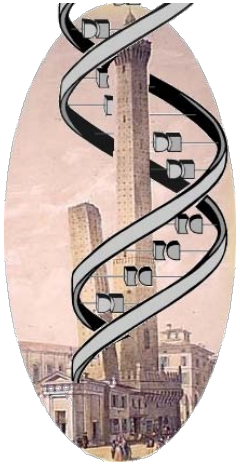
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



# Genomes,Genes,Proteins
# Why Bionformatics?


## Rita Casadio

*BIOCOMPUTING GROUP*
*University of Bologna, Italy*


Bologna Computational Biology Network

WetLab | CompuLab | SME Small Medium Eterprises

AIRBBC

**Syllabus:**

1) The "omic" revolution

2) Next Generation Sequencing Data

3) Omics and data archives

4) The ingredients of biological complexity at the cell level

5) Open problems in the omic era

# The "omic" revolution

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

GA EVANS, Nature Biotechnology 18:127, 2000

# The New Genomics: Global Views of Biology

Landmarks:

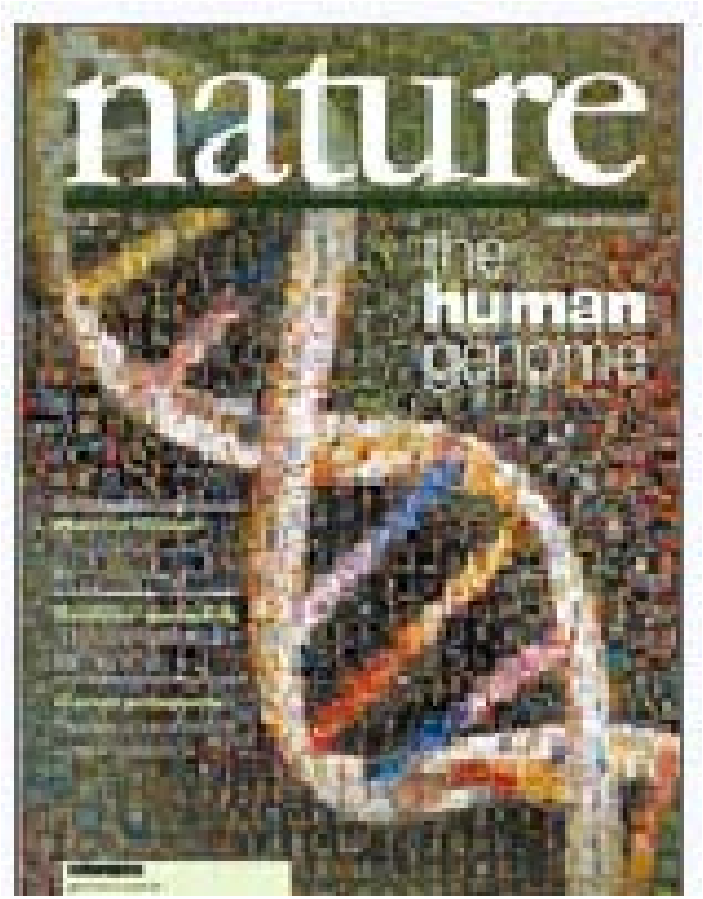
- 1920 Winkler uses the expression "Genome" for the haploid chromosome set

- 1995 The first complete genome of the bacterium *Haemophilus influenzae* (Fleischman et al., Science 269, 496) only one aploid chromosome and all the regoins are coding

- 2001 The first draft of the human genome assembled by Celera and by the International Human Genome Sequencing Consortium

# Sequencing human DNA: some time ago…the draft

15/02/2001




16/02/2001

# Our closest living relative ….*Pan troglodytes*

1/09/2005


alignment

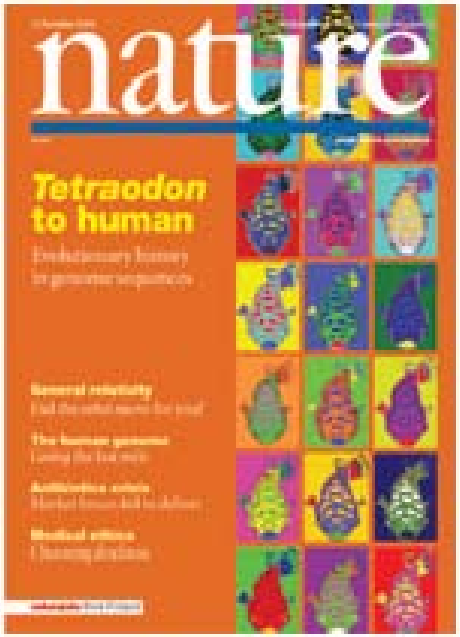
Comparison at a molecular level…..

Divergence: 6 million years ago

Changes: 35 millions single nucleotides

Human DNA: 3,272,187,692
Chimp DNA: 2,733,948,177

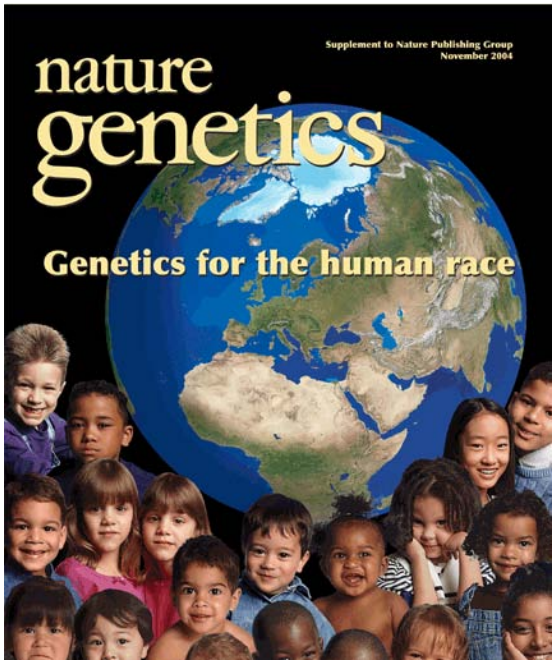
# END OF THE BEGINNING……




October 2004

## Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*


November 2004

The Human Variome

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

SNPs: Single Nucleotide Polymorphisms

Welcome to Human Genome Organisation

Linked In Facebook Twi


# HUGO

## PRESIDENT'S MESSAGE

In 2007, the HUGO Council honored me by selecting me as your President-elect. Since then, our Council and I have sought to enhance the strength of HUGO, to embark on initiatives that will move HUGO into new intellectual territory – conceptual domains that will place HUGO again at the forefront of this new convergence of genomic sciences, medicine, and social policy ...

## THE HUGO JOURNAL


Submission

Home | Search | Contact Us | Lo

Fostering international proteomic initiatives to better understand human disease

# HUPO

**Human Proteome Project**
INFORMATION

**Register for our Newsletter**

First Name:
First name

Last Name:
Last name

Email Address:

Continue

## Welcome to the Human Proteome Organisation's (HUPO) website

The Human Proteome Organisation (HUPO) is an international scientific organization representing and promoting proteomics through international cooperation and collaborations by fostering the development of new technologies, techniques and training. Should you have any questions regarding our activities or how you can become involved in our organization, please click the contact us link in the top right-hand corner and the HUPO Secretariat, based in Montreal Canada, would be happy to assist you.

Results of HUPO Board of Directors election - Click here

**HUPO 11th Annual World Congress, Boston 2012, September 9-13**

**Voting Period for HUPO Board of Directors between August 15 and September 5, 2011**

The voting period for seats that will become vacant

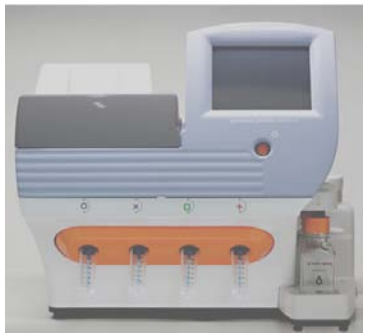
# Next Generation Sequencing Technology allows an unprecedent rate of DNA /RNA sequencing (>4TB per week)

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



Number of sequencing machines by continent

| Name | Number of Machines |
|---|---|
| North America | 854 |
| Europe | 501 |
| Asia | 361 |
| Australia | 71 |
| South America | 16 |
| Africa | 11 |



# Dealing with genomic data….

***Scott D. Kahn***
***Science 331, 728 (2011)***

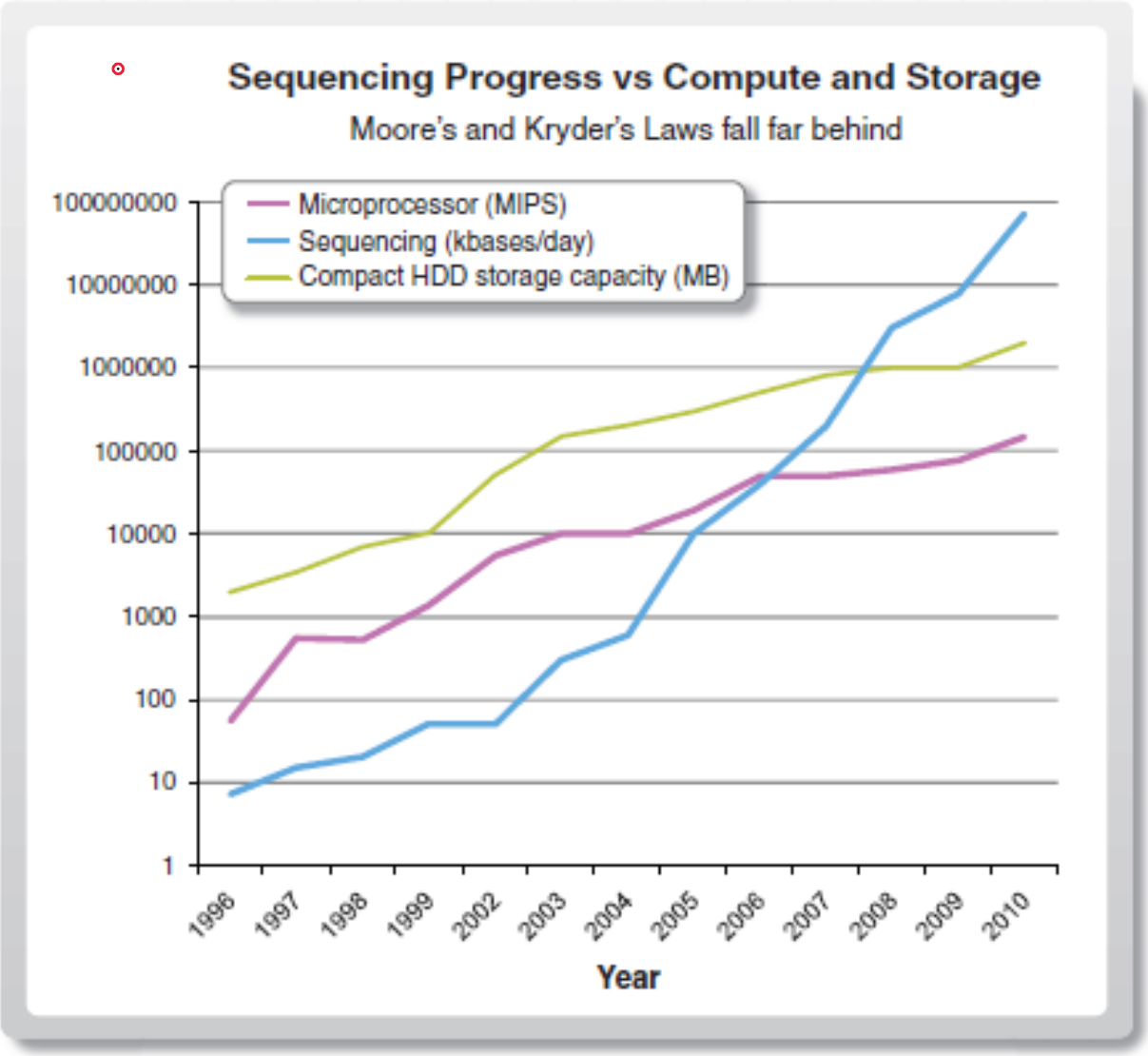
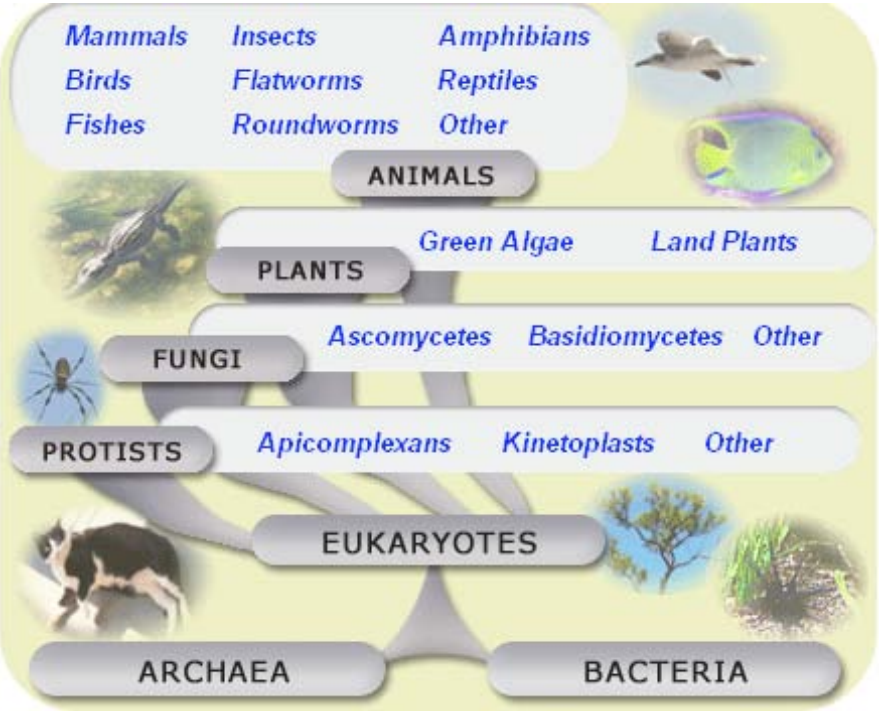



**Fig. 1.** A doubling of sequencing output every 9 months has outpaced and overtaken performance improvements within the disk storage and high-performance computation fields.

# The "omic" era-RESULTS

## Genome Sequencing Projects:




**Archaea : 121 species**

In Progress: 90 species

**Bacteria: 1731 species**

In Progress: 5140 species

**Eukaryotes:**

Complete-150

In Progress-1365

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

Update:
December 2011

Ensembl Genome Browser - Mozilla Firefox

File Edit View History Bookmarks Yahoo! Tools Help

Biocomputing Group - University of Bologna | Ensembl Genome Browser

www.ensembl.org/index.html

ensembl SEARCH

Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search: All species for Go

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

# Eukaryotic Genome archives

## 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
- ıta you want
  database, 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,

# Overall: from Genotype to Phenotype

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

*Genes in DNA…*

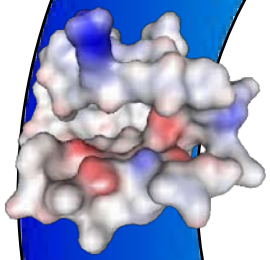
>protein kinase

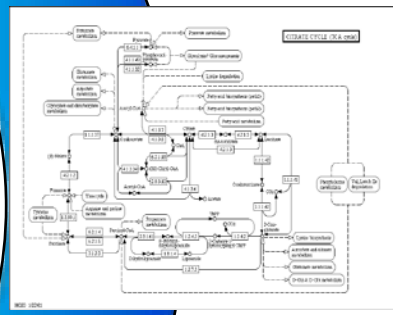
acctgttgatggcgacagggactgtatgctgatct
atgctgatgcatgcatgctgactactgatgtgggg
gctattgacttgatgtctatc....

*…with different effects depending on variability*

Over 50 millions of single mutations are known

*…code for proteins…*



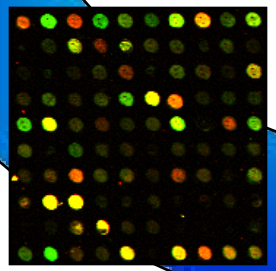


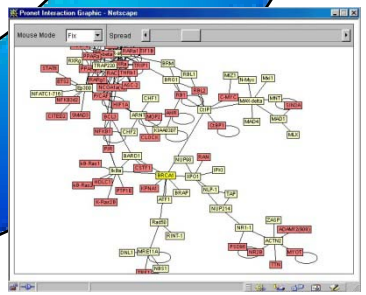
*…proteins correspond to functions…*

*….in methabolic pathways*

*Proteins interact*

From 5000 to 10000 proteins per tissue





*…when they are expressed*

dbSNP Home Page

# Where to search/check for neutral variations

NCBI **Single Nucleotide Polymorphism**

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP

Search for SNP on NCBI Reference Assembly

Search Entrez SNP for Go

**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 |
|---|---|---|---|---|---|
| Homo sapiens | 135 | 37.3 | 178,140,935 | 52,327,221 (41,740,143) | 21,247,880 |
| Mus musculus | 132 | 37.1 | 26,991,031 | 15,522,011 (6,439,098) | 6,696,618 |
| Pongo abelii | 132 | | 10,225,850 | 10,065,309 (0) | |
| Pongo pygmaeus | 127 | | 7,854,083 | 7,854,081 (0) | |
| Rattus norvegicus | 130 | 4.1 | 6,472,989 | 119,436 (1,605) | 1,024,738 |
| Gallus gallus | 131 | 2.1 | 11,318,097 | 3,504,588 (3,269,983) | 1,452,147 |
| Glycine max | 127 | | 6,378,350 | 6,352,034 (234) | |
| Phoenix dactylifera | 133 | | 3,518,029 | 3,429,753 (0) | |
| Oryza sativa | 128 | 4.1 | 5,872,306 | 5,359,569 (21,773) | 1,897,895 |
| Bos taurus | 131 | 4.1 | 4,931,454 | 2,210,557 (13,881) | 677,906 |
| Zea mays | 128 | | 4,556,997 | 4,351,393 (80) | |

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

# Where to check variations for disease association

OMIM - Online Mendelian Inheritanc...

NCBI

OMIM
Online Mendelian Inheritance in Man

Johns Hopkins University

All Databases | PubMed | Nucleotide | Protein | Genome

OMIM Entry Statistics:

Number of Entries in OMIM (12 January 2012) :

| Prefix | Autosomal | X Linked | Y Linked | Mitochondrial | Totals |
|---|---|---|---|---|---|
| * Gene description | 13,052 | 640 | 48 | 35 | 13,775 |
| + Gene and phenotype, combined | 159 | 6 | 0 | 2 | 167 |
| # Phenotype description, molecular basis known | 3,074 | 258 | 4 | 28 | 3,364 |
| % Phenotype description or locus, molecular basis unknown | 1,655 | 136 | 5 | 0 | 1,796 |
| Other, mainly phenotypes with suspected mendelian basis | 1,798 | 129 | 2 | 0 | 1,929 |
| Totals | 19,738 | 1,169 | 59 | 65 | 21,031 |

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

# Where to find disease associated variations in proteins

humsavar.txt - Mozilla Firefox

File Edit View History Bookmarks Yahoo! Tools Help

Biocomputing Group - University of Bologna

www.uniprot.org/docs/humsavar

humsavar

UniProt › Documents

Search Blast

Search in

Documents

humsavar.txt

```
Statistics for single amino acid variants:

        Disease variants: 22023
        Polymorphisms: 36942
        Unclassified variants: 6006
        Total: 64971
```

| Main gene name | Swiss-Prot AC | FTId | AA change | Type of variant | dbSNP | D |
|---|---|---|---|---|---|---|
| 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 | - |
| A2M | P01023 | VAR_000012 | p.Arg704His | Polymorphism | rs1800434 | - |
| A2M | P01023 | VAR_000013 | p.Cys972Tyr | Polymorphism | rs1800433 | - |
| A2M | P01023 | VAR_000014 | p.Ile1000Val | Polymorphism | rs669 | - |
| A2M | P01023 | VAR_026820 | p.Asn639Asp | Polymorphism | rs226405 | - |
| A2M | P01023 | VAR_026821 | p.Leu815Gln | Polymorphism | rs3180392 | - |

```
--------------------------------------------------------------------------
        UniProt - Swiss-Prot Pro
        Swiss Institute of Bioin
        European Bioinformatics Institute (EBI); Hinxton, United Kingdom
        Protein Information Resource (PIR); Washington DC, USA
--------------------------------------------------------------------------

Description: Human polymorphisms and disease mutations: index
Name:        humsavar.txt
Release:     2011_12 of 14-Dec-2011

--------------------------------------------------------------------------
```

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

# BIOINFORMATICS

## Data Bases

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

- Implementation
- Data Mining
- Links

## Computational Biology

Tools for:

- Sequence analysis
- Functional genomics
- Proteomics

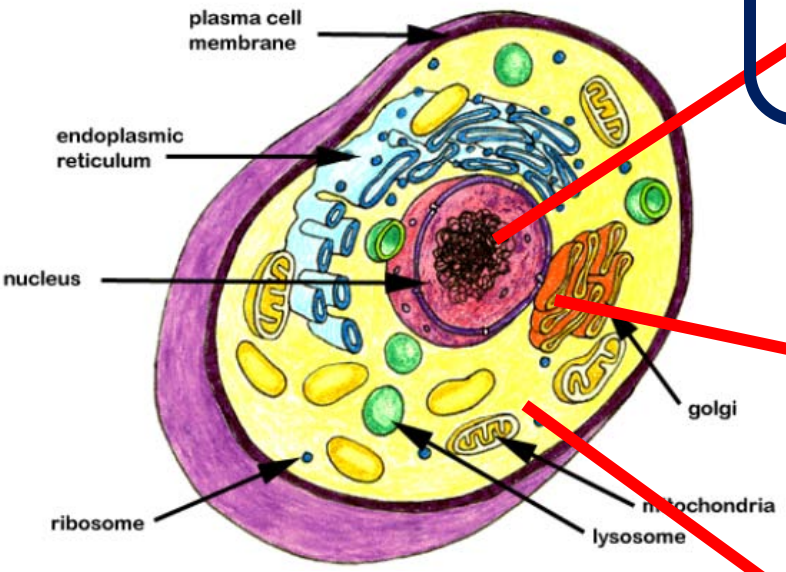
## Systems Biology

Models for:
Interatomics, Methabolomics, Evolving complex biosystems (Cell, Organism,..)

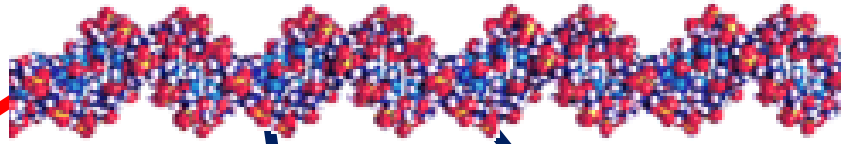
# The ingredients of biological complexity at the cell level

## *From genes to proteins and their interaction*

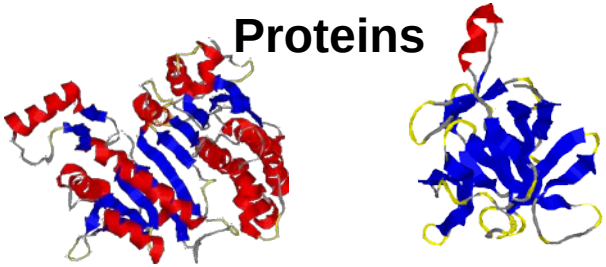
The cell

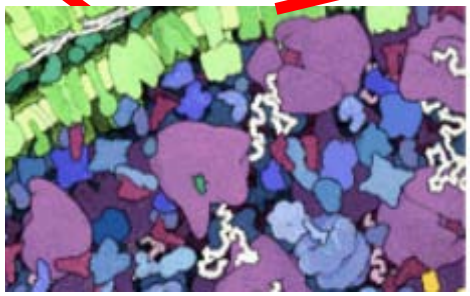



DNA & Genes



Proteins





Macromolecular crawding

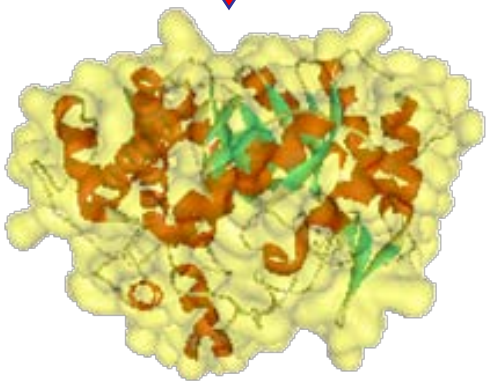
Molecular Interactions and Functions

# The basic information flow: from DNA to proteins

A,T,C,G

cctgttgatggcgacagggactgtatgctgatctatgctgatgcatgcatgctgactactgatgtgggggctat

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

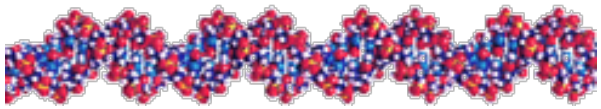


*From genes…*

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

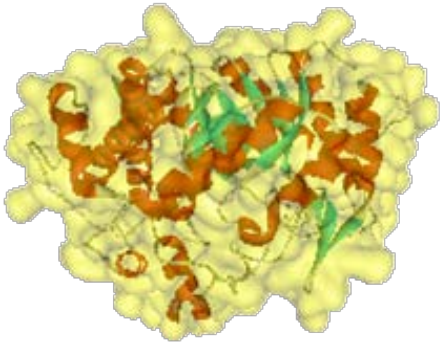


GenBank: 135,440,990 sequences
126,551,501,141 nucleotides

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

UniProt/Tremble:
18,215,214 sequences
5,957,253,786 residues

UniProt/SwissProt:
533,049 sequences
189,064,225 residues



PDB: 75,4633 structures
membrane proteins <2%

≅43 HGE!

Update:
December 2011

RCSB PDB PROTEIN DATA BANK → RCSB PDB-101

An Information
As of Tuesday Dec 06, 2011 at 4 PM PST t

All Categories | Author | Macromolecule | Sequence | Ligand

Search | All Categories: e.g., PDB ID, molecule name, author

Customize This Page

MyPDB Hide
- Login to your Account
- Register a New Account

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

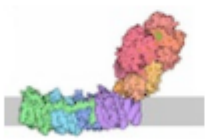
Deposition Hide
- All Deposit Services

# Biological Macromolecular Resource

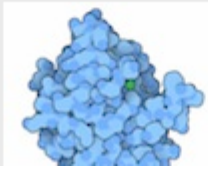
Full Description

Featured Molecules

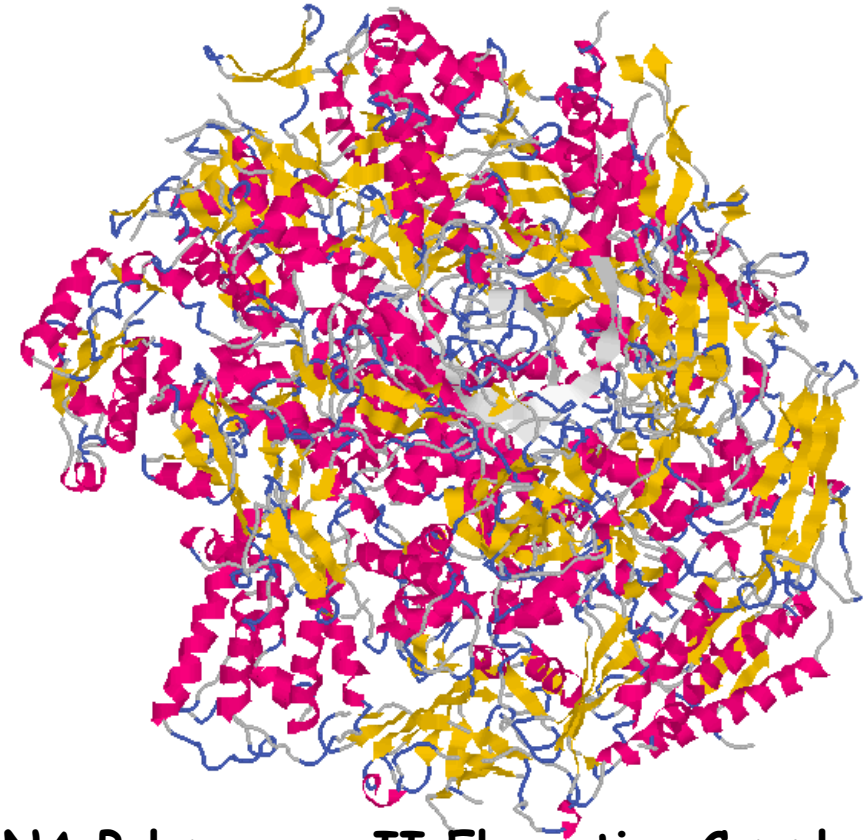
Structural View of Biology



Molecule of
Complex I
Complex
transport
Full Artic



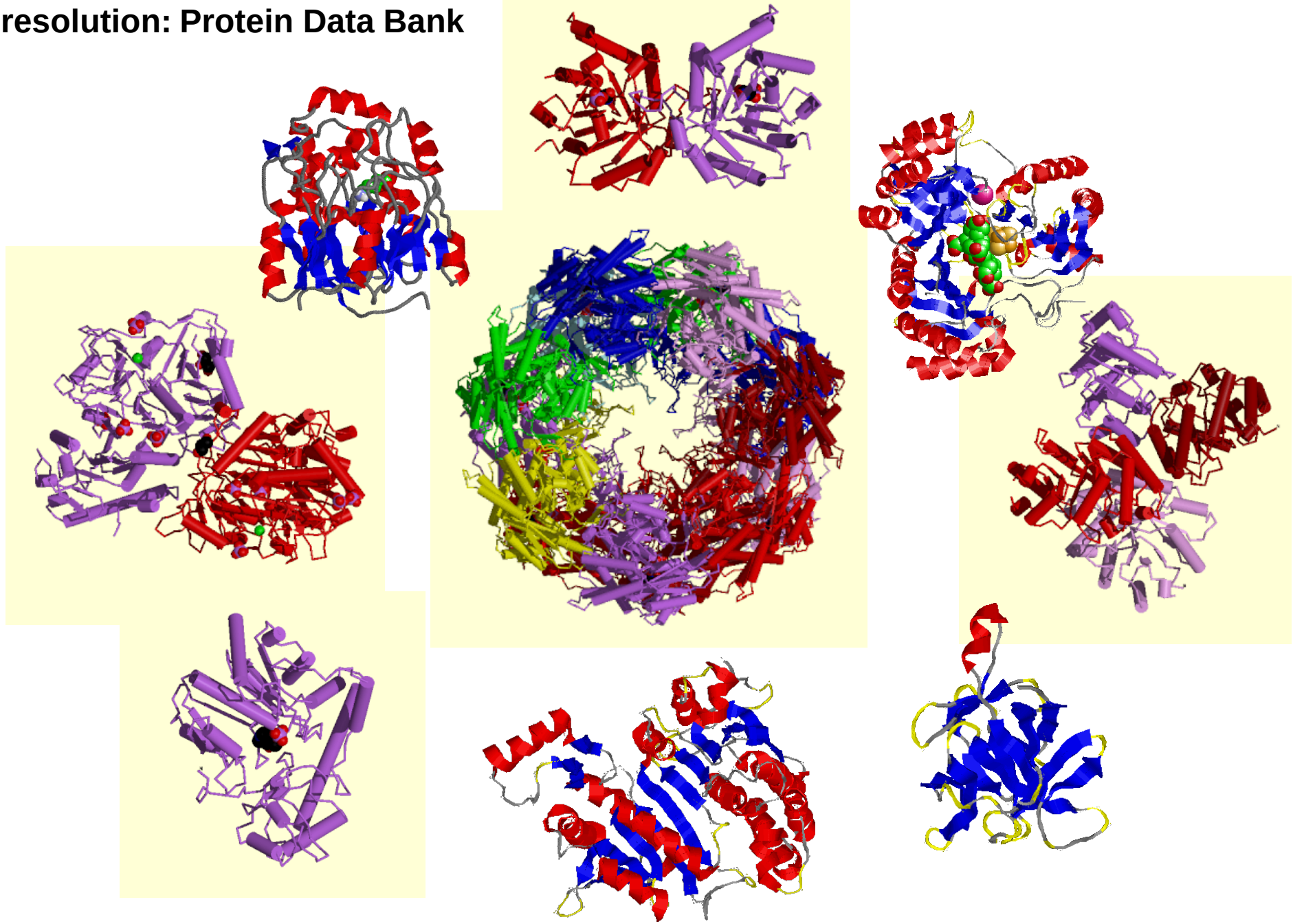
Protein Stru
Superbugs
Antibiotics
Natural a
might exp



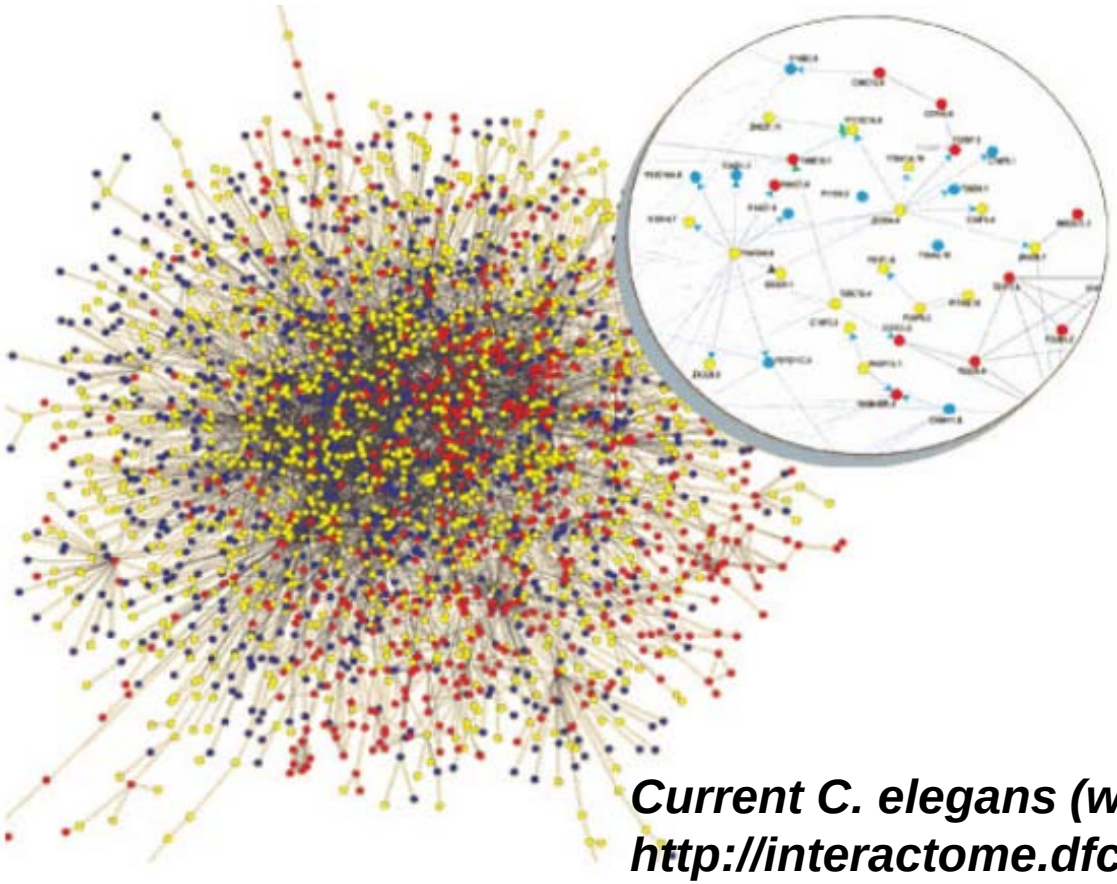
By: T
Biolo
Ener
n res
disco
them

E.G:: RNA Polymerase II Elongation Complex

# BioMolecules/Biomolecular Complexes known with atomic resolution: Protein Data Bank



# In terms of proteomics, interactomics refers to protein-protein interaction networks



***Current C. elegans (worm) interactome***
***http://interactome.dfci.harvard.edu/C_elegans/***

the Gene Ontology

# Protein function

Open menus

**The Ontologies**

- Cellular component
- Biological process
- Molecular function

## GO function vocabulary:

http://www.geneontology.org/

### 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

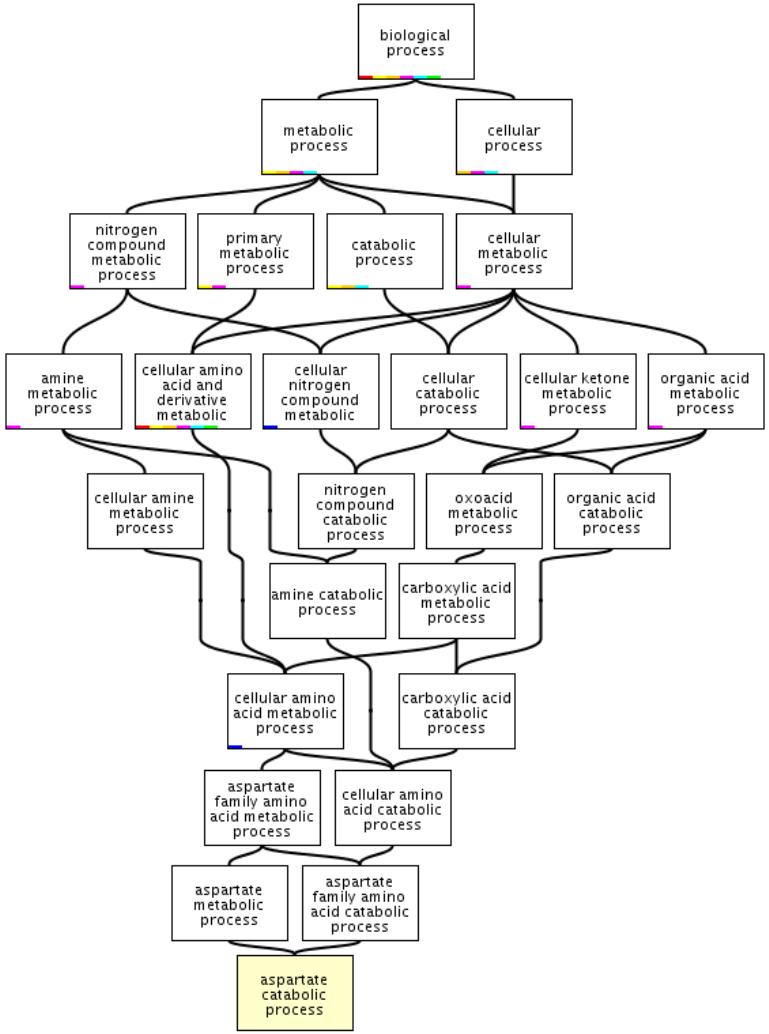
The structure of GO can be described in terms of a graph, where each GO term is a node, and the relationships between the terms are arcs between the nodes. The relationships used in GO are **directed**—for example, a mitochondrion *is an* organelle, but an organelle is not a mitochondrion—and the graph 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.

# Gene Ontology classification:
# E.G: The human cytoplasmic aspartate transaminase

## *Molecular Function*

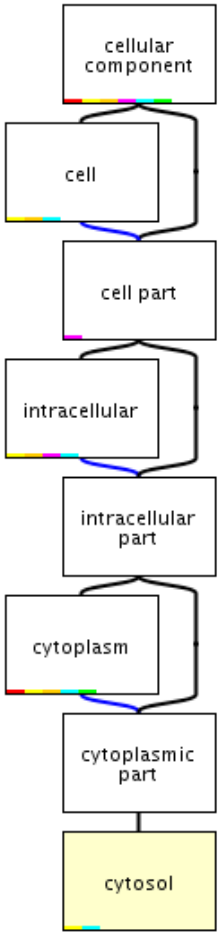

**GO:0004069**

## *Biological Process*




**GO:0006533**

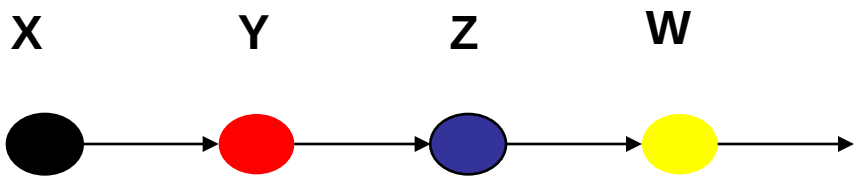
## *Cellular component*



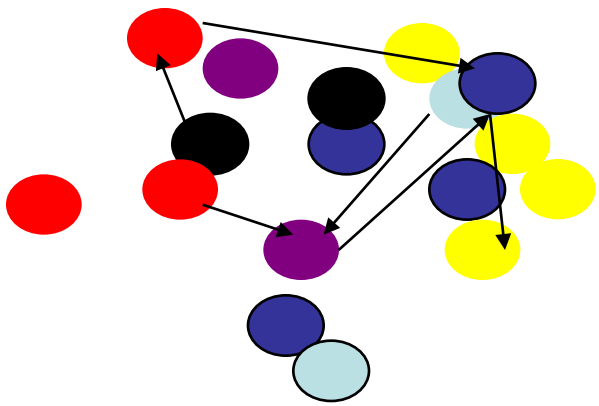

**GO:0005829**

# What did we learn:

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




**from linear**

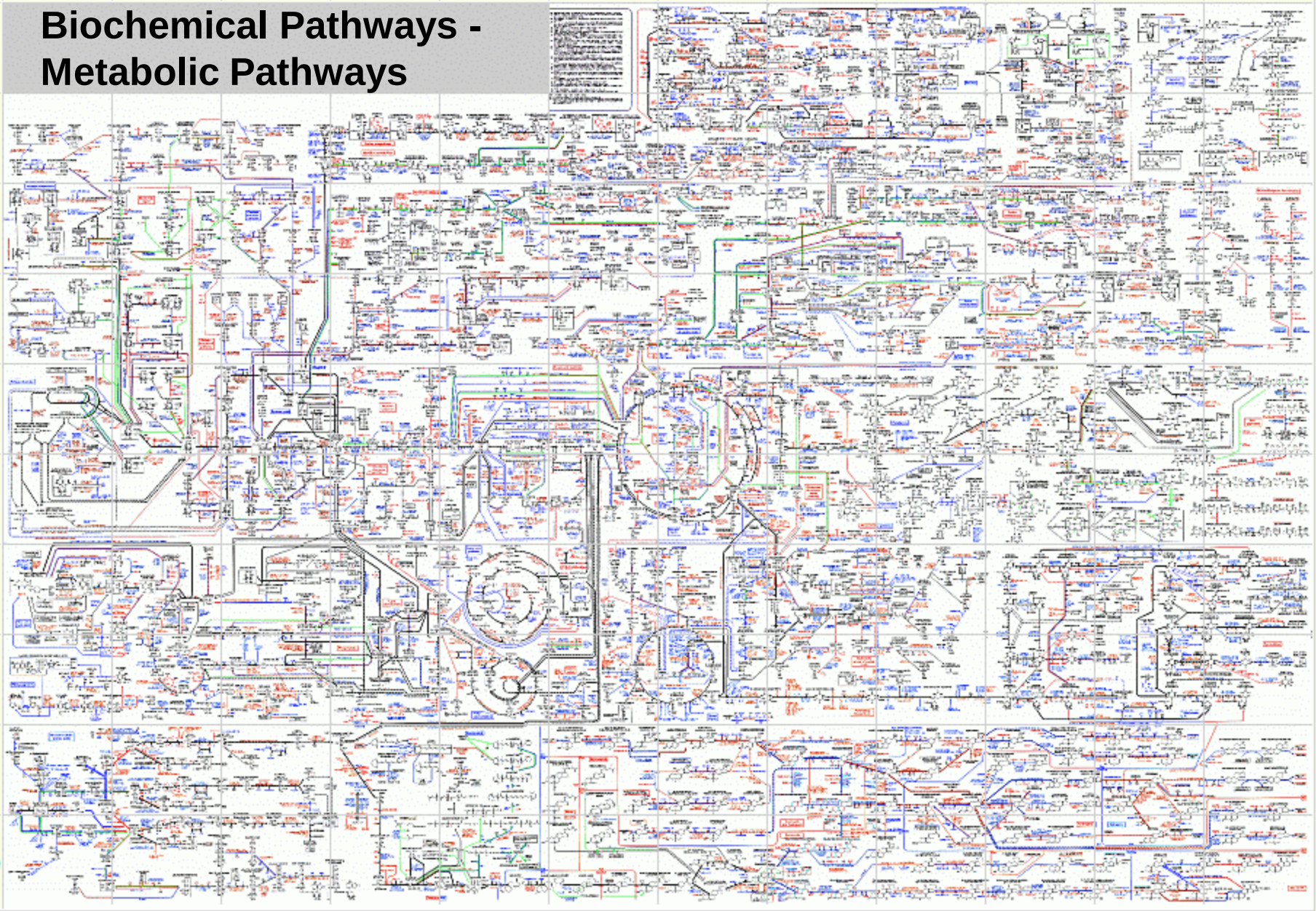


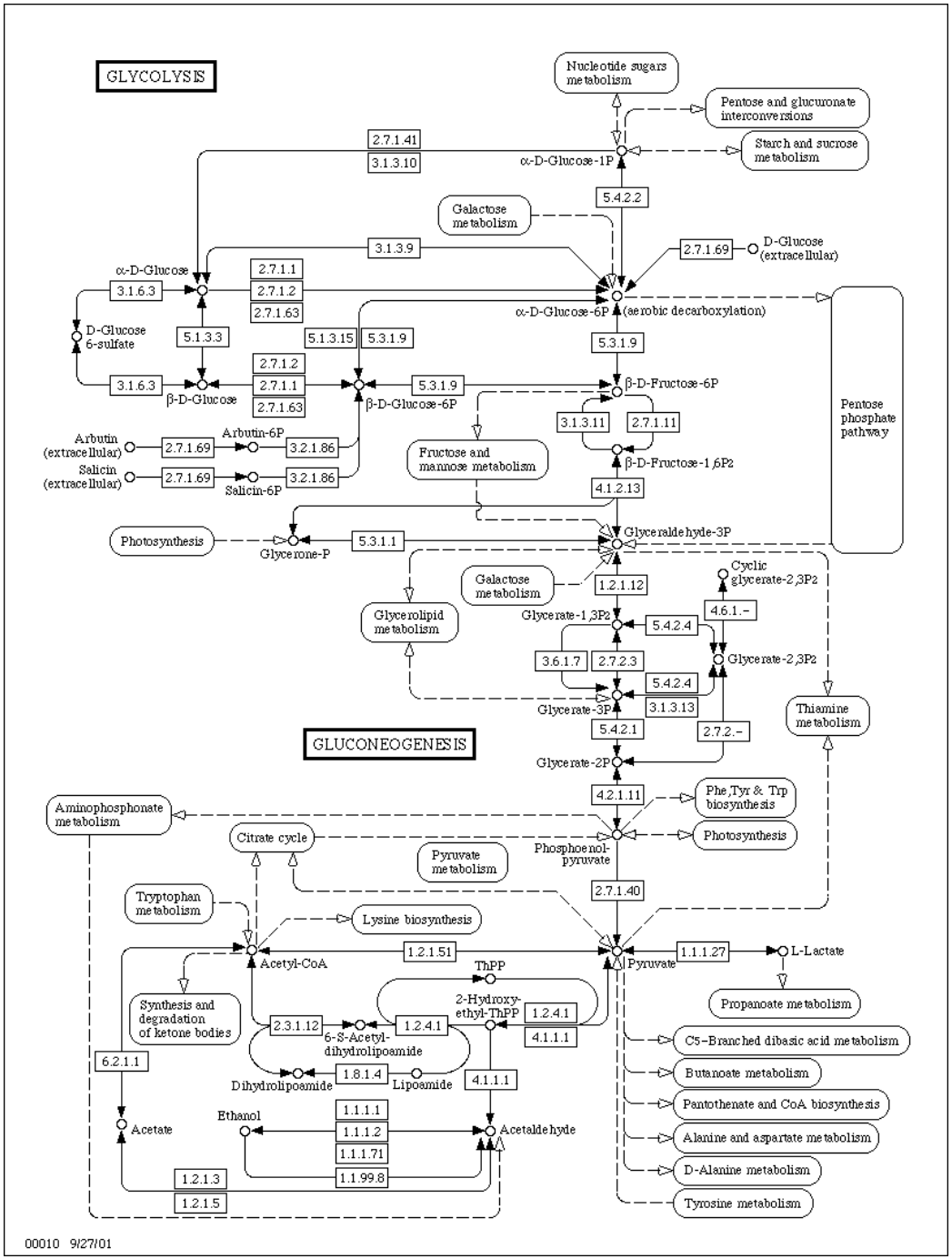
**to network models**

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**)

http://www.expasy.ch/sprot/

# Biochemical Pathways - Metabolic Pathways






# Metabolic Pathways: chemicals and protein interactions

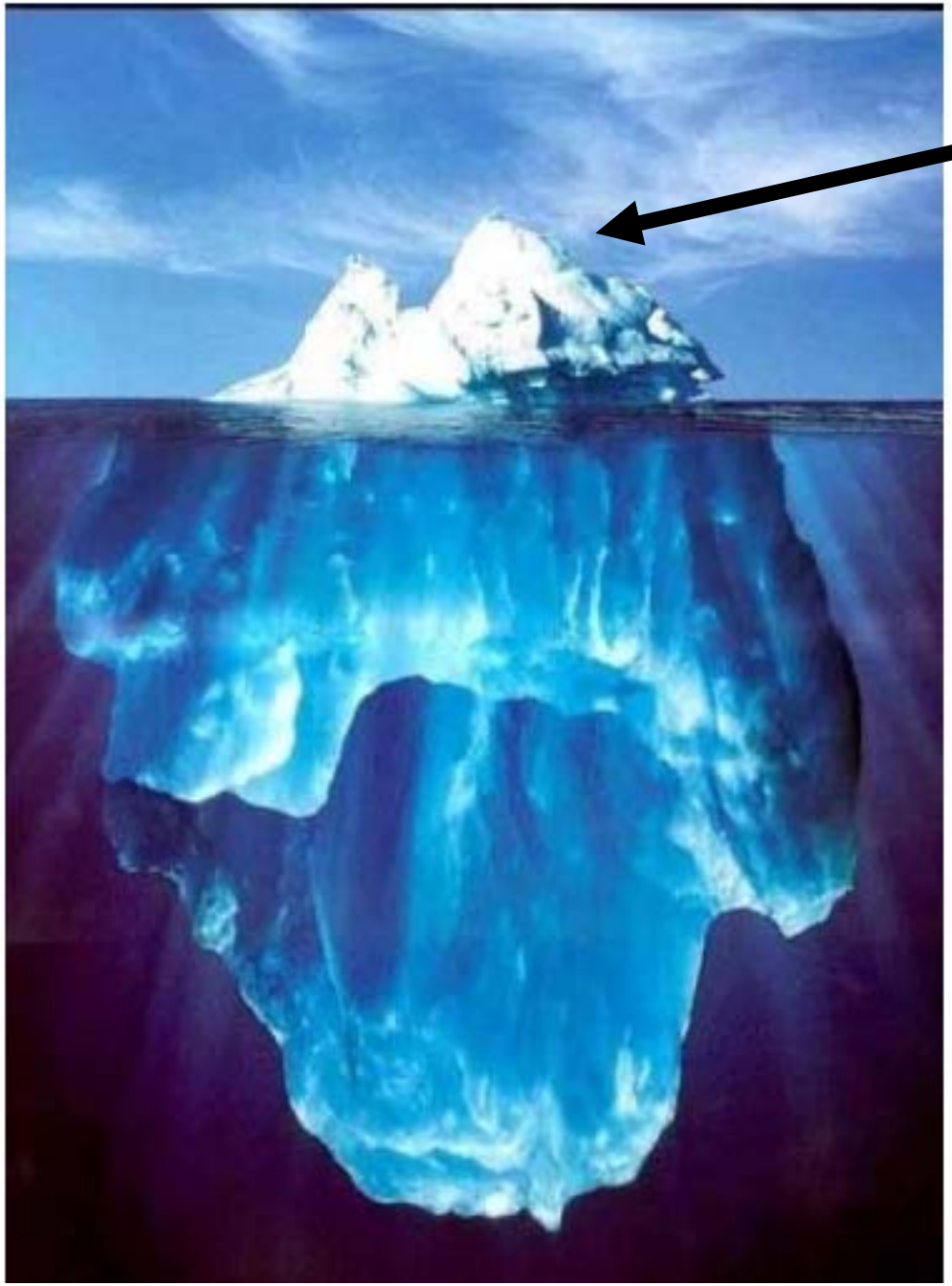
## E.G: Glycolysis and Gluconeogenesis

**KEGG:**

**Kyoto Encyclopedia of Genes and Genomes**

**http://www.genome.jp/kegg/**

# DATA -INTEGRATION



## The "omic" era

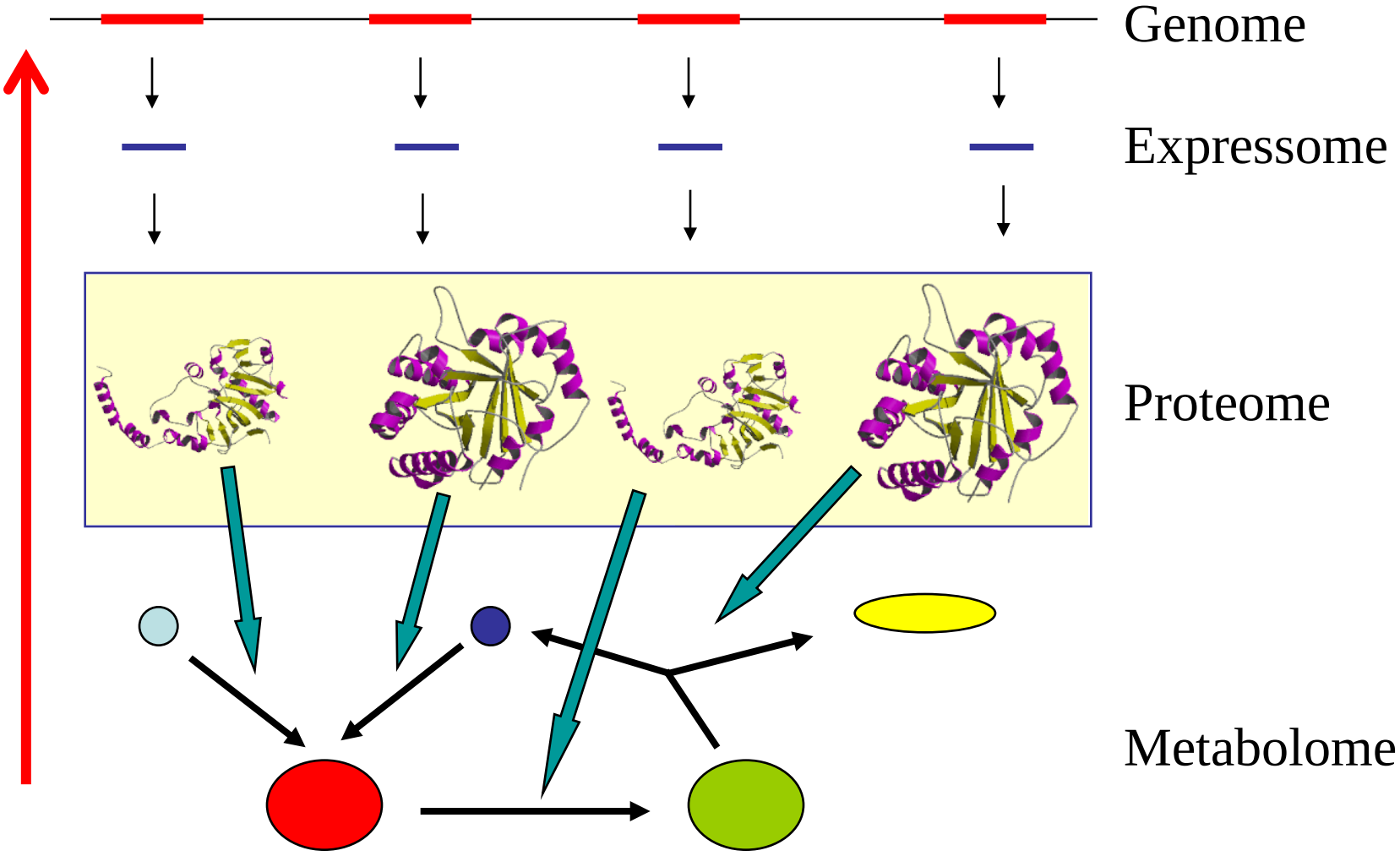
Complexity

- Genomics
- Transcriptomics
- Proteomics
- Metabolomics
- Regulomics

Systems Biology

# Functional Genomics

*From genes to functions and backward*



# Genomic Data Sources:towards cell modelling in silico

## *Vertical Genomics*

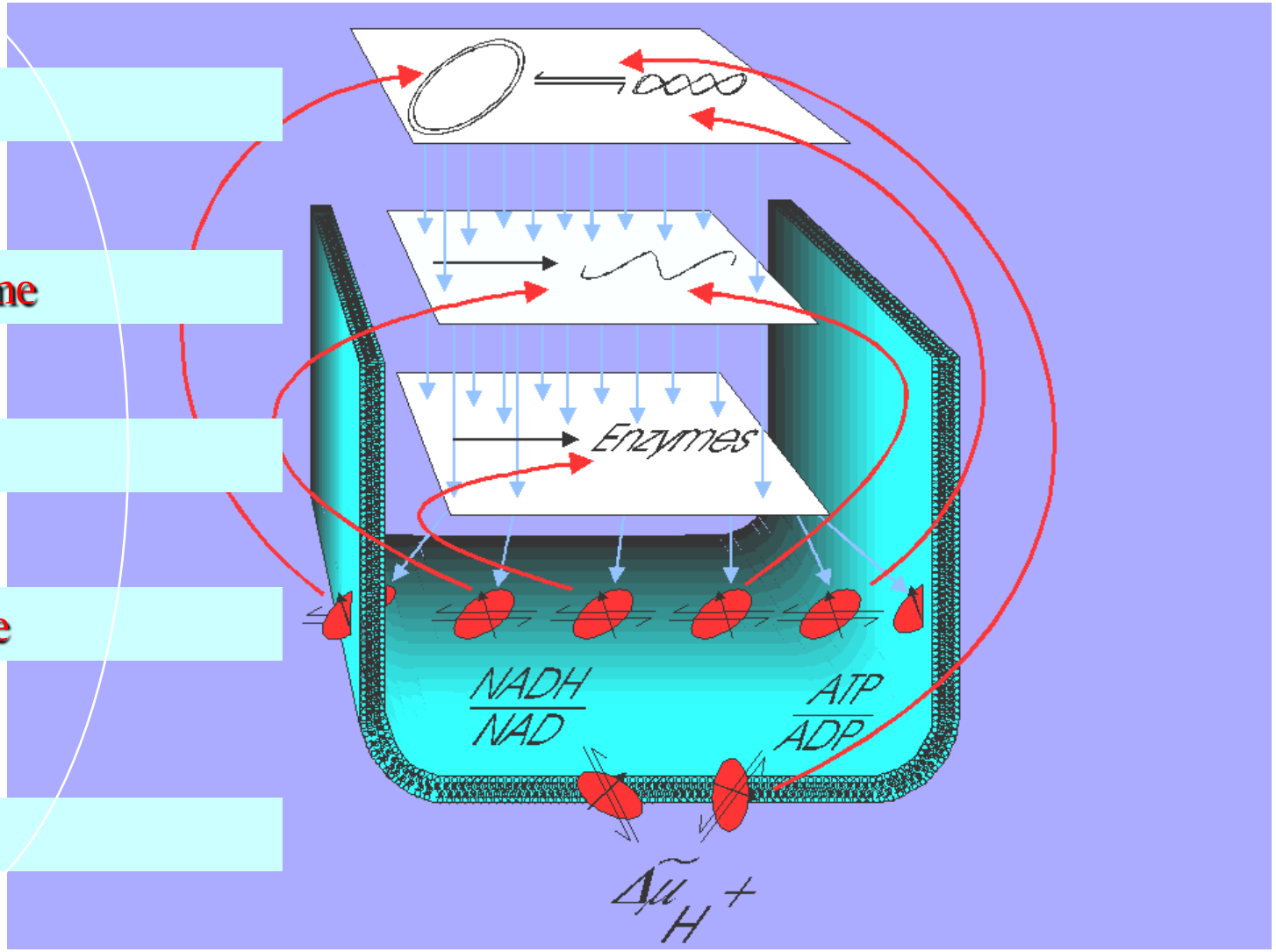
genome

transcriptome

proteome

metabolome

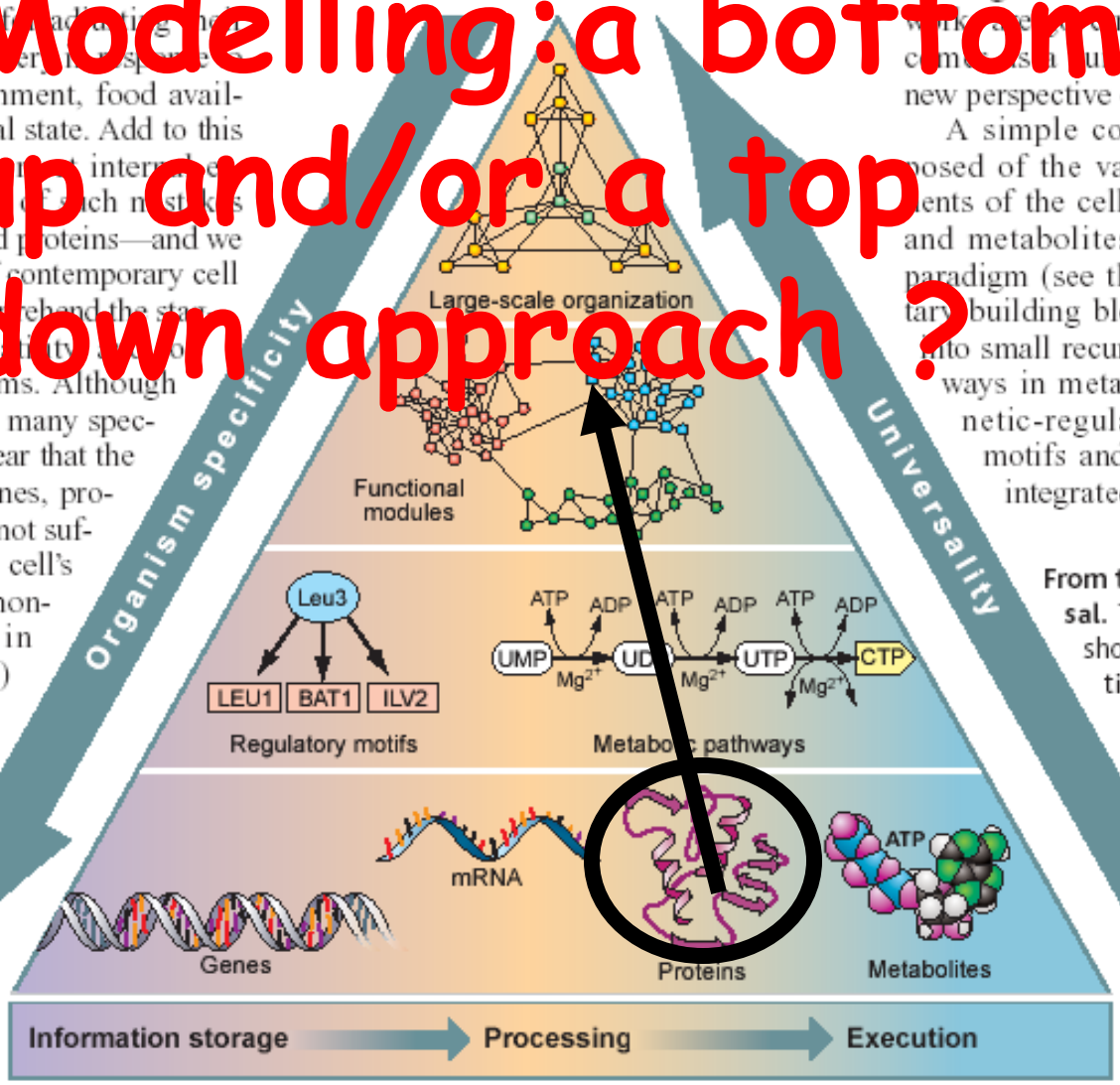
physiome



# Life's Complexity Pyramid

Zoltán N. Oltvai and Albert-László Barabási

Modelling: a bottom-up and/or a top-down approach ?




BOTH BUT IN GENERAL A BOTTOM-UP

# A "BIG" problem of the "omic era" after genome sequencing:

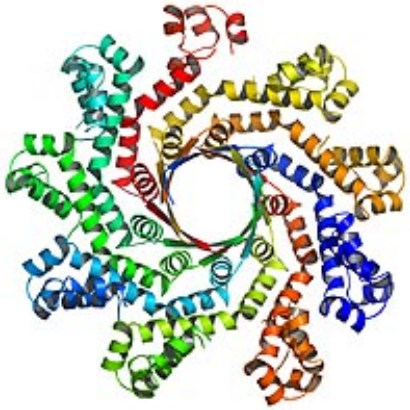
*code for*

*Genomes (~3000)*

```
        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 (~17 millions)*

*that are endowed with*



*Protein structures and functions*

## *Protein sequence Annotation:*

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

# 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***

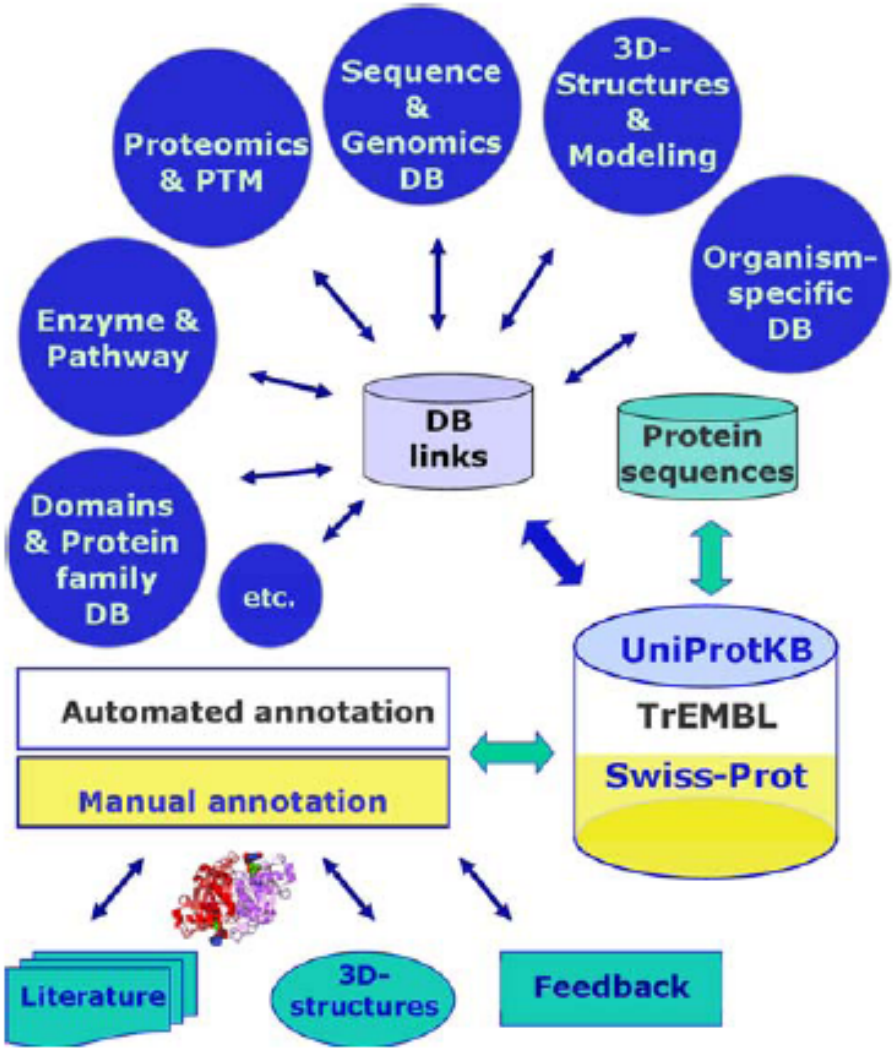



**Fig. 1** UniProtKB serves as a knowledge repository and as a central hub that provides links to numerous other databases. New protein sequences are integrated in UniProtKB/TrEMBL and annotated by an automated procedure. UniProtKB/Swiss-Prot entries are manually annotated, combining carefully checked protein sequences with information from the scientific literature, protein 3D-structures, and specialised databases, together with feedback from the scientific community

# UniProt KB: The largest annotation resource

Ursula Hinz • *The UniProt Consortium*
Cell. Mol. Life Sci. (2010) 67:1049–1064

# The UniProt Universe

UniProt - Mozilla Firefox

File Edit View History Bookmarks Yahoo! Tools Help

Biocomputing Group - University of Bologna | UniProt

www.uniprot.org

UNIPROT SEARCH

UniProt

Search | Blast | Align | Retrieve | ID Mapping

Search in: Protein Knowledgebase (UniProtKB)

Query: Search | Advanced Search » | Clear

## WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

### What we provide

| | |
|---|---|
| UniProtKB | Protein knowledgebase, consists of two sections:<br>Swiss-Prot, which is manually annotated and reviewed.<br>TrEMBL, which is automatically annotated and is **not** reviewed.<br>Includes complete and reference proteome sets. |

## NEWS

**UniProt release 2011_12 - Dec 14, 2011**

Between Charybdis and Cilia | Cross-references to PATRIC and DMDM

› Statistics for UniProtKB:
Swiss-Prot · TrEMBL

› Forthcoming changes

› News archives

Follow @uniprot 167 followers

## SITE TOUR

# 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**

## Release 2011_11 of 16-Nov-2011 of UniProtKB/TrEMBL contains 18,215,214 sequence entries

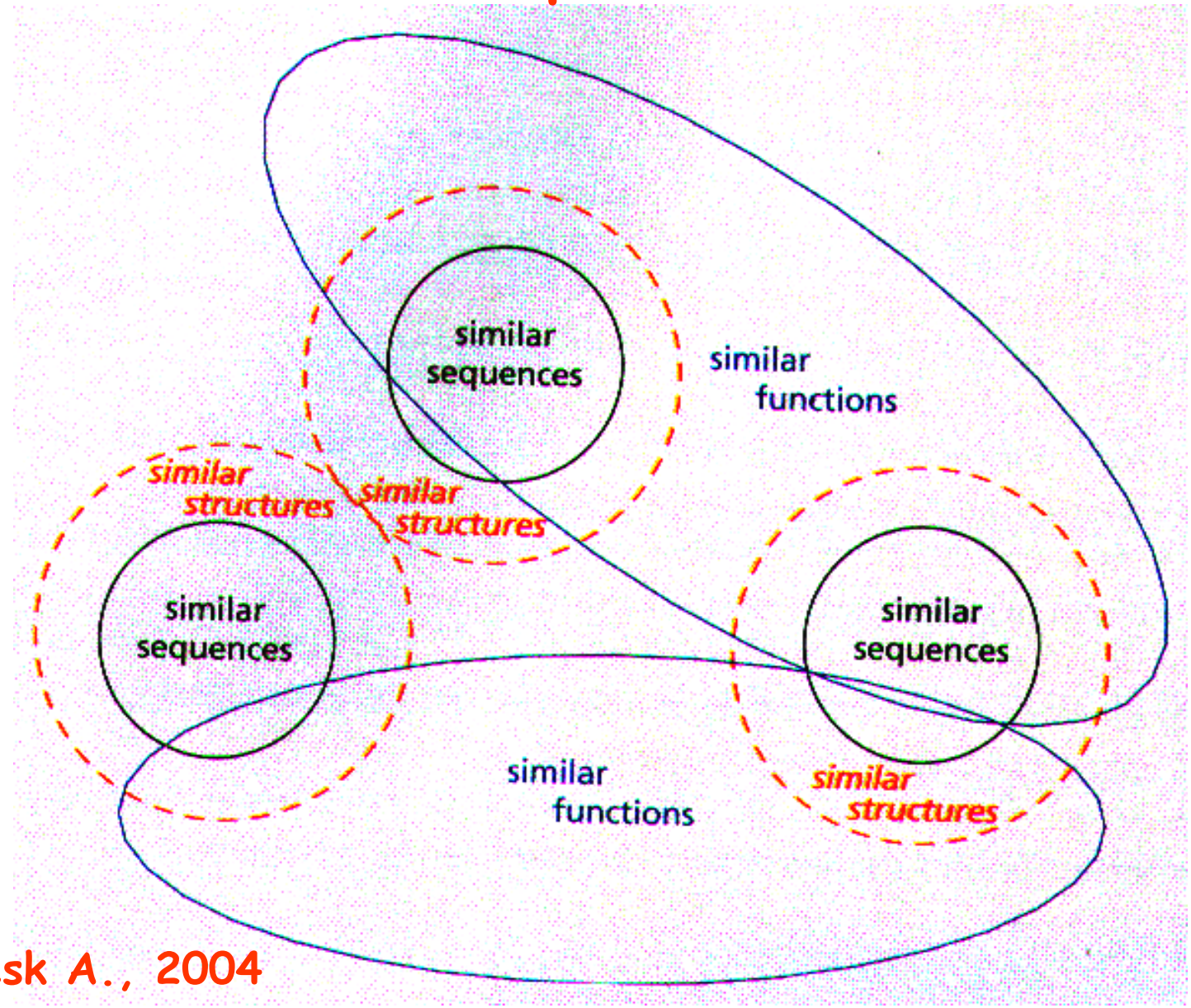
| Protein existence (PE): | entries | |
|---|---|---|
| 1: Evidence at protein level | 13085 | 0.07% |
| 2: Evidence at transcript level | 547306 | 3.00% |
| 3: Inferred from homology | 3857630 | 21.18% |
| 4: Predicted | 13797193 | 75.75% |

## Release 2011_11 of 16-Nov-11 of UniProtKB/Swiss-Prot contains 533,049 sequence entries

| Protein existence (PE): | entries | |
|---|---|---|
| 1: Evidence at protein level | 73298 | 13.8% |
| 2: Evidence at transcript level | 69925 | 13.1% |
| 3: Inferred from homology | 373485 | 70.1% |
| 4: Predicted | 14452 | 2.7% |
| 5: Uncertain | 1889 | 0.4% |

**Only 3.4 % sequences has evidence at the protein and trascript level and only 0.4 % proteins have structures in the Protein Data Bank.**

# How can we infer function and structure from sequence?




Lesk A., 2004

Summing up…..

Open problems in the genome era after DNA sequencing

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