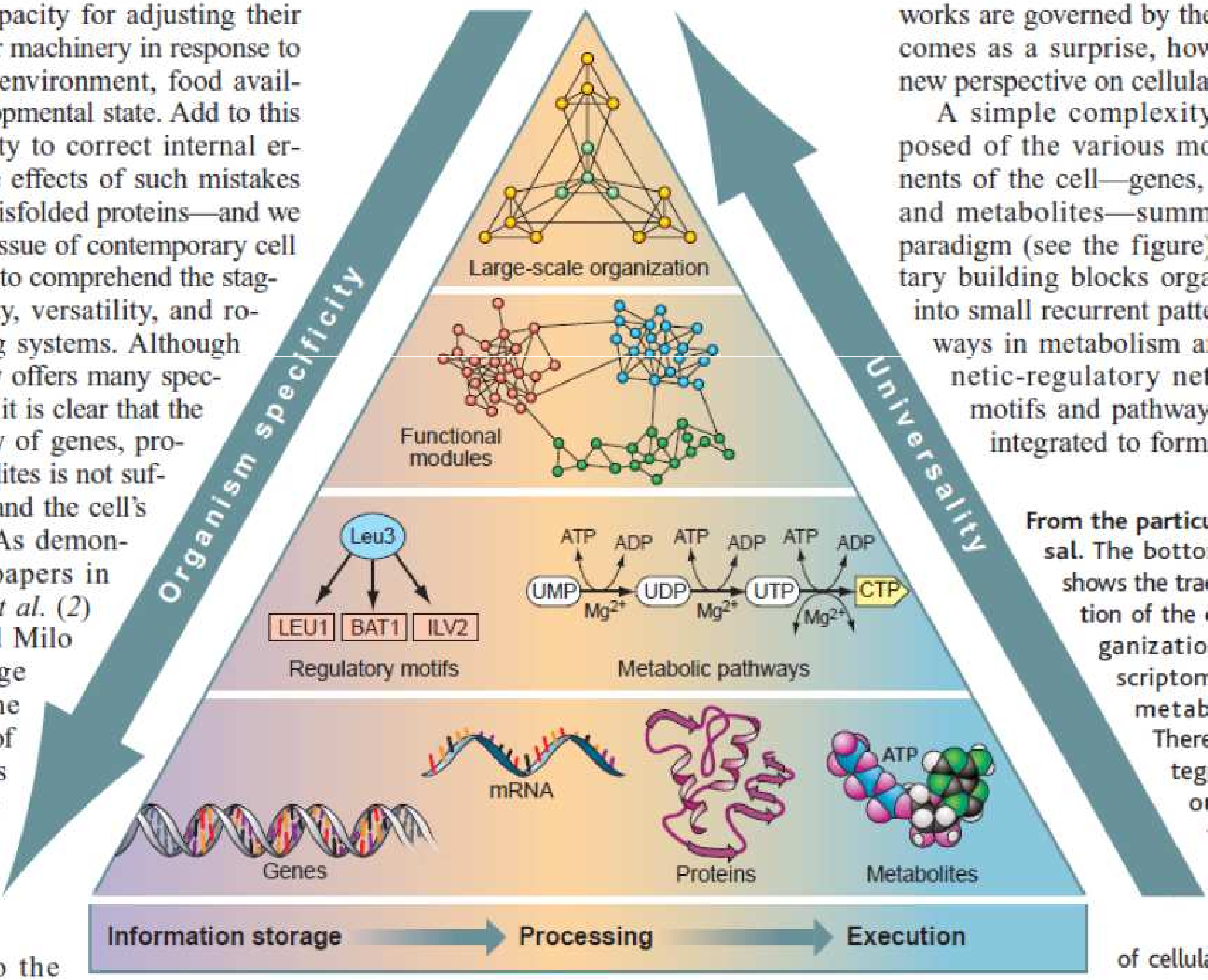
Life's Complexity Pyramid

Zoltán N. Oltvai and Albert-László Barabási

SCIENCE VOL 298 ,2002

Cells and microorganisms have an impressive capacity for adjusting their intracellular machinery in response to changes in their environment, food availability, and developmental state. Add to this an amazing ability to correct internal errors—battling the effects of such mistakes as mutations or misfolded proteins—and we arrive at a major issue of contemporary cell biology: our need to comprehend the staggering complexity, versatility, and robustness of living systems. Although molecular biology offers many spectacular successes, it is clear that the detailed inventory of genes, proteins, and metabolites is not sufficient to understand the cell's complexity (*1*). As demonstrated by two papers in this issue—Lee *et al*. (*2*) on page 799 and Milo *et al*. (*3*) on page 824—viewing the cell as a network of genes and proteins offers a viable strategy for addressing the complexity of living systems.

According to the



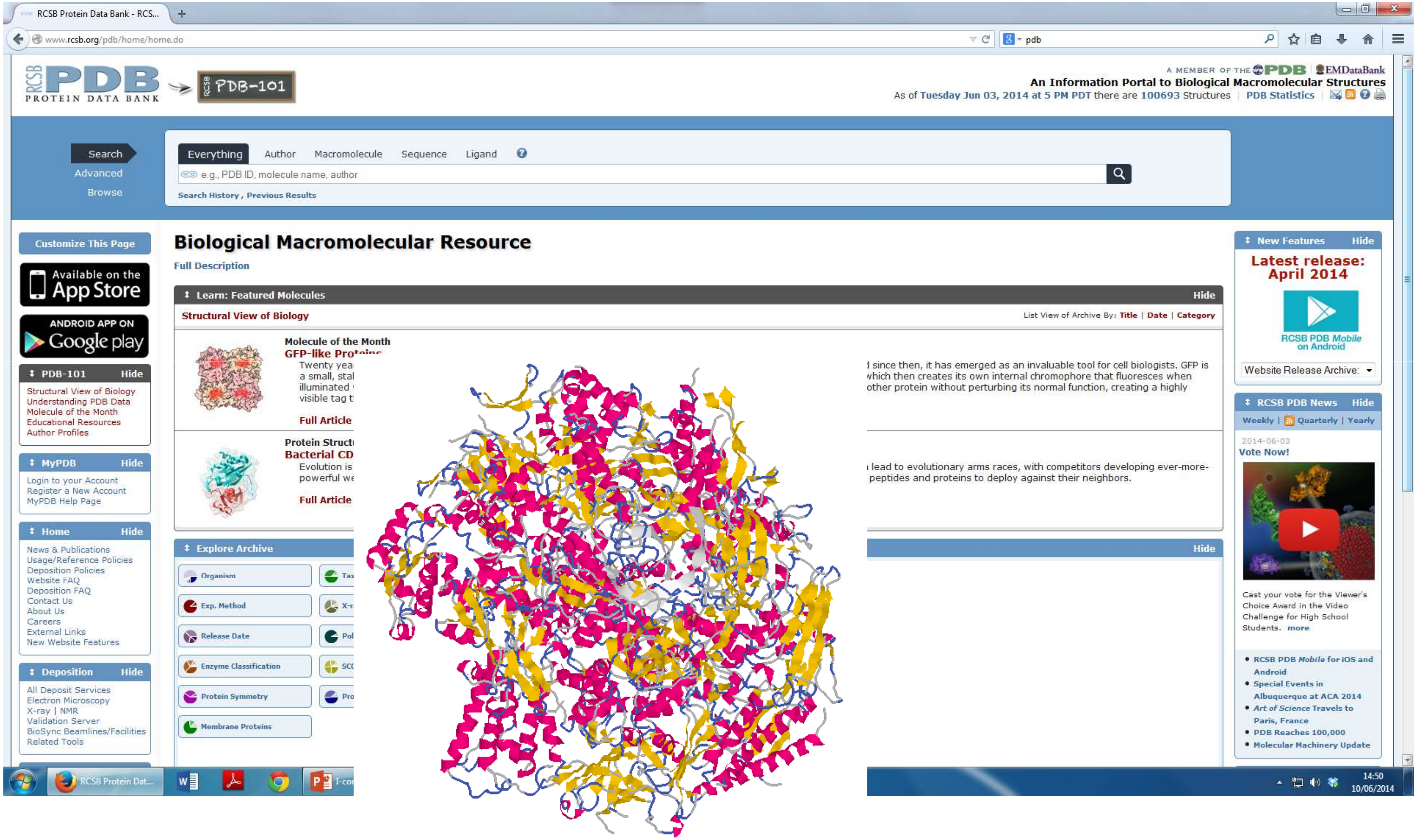
within large networks (*6, 7*).
evidence for the existence o
networks: For example, the p
nizes itself into a protein in
work and metabolites are i
through an intricate metaboli
finding that the structures
works are governed by the s
comes as a surprise, howev
new perspective on cellular o

A simple complexity p
posed of the various mole
nents of the cell—genes, R
and metabolites—summar
paradigm (see the figure). T
tary building blocks organi
into small recurrent pattern
ways in metabolism and
netic-regulatory netw
motifs and pathways a
integrated to form fu

From the particula
sal. The bottom
shows the traditi
tion of the cell
ganization:
scriptome,
metabol
There is
tegrat
ous
the
tl

of cellular c

An Information Portal to Biological Macromolecular Structures



RCSB Protein Data Bank - RCS...

www.rcsb.org/pdb/home/home.do

pdb

A MEMBER OF THE PDB | EMDataBank

An Information Portal to Biological Macromolecular Structures

As of Tuesday Jun 03, 2014 at 5 PM PDT there are 100693 Structures | PDB Statistics

PDB-101
Search
Advanced
Browse
Everything
Author
Macromolecule
Sequence
Ligand
e.g., PDB ID, molecule name, author
Search History, Previous Results
Customize This Page
Available on the App Store
ANDROID APP ON Google play

PDB-101 Hide

Structural View of Biology
Understanding PDB Data
Molecule of the Month
Educational Resources
Author Profiles

MyPDB Hide

Login to your Account
Register a New Account
MyPDB Help Page

Home Hide

News & Publications
Usage/Reference Policies
Deposition Policies
Website FAQ
Deposition FAQ
Contact Us
About Us
Careers
External Links
New Website Features

Deposition Hide

All Deposit Services
Electron Microscopy
X-ray | NMR
Validation Server
BioSync Beamlines/Facilities
Related Tools

Biological Macromolecular Resource

Full Description

Learn: Featured Molecules Hide

Structural View of Biology

List View of Archive By: Title | Date | Category

Molecule of the Month

GFP-like Proteins

Twenty yea … a small, stal … illuminated … visible tag t … l since then, it has emerged as an invaluable tool for cell biologists. GFP is … which then creates its own internal chromophore that fluoresces when … other protein without perturbing its normal function, creating a highly

Full Article

Protein Struct…

Bacterial CD…

Evolution is … powerful we … lead to evolutionary arms races, with competitors developing ever-more- … peptides and proteins to deploy against their neighbors.

Full Article

Explore Archive Hide

New Features Hide

Latest release: April 2014

RCSB PDB Mobile on Android

Website Release Archive:

RCSB PDB News Hide

Weekly | Quarterly | Yearly

2014-06-03

Vote Now!

Cast your vote for the Viewer's Choice Award in the Video Challenge for High School Students. more

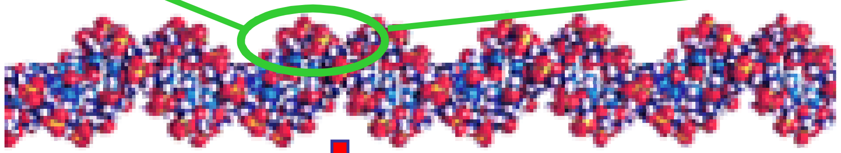
- RCSB PDB *Mobile* for iOS and Android
- Special Events in Albuquerque at ACA 2014
- *Art of Science* Travels to Paris, France
- PDB Reaches 100,000
- Molecular Machinery Update

E.G: RNA Polymerase II Elongation Complex

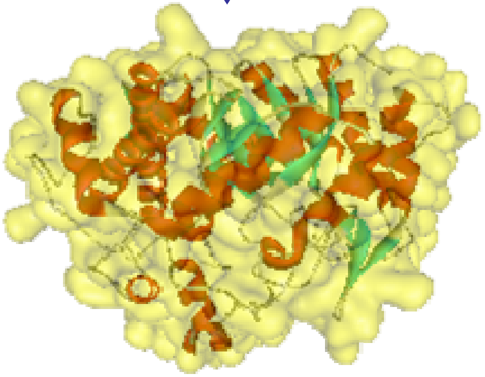
The basic information flow: from DNA to proteins

A,T,C,G

cctgttgatggcgacagggactgtatgctgatctatgctgatgcatgcatgctgactactgatgtgggggctat



```
>BGAL_SULSO BETA-GALACTOSIDASE Sulfolobus solfataricus
MYSFPNSFRFGWSQAGFQSEMGTPGSEDPNTDWYKWVHDPENMAAGLVSG
DLPENGPGYWGNYKTFHDNAQKMGLKIARLNVEWSRIFPNPLPRPQNFDE
SKQDVTEVEINENELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH
WPLPLWLHDPIRVRRGDFTGPSGWLSTRTVYEFARFSAYIAWKFDDLVDE
YSTMNEPNVVGGLGYVGVKSGFPPGYLSFELSRRHMYNIIQAHARAYDGI
```



From genes…

A,C,D,E,F,G,H,I,K,L
M,N,P,Q,R,S,T,V,Y,W

…to Proteins

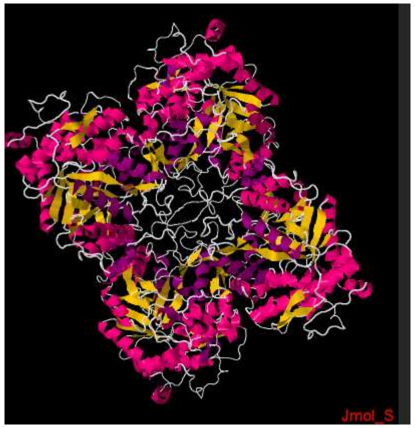
The Data Bases of Biological Sequences and Structures

```
>ENA|M34696|M34696.1 S.solfataricus beta-D-galactosidase (lacS)
gene, complete cds. : Location:1..1000
AAGGAGAAACTTGGCAGTTTATAACTTGACAGTAGGTTGTGGAGTGATGACTGGATCAAT
ACTAGGAGGAGTAGCATATAATTACGTTACACAATTTTATAACCCAATATATTCAATAGA
CCTTATGCTTATCCTATCCTCTATTCTAAGATTCTCGGTATCTCCCCTATTCTTGACCAT
AAAAGATACTCGCTCAAAGCTTAAATAATATTAATCATAAATAAAGTCATGTACTCATTT
CCAAATAGCTTTAGGTTTGGTTGGTCCCAGGCCGGATTTCAATCAGAAATGGGAACACCA
GGGTCAGAAGATCCAAATACTGACTGGTATAAATGGGTTCATGATCCAGAAAACATGGCA
GCGGGATTAGTAAGTGGAGATCTACCAGAAAATGGGCCAGGCTACTGGGGAAACTATAAG
ACATTTCACGATAATGCACAAAAAATGGGATTAAAAATAGCTAGACTAAATGTGGAATGG
TCTAGGATATTTCCTAATCCATTACCAAGGCCACAAAACTTTGATGAATCAAAACAAGAT
GTGACAGAGGTTGAGATAAACGAAAACGAGTTAAAGAGACTTGACGAGTACGCTAATAAA
GACGCATTAAACCATTACAGGGAAATATTCAAGGATCTTAAAAGTAGAGGACTTTACTTT
ATACTAAACATGTATCATTGGCCATTACCTCTATGGTTACACGACCCAATAAGAGTAAGA
AGAGGAGATTTTACTGGACCAAGTGGTTGGCTAAGTACTAGAACAGTTTACGAATTCGCT
AGATTCTCAGCTTATATAGCTTGGAAATTCGATGATCTAGTGGATGAGTACTCAACAATG
AATGAACCTAACGTTGTTGGAGGTTTAGGATACGTTGGTGTTAAGTCCGGTTTTCCCCCA
GGATACCTAAGCTTTGAACTTTCCCGTAGGCATATGTATAACATCATTCAAGCTCACGCA
AGAGCGTATGATGGGATAAAGAGTGTTTCTAAAAAACCAG
```

GenBank

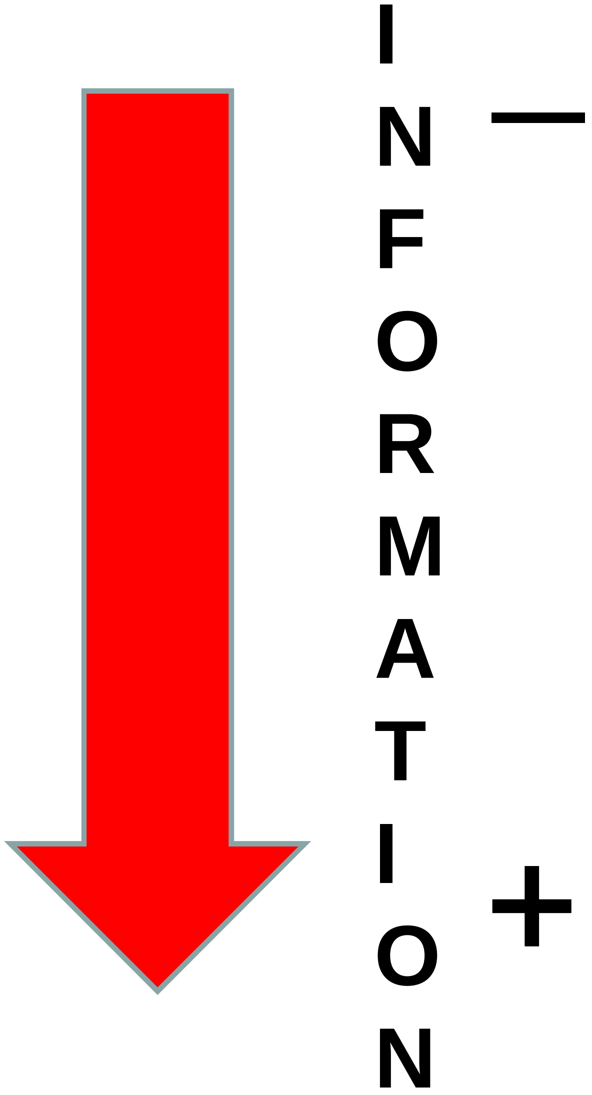
```
>BGAL_SULSO BETA-GALACTOSIDASE Sulfolobus solfataricus.
MYSFPNSFRFGWSQAGFQSEMGTPGSEDPNTDWYKWVHDPENMAAGLVSG
DLPENGPGYWGNYKTFHDNAQKMGLKIARLNVEWSRIFPNPLPRPQNFDE
SKQDVTEVEINENELKRLDEYANKDALNHYREIFKDLKSRGLYFILNMYH
WPLPLWLHDPIRVRRGDFTGPSGWLSTRTVYEFARFSAYIAWKFDDLVDE
YSTMNEPNVVGGLGYVGVKSGFPPGYLSFELSRRHMYNIIQAHARAYDGI
KSVSKKPVGIIYANSSFQPLTDKDMEAVEMAENDNRWWFFDAIIRGEITR
GNEKIVRDDLKGRLDWIGVNYYTRTVVKRTEKGYVSLGGYGHGCERNSVS
LAGLPTSDFGWEFFPEGLYDVLTKYWNRYHLYMYVTENGIADDADYQRPY
YLVSHVYQVHRAINSGADVRGYLHWSLADNYEWASGFSMRFGLLKVDYNT
KRLYWRPSALVYREIATNGAITDEIEHLNSVPPVKPLRH
```

UniProt/SwissProt

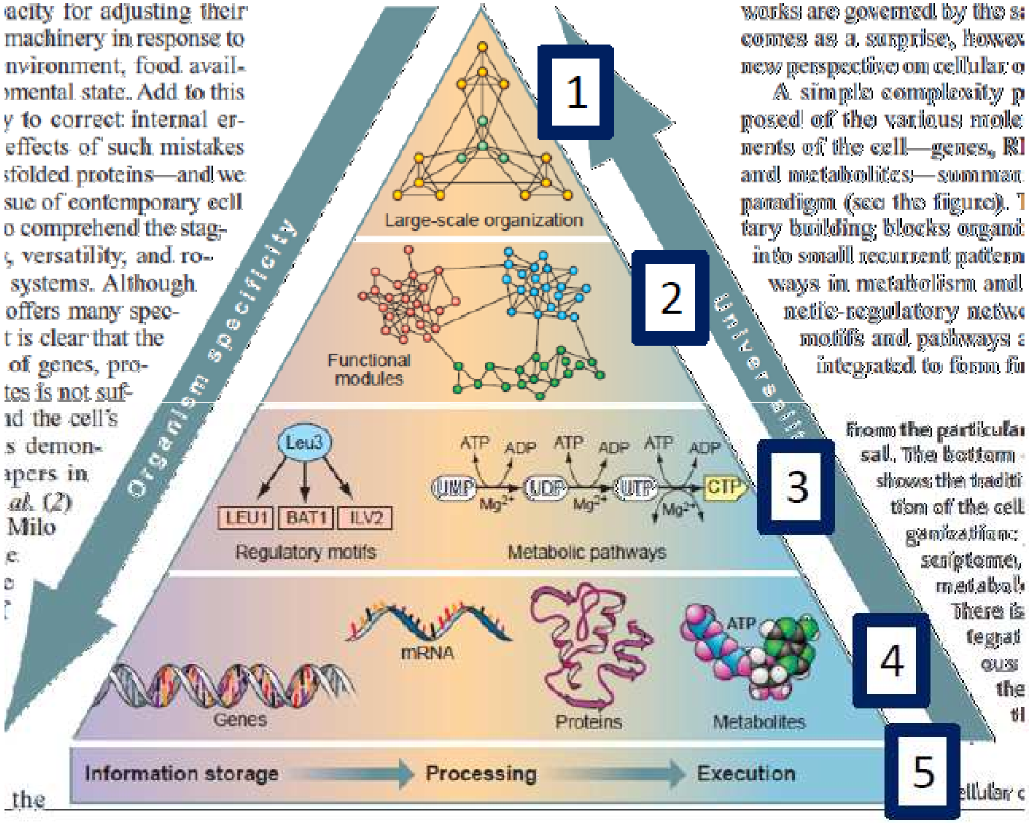


1GOV

PDB



Hierchical levels of cell complexity and our knoweldge



acity for adjusting their
machinery in response to
nvironment, food avail-
omental state. Add to this
y to correct internal er-
effects of such mistakes
sfolded proteins—and we
sue of contemporary cell
o comprehend the stag-
, versatility, and ro-
systems. Although
offers many spec-
t is clear that the
of genes, pro-
tes is not suf-
nd the cell's
s demon-
apers in
al. (2)
Milo
e
e
r
the

works are governed by the sa
comes as a surprise, howev
new perspective on cellular o
A simple complexity p
posed of the various mole
nents of the cell—genes, RI
and metabolites—summar
paradigm (see the figure). T
tary building blocks organi
into small recurrent pattern
ways in metabolism and
netic-regulatory netwo
motifs and pathways a
integrated to form fu

From the particular
sal. The bottom
shows the traditi
tion of the cell
ganization:
scriptome,
metabol
There is
tegrat
ous
the
tl
ellular c

1) Large-scale organisation

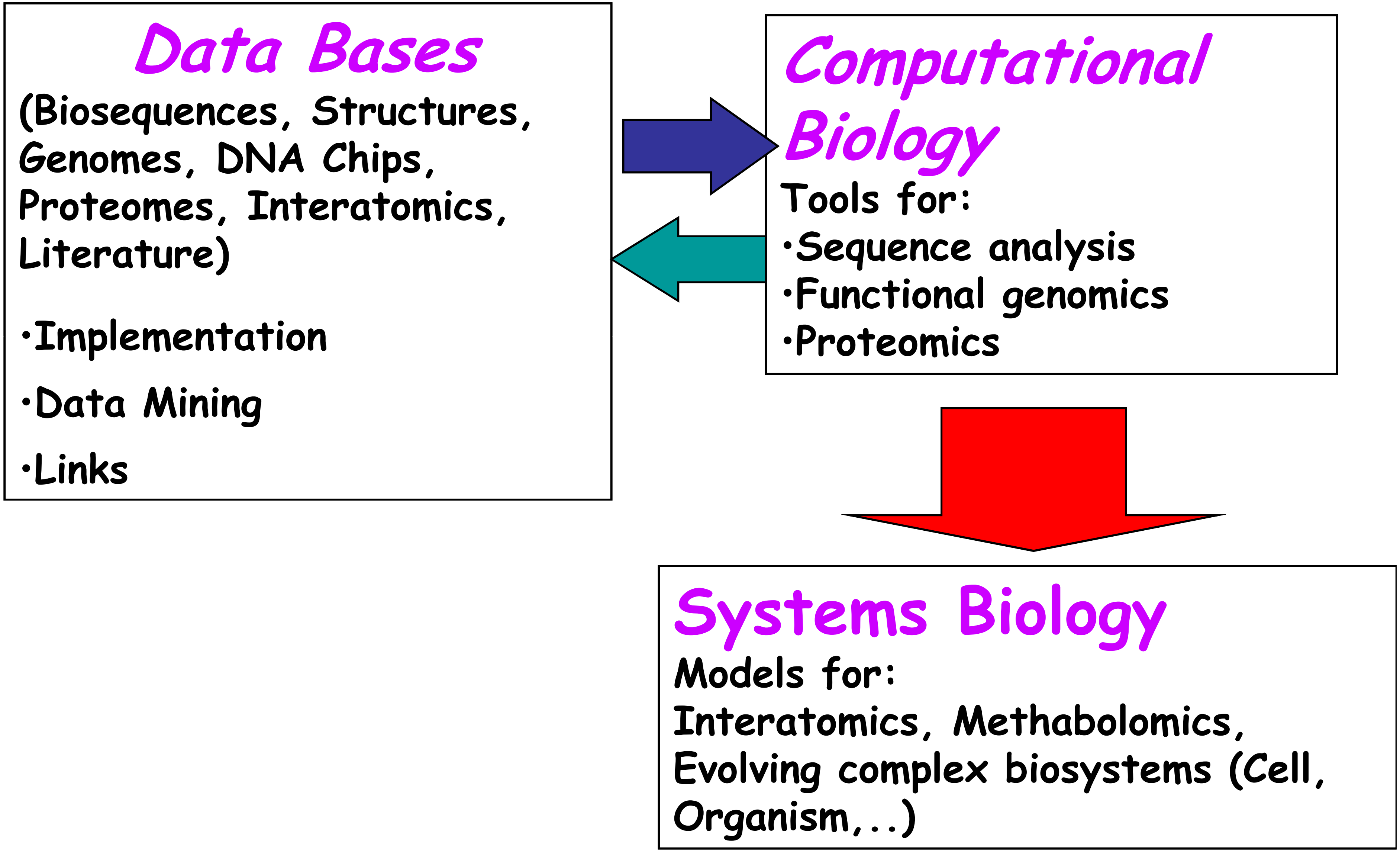
2) Functional modules

3) Regulatory motifs, metabolic pathways

4) Molecules: genes, mRNAs, proteins, metabolites

5) Overall: Information storage, Processing, Execution

BIOINFORMATICS



Data Bases

(Biosequences, Structures, Genomes, DNA Chips, Proteomes, Interatomics, Literature)

- Implementation
- Data Mining
- Links

Computational Biology

Tools for:

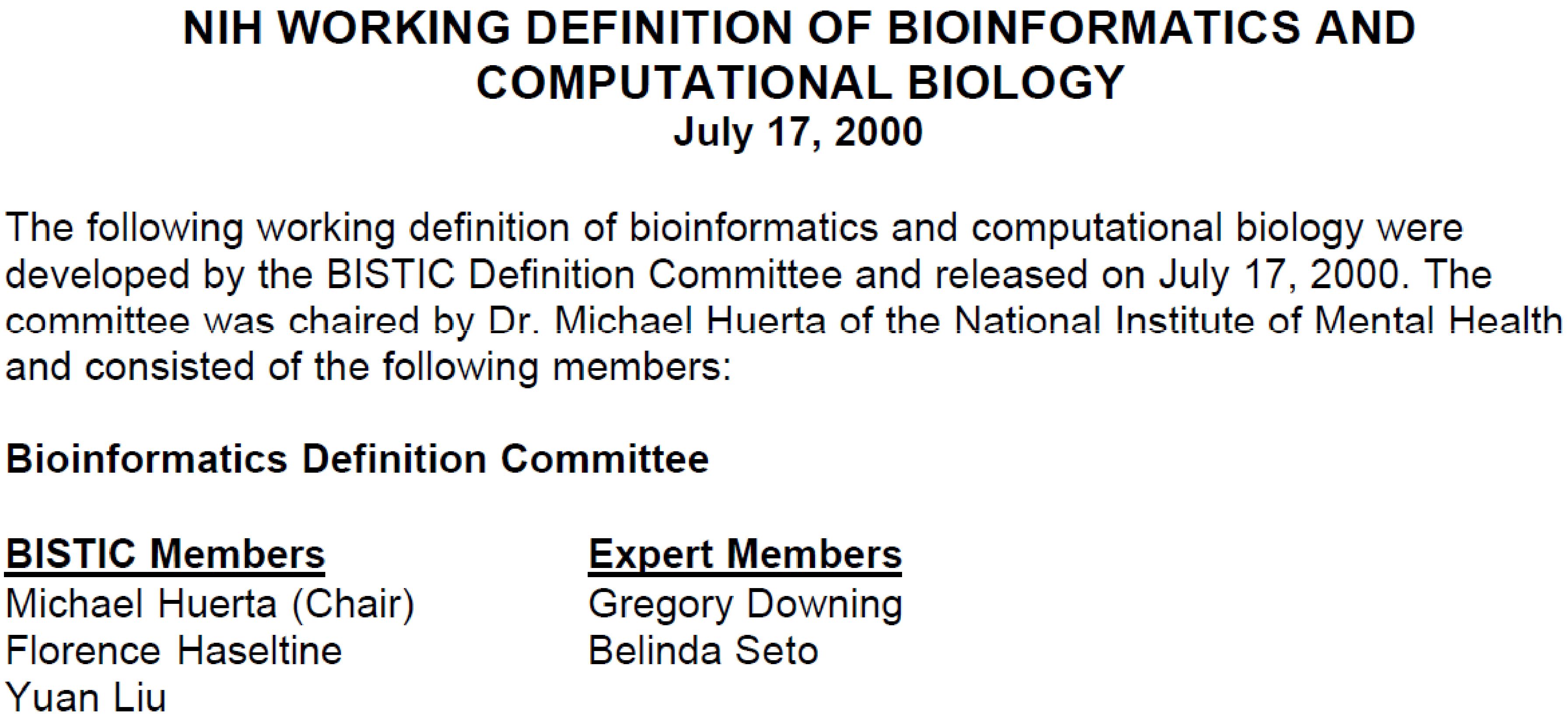
- Sequence analysis
- Functional genomics
- Proteomics

Systems Biology

Models for:

Interatomics, Methabolomics, Evolving complex biosystems (Cell, Organism,..)

Going back to definitions……who said what and when….



NIH WORKING DEFINITION OF BIOINFORMATICS AND COMPUTATIONAL BIOLOGY
July 17, 2000

The following working definition of bioinformatics and computational biology were developed by the BISTIC Definition Committee and released on July 17, 2000. The committee was chaired by Dr. Michael Huerta of the National Institute of Mental Health and consisted of the following members:

Bioinformatics Definition Committee

BISTIC Members	**Expert Members**
Michael Huerta (Chair)	Gregory Downing
Florence Haseltine	Belinda Seto
Yuan Liu	

BISTIC: Biomedical Information Science and Technology Initiative Consortium

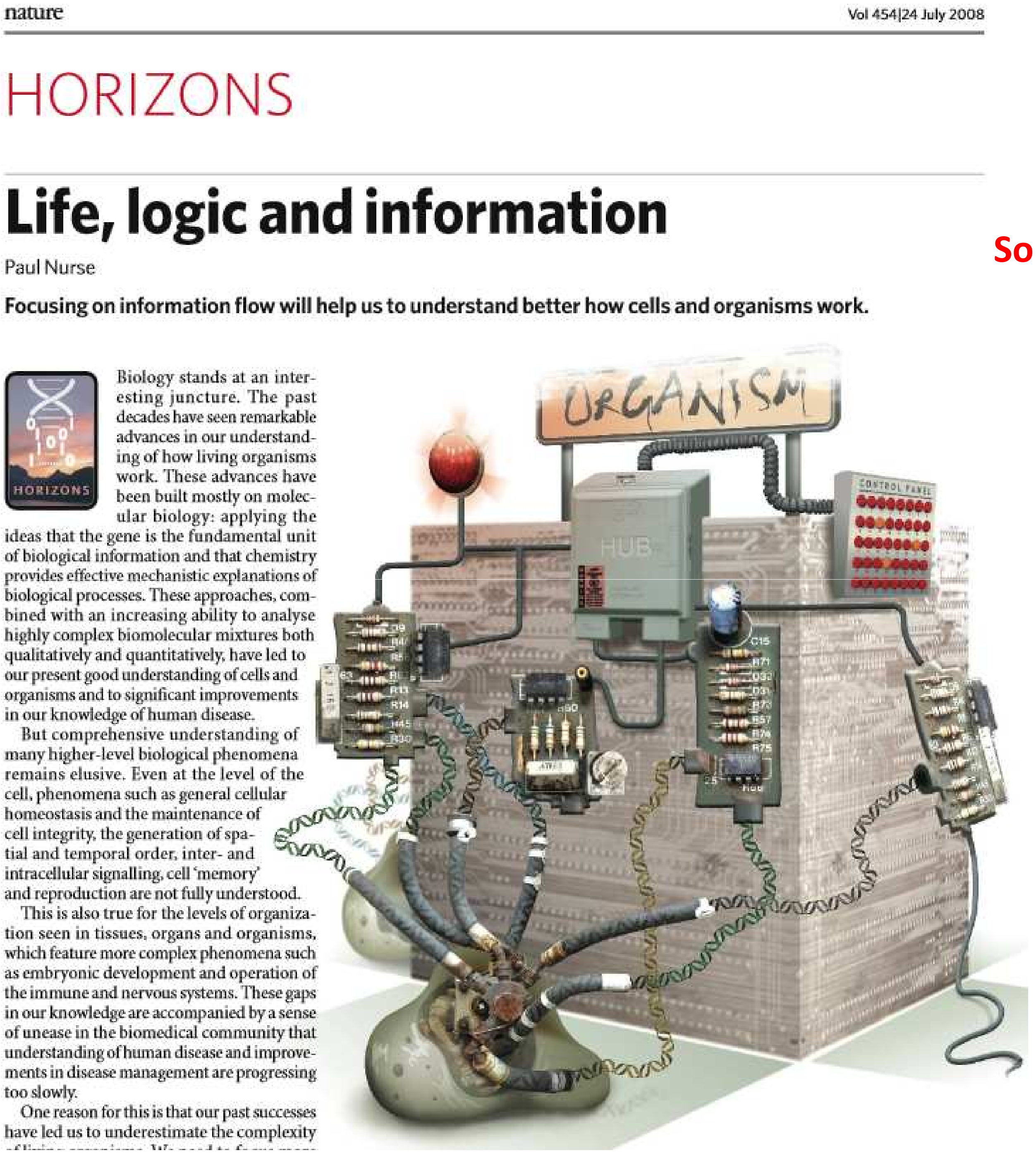
Definition

The NIH Biomedical Information Science and Technology Initiative Consortium agreed on the following definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

Some references



nature

Vol 454|24 July 2008

HORIZONS

Life, logic and information

Paul Nurse

Focusing on information flow will help us to understand better how cells and organisms work.

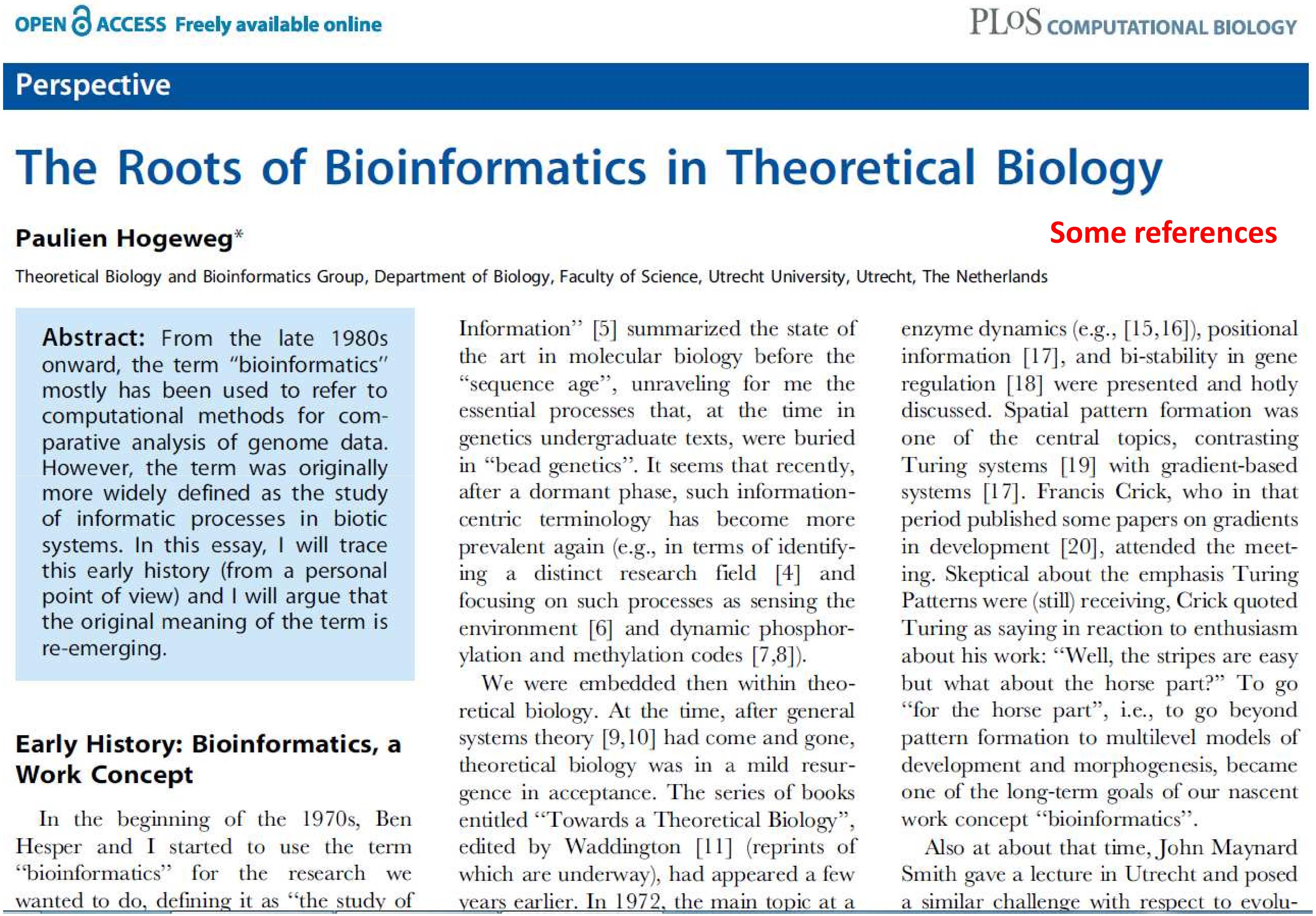
HORIZONS
ORGANISM
HUB
CONTROL PANEL

Biology stands at an interesting juncture. The past decades have seen remarkable advances in our understanding of how living organisms work. These advances have been built mostly on molecular biology: applying the ideas that the gene is the fundamental unit of biological information and that chemistry provides effective mechanistic explanations of biological processes. These approaches, combined with an increasing ability to analyse highly complex biomolecular mixtures both qualitatively and quantitatively, have led to our present good understanding of cells and organisms and to significant improvements in our knowledge of human disease.

But comprehensive understanding of many higher-level biological phenomena remains elusive. Even at the level of the cell, phenomena such as general cellular homeostasis and the maintenance of cell integrity, the generation of spatial and temporal order, inter- and intracellular signalling, cell 'memory' and reproduction are not fully understood.

This is also true for the levels of organization seen in tissues, organs and organisms, which feature more complex phenomena such as embryonic development and operation of the immune and nervous systems. These gaps in our knowledge are accompanied by a sense of unease in the biomedical community that understanding of human disease and improvements in disease management are progressing too slowly.

One reason for this is that our past successes have led us to underestimate the complexity of living organisms. We need to focus more-



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Perspective

The Roots of Bioinformatics in Theoretical Biology

Some references

Paulien Hogeweg*

Theoretical Biology and Bioinformatics Group, Department of Biology, Faculty of Science, Utrecht University, Utrecht, The Netherlands

Abstract: From the late 1980s onward, the term "bioinformatics" mostly has been used to refer to computational methods for comparative analysis of genome data. However, the term was originally more widely defined as the study of informatic processes in biotic systems. In this essay, I will trace this early history (from a personal point of view) and I will argue that the original meaning of the term is re-emerging.

Early History: Bioinformatics, a Work Concept

In the beginning of the 1970s, Ben Hesper and I started to use the term "bioinformatics" for the research we wanted to do, defining it as "the study of

Information" [5] summarized the state of the art in molecular biology before the "sequence age", unraveling for me the essential processes that, at the time in genetics undergraduate texts, were buried in "bead genetics". It seems that recently, after a dormant phase, such information-centric terminology has become more prevalent again (e.g., in terms of identifying a distinct research field [4] and focusing on such processes as sensing the environment [6] and dynamic phosphorylation and methylation codes [7,8]).

We were embedded then within theoretical biology. At the time, after general systems theory [9,10] had come and gone, theoretical biology was in a mild resurgence in acceptance. The series of books entitled "Towards a Theoretical Biology", edited by Waddington [11] (reprints of which are underway), had appeared a few years earlier. In 1972, the main topic at a

enzyme dynamics (e.g., [15,16]), positional information [17], and bi-stability in gene regulation [18] were presented and hotly discussed. Spatial pattern formation was one of the central topics, contrasting Turing systems [19] with gradient-based systems [17]. Francis Crick, who in that period published some papers on gradients in development [20], attended the meeting. Skeptical about the emphasis Turing Patterns were (still) receiving, Crick quoted Turing as saying in reaction to enthusiasm about his work: "Well, the stripes are easy but what about the horse part?" To go "for the horse part", i.e., to go beyond pattern formation to multilevel models of development and morphogenesis, became one of the long-term goals of our nascent work concept "bioinformatics".

Also at about that time, John Maynard Smith gave a lecture in Utrecht and posed a similar challenge with respect to evolu-

Published: June 24, 2010

The “omic” revolution

The analysis of the components of a living organism in its entirety

⇩

Biology becomes a data driven science

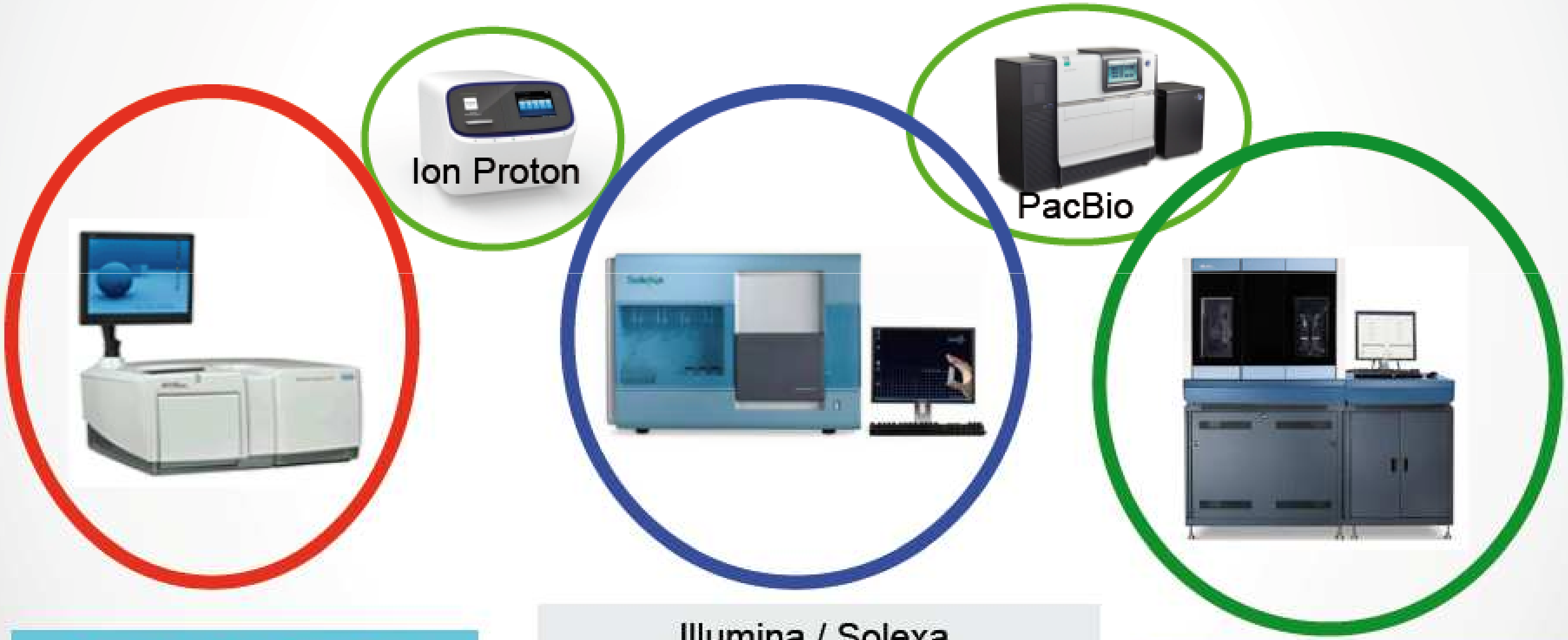
GA EVANS, Nature Biotechnology 18:127, 2000

NGS technology allows an unprecedent rate of DNA/RNA sequencing (>4TB per week)

Next-Generation Sequencing

A large number of platforms using different strategies and chemistries, and with a different throughput are entering the market.

Results:> 3000 fully sequenced genomes; >1000 human genomes; >10,000 human exoms



Roche / 454 Genome
Sequencer FLX titanium
(800 bp, 800 Mb / run)

Illumina / Solexa
Genetic Analyzer HiSeq 2000
(150x2 bp, 600 Gb / run)

Applied Biosystems
SOLiD 4 System^{TM}
(100x2 bp, 400 Gb / run)

Dealing with genomic data…makes modern biology a BIG science

Scott D. Kahn
Science 331, 728 (2011)

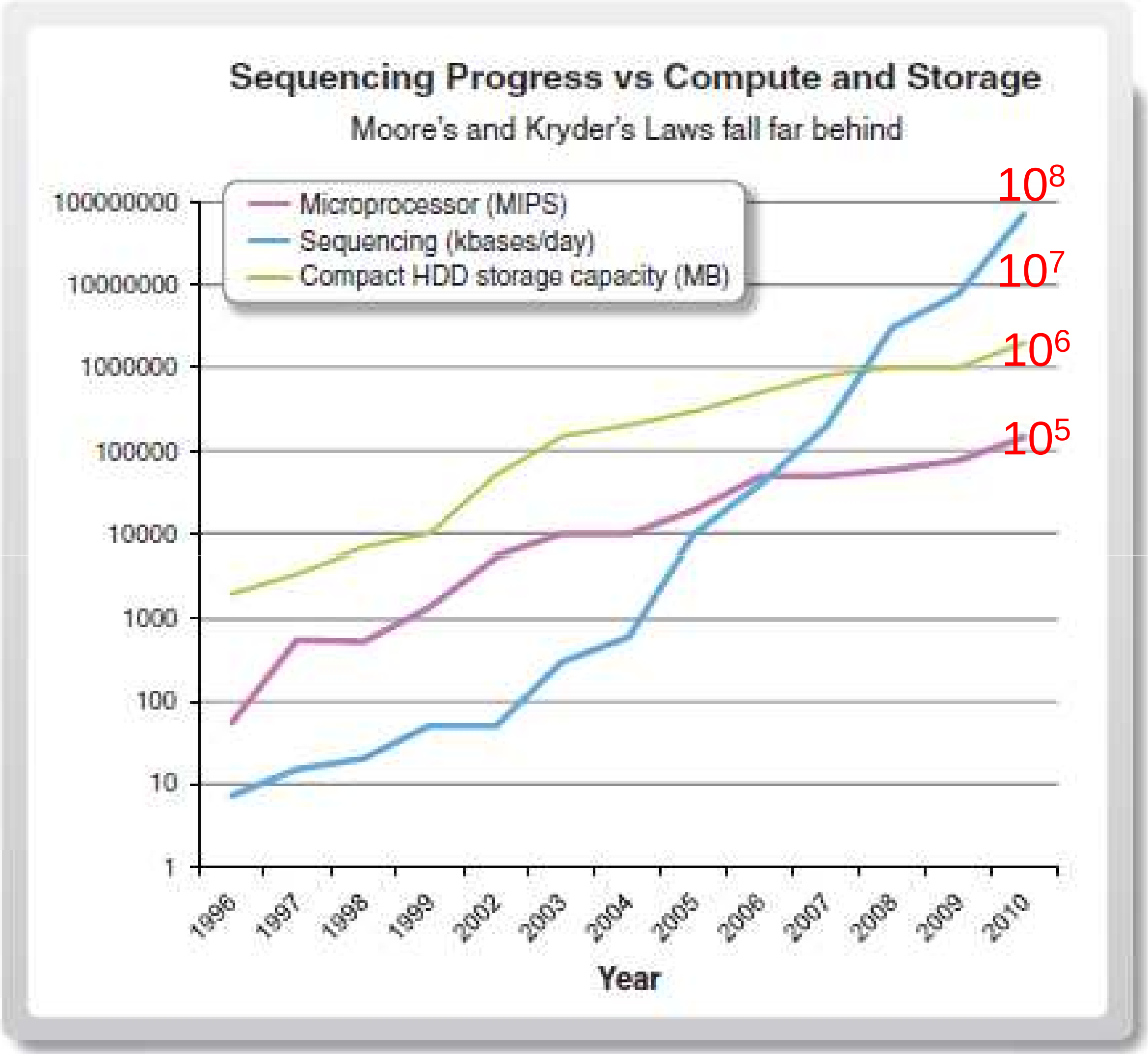
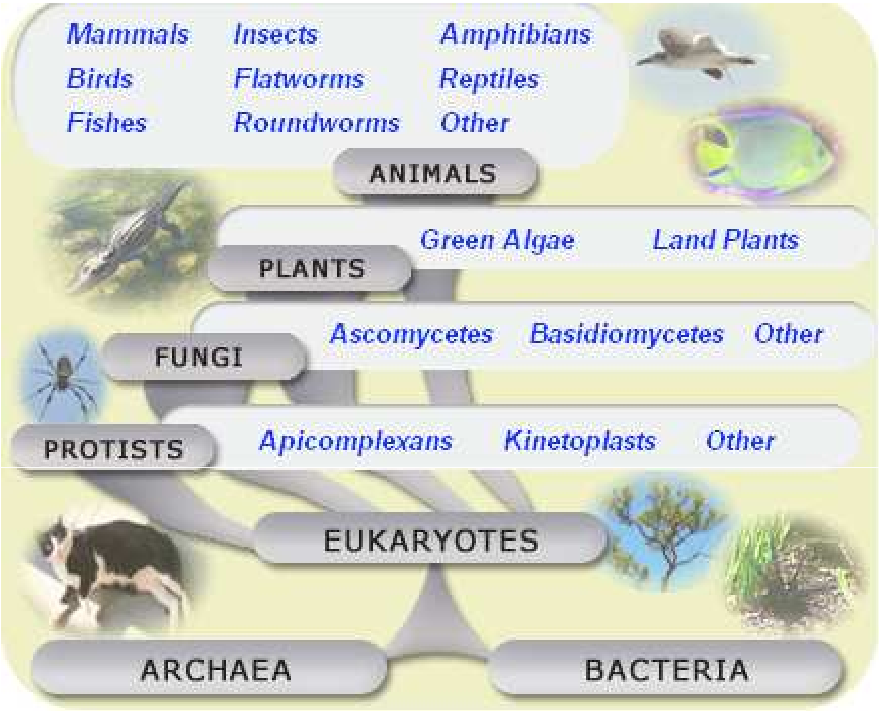


Fig. 1. A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields.

The World's Most Powerful Supercomputer Is in China: the Tianhe-2. It's a system developed by China's National University of Defense Technology, and it is capable of running at 33.86 petaflops. (A petaflop is a quadrillion calculations per second.)

***million instructions per second (MIPS)**

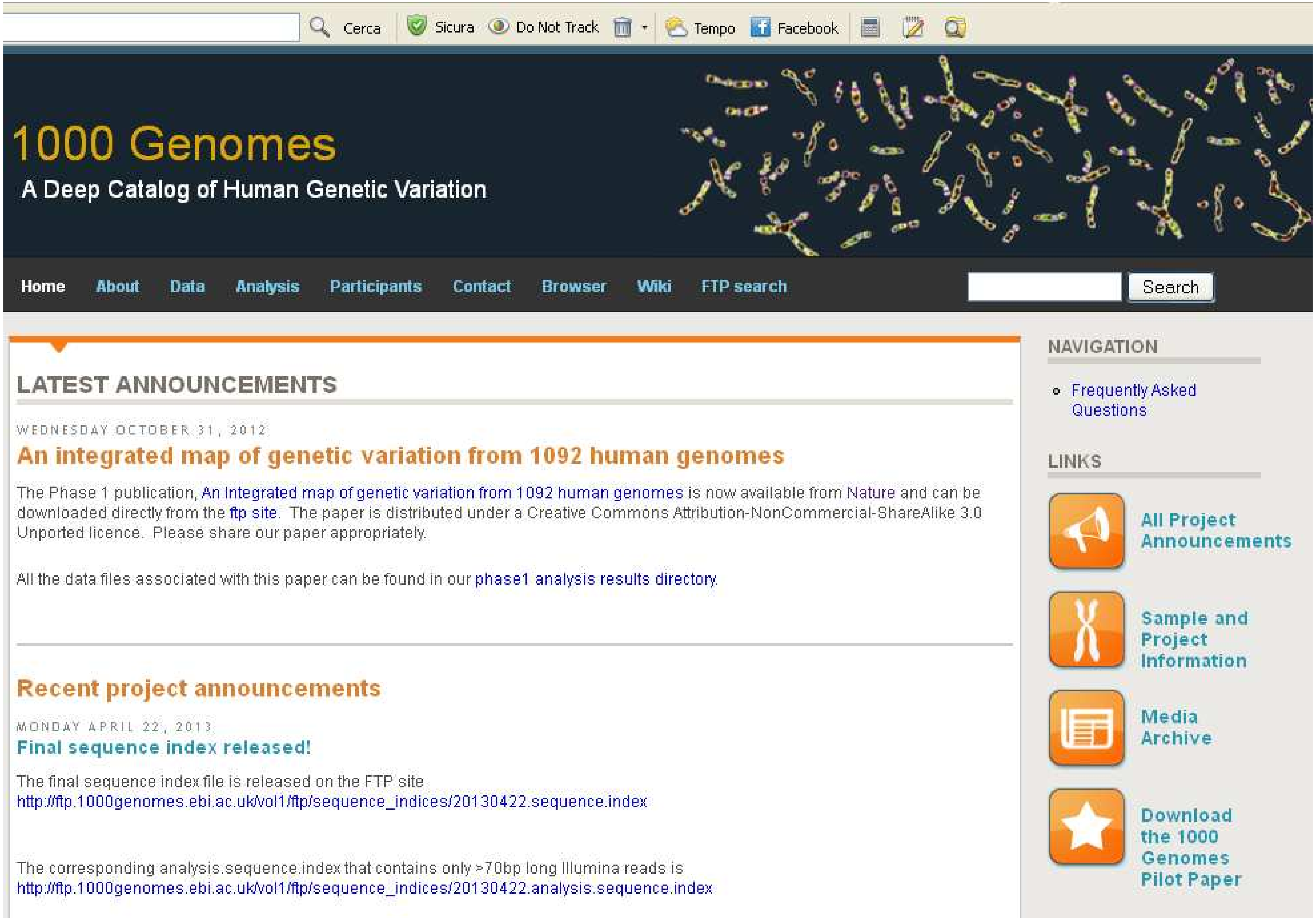
The "omic" era-RESULTS



Complete Genomes

Prokaryotes:	2975
Eukaryotes:	213
Viruses:	4101

http://www.ncbi.nlm.nih.gov/ Update: May 2014



1000 Genomes

A Deep Catalog of Human Genetic Variation

LATEST ANNOUNCEMENTS

WEDNESDAY OCTOBER 31, 2012

An integrated map of genetic variation from 1092 human genomes

The Phase 1 publication, An Integrated map of genetic variation from 1092 human genomes is now available from Nature and can be downloaded directly from the ftp site. The paper is distributed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported licence. Please share our paper appropriately.

All the data files associated with this paper can be found in our phase1 analysis results directory.

Recent project announcements

MONDAY APRIL 22, 2013

Final sequence index released!

The final sequence index file is released on the FTP site
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/sequence_indices/20130422.sequence.index

The corresponding analysis.sequence.index that contains only >70bp long Illumina reads is
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/sequence_indices/20130422.analysis.sequence.index

NAVIGATION

- Frequently Asked Questions

LINKS

All Project Announcements

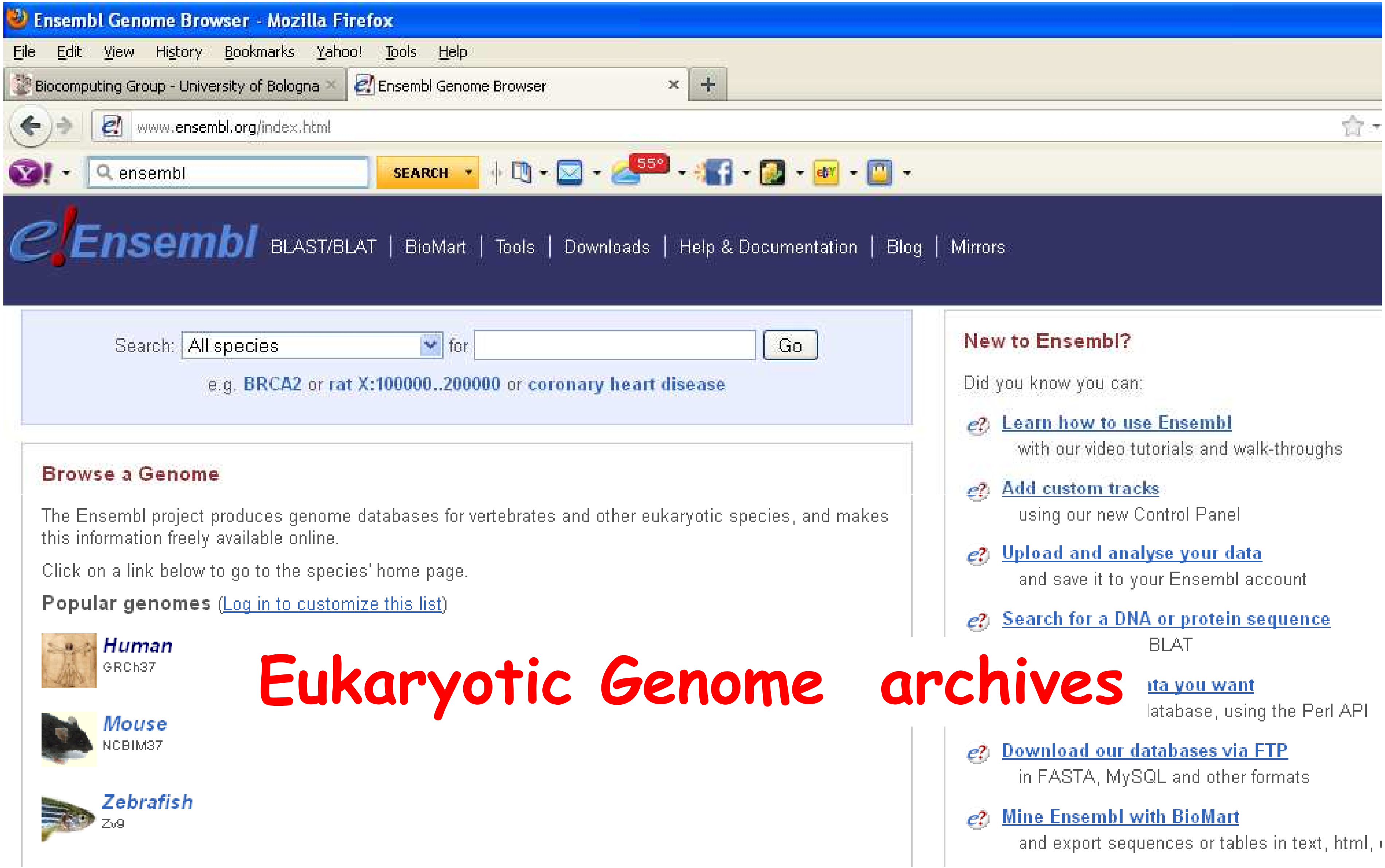
Sample and Project Information

Media Archive

Download the 1000 Genomes Pilot Paper

http://www.1000genomes.org/

Eukaryotic Genome archives



Ensembl Genome Browser - Mozilla Firefox

www.ensembl.org/index.html

ensembl

Search: All species for

e.g. BRCA2 or rat X:100000..200000 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Click on a link below to go to the species' home page.

Popular genomes (Log in to customize this list)

Human GRCh37

Mouse NCBIM37

Zebrafish Zv9

New to Ensembl?

Did you know you can:

Learn how to use Ensembl with our video tutorials and walk-throughs

Add custom tracks using our new Control Panel

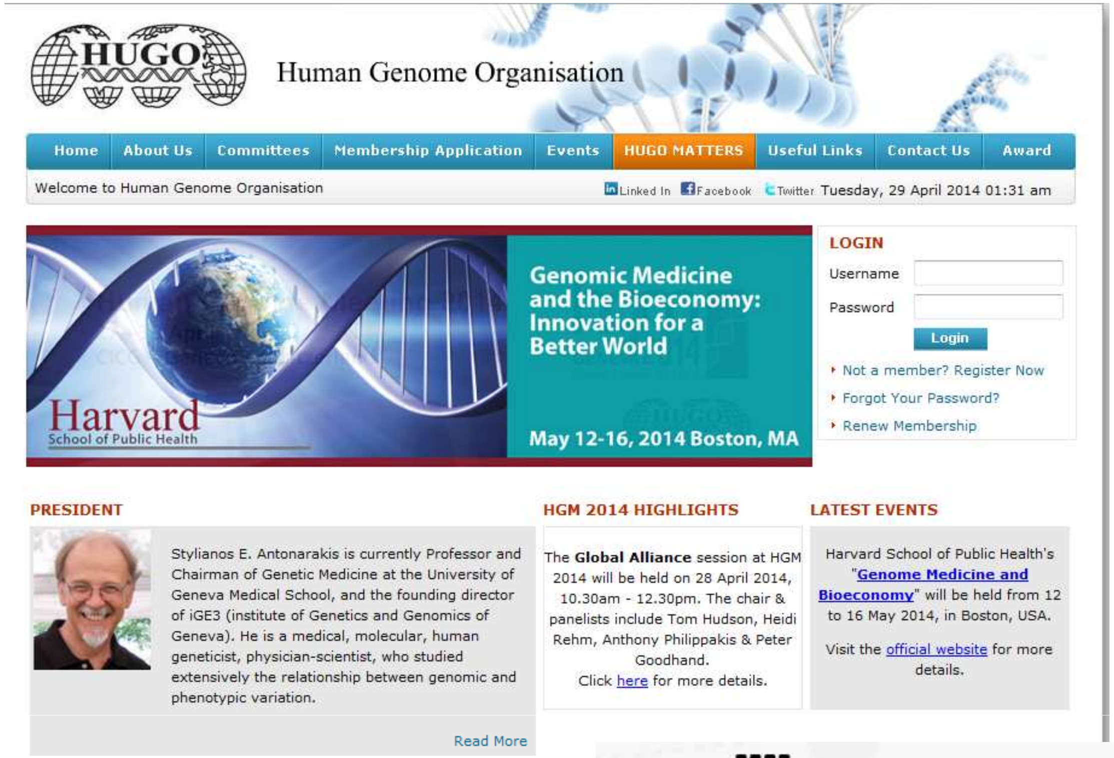
Upload and analyse your data and save it to your Ensembl account

Search for a DNA or protein sequence BLAT

ta you want atabase, using the Perl API

Download our databases via FTP in FASTA, MySQL and other formats

Mine Ensembl with BioMart and export sequences or tables in text, html,



Welcome to Human Genome Organisation

Tuesday, 29 April 2014 01:31 am

Genomic Medicine and the Bioeconomy: Innovation for a Better World

May 12-16, 2014 Boston, MA

PRESIDENT

Stylianos E. Antonarakis is currently Professor and Chairman of Genetic Medicine at the University of Geneva Medical School, and the founding director of iGE3 (institute of Genetics and Genomics of Geneva). He is a medical, molecular, human geneticist, physician-scientist, who studied extensively the relationship between genomic and phenotypic variation.

HGM 2014 HIGHLIGHTS

The **Global Alliance** session at HGM 2014 will be held on 28 April 2014, 10.30am - 12.30pm. The chair & panelists include Tom Hudson, Heidi Rehm, Anthony Philippakis & Peter Goodhand.

Click here for more details.

LATEST EVENTS

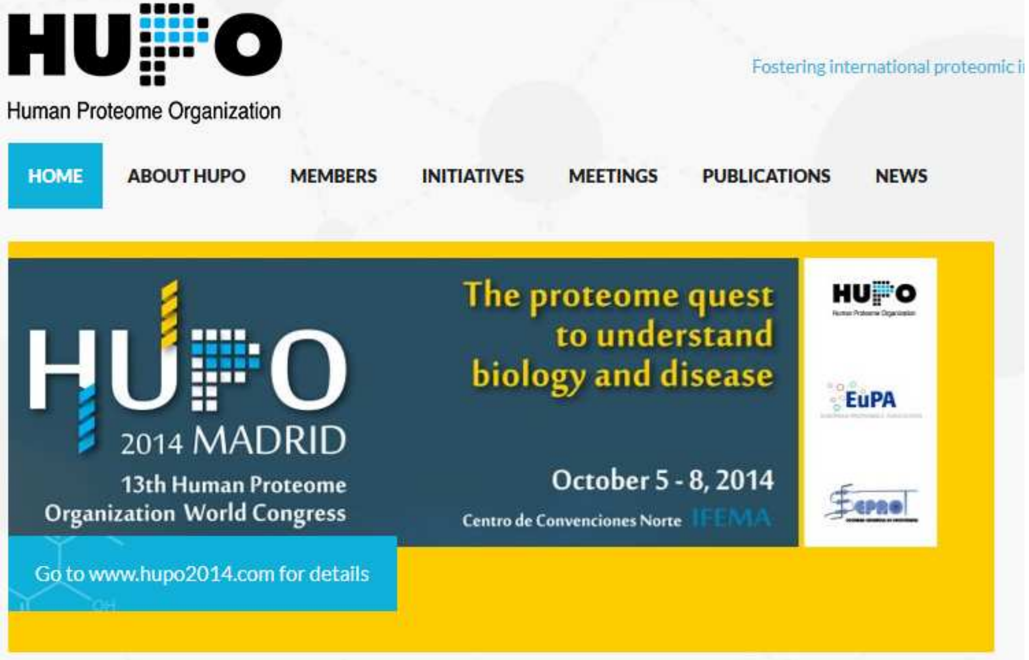
Harvard School of Public Health's "**Genome Medicine and Bioeconomy**" will be held from 12 to 16 May 2014, in Boston, USA.

Visit the official website for more details.

http://www.hugo-international.org

HUGO

Worldwide consortia



13th Human Proteome Organization World Congress

The proteome quest to understand biology and disease

October 5 - 8, 2014

Centro de Convenciones Norte

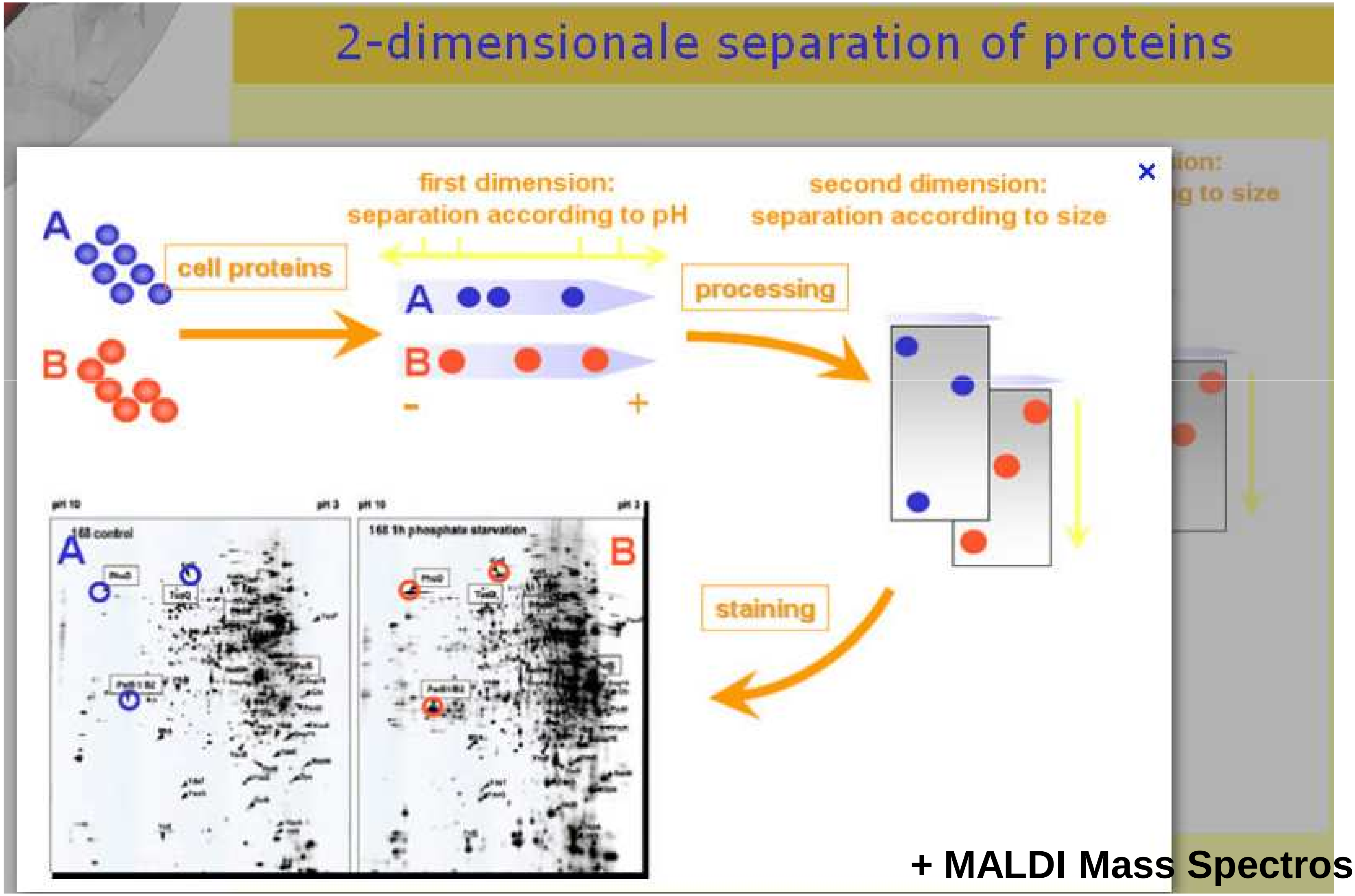
Go to www.hupo2014.com for details

HUPO

http://www.hupo.org/

Proteomics

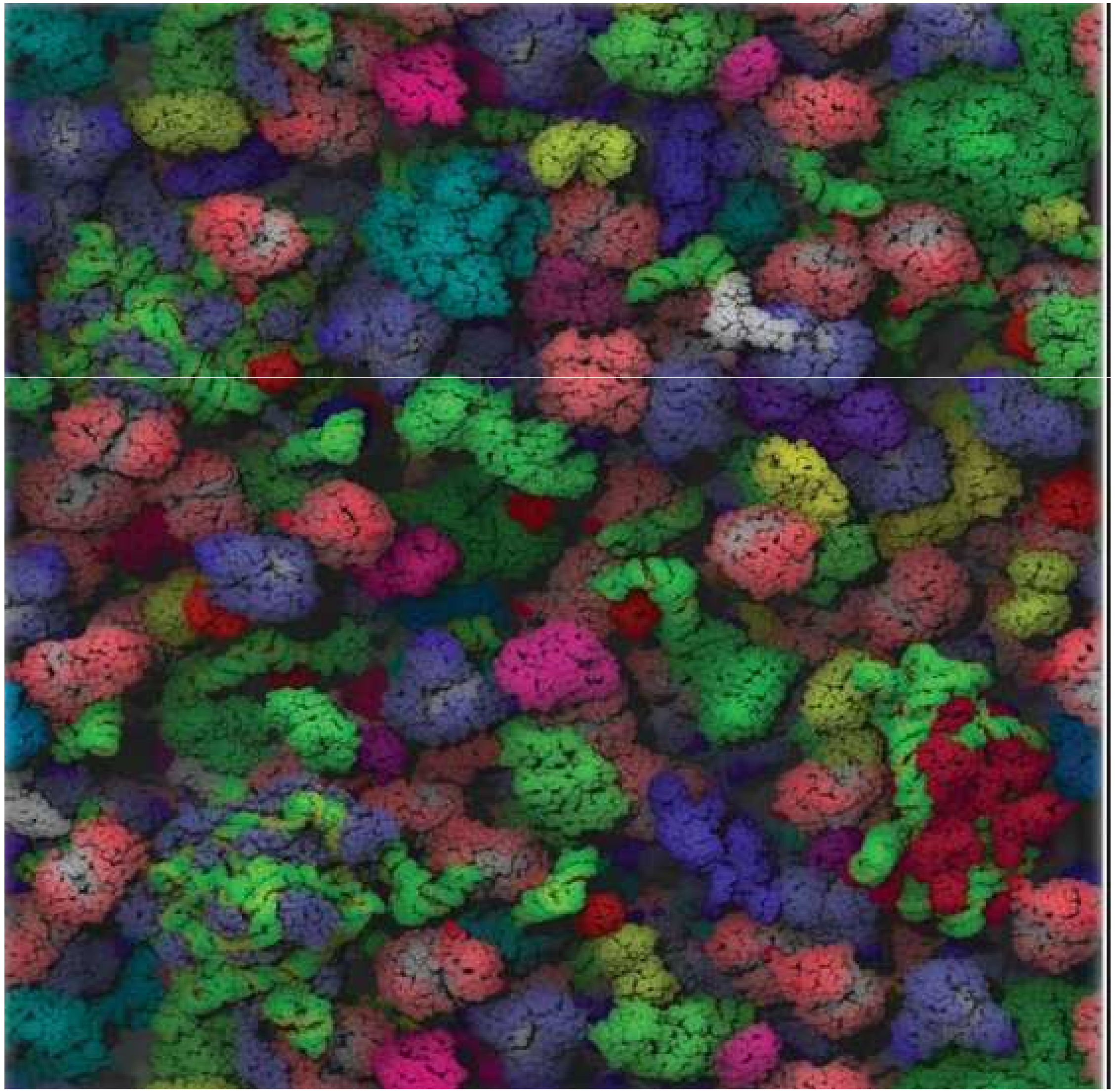
the large-scale study of proteins, particularly their expression level, structures and functions



2-dimensionale separation of proteins

+ MALDI Mass Spectroscopy

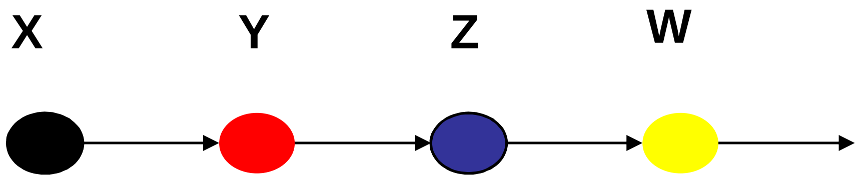
LIFE IS CROWDED: Macromolecular crowding is under-appreciated



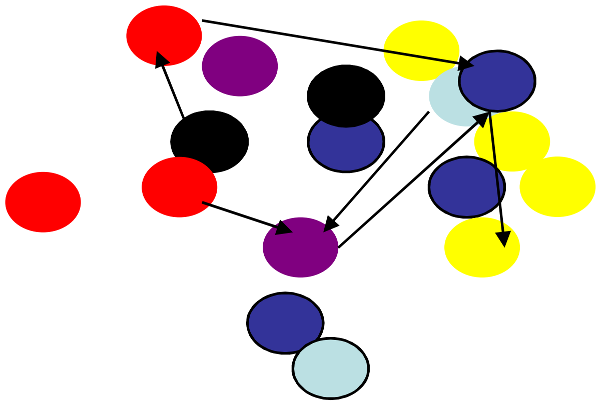
The Crowded Cell: This picture shows an atomically detailed model of the crowded E. coli cytoplasm, including the 50 most abundant macromolecules. RNA is shown as green and yellow. Reprinted from: McGuffee SR, Elcock AH (2010) Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. PLoS Comput Biol 6(3): e1000694.

What did we learn:

A shift of paradigm….to describe protein-protein and protein/DNA/RNA interactions



from linear

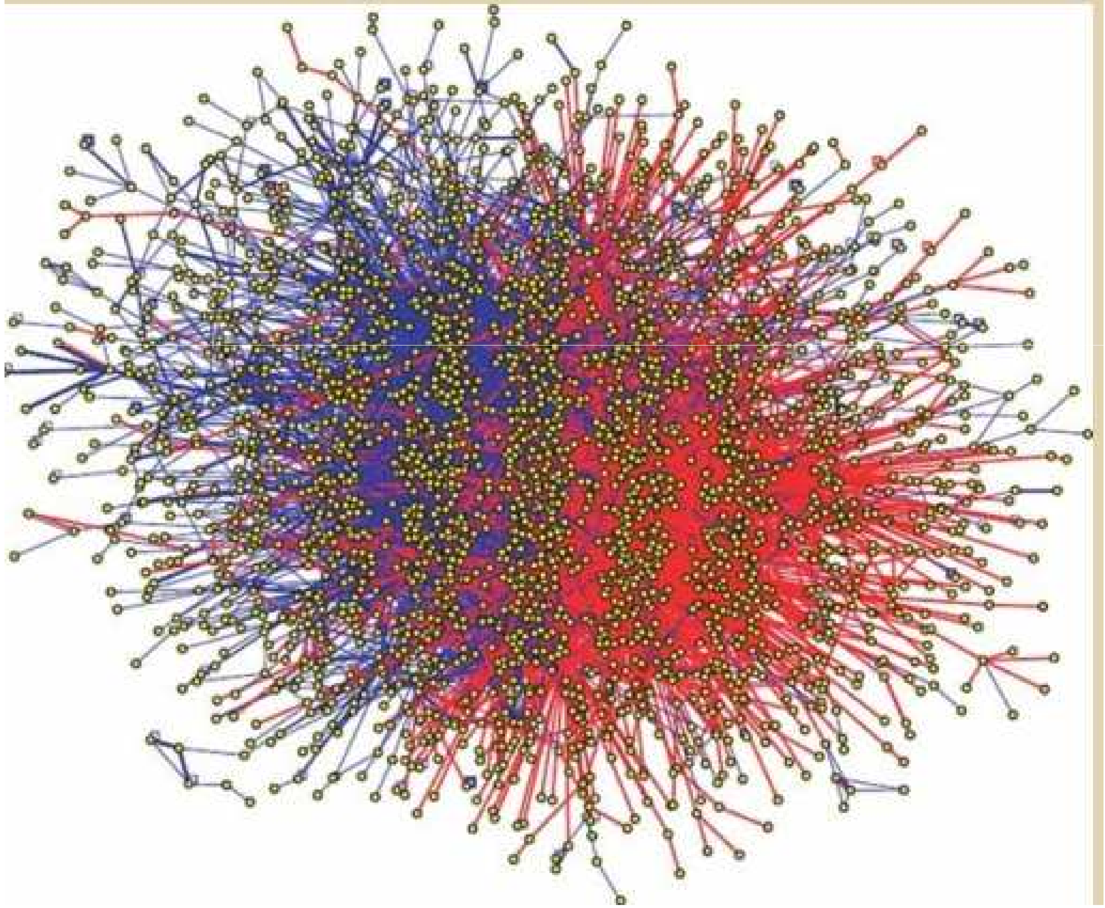


to network models

A protein is a node characterised by a degree of connections (**number of possible interactions or number of other proteins/molecules with which it can interact**)

Interactomics

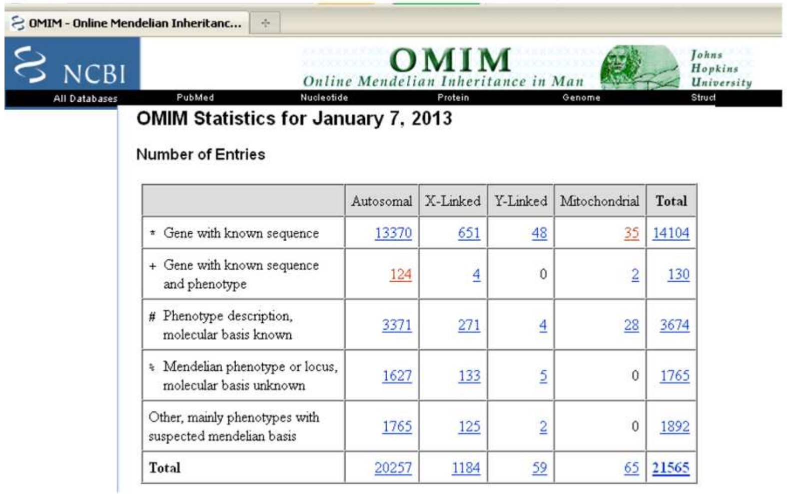
In terms of proteomics, interactomics refers to protein-protein interaction networks (or protein/DNA/RNA interactions)



Vidal, M et al. Nature 2005. 437: 1173–1178,

Search for human genetic variability…..

Where to check variations for disease association



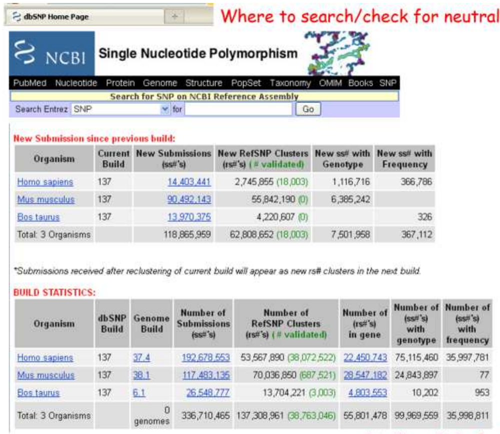
OMIM Statistics for January 7, 2013

Number of Entries

	Autosomal	X-Linked	Y-Linked	Mitochondrial	Total
* Gene with known sequence	13370	651	48	35	14104
+ Gene with known sequence and phenotype	124	4	0	2	130
# Phenotype description, molecular basis known	3371	271	4	28	3674
% Mendelian phenotype or locus, molecular basis unknown	1627	133	5	0	1765
Other, mainly phenotypes with suspected mendelian basis	1765	125	2	0	1892
Total	20257	1184	59	65	21565

http://www.ncbi.nlm.nih.gov/Omim/mimstats.html

Where to search/check for neutral variations



dbSNP Home Page
NCBI
Single Nucleotide Polymorphism
PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP
Search for SNP on NCBI Reference Assembly
Search Entrez
for
Go

Search Entrez: SNP

New Submission since previous build:

Organism	Current Build	New Submissions (ss#'s)	New RefSNP Clusters (rs#'s) (# validated)	New ss# with Genotype	New ss# with Frequency
Homo sapiens	137	14,403,441	2,745,855 (18,003)	1,116,716	366,786
Mus musculus	137	90,492,143	55,842,190 (0)	6,385,242	
Bos taurus	137	13,970,375	4,220,607 (0)		326
Total: 3 Organisms		118,865,959	62,808,652 (18,003)	7,501,958	367,112

*Submissions received after reclustering of current build will appear as new rs# clusters in the next build.

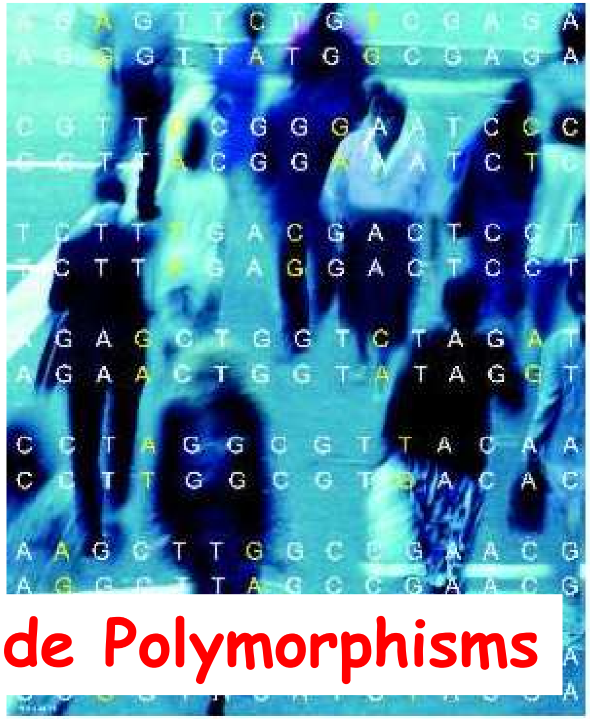
BUILD STATISTICS:

Organism	dbSNP Build	Genome Build	Number of Submissions (ss#'s)	Number of RefSNP Clusters (rs#'s) (# validated)	Number of (rs#'s) in gene	Number of (ss#'s) with genotype	Number of (ss#'s) with frequency
Homo sapiens	137	37.4	192,678,553	53,567,890 (38,072,522)	22,450,743	75,115,460	35,997,781
Mus musculus	137	38.1	117,483,136	70,036,850 (687,521)	28,547,182	24,843,897	77
Bos taurus	137	6.1	26,548,777	13,704,221 (3,003)	4,803,553	10,202	953
Total: 3 Organisms		0 genomes	336,710,465	137,308,961 (38,763,046)	55,801,478	99,969,559	35,998,811

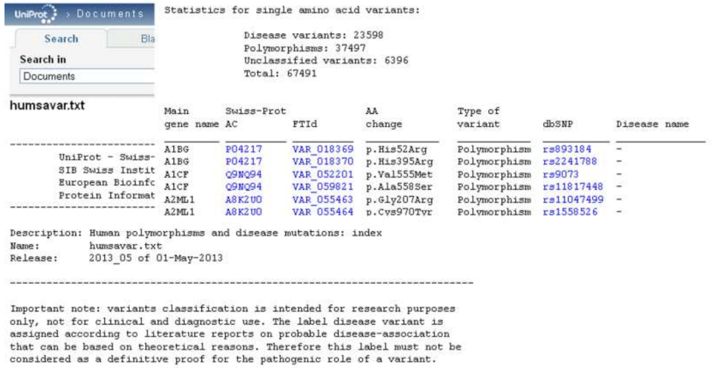
http://www.ncbi.nlm.nih.gov/projects/SNP/

The Human Variome

http://www.ornl.gov/hgmis.



Where to find disease associated variations in proteins



```
Statistics for single amino acid variants:

          Disease variants: 23598
          Polymorphisms: 37497
          Unclassified variants: 6396
          Total: 67491

Main      Swiss-Prot                   AA             Type of
gene name AC         FTId              change         variant        dbSNP       Disease name
_________ __________ ______________    ____________   ____________   _________   ______________
A1BG      P04217     VAR_018369        p.His52Arg     Polymorphism   rs893184    -
A1BG      P04217     VAR_018370        p.His395Arg    Polymorphism   rs2241788   -
A1CF      Q9NQ94     VAR_052201        p.Val555Met    Polymorphism   rs9073      -
A1CF      Q9NQ94     VAR_059821        p.Ala558Ser    Polymorphism   rs11817448  -
A2ML1     A8K2U0     VAR_055463        p.Gly207Arg    Polymorphism   rs11047499  -
A2ML1     A8K2U0     VAR_055464        p.Cys970Tyr    Polymorphism   rs1558526   -
```

```
------------------------
        UniProt - Swiss-
        SIB Swiss Instit
        European Bioinfc
        Protein Informat
------------------------

Description: Human polymorphisms and disease mutations: index
Name:        humsavar.txt
Release:     2013_05 of 01-May-2013

--------------------------------------------------------------------------

Important note: variants classification is intended for research purposes
only, not for clinical and diagnostic use. The label disease variant is
assigned according to literature reports on probable disease-association
that can be based on theoretical reasons. Therefore this label must not be
considered as a definitive proof for the pathogenic role of a variant.
```

http://www.uniprot.org/docs/humsavar

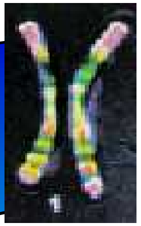
SNPs: Single Nucleotide Polymorphisms

(about 20,876 genes and 181,744 transcripts in the human genome)

Genes in DNA…

>protein kinase

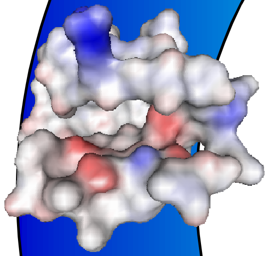
acctgttgatggcgacagggactgtatgctgatct
atgctgatgcatgcatgctgactactgatgtgggg
gctattgacttgatgtctatc....



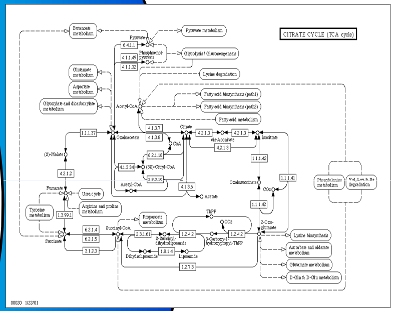
…with different effects depending on variability

Over 50 millions of single mutations are known

…code for proteins…



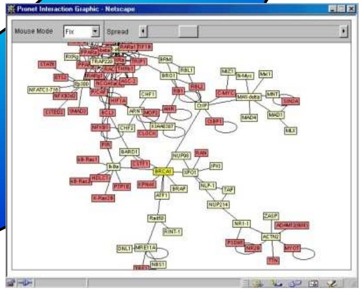
Overall: from Genotype to Phenotype

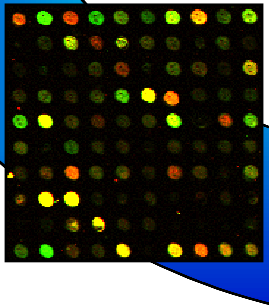


….in methabolic pathways

…proteins correspond to functions…

Proteins interact

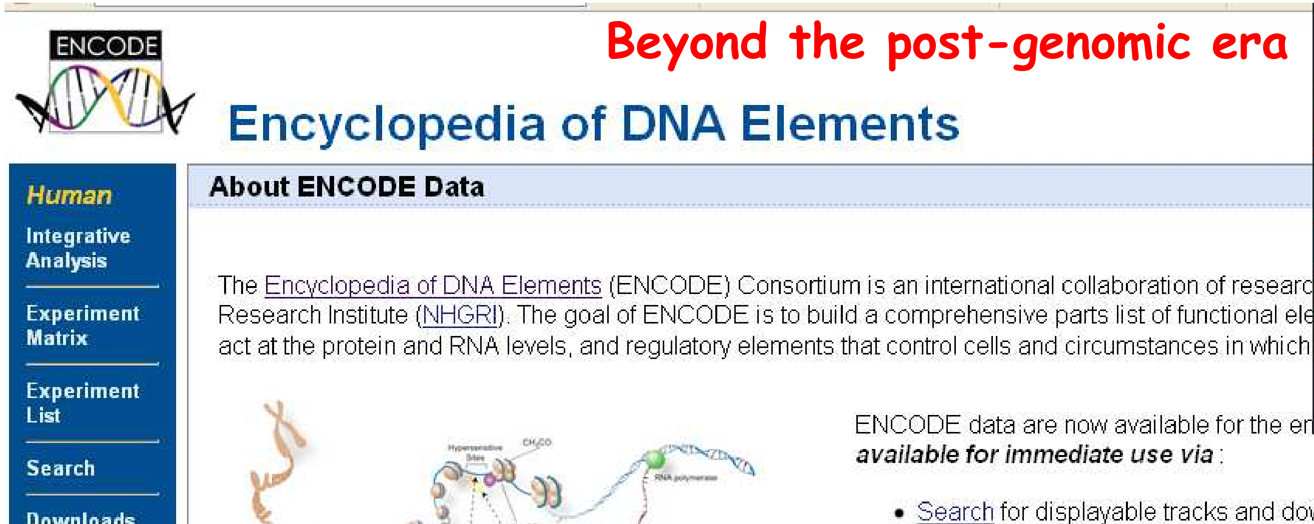




From 5000 to 10000 proteins per tissue

…when they are expressed

Beyond the post-genomic era



Encyclopedia of DNA Elements

About ENCODE Data

The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of researc
Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional ele
act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which

ENCODE data are now available for the en
available for immediate use via :

- Search for displayable tracks and downloadable files

To sea
feature

All ENC
Data R
UCSC

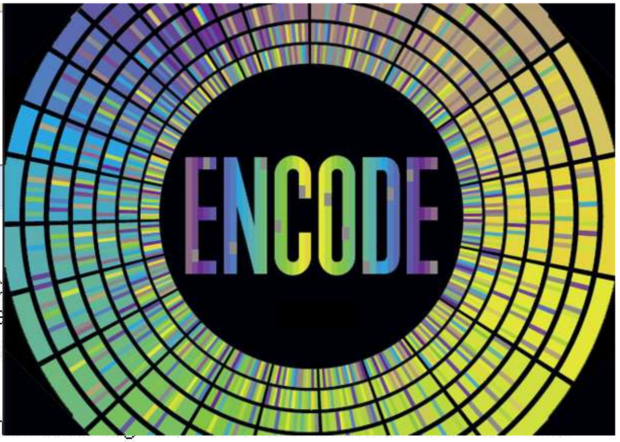
l

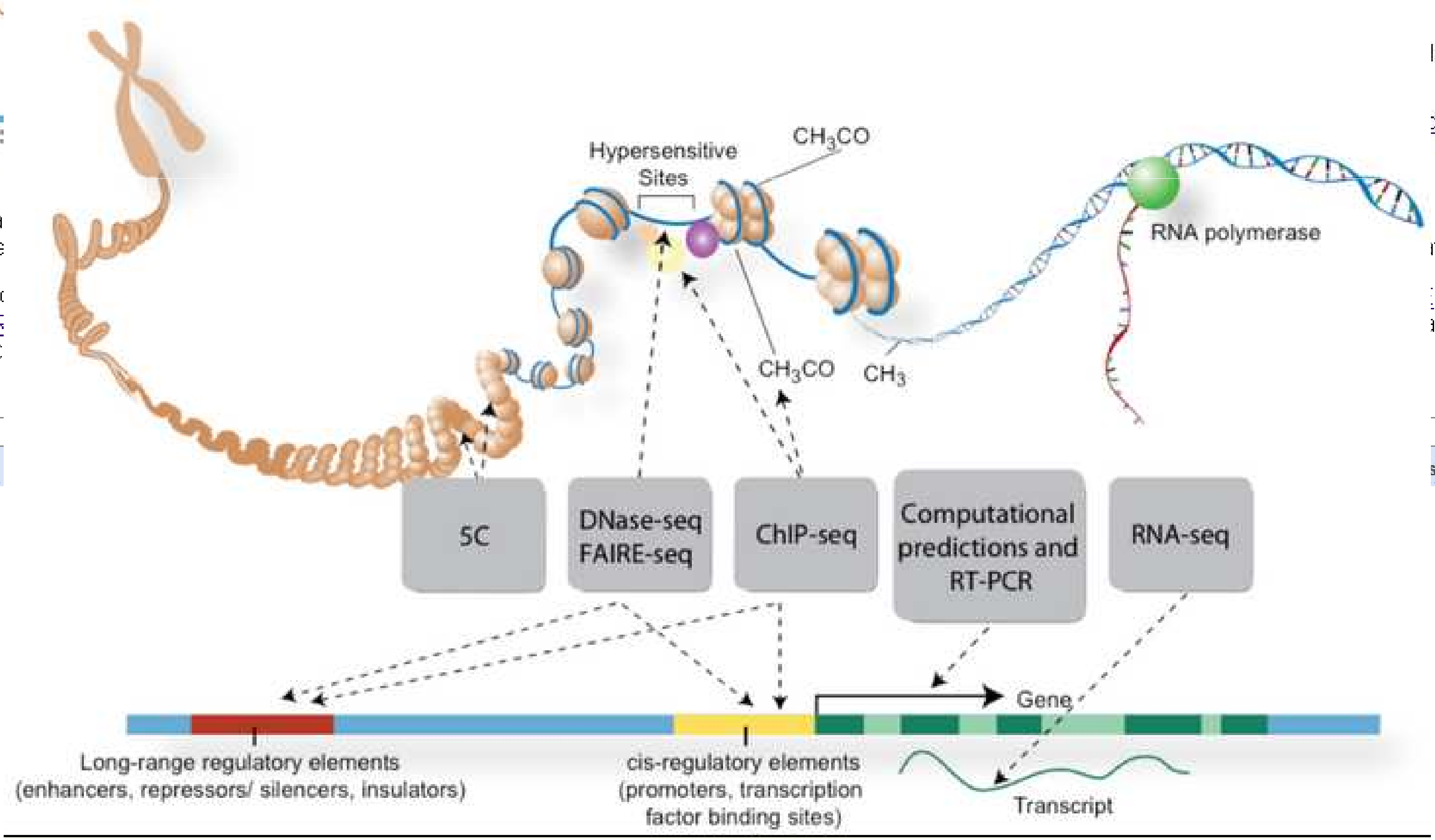
ols

ition.

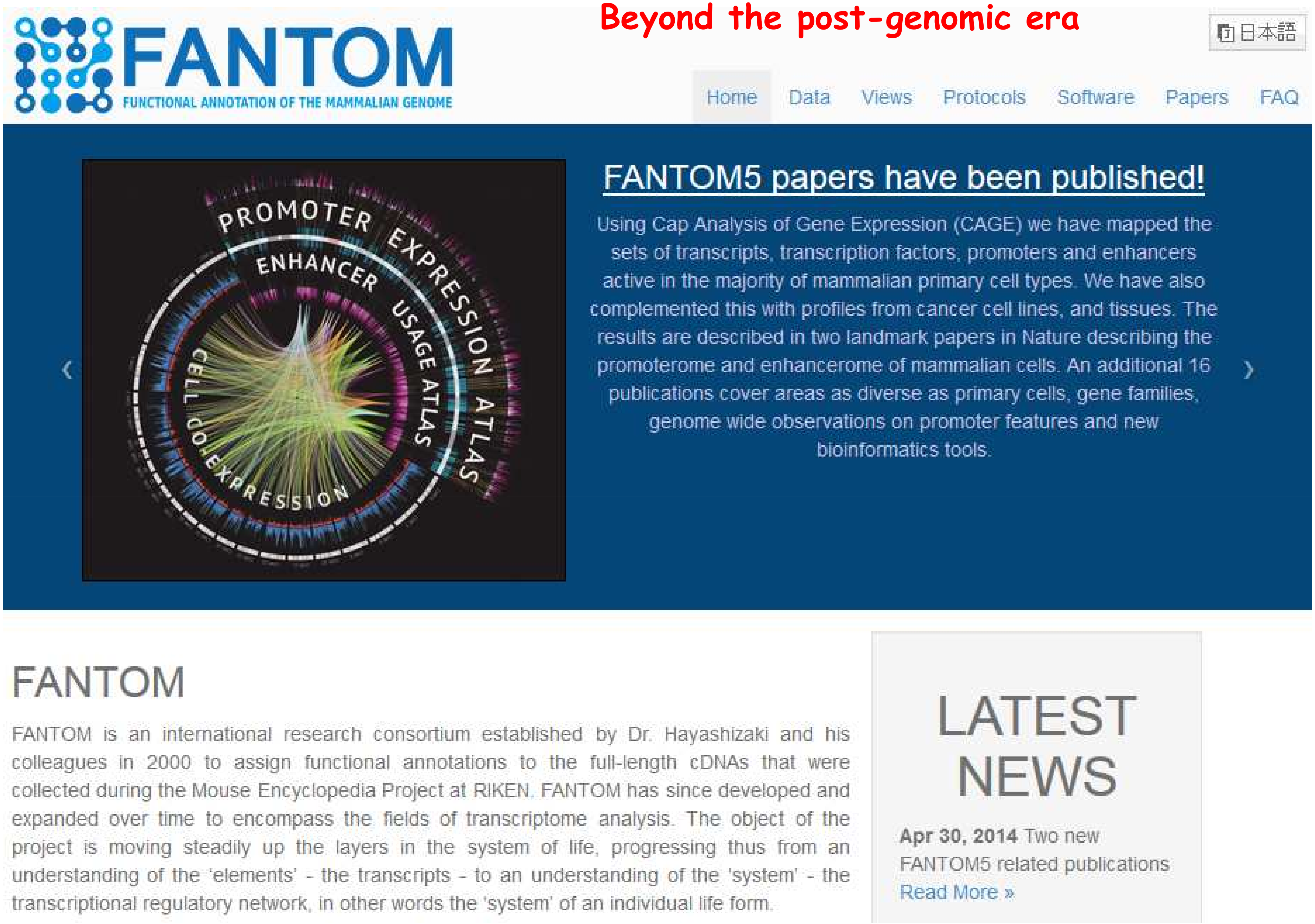
a at

s ▶




Hypersensitive Sites
CH_3CO
RNA polymerase
CH_3CO
CH_3
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

Beyond the post-genomic era



FANTOM5 papers have been published!

Using Cap Analysis of Gene Expression (CAGE) we have mapped the sets of transcripts, transcription factors, promoters and enhancers active in the majority of mammalian primary cell types. We have also complemented this with profiles from cancer cell lines, and tissues. The results are described in two landmark papers in Nature describing the promoterome and enhancerome of mammalian cells. An additional 16 publications cover areas as diverse as primary cells, gene families, genome wide observations on promoter features and new bioinformatics tools.

FANTOM

FANTOM is an international research consortium established by Dr. Hayashizaki and his colleagues in 2000 to assign functional annotations to the full-length cDNAs that were collected during the Mouse Encyclopedia Project at RIKEN. FANTOM has since developed and expanded over time to encompass the fields of transcriptome analysis. The object of the project is moving steadily up the layers in the system of life, progressing thus from an understanding of the 'elements' - the transcripts - to an understanding of the 'system' - the transcriptional regulatory network, in other words the 'system' of an individual life form.

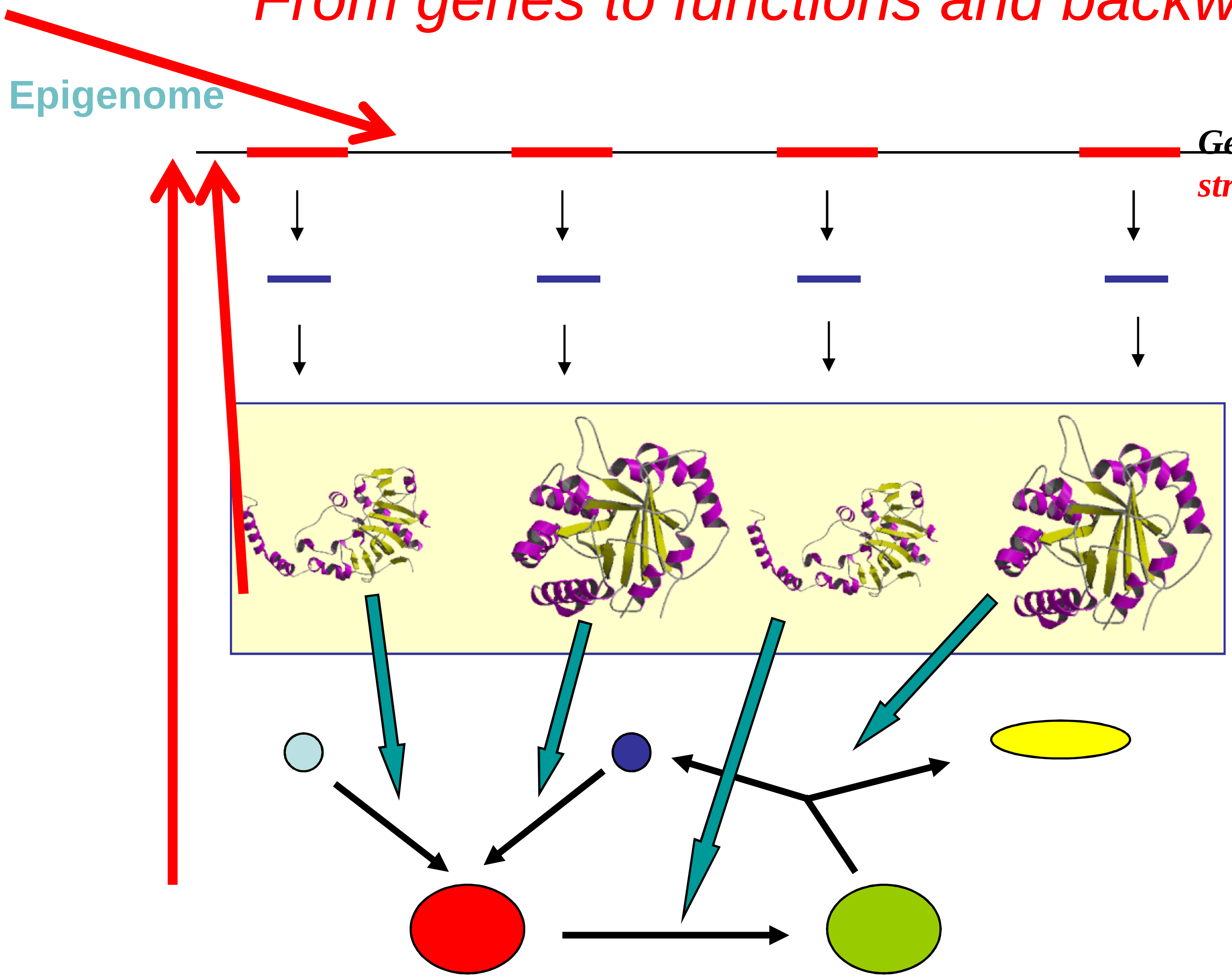
LATEST NEWS

Apr 30, 2014 Two new FANTOM5 related publications

Read More »

Functional Genomics/Epigenomics

From genes to functions and backward



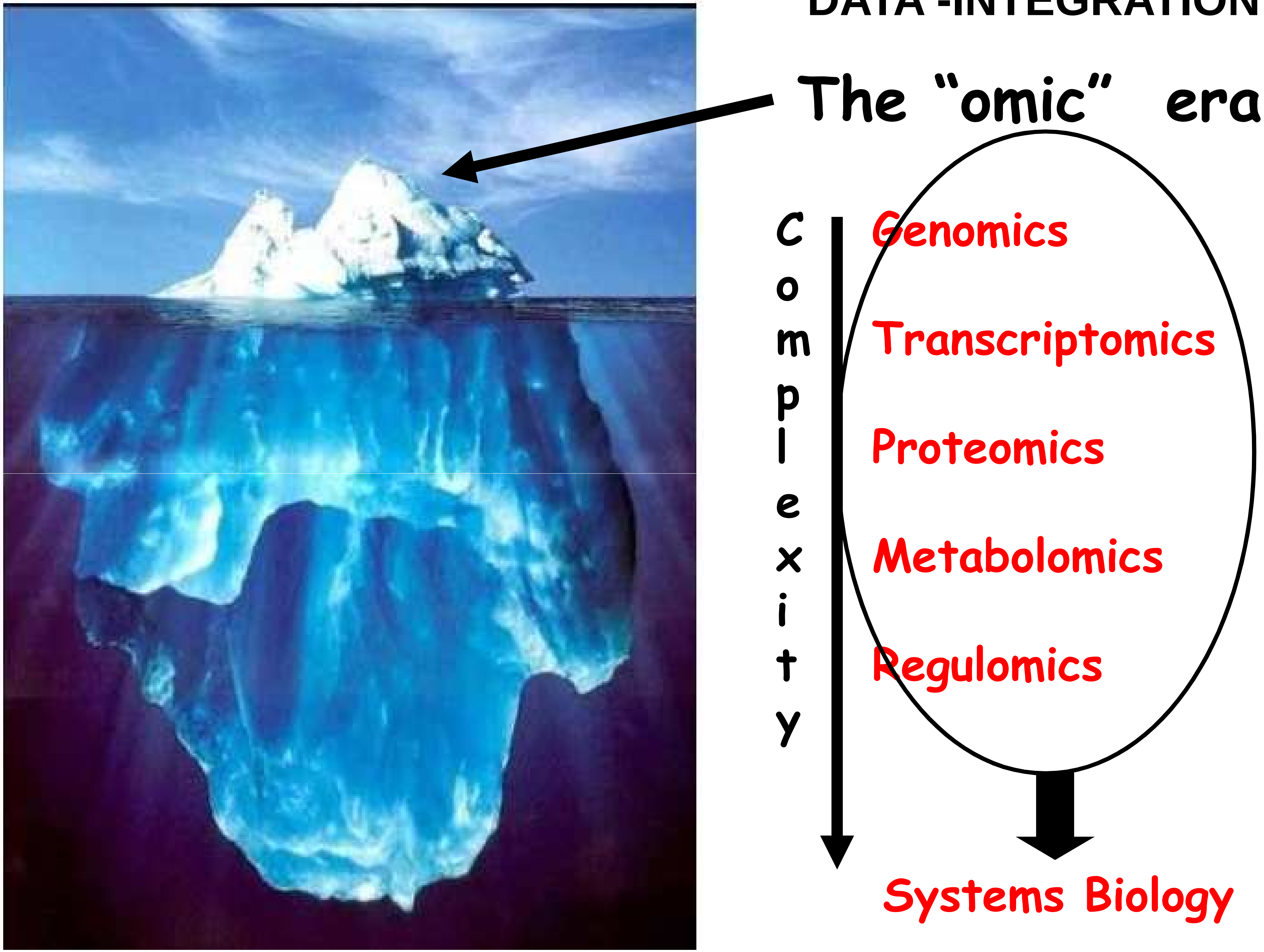
Genome [chromatin structure and dynamics]

***Expressome* [transcriptomics, level and regulation of expression]**

***Proteome*[*dynamic*]**

Metabolome

DATA -INTEGRATION



The “omic” era

Complexity

Genomics

Transcriptomics

Proteomics

Metabolomics

Regulomics

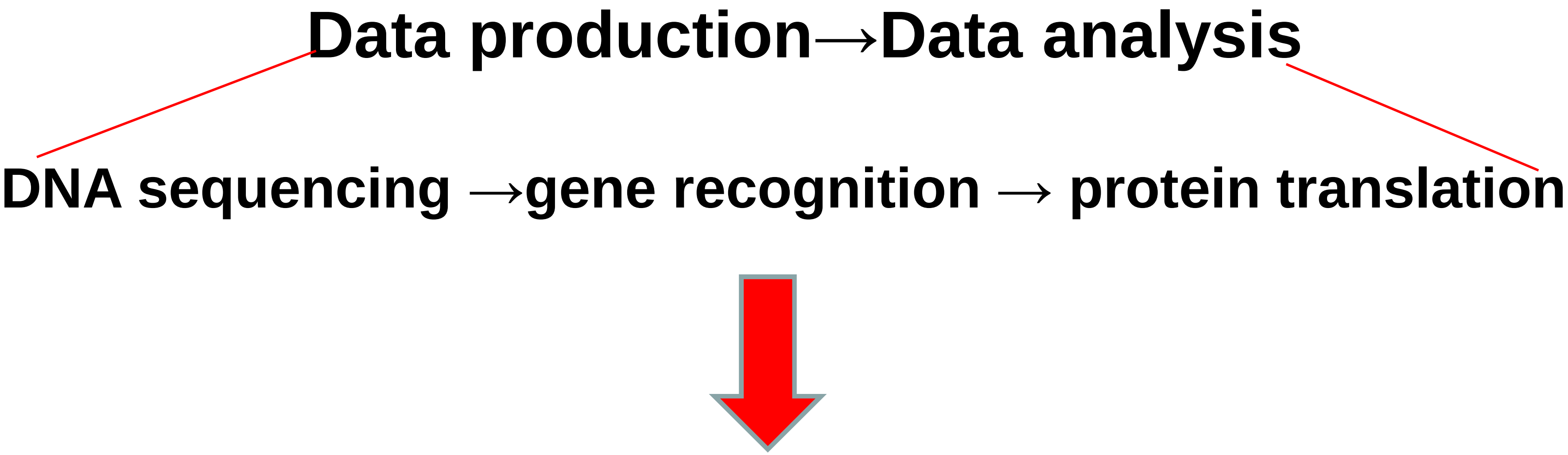
Systems Biology

Summing up…..

Open problems in the post-genomic era after DNA/RNA sequencing

1) Genome assembly & Genome annotation (e.g. exon/intron boundaries)
2) Chromatin dynamics
3) Finding alternative splicing variants
4) Protein structural and functional annotation
5) Annotation of SNP variants & Correlation among SNPs and diseases
6) Simulation of cell complexity

Genomic data and the problem of protein validation



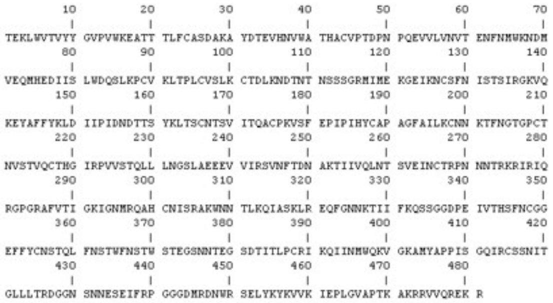
Data production→Data analysis

DNA sequencing →gene recognition → protein translation

Experiments to validate protein structure and function produce data in a time >> than that required to deposit putative protein sequences into data bases

A “BIG” problem of the “omic era” after genome sequencing:

code for



```
        10         20         30         40         50         60         70
         |          |          |          |          |          |          |
TEKLWVTVYY GVPVWKEATT TLFCASDAKA YDTEVHNVWA THACVPTDPN PQEVVLVNVT ENFNMWKNDM
        80         90        100        110        120        130        140
         |          |          |          |          |          |          |
VEQMHEDIIS LWDQSLKPCV KLTPLCVSLK CTDLKNDTNT NSSSGRMIME KGEIKNCSFN ISTSIRGKVQ
       150        160        170        180        190        200        210
         |          |          |          |          |          |          |
KEYAFFYKLD IIPIDNDTTS YKLTSCNTSV ITQACPKVSF EPIPIHYCAP AGFAILKCNN KTFNGTGPCT
       220        230        240        250        260        270        280
         |          |          |          |          |          |          |
NVSTVQCTHG IRPVVSTQLL LNGSLAEEEV VIRSVNFTDN AKTIIVQLNT SVEINCTRPN NNTRKRIRIQ
       290        300        310        320        330        340        350
         |          |          |          |          |          |          |
RGPGRAFVTI GKIGNMRQAH CNISRAKWNN TLKQIASKLR EQFGNNKTII FKQSSGGDPE IVTHSFNCGG
       360        370        380        390        400        410        420
         |          |          |          |          |          |          |
EFFYCNSTQL FNSTWFNSTW STEGSNNTEG SDTITLPCRI KQIINMWQKV GKAMYAPPIS GQIRCSSNIT
       430        440        450        460        470        480
         |          |          |          |          |          |
GLLLTRDGGN SNNESEIFRP GGGDMRDNWR SELYKYKVVK IEPLGVAPTK AKRRVVQREK R
```

Protein sequences (~56millions)

that are endowed with

Genomes (~3000)

Protein structures and functions

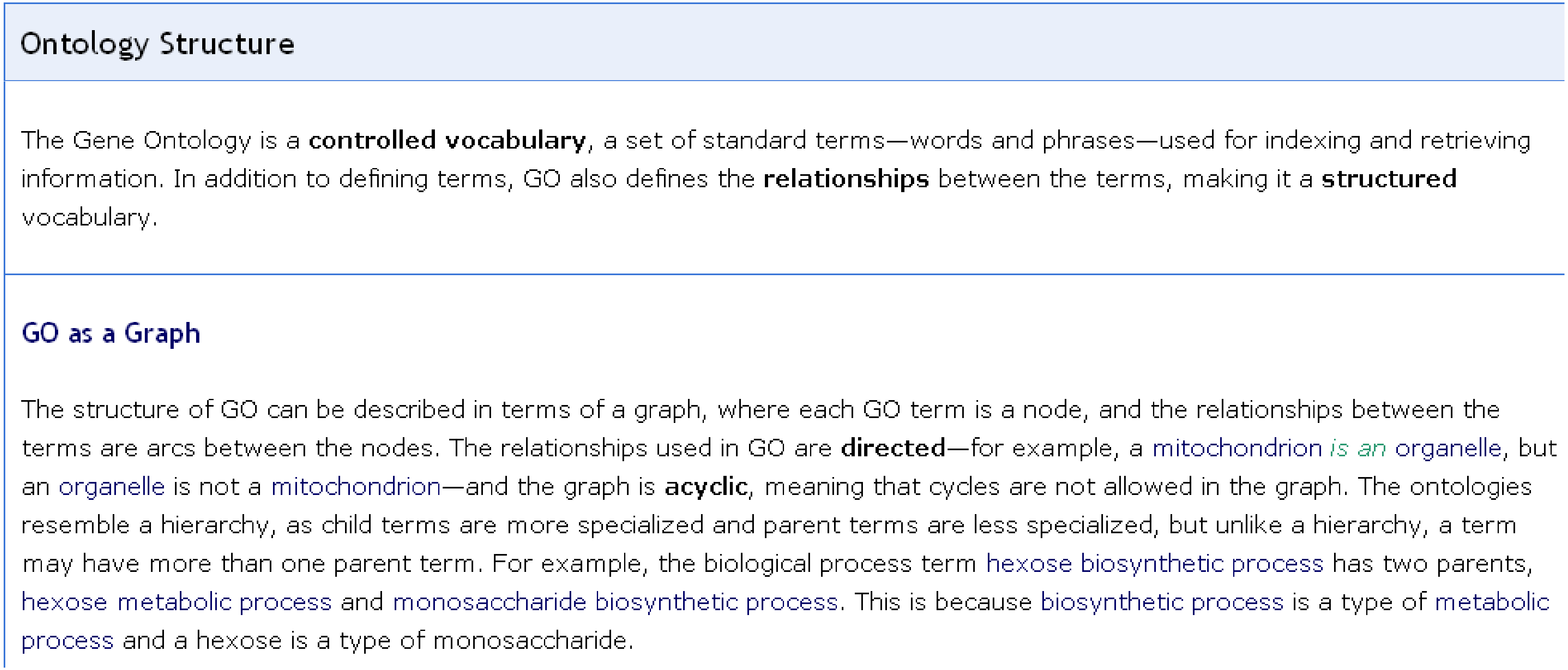
Protein sequence Annotation:

to endow with structural and functional features protein sequences after gene translation

Protein function

GO function vocabulary:
http://www.geneontology.org/

The Ontologies
- Cellular component
- Biological process
- Molecular function



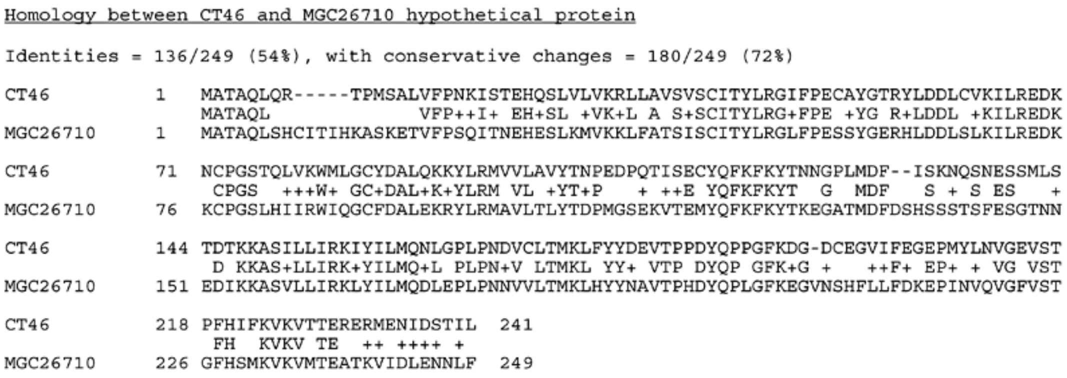
Ontology Structure

The Gene Ontology is a **controlled vocabulary**, a set of standard terms—words and phrases—used for indexing and retrieving information. In addition to defining terms, GO also defines the **relationships** between the terms, making it a **structured** vocabulary.

GO as a Graph

The structure of GO can be described in terms of a graph, where each GO term is a node, and the relationships between the terms are arcs between the nodes. The relationships used in GO are **directed**—for example, a mitochondrion *is an* organelle, but an organelle is not a mitochondrion—and the graph is **acyclic**, meaning that cycles are not allowed in the graph. The ontologies resemble a hierarchy, as child terms are more specialized and parent terms are less specialized, but unlike a hierarchy, a term may have more than one parent term. For example, the biological process term hexose biosynthetic process has two parents, hexose metabolic process and monosaccharide biosynthetic process. This is because biosynthetic process is a type of metabolic process and a hexose is a type of monosaccharide.

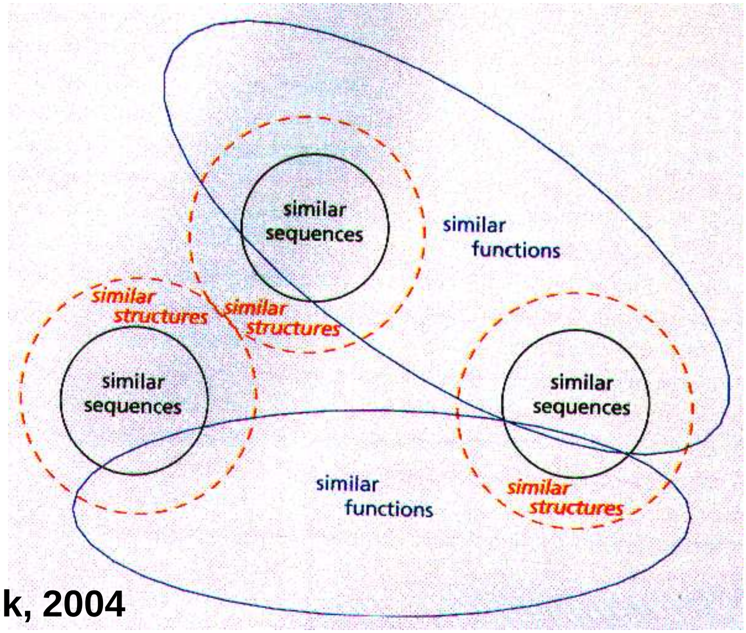
Protein annotation by sequence similarity



```
Homology between CT46 and MGC26710 hypothetical protein

Identities = 136/249 (54%), with conservative changes = 180/249 (72%)

CT46         1   MATAQLQR-----TPMSALVFPNKISTEHQSLVLVKRLLAVSVSCITYLRGIFPECAYGTRYLDDLCVKILREDK
                 MATAQL             VFP++I+ EH+SL +VK+L A S+SCITYLRG+FPE +YG R+LDDL +KILREDK
MGC26710     1   MATAQLSHCITIHKASKETVFPSQITNEHESLKMVKKLFATSISCITYLRGLFPESSYGERHLDDLSLKILREDK

CT46         71  NCPGSTQLVKWMLGCYDALQKKYLRMVVLAVYTNPEDPQTISECYQFKFKYTNNGPLMDF--ISKNQSNESSMLS
                  CPGS  +++W+ GC+DAL+K+YLRM VL +YT+P   + ++E YQFKFKYT  G  MDF   S + S ES   +
MGC26710     76  KCPGSLHIIRWIQGCFDALEKRYLRMAVLTLYTDPMGSEKVTEMYQFKFKYTKEGATMDFDSHSSSTSFESGTNN

CT46         144 TDTKKASILLIRKIYILMQNLGPLPNDVCLTMKLFYYDEVTPPDYQPPGFKDG-DCEGVIFEGEPMYLNVGEVST
                  D KKAS+LLIRK+YILMQ+L PLPN+V LTMKL YY+ VTP DYQP GFK+G +   ++F+ EP+ + VG VST
MGC26710     151 EDIKKASVLLIRKLYILMQDLEPLPNNVVLTMKLHYYNAVTPHDYQPLGFKEGVNSHFLLFDKEPINVQVGFVST

CT46         218 PFHIFKVKVTTERERMENIDSTIL  241
                  FH  KVKV TE  ++ ++++ +
MGC26710     226 GFHSMKVKVMTEATKVIDLENNLF  249
```

If sequences share more than 30/40% sequence identity they can share similar structure and function

But the problem is much more complex



Lesk, 2004

Transfer of annotation *in silico* by homology search

```
ADH1_SULSO  ----------MRAVRLVEIGKP--LSLQEIGVPKPKGPQVLIKVEAAGVCHSDVHMRQGRFGNLRIVE
ADH_CLOBE   ----------MKGFAMLGINKLG---WIEKERPVAGSYDAIVRPLAVSPCTSDIHTVFEGA-------
ADH_THEBR   ----------MKGFAMLSIGKVG---WIEKEKPAPGPFDAIVRPLAVAPCTSDIHTVFEGA-------
ADH1_SOLTU  MSTTVGQVIRCKAAVAWEAGKP--LVMEEVDVAPPQKMEVRLKILYTSLCHTDVYFWEAKG-------
ADH2_LYCES  MSTTVGQVIRCKAAVAWEAGKP--LVMEEVDVAPPQKMEVRLKILYTSLCHTDVYFWEAKG-------
ADH1_ASPFL  ----MSIPEMQWAQVAEQKGGP--LIYKQIPVPKPGPDEILVKVRYSGVCHTDLHALKGDW-------
```

Sequence comparison is performed with alignment programs

Sequence identity ≥ 30 % ⇨ 3D ?; Similar function ??

Methods for similarity searches:

BLAST, Psi-BLAST (http://www.ncbi.nlm.nih.gov/BLAST/)

Altschul et al., (1990) J Mol Biol 215:403-410

Altschul et al., (1998) Nucleic Acids Res. 25:3389-3402

Pfam (http://pfam.wustl.edu/hmmsearch.shtml)

Bateman et al., (2000) Nucleic Acids Research 28:263-266

The little we know (SwissProt)……is expanded to annotate all the protein sequences (TrEMBL)

Release 2014_03 (19-Mar) of UniProtKB/Swiss-Prot contains 542,782 sequence entries:

Protein existence (PE):	entries	%
1: Evidence at protein level	82,087	15.1
2: Evidence at transcript level	62,227	11.5
3: Inferred from homology	380,832	70.2
4: Predicted	15,705	2.9

Release 2014_03 (19-Mar) of UniProtKB/TrEMBL contains 54,247,468 sequence entries:

Protein existence (PE):	entries	%
1: Evidence at protein level	22,013	0.04
2: Evidence at transcript level	931,313	1.72
3: Inferred from homology	13,573,938	25.02
4: Predicted	39,720,204	73.22

Automatic annotation at UNIPROTKB

http://www.uniprot.org/program/automatic_annotation

UniProt has developed two prediction systems, UniRule and the Statistical Automatic Annotation System (SAAS) to automatically annotate UniProtKB/TrEMBL in an efficient and scalable manner with a high degree of accuracy:

- Based on rules
- Rules are created, tested and validated against published experimental data in UniProtKB/Swiss-Prot
- Rules are linked to InterPro member database signatures
- Rules have annotations and conditions
- Rules are reapplied to UniProtKB/TrEMBL every four-weekly release with both automatic and manual QA procedures ensuring they are still valid