



PERSPECTIVES OF DATABASES AND PROGRAM TOOLS DEVELOPMENT IN BIOINFORMATICS

Tetyana V. Barannik

V. N. Karazin Kharkiv National University

Biochemistry department

tbarannik@univer.kharkov.ua

Outline of talk



- ▣ **Biological data growth in post-genomic era**
- ▣ **Bioinformatics resources development**
- ▣ **New IT solutions for bioinformatics**



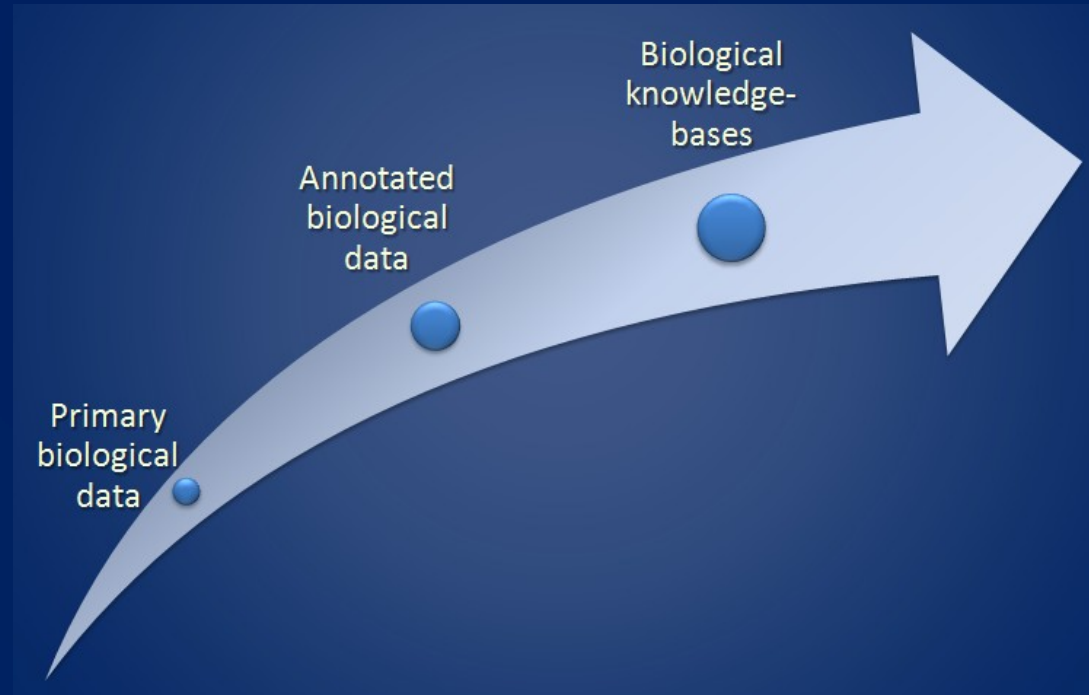
*“Bioinformatics is the science of managing, mining, and interpreting information from biological data” **

*- <http://bio.informatics.iupui.edu/biokdd10/>

BIOLOGICAL DATA GROWTH IN POST-GENOMIC ERA

Biological data growth in post-genomic era

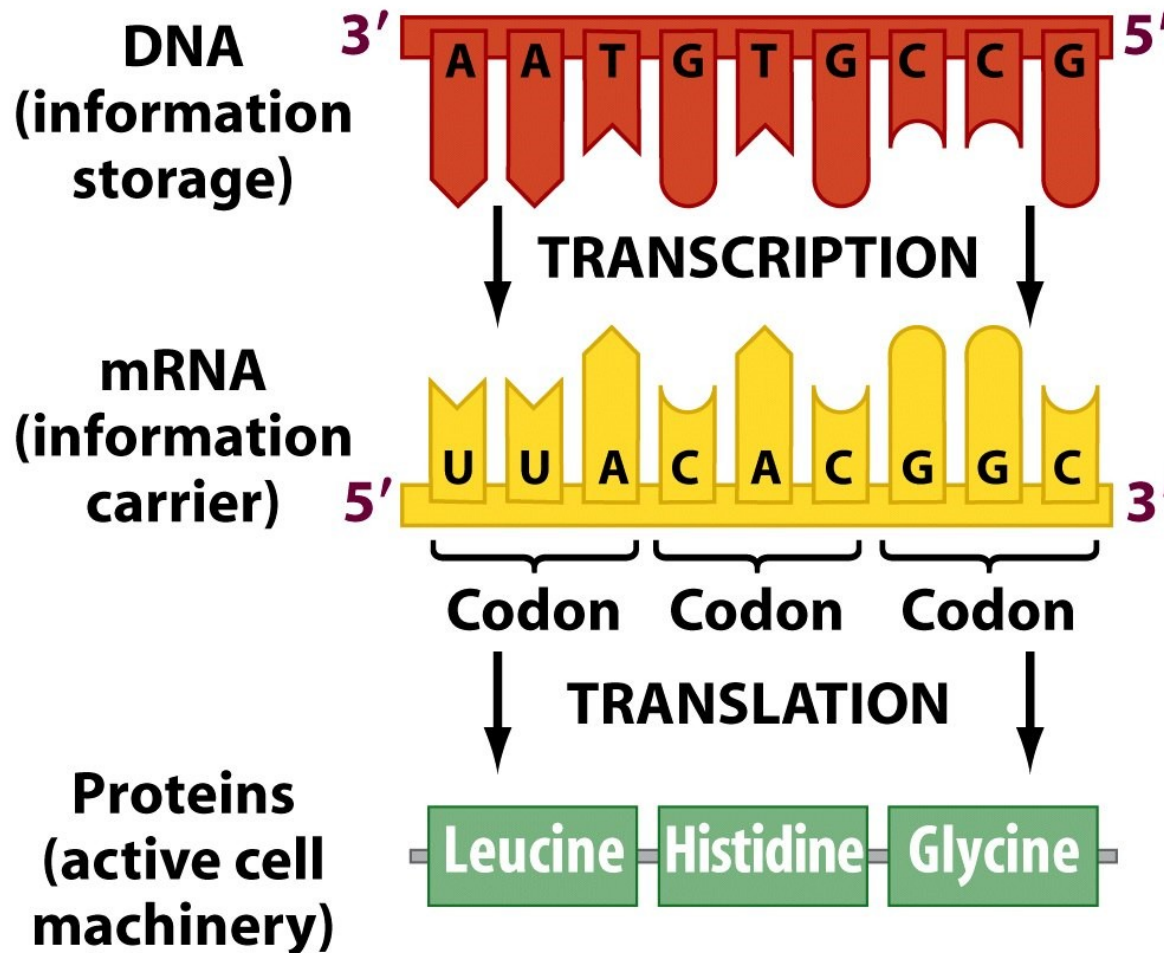
- Information flow
- New 'omics' era
- Worldwide projects



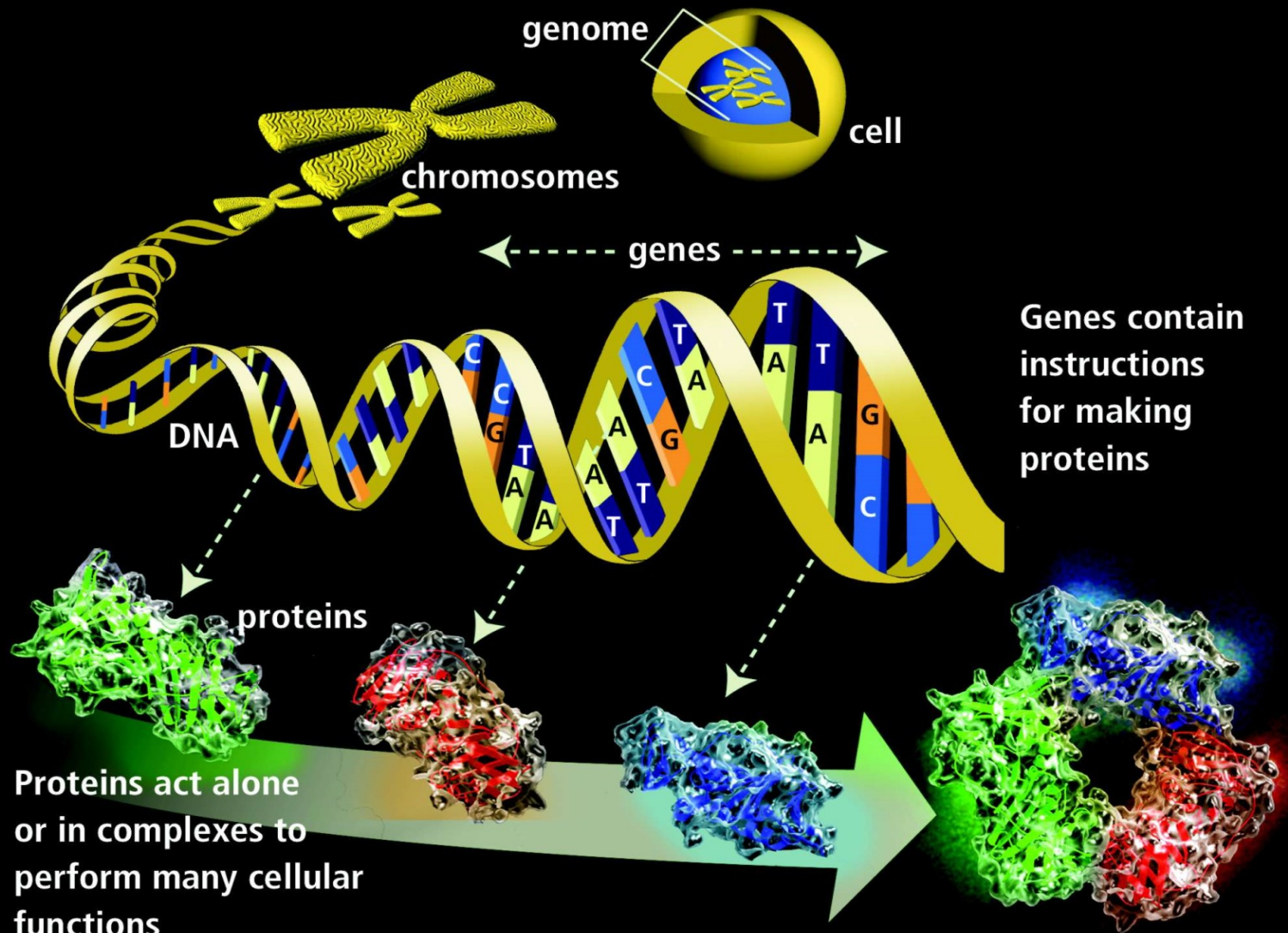
Biological information carriers

Object	Polymer	Monomers	Types of mono- mers	Approx. length (nbases)	Biological function	Template for synthesis
Gene	DNA	(deoxyribo)- nucleotides	4 (AGCT)	10,000 - 100,000	Information storage	Compleme ntary DNA strand
Transcript	RNA	(ribo)- nucleotides	4 (AGCU)	?<50 - 10,000	Information messenger, catalysis, regulation	Gene (1 strand of DNA region)
Protein	Protein = poly- peptide	Amino acids	20 + modified	50-1,000 or more	Catalysis, transport, regulation, movement, cytoskeleton, etc.	mRNA (processed transcript)

Information flows from DNA to RNA to proteins.

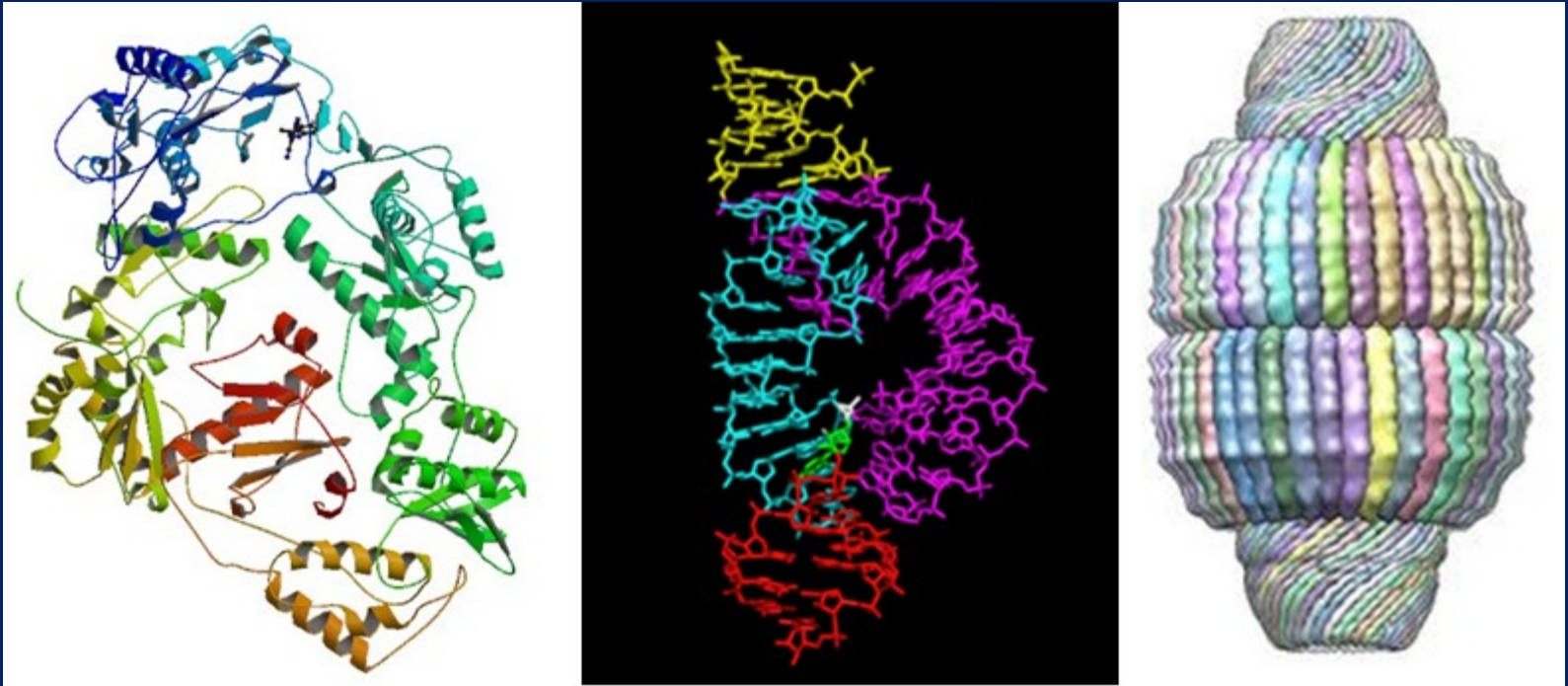


4^3 types of triplets
(codons)=
61 coding (for 20
types of AAs)+
3 stops



>gi|11321561|ref|NP_000604.1| hemopexin precursor [Homo sapiens]
MARVLGAPVALGLWSLCWSLA IATPLPPTSAHG NVAEGETKPD PDVTERCSDGWSFDATTLDDNGTMLFF
KGEFVWKSHKWDRELISERWKNFPSPVDA AFRQGHNSVFLIKGDKVWVYPPEKKEKGYPKLLQDEFFGIP
SPLDAAVECHRGECQAEGVLFFQGDREWFWDLATGTMKERSWPAVGNCSALRWLG RYYCFQGNQFLRF
PVRGEVPPRYPRDVRDYFMPCPGRGHGHRNGTGHGNS THHGPEYMRCSPHLVLSALTSDNHGATYA FSGT
HYWRLDTSRDGWSWPIAHQWPQGSAVDAA FSWEEKLYLVQGTQVYVFLT KGGYTLVSGYPKRLEKEVG
TPHGIILDSVDAAFICPGSSRLHIMAGRRLWMLDLKSGA QATWTLPWPHEKVDGALCMEKSLGPNSCSA
NGPGLYL IHGPNLYCYSDVEKLNAAKALPOPQNVTSLLGCTH

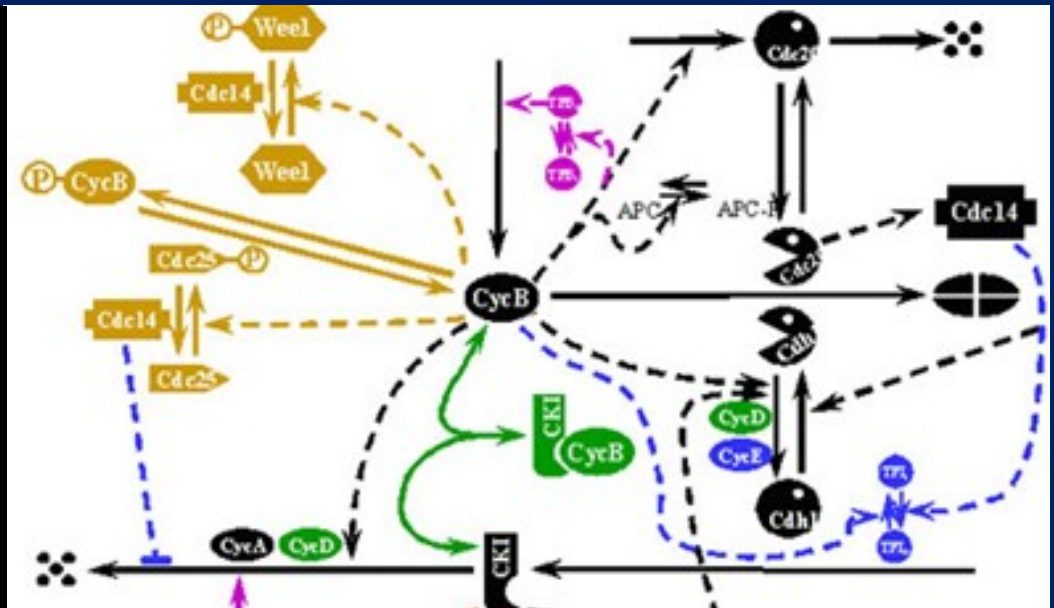
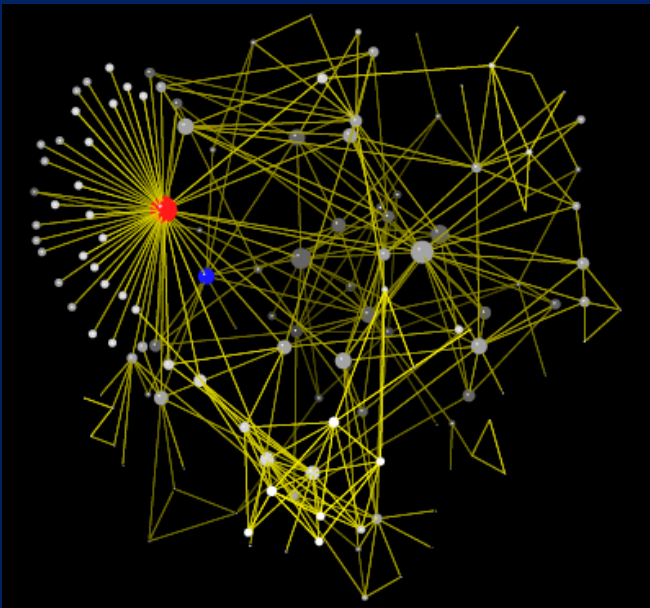
3D biological Information : structure



Proteins and RNAs

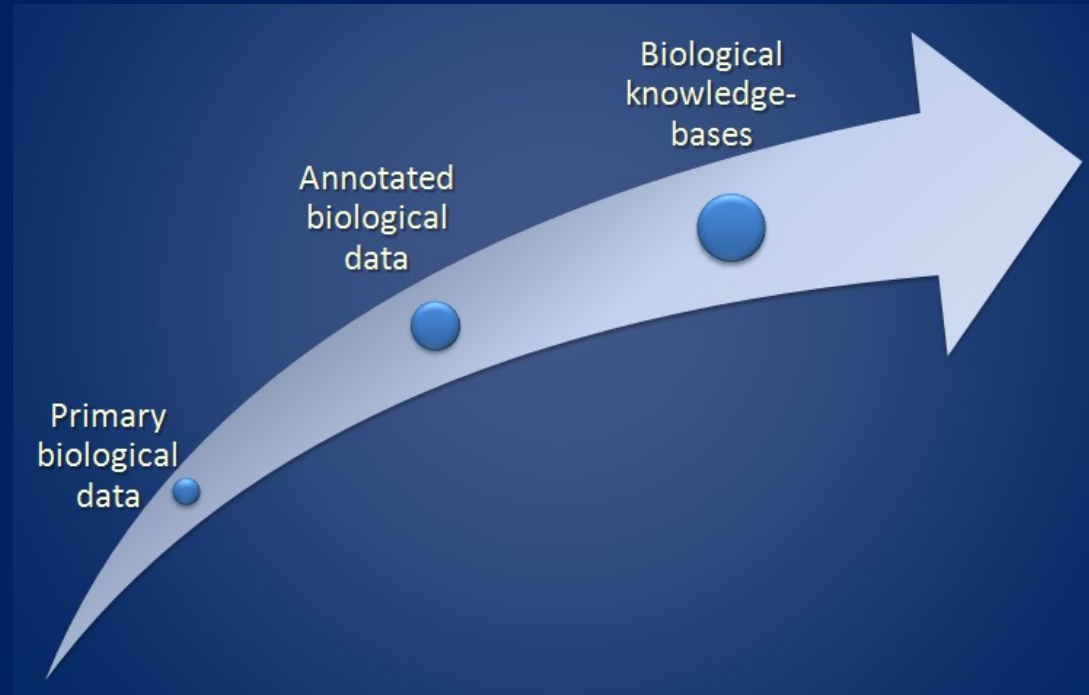
Sequence determines structure

4D biological Information : networks, metabolites flows

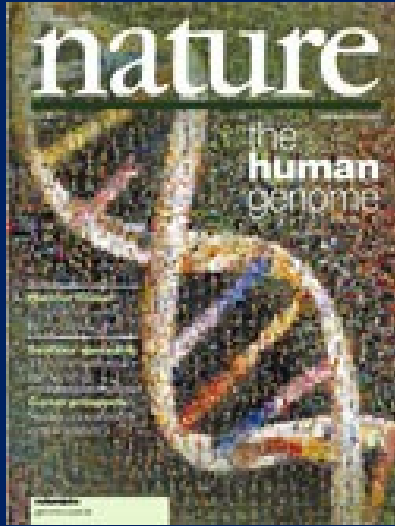


Biological data growth in post-genomic era

- Information flow
- New 'omics' era
- Worldwide projects



2001: publication of human genome



Nature **409**, 745
(15 February 2001)



Science 16 February 2001:
Vol. 291, pp. 1304 - 1351

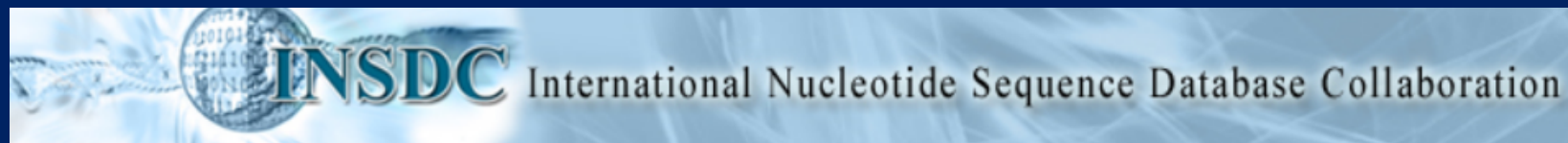
Pre-genomic → Post-genomic era

Bioinformatics analysis for 'omics' era in biology

'omics'	Object for study	Types of analysis
Genomics	Genome (full set of genes)	Genome assembly, gene finding, gene structure analysis, regulators identification, finding of sites of regulators binding with DNA
Transcriptomics	Transcriptome (full set of transcripts: mRNAs, rRNAs, tRNAs, ncRNAs)	Analysis of gene expression, RNA processing, splicing isoforms , RNA structure, ncRNAs functions
Proteomics	Proteome (entirety of proteins)	Proteome analysis, identification of proteins, analysis and prediction of protein structure and interactions
Metabolomics	Metabolome (set of metabolites, regulators)	Small molecules identification, analysis of their transformations

International Nucleotide Sequence Database Collaboration

- GenBank (USA)
- EMBL-Bank (Europe)
- DDBJ (DNA Databank of Japan)



ABOUT INSDC

POLICY

ADVISORS

DOCUMENTS

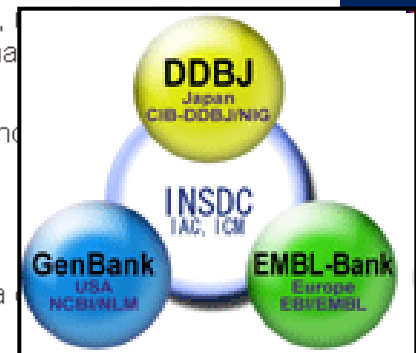
<http://www.insdc.org/>

International Nucleotide Sequence Database Collaboration

- The International Nucleotide Sequence Databases (INSD) have been developed and maintained collaboratively between [DDBJ](#), [ENA](#), and [GenBank](#) for over 18 years.
- The INSDC advisory board, the [International Advisory Committee](#), is made up of members of each of the databases' advisory bodies. At their most recent meeting, committee unanimously endorsed and reaffirmed the existing data-sharing policy of the databases that make up the INSDC, which is stated below.
- Individuals submitting data to the international sequence databases should follow the [policy](#).

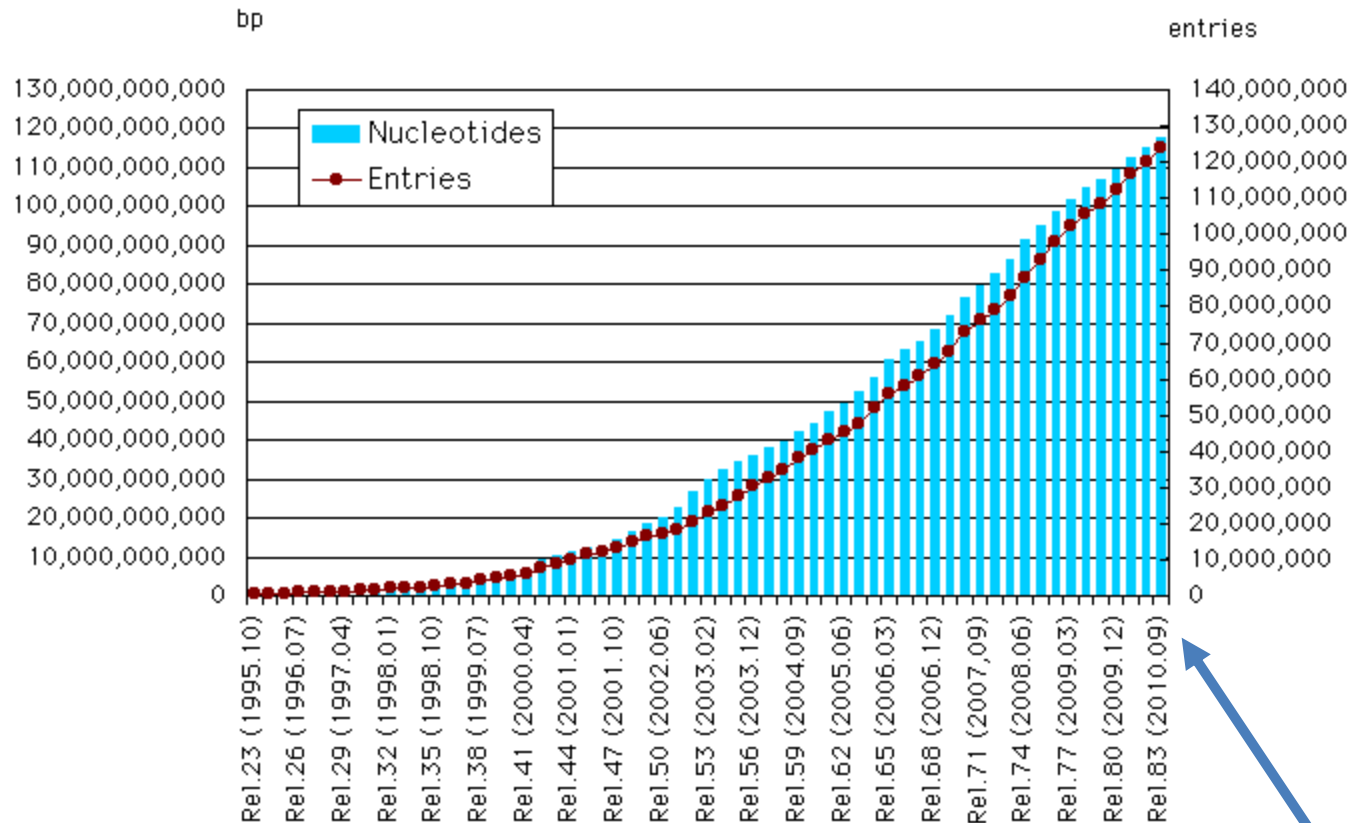
How to submit data

- For full details of how to submit data to the databases, please select a database from the [list](#).
- [DDBJ](#), [ENA](#), [GenBank](#)
- The INSDC Feature Table Definition Document is available [here](#).



http://www.ddbj.nig.ac.jp/images/breakdown_stats/DBGrowth-e.gif

DDBJ/EMBL/GenBank database growth

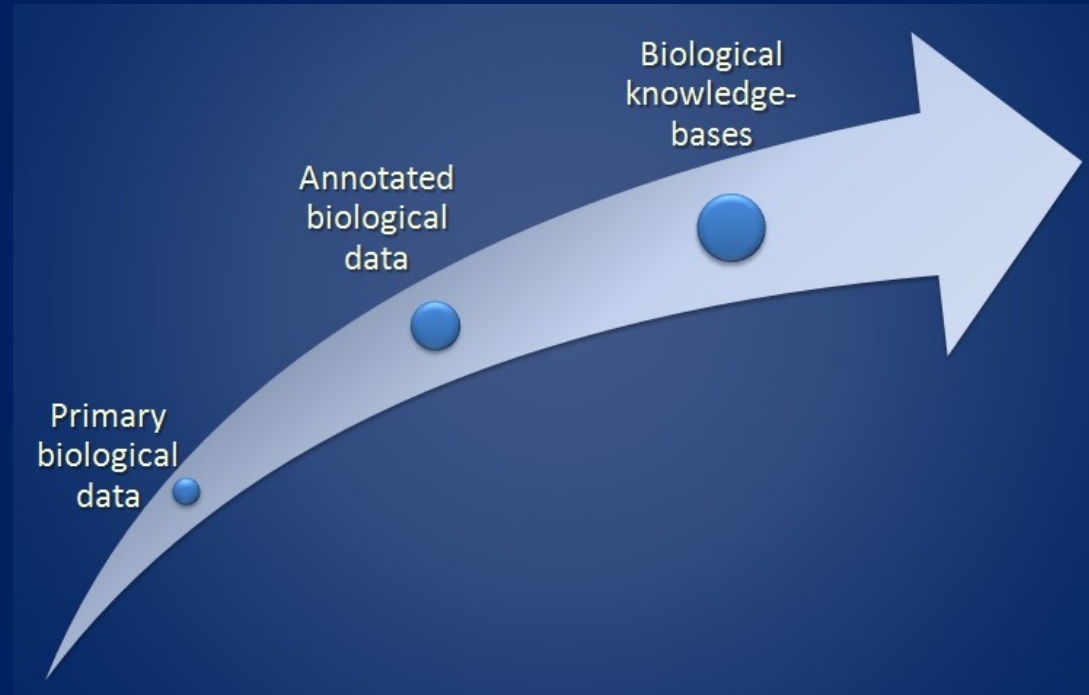


* Note : CON division is not counted in statistics of DDBJ periodical releases.

September 2010

Biological data growth in post-genomic era

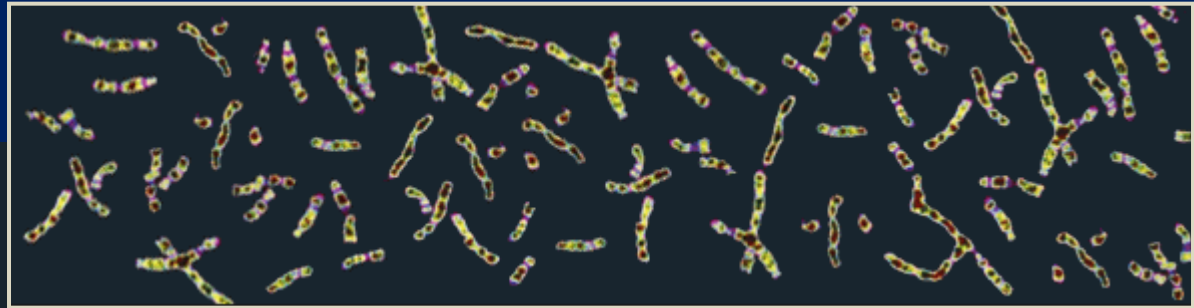
- Information flow
- New 'omics' era
- Worldwide projects



http://www.1000genomes.org/page.php

1000 Genomes

A Deep Catalog of Human Genetic Variation



[Home](#) [About](#) [Data](#) [Analysis](#) [Participants](#) [Contact](#) [Browser](#) [Wiki](#)

LATEST ANNOUNCEMENTS

July 2010 Data Release

20 JULY 2010

Pilot Project Variant call release

Variant Calls from the three pilot projects are now available in VCF 4.0 format. This release includes SNPs, short indels and large scale structural variants. All 1000 genomes pilot project files reference the NCBI build 36 assembly of the human genome

Data access links: [EBI](#) / [NCBI](#)

Link to additional information: [README file](#)

Recent project announcements

4 AUGUST 2010 [New sequence data is available](#)

The latest release of sequence data from the 1000 Genomes full project is now available. The new sequence.index file can be found at: [20100804.sequence.index](#)

Data access links: [EBI](#) / [NCBI](#) / [Instructions for data download and Aspera](#)

Links to additional information: [List of new index and statistics files](#) / [Sequence index file format](#)

LOG IN

Username:

Password:

([Send me my password](#))

LINKS



[All Project Announcements](#)




[Sample and Project Information](#)




[Media Archive](#)

<http://www.sanger.ac.uk/genetics/CGP/>



wellcome trust
sanger
institute



RSS

[Information](#) | [Projects](#) | [Other Services](#)


Cancer Genome Project
Genomics & Genetics

The Cancer Genome Project


Summary

All cancers occur due to abnormalities in DNA sequence. Throughout life, the genome within cells of the human body is exposed to mutagens and suffers mistakes in replication. These corrosive influences result in progressive subtle divergence from the normal genome. The accumulation of additional mutations, and consequent waves of clonal expansion result in the evolution of the mutinous cells that invade surrounding tissues and metastasise. One in three people in the Western world develop cancer and one in five die of the disease. Cancer is therefore the commonest genetic disease.


Data Resources




Cancer Gene Census:
Mutated genes causally implicated in human cancer.




COSMIC:
Catalogue Of Somatic Mutations In Cancer



CGP Resequencing Studies:
Somatic mutations from systematic large scale resequencing of genes in human cancers.



CGP Cancer Cell Line Project:
Resequencing of known cancer genes and other analyses of human cancer cell lines.



CGP Copy Number Analysis in Cancer:
Analysis of copy number and loss of heterozygosity in cancer cell lines and primary tumours.

Projects

AutoCSA	Mutation detection software
DbCon	Database pooling, distributed configuration and SQL Libraries for Java
PICNIC	SNP6 Copy number Segmentation Tool
GRAFT	Rearrangement Phylogeny Tool

<http://www.hupo.org/>



Human Proteome Organisation

[Home](#) | [Search](#) | [Contact Us](#) | [Login](#)

☐ Overview

☐ HUPO Initiatives

☐ Meetings

☐ Educational Programs

☐ News & Highlights

☐ HUPO Journals

Fostering international proteomic
initiatives to better understand
human disease

Human Proteome Project
INFORMATION

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.

[Register for our Newsletter](#)

First Name:

<http://www.metabolomics.ca/>

Metabolomics Toolbox

[Contact](#) | [About](#) | [Personnel](#) | [Publications](#) | [SOPs](#) | [Jobs](#) | [Partners](#) | [News & Links](#) | [MetaboDatabase](#) | [MetaboLibrary](#) | [DrugBank](#)

[Home](#)

Welcome

Welcome to the official website of the Human Metabolome Project.

Our goal is to be the first group in the world to complete the human metabolome.

This large-scale and integrated effort will involve identifying and quantifying several hundred unknown metabolites in both human tissues and

The Human Metabolome Project

Metabolomics is a newborn cousin to genomics and proteomics. Specifically, metabolomics involves the rapid, high throughput analysis of the metabolites found in an organism. Since the metabolome is closely tied to the genotype of an organism, its physiology and its environment offers a unique opportunity to look at genotype-phenotype as well as genotype-environment relationships. Metabolomics is increasingly being applied in fields including pharmacology, pre-clinical drug trials, toxicology, transplant monitoring, newborn screening and clinical chemistry. However, the human metabolome is not at all well characterized.

Unlike the situation in genomics, where the human genome is now fully sequenced and freely accessible, metabolomics is not. Only a few endogenous or common metabolites that are detectable in the human body. Not all of these metabolites can be found in all tissues/biofluids serve different functions or have different metabolic roles. To date, the HMP has identified and quantified (i.e. characterized) 1122 metabolites in CSF, 1122 metabolites in serum, 458 metabolites in urine and approximately 300 metabolites in other tissues and biofluids. This is a desirable and this is one of the long term goals of the HMP and other affiliated metabolomic projects around the world.



Computational Biology needs HPC

Problems of scale

- Genomes with millions to billions of nucleotides
- Profiling experiments with tens of thousands of data points measured on hundreds or thousands of samples
- Thousands of protein mass spectra representing GigaBytes of data/experiment

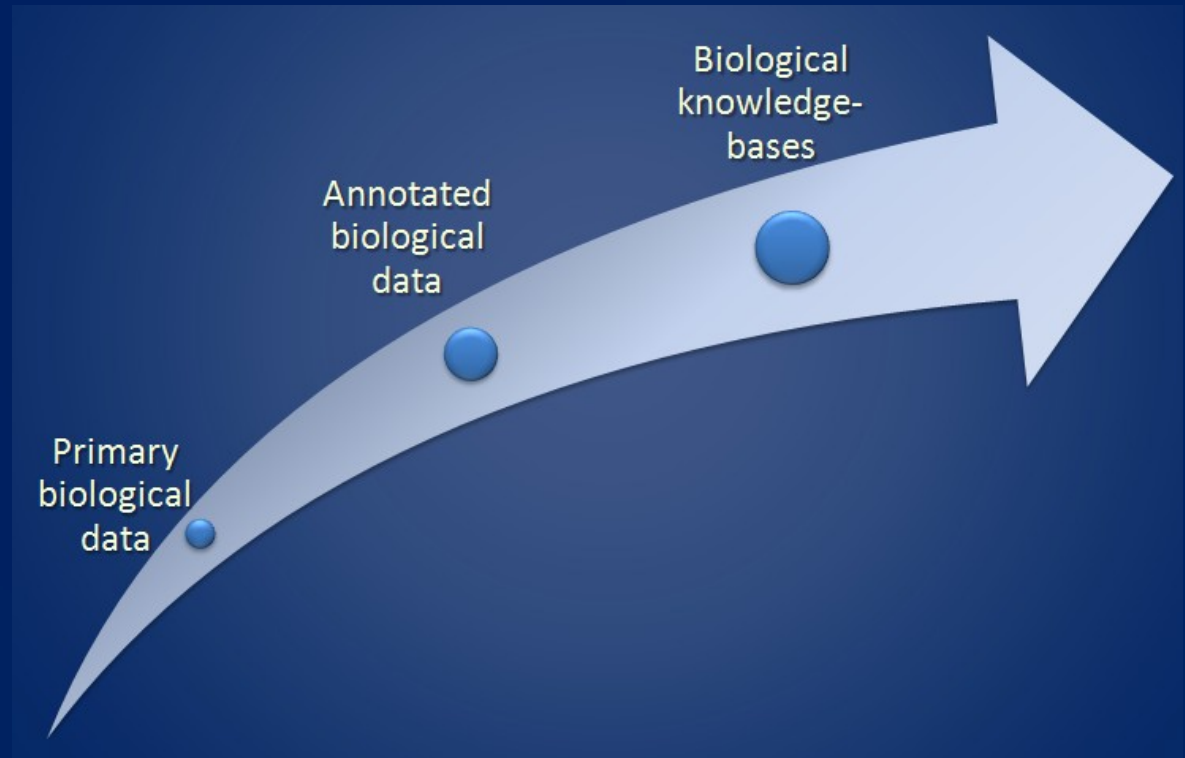
Problems of complexity

- ❖ Combinatorial: $>3 \cdot 10^4$ interacting gene products can create more functions than there are atoms in the Universe
- ❖ Structural: $>10^5$ dynamically interacting atoms make up the smallest of molecular machines

BIOINFORMATICS RESOURCES DEVELOPMENT

Bioinformatics resources development

- Resources variety
- Databases
- Program tools
- Search and workflows
- Towards knowledgebases



Bioinformatics resources

Databases

Annotated
genomic DBs
(on-line only)

DBs
specialized
by object
(organism,
organelle,
molecule),
disease, etc

(on-line &
downloads)

Genome
browsing,
sequences,
annotations
alignments,
references,
cross-links,
phylo-
genetics

Program tools

(on-line & downloads)

Linear
bio-
informatics

Structural
bio-
informatics

System
biology tools

'Omics'
raw data
analysis

Sequence
analysis:
alignments,
similarity
search,
phylogene-
tics trees

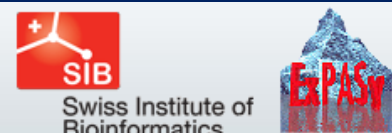
Structure
(proteins,
RNA)
analysis &
prediction,
modeling,
docking,
drug design

Metabolic
pathways
analysis,
networks
modeling,
virtual cell
simulations

Data
clustering,
identifi-
cation,
normali-
zation

Thematic collections of DB and tools

<http://www.expasy.org/>



ExPASy Proteomics Server

You are here: [ExPASy CH](#)

The ExPASy (**Expert Protein Analysis System**) [proteomics](#) server of the [Swiss Institute of Bioinformatics](#) (SIB) is dedicated to the analysis of protein sequences and [ExPASy](#).

Databases

[UniProtKB](#), [PROSITE](#), [HAMAP](#), [SwissVar](#), [ViralZone](#), [SWISS-MODEL Repository](#),
[SWISS-2DPAGE](#), [World-2DPAGE Repository](#), [MIAPEGelDB](#), [ENZYME](#),
[GlycoSuiteDB](#), [UniPathway](#)
[\[details\]](#) [\[full list\]](#)

Education & services

[Downloads](#), [Protein Spotlight](#), [Protéines à la «Une»](#), [e-proxemis](#), [Bioinformatics core facility for Proteomics](#), [Click2Drug - in silico Drug Design tools](#)
[\[full list\]](#)

Tools & Software

[Proteomics tools](#), [Blast](#), [ScanProsite](#), [Melanie](#), [MSight](#), [Make2D-DB](#),
[SWISS-MODEL](#), [Swiss-PdbViewer](#), [SwissDock](#), [SwissParam](#)
[\[full list\]](#)

Documentation

[What's New?](#), [E-mail alerts](#), [UniProtKB documentation](#), [How to link to ExPASy](#), [Advanced search](#)
[\[full list\]](#)

Web servers annual collection

http://nar.oxfordjournals.org/content/38/suppl_2

Contents

Volume 38, Web Server issue, July 1, 2010

Editorial: *Nucleic Acids Research* Annual Web Server Issue in 2010

G.Benson

W1–W2

Providing web servers and training in Bioinformatics: 2010 update on the Bioinformatics Links Directory

S TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations

S ALTER: program-oriented conversion of DNA and protein alignments

DIALIGN-TX and multiple protein alignment using secondary structure information at GOBICS

GUIDANCE: a web server for assessing alignment confidence scores

SATCHMO-JS: a webserver for simultaneous protein multiple sequence alignment and phylogenetic tree construction

Multi-Harmony: detecting functional specificity from sequence alignment

ALADYN: a web server for aligning proteins by

D.Glez-Peña, D.Gómez-Blanco, M.Reboiro-Jato, F.Fdez-Riverola and D.Posada

W14–W18

A.R.Subramanian, S.Hiran, R.Steinkamp, P.Meinicke, E.Corel and B.Morgenstern

O.Penn, E.Privman, H.Ashkenazy, G.La D.Graur and T.Pupko

R.Hagopian, J.R.Davidson, R.S.Datta, G.R.Jarvis and K.Sjölander

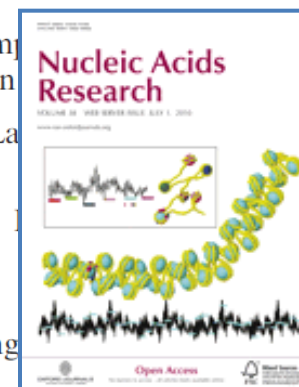
B.W.Brandt, K.A.Feenstra and J.Hering

R.Potestio, T.Aleksiev, F.Pontiggia, S.Cozzini and

W41–W45


Nucleic Acids Research

NEW IMPACT FACTOR OF 7.479



Annual bioinformatics links collections:

http://bioinformatics.ca/links_directory/



bioinformatics.ca
links directory

Bioinformatics Links Directory

The Bioinformatics Links Directory features curated links to molecular resources, tools and databases. The links listed in this directory are selected on the basis of recommendations from bioinformatics experts in the field. We also rely on input from our community of bioinformatics users for suggestions. Starting in 2003, we have also started listing all links contained in the NAR Webserver issue.

Bioinformatics Links Directory

Computer Related (76)

This category contains links to resources relating to programming languages often used in bioinformatics. Other tools of the trade, such as web development and database resources, are also included here.

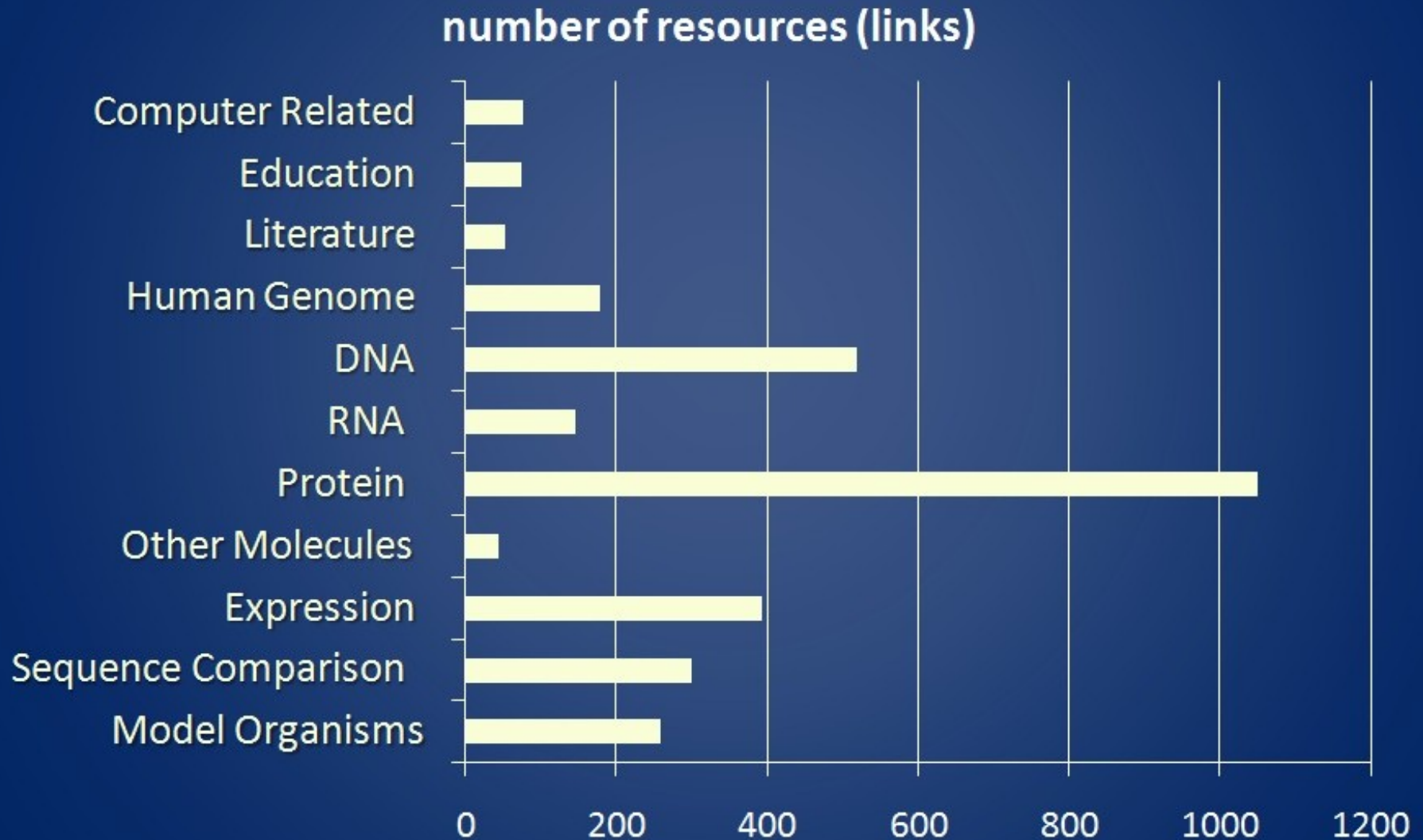
DNA (520)

This category contains links to useful resources for DNA sequence analyses such as tools for comparative sequence analysis and sequence assembly. Links to programs for sequence manipulation, primer design, and sequence retrieval and submission are also listed here.

Education (74)

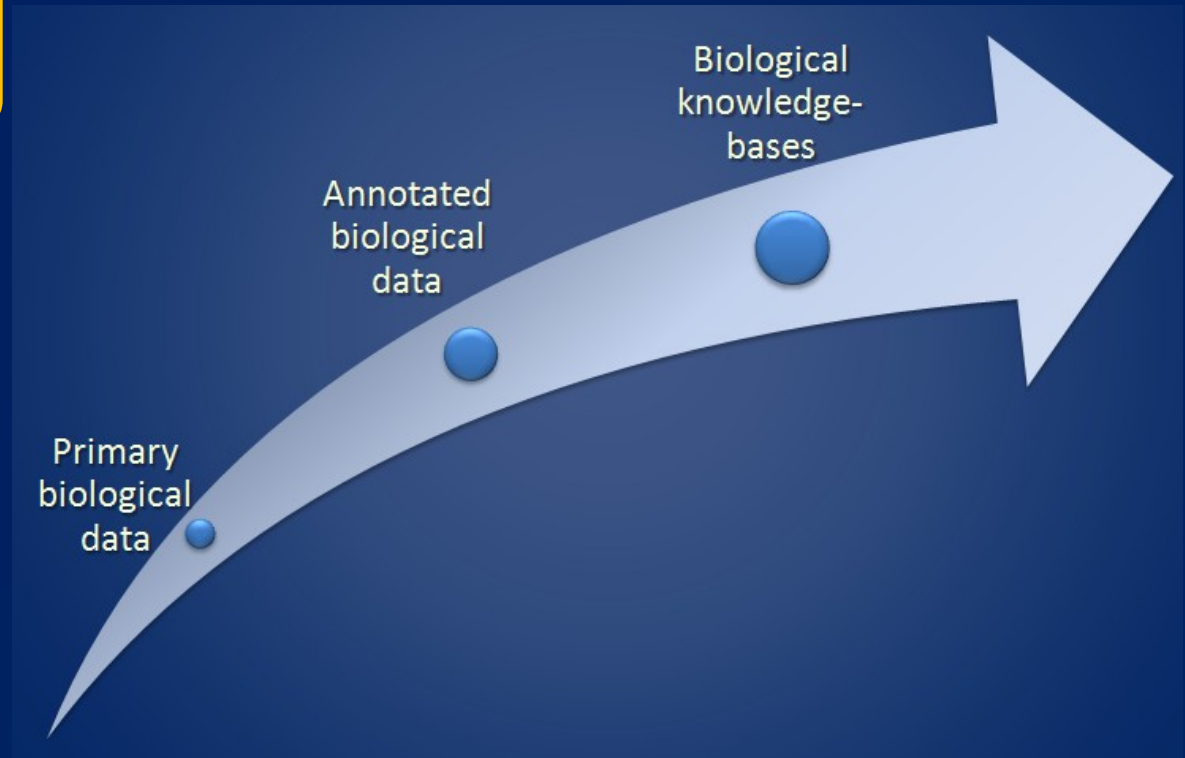
Expression (394)

Thematic analysis of 2010 resources collection on Bioinformatics links directory



Bioinformatics resources development

- Resources variety
- Databases
- Program tools
- Search and workflows
- Towards knowledgebases



Annotated genomic databases: <http://www.ensembl.org/>

e!Ensembl Login | Register | BLAST/BLAT | BioMart | Tools | Downloads | Help | Documentation | Mirrors


Human (GRCh37) Location: 11:6,452,279-6,462,294 Gene: HPX Transcript: HPX-001

Location-based displays

- Whole genome
- Chromosome summary
- Region overview
- Region in detail**
- Comparative Genomics
 - Alignments (image) (51)
 - Alignments (text) (51)
 - Multi-species view (47)
 - Synteny (13)
- Genetic Variation
 - Resequencing (2)
 - Linkage Data
- Markers
- Other genome browsers
 - UCSC
 - NCBI
 - Vega

Configure this page | Manage your data | Export data | Bookmark this page

Chromosome 11: 6,452,279-6,462,294

chromosome 11  Export Image

Region in detail [help](#)

« Region overview

Chromosome bands
Contigs
Ensembl/Havana g...

1.00 Mb
6.00 Mb 6.10 Mb 6.20 Mb 6.30 Mb 6.40 Mb 6.50 Mb 6.60 Mb 6.70 Mb 6.80 Mb 6.90 Mb
p15.4
< AC111177.15 AC091564.12 >

TRIM5 OR56A1 C11orf42 HPX DNHD1 ILK MRPL17 OR2AG1 OR2
OR56A3 OR52L2P FAM160A2 TRIM3 TAF10 OR2AG2 OR10
AC025016.1 CCKBR ARFIP2 TPP1 OR6A2 O
OR52L1 OR56B4 PRKCDBP FXC1 DCHS1 OR10A5
OR56A4 AC111177.2 SMPD1 RRP8 RP11-732A19.1 OR10
AC022762.1 OR52B2 APBB1
OR52X1P OR52W1 CNGA4

ncRNA gene
All Structural varia...
5S rRNA

Ensembl Homo sapiens version 59.37d (GRCh37) Chromosome 11: 5,957,287 - 6,957,286
protein coding merged Ensembl/Havana

e!Ensembl Login | Register | BLAST/BLAT | BioMart | Tools | Downloads | Help

Human (GRCh37) Location: 11:6,452,279-6,462,294 Gene: HPX Transcript: HPX-001

Gene-based displays

Gene: HPX (ENSG00000110169)

Annotated genomic databases:

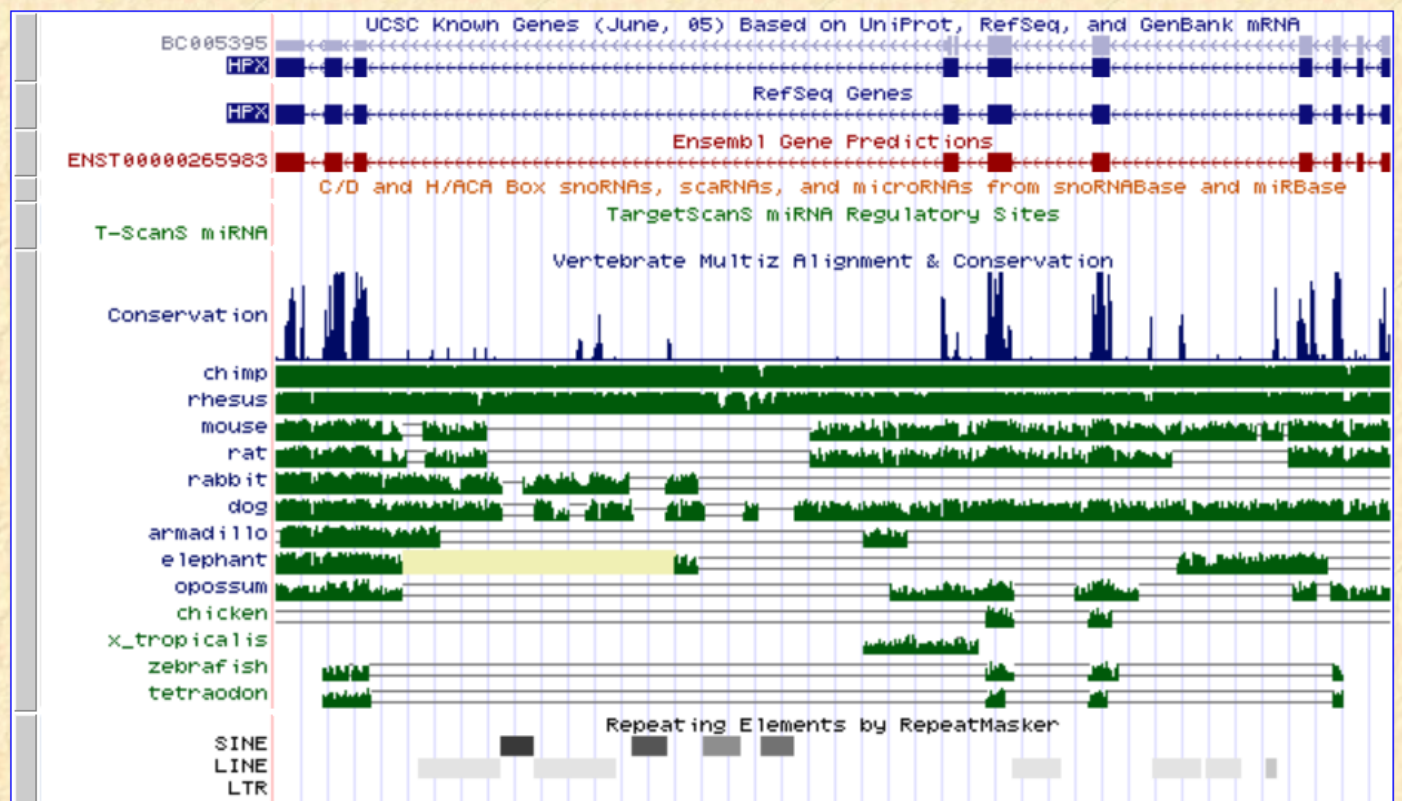
<http://www.genome.ucsc.edu/>

Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl PDF/PS Session Help

UCSC Genome Browser on Human May 2004 (NCBI35/hg17) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr11:6,409,017-6,418,769 [gene](#) jump clear size 9,753 bp. configure



On-line collections of databases:
<http://www.oxfordjournals.org/nar/database/c/>

OXFORD JOURNALS

Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS

[Oxford Journals](#) > [Life Sciences](#) > [Nucleic Acids Research](#) > Database Summary Paper

2010 NAR Database Summary Paper Category List

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- [Human and other Vertebrate Genomes](#)
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases

1230 selected
databases



Database catalogue

<http://lifesciencedb.jp/?lng=en&pg=1>

[Home](#) **Database catalog** [About us](#)

Search:

[Tree](#) [List](#)

Type: Category

Top(822)

DB Type(460)

Analysis Service(18)

Annotation(40)

BioResource(38)

Catalog(18)

Databank(40)

Dictionary(46)

Knowledge Model(36)

Program(82)

Project(142)

NAR category(65)

Target(603)

organism species(364)

No Category(205)


[Download all the Database Catalog](#)

System Information


Top Page

Database catalog is a collection of life science database informations, it has been collected primarily in Japanese database.


What's New




ChEBI - Chemical Ent...
2010/08/06 09:19:24




inPrep
2010/08/04 09:33:58




Cell System Markup L...
2010/07/16 06:14:57




KEGG - kyoto
2010/07/16 05:58:05




2010/07/16 05:53:29




2010/07/16 05:43:01




2010/07/16 05:31:11



Pathogenic microbes
2010/07/16 05:26:15



2010/07/16 05:18:26



KOMUGI
2010/07/16 05:05:35

DBs standardization

http://casimir1.pdn.cam.ac.uk/casimir_ddf/



CASIMIR Database Description Framework

Navigation

- **DDF summary**
- [Download](#)
- [Web services](#)

User login

Username: *

Password: *

CAPTCHA

This question is for testing whether you are a human visitor and to prevent automated spam submissions.



What code is in the image?: *

Enter the characters (without spaces) shown in the image.

Log in

- [Create new account](#)
- [Request new password](#)

The CASIMIR Database Description Framework (DDF) allows resources to describe key technical metadata in a formalised way. The aim of the DDF is to support the standards and interfaces they require. This is a vital component for the online registries of resources currently being developed for many of our users. This deployment displays the DDF annotation performed by resources as part of the MRB project. Other communities can follow the **Download** link in the top right for instructions on how to install this site for their own curation requirements. The DDF annotation is also available through RESTful **Web services**.

Please feel free to **create an account** and try out annotating your own resource using the **Add a new resource** link.

Legend

Yellow blocks represent resources assessed as being at level 1 of the DDF category

Light blue blocks represent resources assessed as being at level 2 of the DDF category

Dark blue blocks represent resources assessed as being at level 3 of the DDF category

Click on individual blocks to see full descriptions of each level per category

Accessibility

- 1 - Access via browser only
- 2 - Access via browser + database reports or dumps
- 3 - Access via browser + API, SQL access or web services

Data representation standards

- 1 - Data coded by local formalism only
- 2 - Some use of controlled vocabs, ontologies or MIBBI
- 3 - General use of controlled vocabs, ontologies or MIBBI

Output

- 1 - HTML or similar to browser only
- 2 - HTML + sparse standard file formats e.g. FASTA
- 3 - HTML + rich standard file formats e.g. XML, SBML

Technical documentation

- 1 - Written text only
- 2 - Written text + formal docs (API docs, schema, UML etc)
- 3 - Written text + formal docs + tutorials/demos

Versioning

- 1 - No provision
- 2 - Old versions available but no tracking between versions
- 3 - Old versions available and tracking between versions

Currency

- 1 - Closed legacy database
- 2 - Updates or versions more than once a year
- 3 - Updates or versions more than once a month

Data structure standards

- 1 - Data structured with local model only
- 2 - Data structured with formal model e.g. an XML schema
- 3 - Use of recognised standard model e.g. FUGE

Quality and Consistency

- 1 - No explicit process for assuring consistency
- 2 - Process for assuring consistency, automatic curation only
- 3 - Process for assuring consistency with manual curation

User support

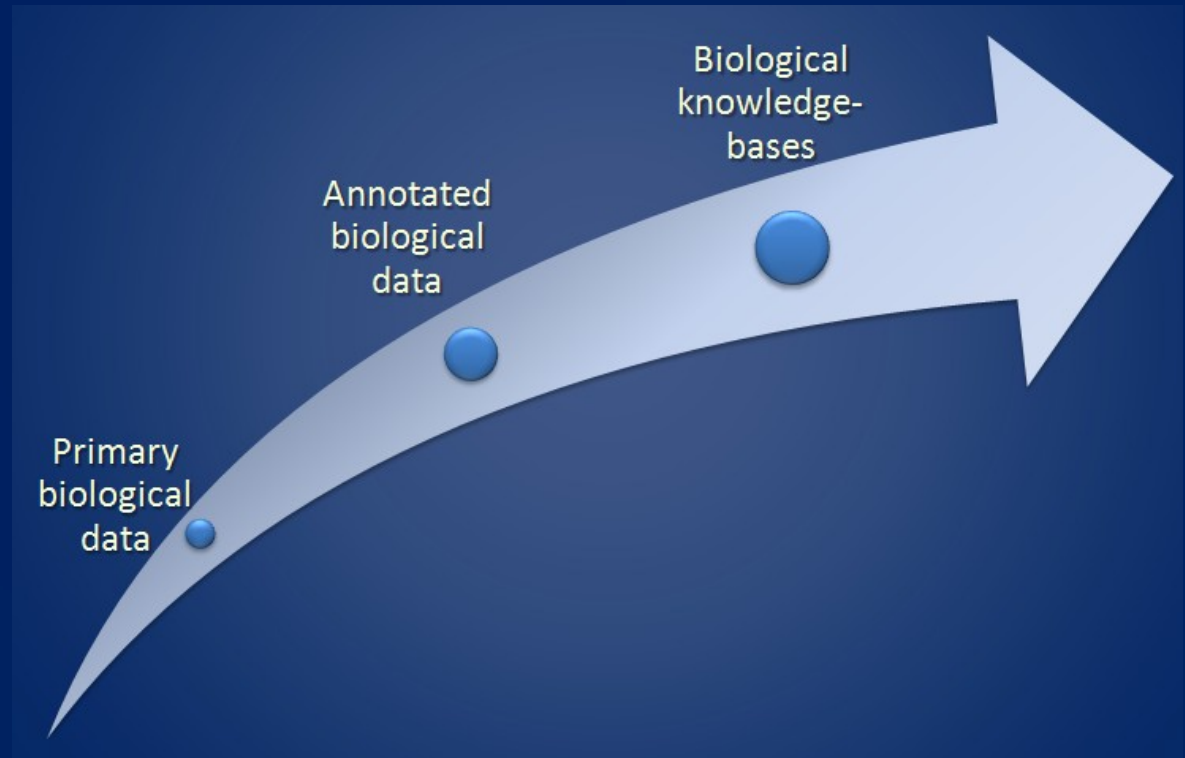
- 1 - User docs only
- 2 - User docs + Email/web form help desk function
- 3 - User docs + personal contact help desk function/training

Resource name contains

Apply

Bioinformatics resources development

- Resources variety
- Databases
- Program tools
- Search and workflows
- Towards knowledgebases



<http://www.ncbi.nlm.nih.gov/Tools/>



Tools for Data Mining

PubMed

Entrez

BLAST

OMIM

Books

TaxBrowser

Structure

Search

Entrez



for

Go

Nucleotide Sequence Analysis

Protein Sequence Analysis

Structures

Genome Analysis

Gene Expression

NCBI

Site Map

Guide to NCBI
resources

Tools for
Programmers

Tools - Nucleotide Sequence Analysis

BLAST

The **Basic Local Alignment Search Tool (BLAST)** for comparing gene and protein sequences against others in public databases, now comes in several types including PSI-BLAST, PHI-BLAST, and BLAST 2 sequences. Specialized BLASTs are also available for human, microbial, malaria, and other genomes, as well as for vector contamination, immunoglobulins, and tentative human consensus sequences.

electronic
PCR
001101011AG

Electronic PCR - allows you to search your DNA sequence for sequence tagged sites (STSs) that have been used as landmarks in various types of genomic maps. It compares the query sequence against data in NCBI's **UniSTS**, a unified, non-redundant view of STSs from a wide range of sources.

<http://pbil.univ-lyon1.fr/alignment.html>

Tools for Multiple Alignments

Search for sequence similarities in databases

- [WU-BLAST at ISREC](#) (Lausanne, Switzerland)
- [BLAST2 Search at EMBL](#) (Heidelberg, Germany) *Performs multiple alignment on homologous sequences detected by BLAST.*
- [FASTA at EBI](#) (Hinxton, UK)
- [Smith-Waterman search at EBI](#) (Hinxton, UK)
- [BCM search launcher](#) (Houston, USA)
- [GeneStream at CRBM](#) (Montpellier, France)
- [BLAST search at PBIL](#) (Lyon, France) *Possibility to select BLAST output results by taxa or keyword.*

Web Sites for Pairwise Alignments

- [LFASTA - Local alignment tool at PBIL](#) (Lyon, France)
- [SIM4 - align cDNA and genomic DNA at PBIL](#) (Lyon, France)
- [WISE - align protein and genomic DNA at Pasteur](#) (Paris, France)
- [SIM - Alignment Tool at ExPASy](#) (Geneva, Switzerland)
- [BLAST two sequences at NCBI](#) (Bethesda, USA)
- [LALIGN at CRBM](#) (Montpellier, France)
- [SIM, GAP, NAP, LAP](#) (Michigan Tech. Univ., USA)
- [JAligner: open source Java implementation of the Smith-Waterman algorithm](#) (Alexandria, Egypt)

Toolkits


<http://toolkit.tuebingen.mpg.de/>

[HOME](#)

[Login](#)

[PDBalert](#)

[Personal](#)



MAX-PLANCK-GESELLSCHAFT

Show results of job:

Recent jobs:

[Select all](#) [Deselect all](#)

queued

running

done

error

Bioinformatics Toolkit

Max-Planck Institute for Developmental Biology

[Search](#) [Alignment](#) [Sequence Analysis](#) [2ary Structure](#) [3ary Structure](#) [Classification](#) [Utils](#)

Welcome to the Bioinformatics Toolkit

The Bioinformatics Toolkit is a platform that integrates a great variety of tools for protein sequence analysis. Many tools are developed in-house, and several public tools are offered with extended functionality.

The toolkit includes, among others: NucleotideBLAST, ProteinBLAST, PSI-BLAST, fastHMMER, HHsenser; ClustalW, MUSCLE, Mafft, ProbCons; HHrep, PCOILS, REPPER; Quick2D; HHpred, Modeller; CLANS, ANCESCON, PHYLIP; Reformat, RetrieveSeq, gi2promoter. For a short description of the tools, click the section tabs.

Job submission

Each tool has a separate input page with a web form in which the user can input sequence data, upload sequence files, and specify options. All tools that take alignments as input accept the most widely used formats (FASTA, CLUSTAL, Stockholm and A3M). You may also choose your own job-names to organize your work. Snail symbols inside the submit buttons inform you about tools that typically run for more than 10 minutes.



Welcome to the Integrated Gene Analysis System MIGenAS of the Max-Planck-Society

MIGenAS (Max-Planck Integrated Gene Analysis System) provides an integrated software environment for bioinformatics applications

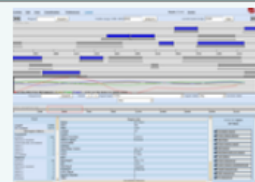
▶ MIGenAS workflow engine: Integrated bioinformatics toolkit for web-based sequence analysis

- facilitates similarity searches in public or user-supplied sequence databases, computation and validation of multiple sequence alignments, phylogenetic analysis, protein structure prediction.
- allows seamless chaining of different tools into pipelines.
- no need for format conversions or parsing of intermediate results.
- supports efficient processing of predefined workflows.
- offers programmatic access via webservice.



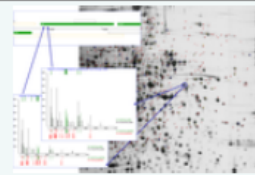
▶ GenDB: Annotation system for prokaryotic genomes (provided by University of Bielefeld)

- Software system for automatic identification, classification and annotation of genes.
- Web interface allows manual annotation with geographically dispersed teams of experts.
- Local installation of GenDB 2.2 available at RZG with connection to dedicated computing facilities.



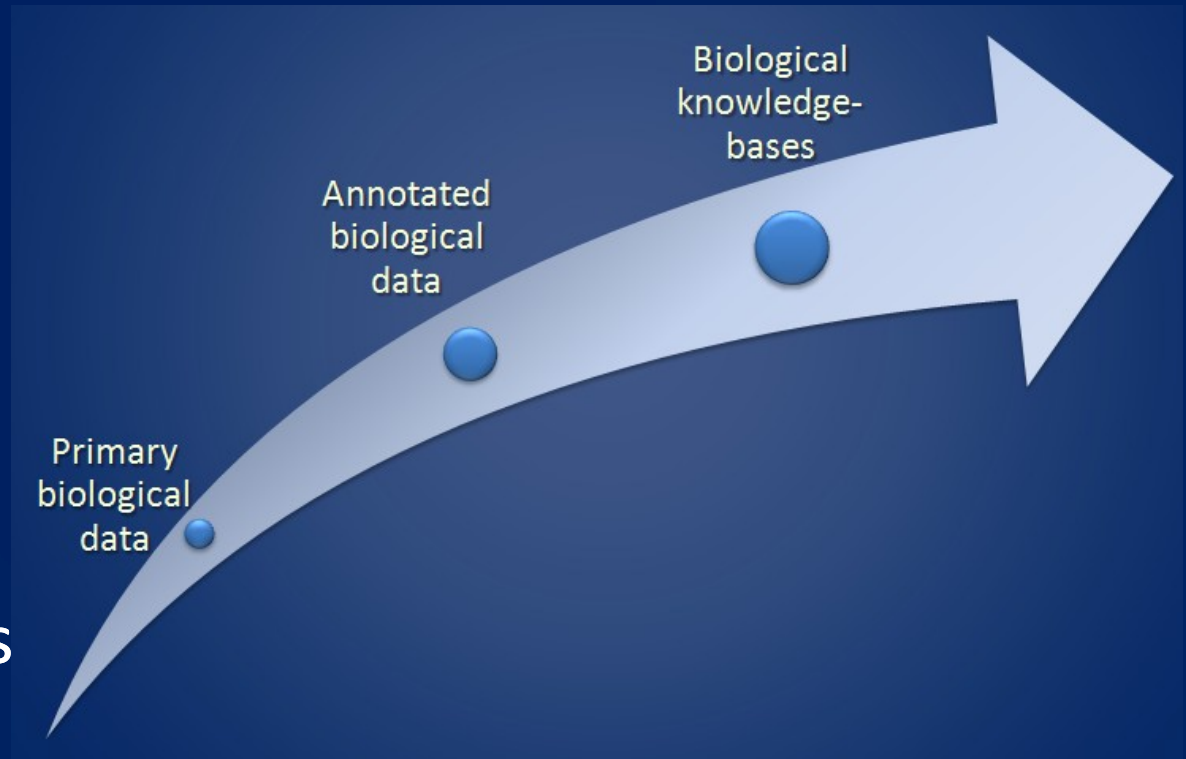
▶ HaloLex: Genome information system for archaea and other prokaryotic genomes



- Data management and analysis platform for microbial genomes and related *omics* data.
- Web interface supports browsing and versatile searches in annotated (public) microbial genomes.
- Provides access to high-quality, expert-curated annotations of a number of halophilic archaea.



Bioinformatics resources development

- Resources variety
- Databases
- Program tools
- Search and workflows
- Towards knowledgebases





 **Entrez, The Life Sciences Search Engine**


HOME | SEARCH | SITE MAP | PubMed | All Databases | Human Genome | GenBank | Map Viewer


Search across databases Help


Welcome to the Entrez cross-database search page


 **PubMed:** biomedical literature citations and abstracts


 **PubMed Central:** free, full text journal articles


 **Site Search:** NCBI web and FTP sites


 **Books:** online books


 **OMIM:** online Mendelian Inheritance in Man


 **OMIA:** online Mendelian Inheritance in Animals


 **Nucleotide:** Core subset of nucleotide sequence records


 **EST:** Expressed Sequence Tag records


 **GSS:** Genome Survey Sequence records


 **Protein:** sequence database


 **Genome:** whole genome sequences


 **Structure:** three-dimensional macromolecular structures


 **Taxonomy:** organisms in GenBank


 **SNP:** single nucleotide polymorphism


 **dbVar:** Genomic structural variation


 **Gene:** gene-centered information


 **SRA:** Sequence Read Archive


 **dbGaP:** genotype and phenotype


 **UniGene:** gene-oriented clusters of transcript sequences


 **CDD:** conserved protein domain database


 **3D Domains:** domains from Entrez Structure


 **UniSTS:** markers and mapping data


 **PopSet:** population study data sets

 **GEO Profiles:** expression and molecular abundance profiles

 **GEO DataSets:** experimental sets of GEO data

 **Epigenomics:** Epigenetic maps and data sets

 **Cancer Chromosomes:** cytogenetic databases

 **PubChem BioAssay:** bioactivity screens of chemical substances

Cross-links and references

Entrez-gene: <http://www.ncbi.nlm.nih.gov/gene/>

NCBI Resources How To

Entrez Gene

Genes and mapped phenotypes

Search: Limits Advanced search Help

Display Settings: ☒ Full Report

HPX hemopexin [*Homo sapiens*]

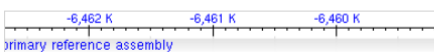
Gene ID: 3263, updated on 19-Sep-2010

Summary

Official Symbol HPX provided by [HGNC](#)
Official Full Name hemopexin provided by [HGNC](#)
Primary source [HGNC:5171](#)
See related [Ensembl:ENSG00000110169](#), [UniProt:Q0793](#), [MIM:142290](#)
Gene type protein coding
RefSeq status REVIEWED
Organism [Homo sapiens](#)
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Cat
Also known as HX; FLJ56652; HPX
Summary This gene encodes a plasma glycoprotein that binds heme with high affinity. The encoded protein is an acute phase protein that trans be involved in protecting cells from oxidative stress. [provided by RefSeq]

Genomic regions, transcripts, and products

(minus strand) Go to [reference sequence details](#)



Links to external resources

[HGNC](#)

[Ensembl](#)

[UniProt](#)

[Evidence Viewer](#)

[ModelMaker](#)

[AceView](#)

[UCSC](#)

[MGC](#)

[HuGE Navigator](#)

[KEGG](#)

Related sites

[BLAST](#)

[Entrez Genome](#)

[Genome Project](#)

[Genomic Biology](#)

[GEO](#)

[HomoloGene](#)

[Map Viewer](#)

[OMIM](#)

[Probe](#)

[RefSeq](#)

[UniGene](#)

[UniSTS](#)

Links

[Order cDNA clone](#)

[BioAssay, by Gene target](#)

[CCDS](#)

[Conserved Domains](#)

[EST](#)

[Full text in PMC](#)

[GEO Profiles](#)

[Genome](#)

[HomoloGene](#)

[Map Viewer](#)

[Nucleotide](#)

[OMIM](#)

[Peptidome](#)

[Probe](#)

[Protein](#)

[PubChem Compound](#)

[PubChem Substance](#)

[PubMed](#)

[PubMed \(GeneRIF\)](#)

[PubMed \(OMIM\)](#)

[RefSeq Proteins](#)

[RefSeq RNAs](#)

[SNP](#)


[SNP: GeneView](#)

[SNP: Genotype](#)

[Taxonomy](#)

DBs/tools cross-search

<http://www.hslls.pitt.edu/guides/genetics/obrc/>

**Health Sciences Library System**
Serving the University of Pittsburgh and UPMC

About HSLs - Contact Us - Remote Access

Journals & Articles ▪ Books ▪ More Resources ▪ Library Services ▪ How Do I?

[HSLs Home](#) > [Guides](#) > [Molecular Biology](#) >

OBRC: Online Bioinformatics Resources Collection

OBRC

- [Email Suggestions](#)
- [Recommend a New Resource](#)

search.HSLs.OBRC[About search.HSLs.OBRC](#)

Databases/Tools | Articles on Databases/Tools | Web

Databases/Tools

Find molecular databases & software tools with a combined search of the **HSLs Online Bioinformatics Resource Collection (OBRC)** & the [BioMed Central Databases](#) collection.

Search

Search Examples: keyword ([HapMap](#), [SNP](#)); phrase ([protein structure prediction](#))

The Online Bioinformatics Resources Collection (OBRC) contains annotations and links for 2724 bioinformatics databases and software tools.

Browse:

- [DNA Sequence Databases and Analysis Tools](#) (493)
- [Enzymes and Pathways](#) (266)
- [Gene Mutations, Genetic Variations and Diseases](#) (246)
- [Genomics Databases and Analysis Tools](#) (674)
- [Immunological Databases and Tools](#) (61)

2724 links

Harvesting engines

<http://harvester.fzk.de/harvester/>



Harvester crawls and crosslinks the following bioinformatic sites:

[4DXp](#) - [AceView](#) - [BLAST](#) - [BLINK](#) - [Biocompare](#) - [CDART](#) - [CDD](#) - [EB-eye](#) - [ensEMBL](#) - [Entrez](#) - [FishMap](#) - [Galaxy](#) - [GeneSorter](#) - [UCSC GenomeBrowser](#) - [gfp-cDNA](#) - [Google-Scholar](#) - [gpubmed](#) - [Harvester42](#) - [H-Inv](#) - [HomoloGene](#) - [iHOP](#) - [IPI](#) - [LOCATE](#) - [MapView](#) - [MGI](#) - [MINT](#) - [Mitocheck](#) - [OMIM](#) - [PolyMeta](#) - [ProteinAtlas](#) - [PSORT II](#) - [RGD](#) - [SMART](#) - [SOSUI](#) - [STITCH](#) - [STRING](#) - [TAIR](#) - [Unigene](#) - [UniprotKB](#) - [Wikipedia](#) - [WikiProtein](#) - [YIF](#) - [ZFIN](#)
[Harvester Sequence Search device](#) - [YaCy-Sciencenet p2p search engine](#)
See our [H-Wiki](#) for latest activities... [Try YaCy Harvester Search \(all genomes\)](#) - Have fun...

Bioinformatic Harvester IV

...serving 10.000s of pages every day - Note that '*' and '?' wildcards are supported.

<input type="text"/>	<input type="button" value="Search human"/>
<input type="text"/>	<input type="button" value="Search mouse"/>
<input type="text"/>	<input type="button" value="Search rat"/>
<input type="text"/>	<input type="button" value="Search zebrafish"/>
<input type="text"/>	<input type="button" value="Search arabidopsis"/>

Organization of frameworks

<https://projets.pasteur.fr/wiki/mobyle>

Mobyle

[Просмотр](#) [Активность](#) [Оперативный план](#) [Задачи](#) [Новости](#) [Документы](#) [Wiki](#) [Download](#) [Файлы](#) [Хранилище](#)

Welcome to the Mobyle Project Website!

Mobyle is a framework and web portal specifically aimed at the integration of bioinformatics software and databanks.

Mobyle is the successor of [Pise](#) and the [RPBS server](#), previous systems that provided web environments to define and execute bioinformatics analyses.

Functionalities:

- **data reusability**: the tagging of the user data facilitates the reuse of input values or results between different programs.
- **automatic data validation and format conversion**: the description of the expected data and their format allows to verify and convert input values if necessary.
- **service discovery and workflow authoring assistance**: services are provided through a searchable menu; furthermore, type compatibility mechanisms between and potential program inputs let users either interactively pipe tasks or build complete workflows before to run them.

Based on extensive user studies, we developed the end-user interface as a Web Portal that provides a global and integrated view of all the elements needed to perform such as the available programs, the submitted jobs and the data of interest.

Users

Life scientists

Want to run bioanalyses through a web interface without ins

See our [User guide](#) - [Mobyle Public Servers](#)

Administrators and developers

Functionalities:

Users

Life scientists

Administrators and developers

[Learn more ...](#)

[Contact us!](#)

B.Néron, H.Ménager, C.Maufrais et al.

Mobyle: a new full web bioinformatics framework

Bioinformatics (2009) 25(22): 3005-3011

Programs

- ▶ alignment
- ▶ assembly
- ▶ database
- ▶ display
- ▶ genetics
- ▶ hmm
- ▶ nucleic
- ▶ phylogeny
- ▶ protein
- ▶ sequence
- ▶ structure

- ▶ LIPM
- ▶ RPBS

Data Bookmarks

Jobs refresh

Tutorials refresh

[How to use Mobyle? A step by step tutorial](#)
[Registration information](#)
[Sequence formats](#)
[Alignment formats](#)

Welcome
Programs
Data Bookmarks
Jobs
Tutorials

Welcome to Mobyle,


Select an analysis in the **Programs** [menu](#).


Tutorials are available. See our [interactive guided tour](#).

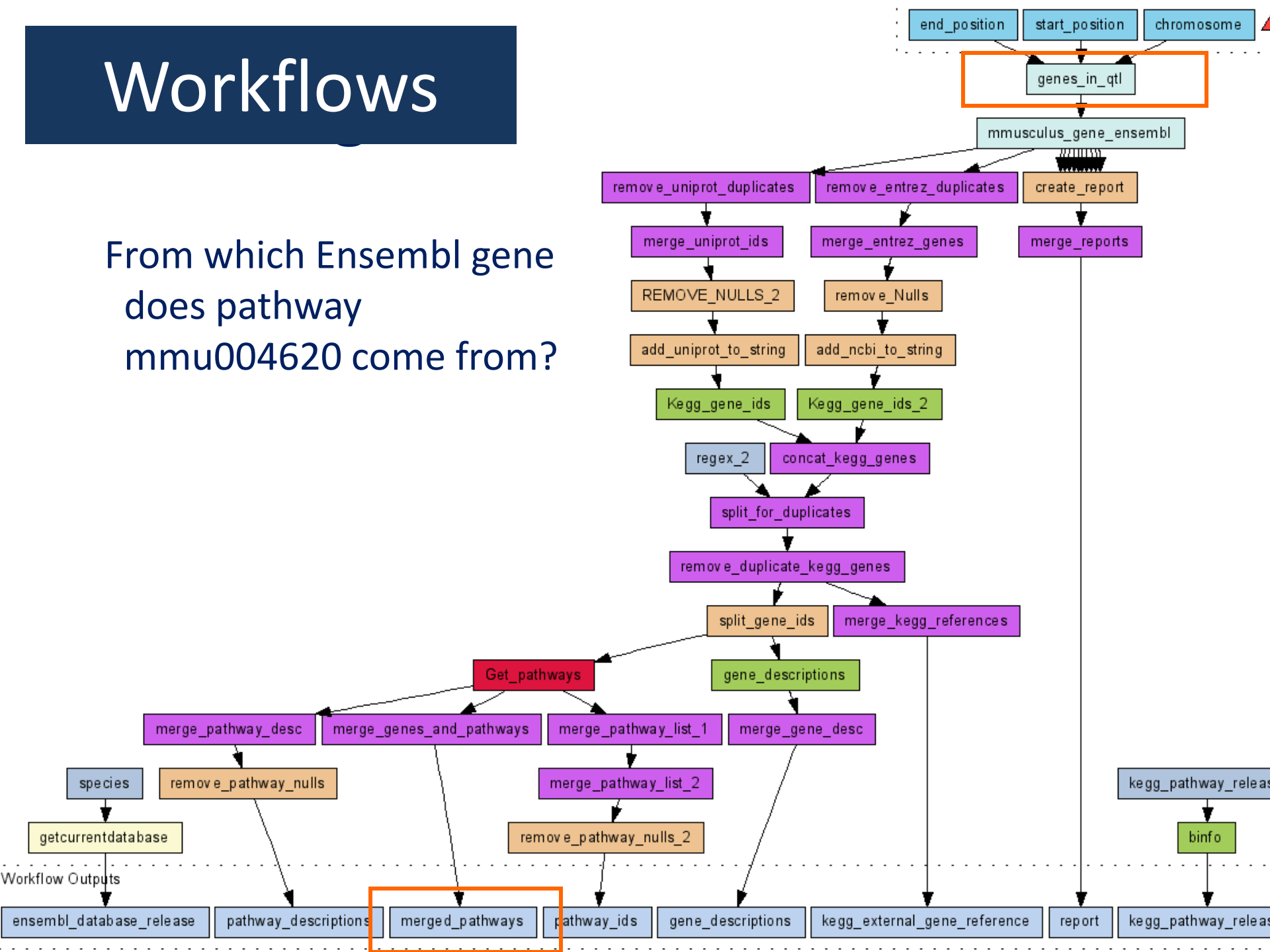
Credits

Mobyle is a platform developed jointly by the [Institut Pasteur](#) "Logiciels et More information about this project can be found [here](#).

The end-user interface as a Web Portal provides a global and integrated view of all the elements needed to perform analysis, such as the available programs, the submitted jobs and the data of interest.

Workflows

From which Ensembl gene
does pathway
mmu004620 come from?



Taverna Workflow Workbench

Taverna Workflow Workbench v1.5.1.6

File Tools Workflows Advanced

Design Results Discover + Add

Search

www.visualgenomics.ca
ConvertAAtoFASTA_AA
www.weigelworld.org
ATH_Deletion_Prediction
ATH_SNP_Coding
ATH_SNP_Range
MapSequence
Soaplab @ http://www.ebi.ac.uk/soaplab/emboss4/services/
alignment_consensus
cons
megamerger
merger
alignment_differences

Advanced model explorer

Workflow Object properties

Add Nested Workflow

Workflow object	Retries	Delay	Back...	Thr
Compare functions of genes on t				
Workflow inputs				
Workflow outputs				
Graph				
Processors				
GetUniqueIDs	0	0	1	
GenericSetOperations	0	0	1	
Yellow : gold	0	0	1	
PassUniqueTerms	0	0	1	
Fail_if_false	0	0	1	
FlattenList	0	0	1	
Fail_if_true	0	0	1	
Red : crimson	0	0	1	

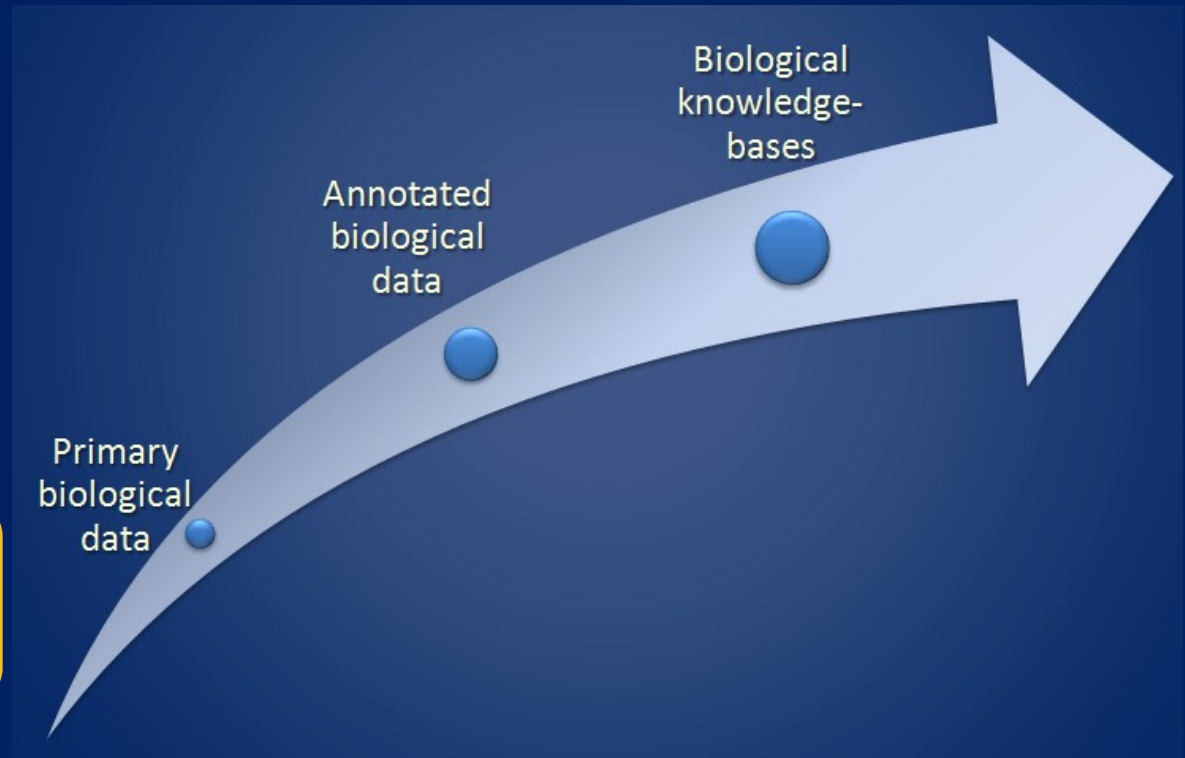
<http://www.omii.ac.uk/wiki/Taverna>

Save diagram Refresh Configure diagram

In perspective bioinformatics research will be performed by certain workflow as molecular biology experiment is performed by certain protocol

Bioinformatics resources development

- Resources variety
- Databases
- Program tools
- Search and workflows
- Towards knowledgebases



Integrated Knowledgebases as Web environment

<http://genomicscience.energy.gov/compbio/>

The Systems Biology Knowledgebase is a **cyberinfrastructure** to facilitate a new level of scientific inquiry by serving as a central component for the integration of modeling, simulation, experimentation, and bioinformatic approaches.

genomics.energy.gov

Human Genome Project Information • Genomic Science

U.S. Department of Energy Office of Science

Genomic Science Program

Systems Biology for Energy and Environment

ABOUT RESEARCH TECHNOLOGIES MISSIONS COMPUTING EDUCATION BIOFUELS BIOENERGY RESEARCH CENTERS

NEWS

DOE Systems Biology Knowledgebase Implementation Plan

Now available.
(September 2010)

[PDF](#)

Contribute to Defining Knowledgebase Requirements and Specifications

A Wiki has been created to provide open access to an interactive resource for viewing, providing comments, and contributing to this community-driven effort to specify the requirements for the DOE Systems Biology Knowledgebase. Go to the [Wiki](#) and select "How to Participate" to join this effort.

[Wiki](#)

DOE Systems Biology Knowledgebase

Community-Driven

DOE Systems Biology Knowledgebase for a New Era in Biology

Community-Driven Cyberinfrastructure for Sharing and Integrating Data and Analytical Tools

DOE Systems Biology Knowledgebase

Biological Principles

Genomes

Experiments by investigators

- Individual labs
- Distributed collaborations
- Research centers

Modeling and simulation

Data analysis and reduction

Shared tools for data analysis and visualization

- Expression analysis
- Regulatory network modeling tools
- Metabolic network modeling tools

Framework for storing, integrating, and sharing massive datasets

- Expression
- Image
- Mass spec
- Molecular dynamics simulation library

Data generators



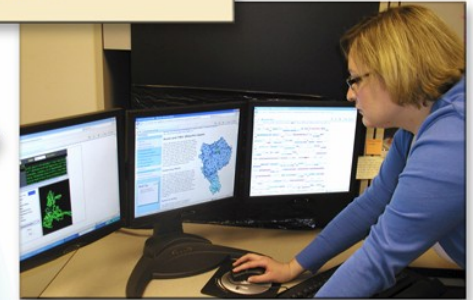
Seamless Submission and Incorporation of Diverse Data

- Standards for data and metadata representation
- Quality control and assurance capabilities
- Automated systems for depositing and updating bulk data
- Tracking and evaluation of data use

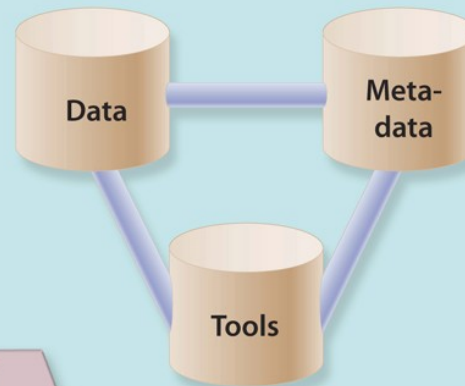
Open-Access Data and Information Exchange

- Access through several flexible interfaces
- Retrieval of experimental data and products of modeling and simulation
- Working environment for testing hypotheses by *in silico* experimentation
- Provision and sharing of user feedback

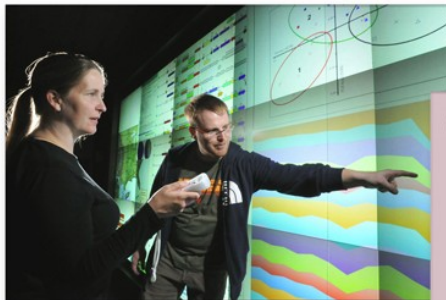
Data users



DOE Systems Biology Knowledgebase



Software and tool developers



Open Development of Open-Source Software and Tools

- Data analysis and visualization tools
- Resources for *in silico* experimentation
- Modeling and simulation tools
- Customizable tools with layers of functionality
- Tracking and evaluation of tool use

Community-Wide Stewardship

- User committee
- Standards committee
- Advisory committee
- Operations
- Development
- Value-added analysis
- Training, tutorials, support staff

The Systems Biology Knowledgebase (Kbase)

In addition to supporting data storage, retrieval, and management capabilities, Kbase also would enable *new knowledge acquisition and management*, through *free and open access* to data, analysis tools, and information for the scientific research community.

Kbase, therefore, must serve multiple roles, *including*:

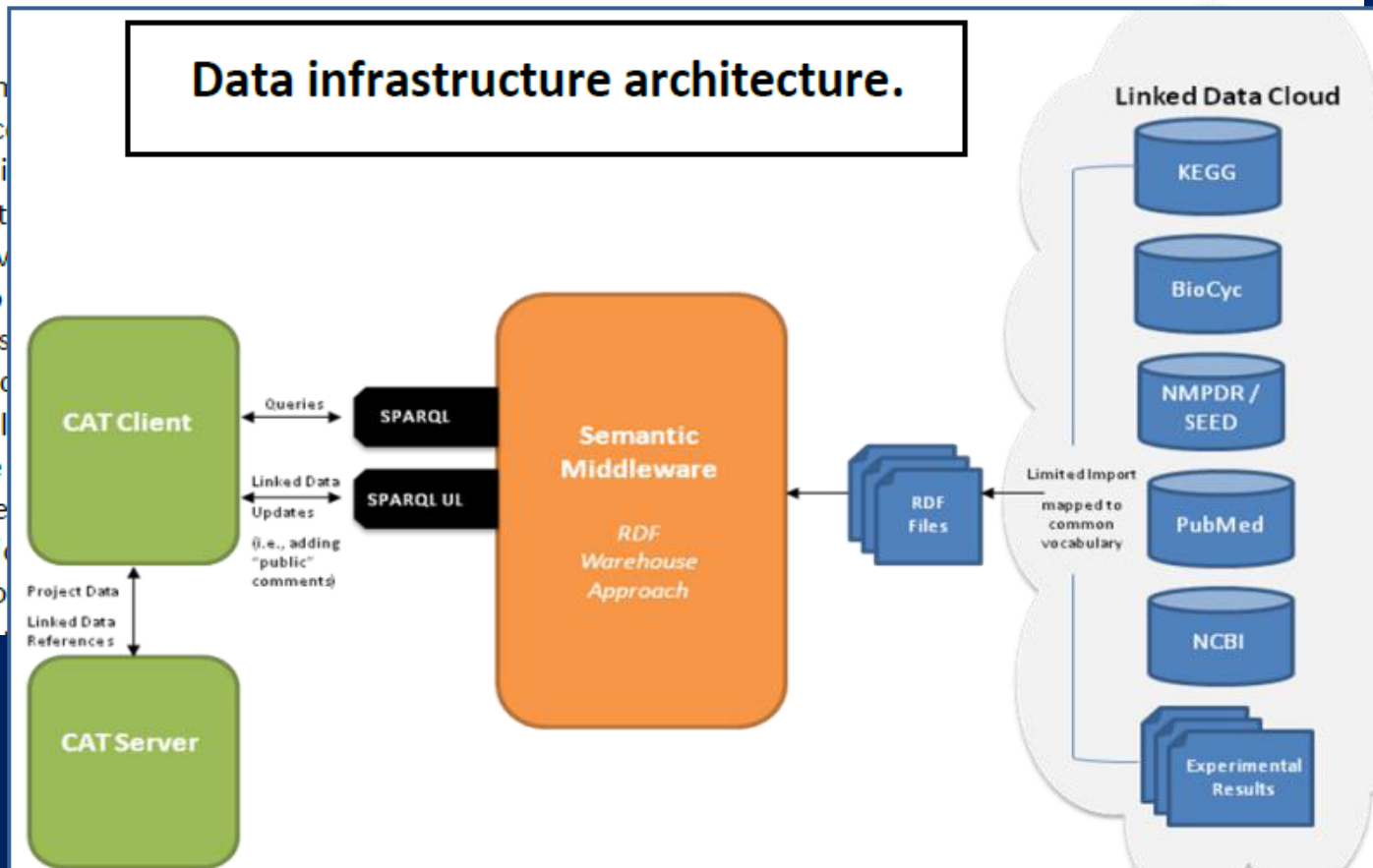
- A repository of data and results from high-throughput experiments.
- A collection of tools to derive new insights through data synthesis, analysis, and comparison.
- A foundation for prediction, design, manipulation, and ultimately, engineering of biological systems.

Pilot projects

Semantic Driven Knowledge Discovery and Integration in the System Biology Knowledgebase Project

Kerstin Kleese van Dam, Cliff Joslyn, Lee Ann McCue, Bill Cannon, Carina Lansing, Zoe Guillen, Margaret Romine,
Gordon Anderson, Abigail Corrigan
Pacific Northwest National Laboratory

It is the goal of the DOE System Knowledgebase to become a core infrastructure for sharing and integrating Analysis tools. This new infrastructure enables the science community to move from Biology, where it is possible to gain a basic understanding of biological systems, to address core DOE Missions and the community wide accessibility and the capability to integrate within its environmental context technical functionalities the Biology enable. The ultimate success of the Knowledgebase will however



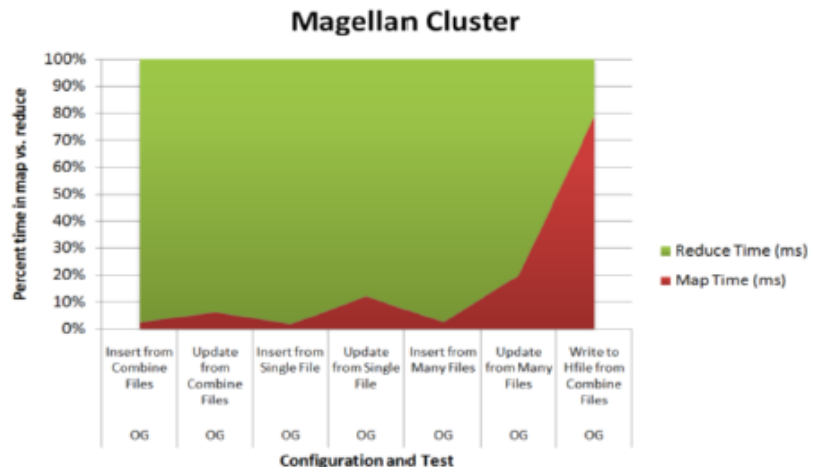
Database Management Systems Technologies for Computational Biology & Bioinformatics Applications

Knowledgebase R&D Pilot Project

Victor Markowitz, Keith Jackson, Ernest Szeto, Konstantinos Mavrommatis
Lawrence Berkeley National Laboratory (LBNL)

The aim of this project was to examine new **database management system technologies** for supporting efficient analysis of very large genome and metagenome sequence datasets.

Comparative analysis of genomic and metagenomic datasets is usually based on integrating these datasets in the context of databases implemented using relational commercial database management systems (DBMS) such as Oracle or open source DBMS such as MySQL. The rapid increase in the number and size of these datasets results in a decrease in performance of typical comparative analysis tools, such as examining putative operons across microbial genomes. A recent benchmark of relational DBMS¹ indicates that new database management technologies are better suited for scientific data management applications. We set out to evaluate the usage of cloud based data management technologies for handling large genome and metagenome datasets, in particular Hadoop data management components for data storage and querying. Hbase² is a distributed, column-oriented data store that supports real-time access to extremely large data.

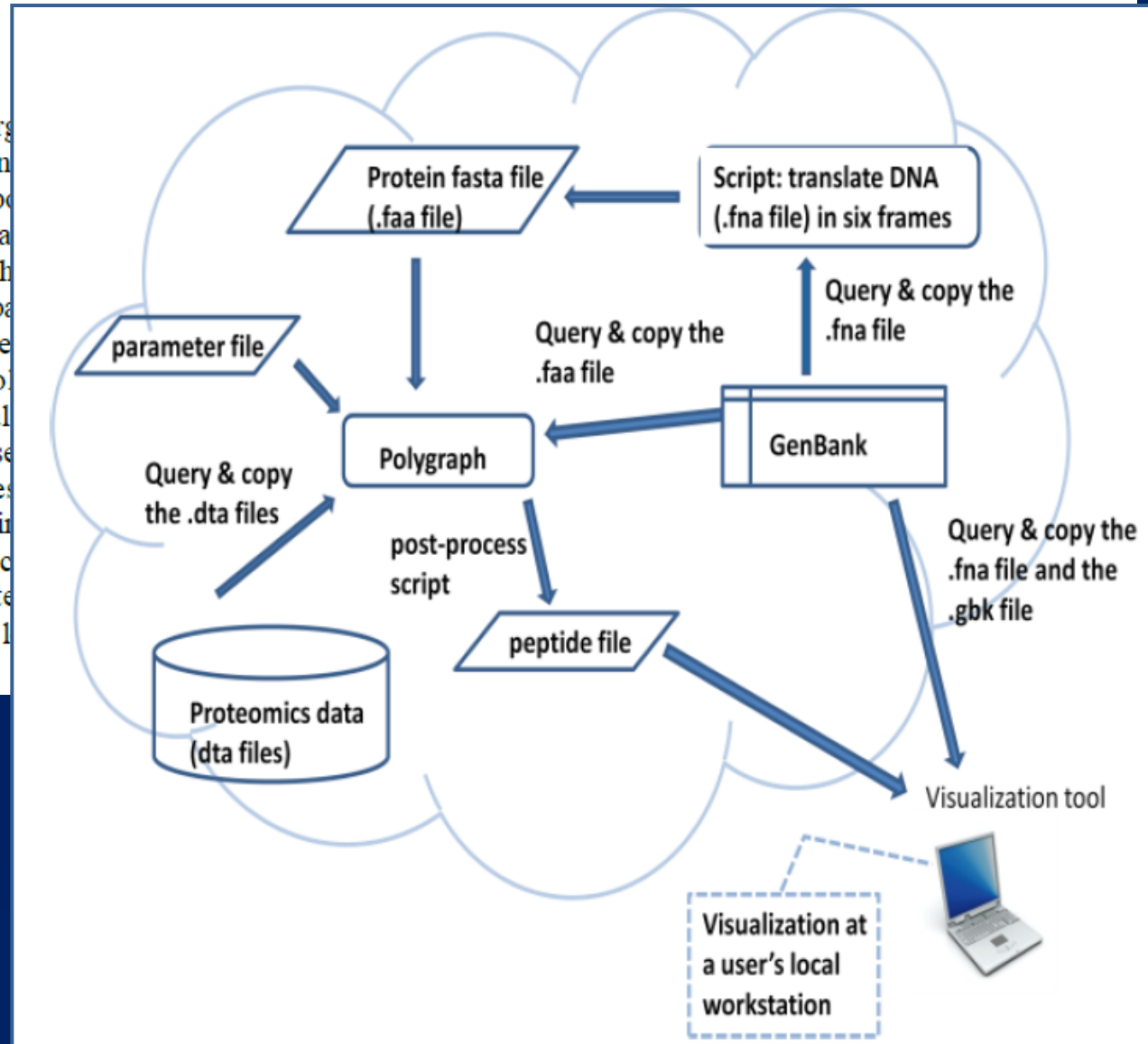


Exploring Architecture Options for Workflows in a Federated, Cloud-based Systems Biology Knowledgebase

Ian Gorton, Yan Liu, Jian Yin, Leeann McCue, Bill Cannon, Gordon Anderson

Systems biology is characterized by a large number of researchers who use a wide variety of fragmented and heterogeneous computational tools of all scales to support their research. To provide a more coherent computational environment for systems biology, we are working as part of the Systems Biology Knowledgebase (Kbase) to develop a federated cloud-based system architecture that will eventually host massive amounts of biological data, provide high performance and scalable computational resources, and serve a large user community with tools and services. We intend to utilize the Kbase resources. We invest in developing workflow infrastructure suitable for use in the cloud, which utilizes standards-based workflow description languages, integration technologies, and incorporates a distributed execution layer for exploiting data locality in the architecture.

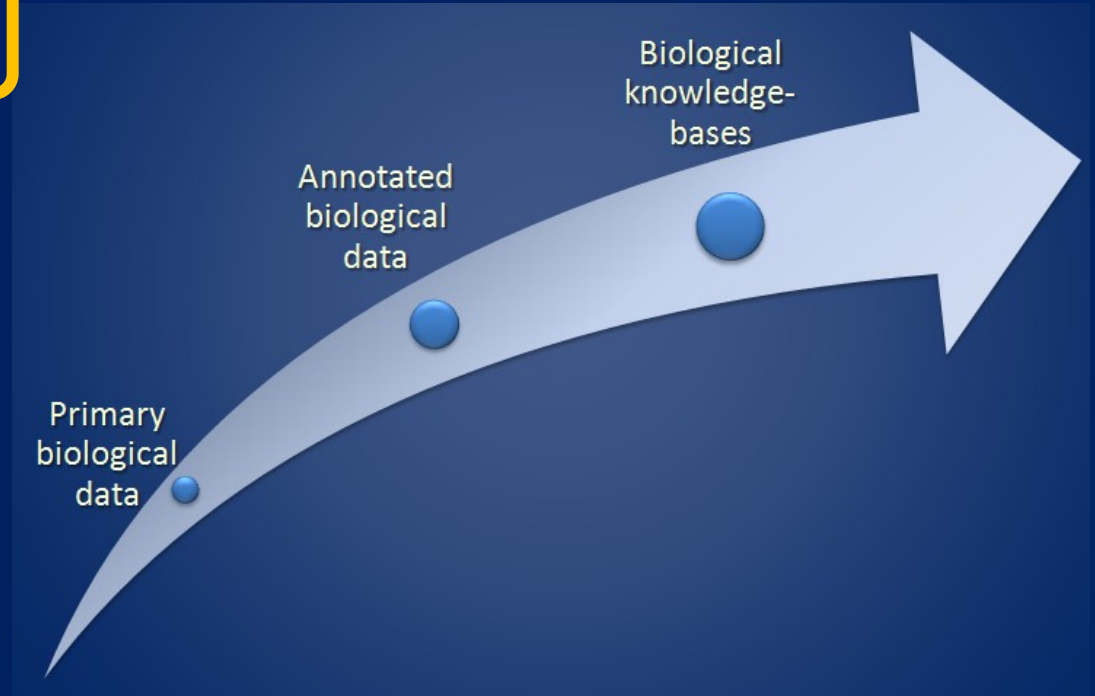
Pilot projects



NEW IT SOLUTIONS FOR BIOINFORMATICS

New IT Solutions for bioinformatics

- Data compression
- Web semantics
- Web services
- Gridification
- Biocomputing



Textual data compression in post-genomics era

- **DNAzip: DNA sequence compression using a reference genome** [Christley S., Lu Y., Li C., and Xie X. Human genomes as email attachments //Bioinformatics (2009) 25:274-5] <http://silver.ics.uci.edu/~dnazip/index.html>
- **G-SQZ: compact encoding of genomic sequence and quality data** [W.Tembe, J.Lowey, and E.Suh // Bioinformatics (2010) 26(17): 2192-2194]
- **Compression of whole genome alignments** [P.Hanus, J.Dingel, G.Chalkidis, J.Hagenauer//IEEE Transactions on Information Theory.- Vol.56, No.2 (Feb2010) *Special issue on information theory in molecular biology and neuroscience*. P.696-705]

```
HIT000291102_Homo.      35 VAEG--ETKPDPDVTERCSDGWSFDATTLDDNGTMLFFK-GEFVWKSHKW 81
ENSPTRT00000006246_Pan.  VAEG--ETKPDPDVTERCSDGWSFDATTLDDNGTMLFFK-GEFVWKSHKW
XM_001109797_Macaca.    VAEG--ETKPDPDVTERCSDGWSFDATTLDDNGTMLFFK-GEFVWKSHKW
AK145928_Mus.          VAEVENGTKPDSDVPEHCLDTWSFDAAATMDHNGTMLFFK-GEFVWRGHSG
BC091137_Rattus.       VAKGENGTKPDSDVIEHCSDAWSFDATTMDHNGTMLFFK-GEFVWRGHSG
ENSCAFT00000010394_Canis. GTEGGSGARVKPDVTELCLDGWSFDATTLDEHGAMLFFK-GEFVWKSHRW
XM_001504590_Equus.    GAEGGNGVKQDPDVIERCSDGWSSDATTLDEHGAMVFFK-GEFMWKSPNW
BC102687_Bos.         GVEGGNVAKPDPEVTERCSDGWGFDAATLDEHGNMFLK-GEFVWKGHAW
ENSDART00000073621_Danio. ---MLKDAPEDHHEDRCKG-IEFDAIAPDEKGNTEFFKVGDRLLWKGLTG
ENSORLT00000004740_Oryzias. SVILISHPDGDSALPDRGAG-IEFDAITPDDKGQTFFFK-GDHVWKGFDDG
AB075198_Oryzias.     ---HHEHRRKGAVRDRCKG-IEMDAVAVNEEGIPYFFK-EDHLFKGFHG
CR647288_Tetraodon.   NVSEMRDEDSPGALPDRGAG-IEFDAITPDEKGTFFFK-GAYMWKGFQW
CR635722_Tetraodon.   ---GDSHG--LAKLDRGAG-LEMDAVAVNEIGIPYFFK-GDHLFKGFHG
AB125933_Takifugu.    NISEVKKEEDSGPALPDRGAG-IEFDAITPDEKGTTLFFK-GAYMWKDFHG
                        : * . ** : :. * *: * :..
```

<http://www.tgen.org/research/gsqueeze.cfm>



THE TRANSLATIONAL GENOMICS RESEARCH INSTITUTE
A Non-profit Biomedical Research Institute

October 2, 2010

Genomic Sequence-Quality Data Encoding & Compression

G-SQueueZ offers indexed, order preserving,
compressed format for genomic sequence reads.



WHAT IS G-SQUEEZ™?

Genomic Squeeze (G-SQueueZ™) is a technique to encode genomic sequence-quality data into an indexed, compact binary format, and that can result in substantial savings in storage and processing over conventional plain text formats (such as FASTQ, CSFASTA/QUAL formats).

In G-SQueueZ™ encoding, order of the data is preserved and the indexed structure directly allows selective access to various parts of the file. In addition, resulting binary files can be queried to obtain useful information about the file, such as number of reads, base composition, platform, etc.

For an in-depth description, please read the manuscript [[Link to the paper on the Bioinformatics Journal Website](#)]

CONTACT INFORMATION

Questions, comments and usage inquiries should be directed to:

gsgz-admin@tgen.org.

F.A.Q.

Which platforms are supported? G-SQueueZ™ version 0.5 supports SOLEXA



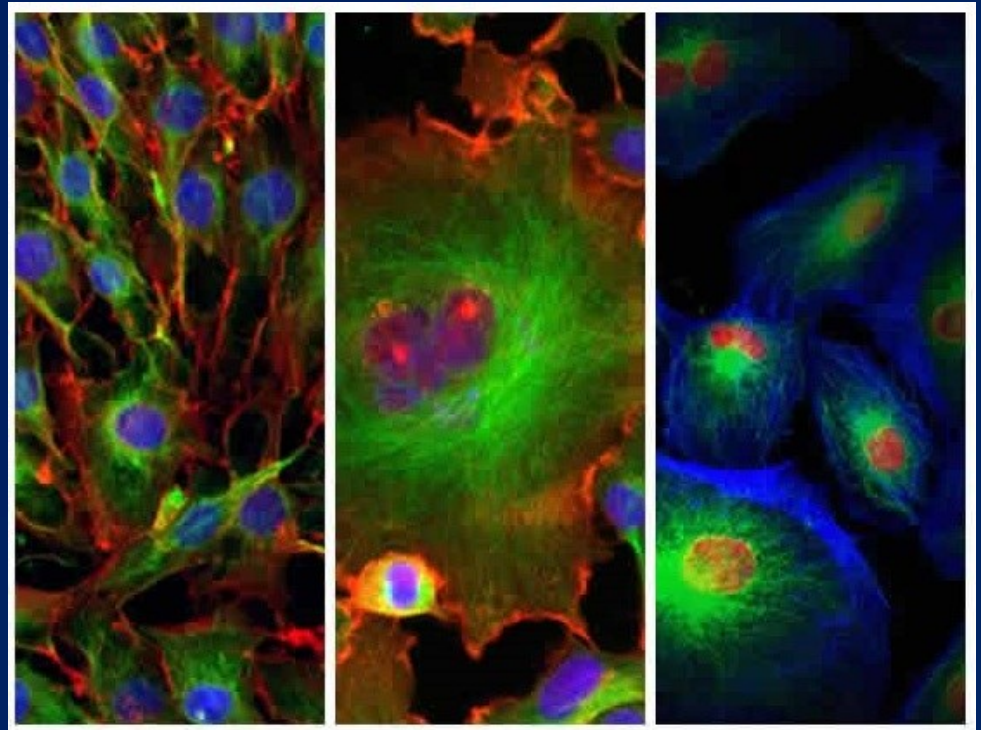
AVAILABILITY

W.Tembe, J.Lowey, and E.Suh

G-SQZ: compact encoding of genomic sequence and quality data
Bioinformatics (2010) 26(17): 2192-2194

Compression techniques for images processing

- **CATMAID: collaborative annotation toolkit for massive amounts of image data** [S.Saalfeld, A.Cardona, V.Hartenstein, and P.Tomančák// Bioinformatics (2009) 25(15): 1984-1986]
- **Bisque: a platform for bioimage analysis and management** [K.Kvilekval, D.Fedorov, B.Obara et al. //Bioinformatics (2010) 26(4): 544-552]
- **Bioimage informatics: a new area of engineering biology** [H.Peng//Bioinformatics (2008) 24(17): 1827-1836]



<http://www.jatit.org/volumes.php>

Journal of Theoretical and Applied Information Technology

P.R.Rajeswari & A.Apparao **Genbit compress – algorithm for repetitive and non-repetitive DNA sequences**// 2010, vol 11, no. 1

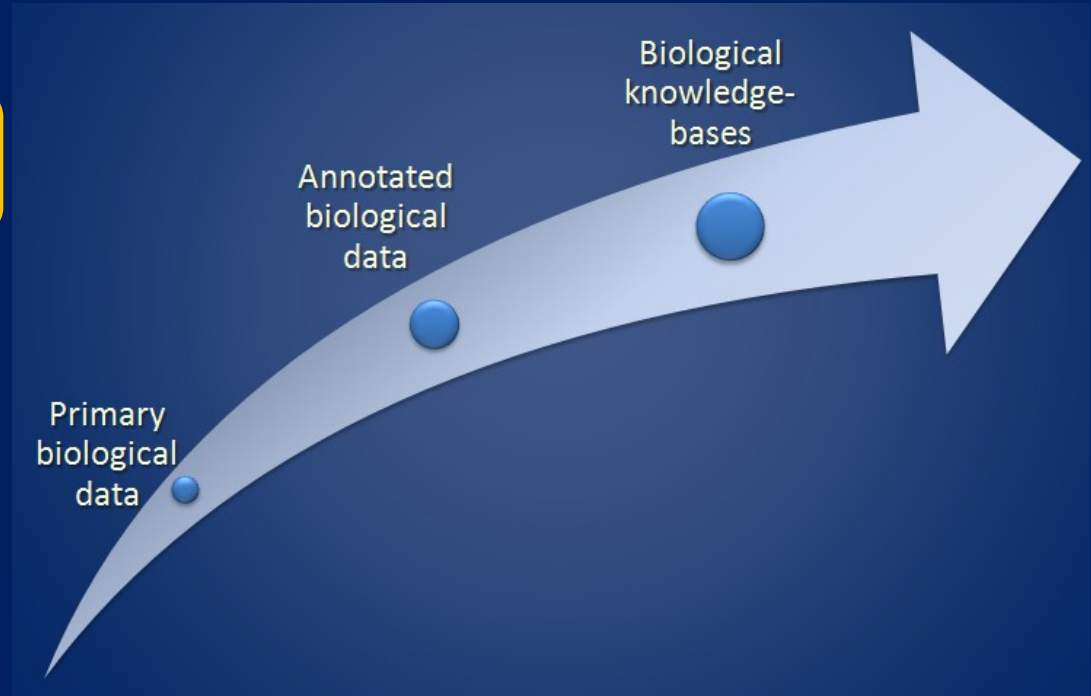
P.R.Rajeswari, A.Apparao, R.K.Kumar **Huffbit compress algorithm to compress DNA sequences using extended binary trees**// 2010, vol 13, no.2.

J.H. Pujar, I.M. Kadlaskar **A new lossless method of image compression and decompression using Huffman coding techniques**//2010, vol 15, no.1

G.Bhopale **Image noise reduction using mathematical morphology size distributions. A new image noise reduction and compression algorithm for grayscale images** //2010, vol 16. No. 1

New IT Solutions for bioinformatics

- Data compression
- Web semantics
- Web services
- Gridification
- Biocomputing



<http://hackathon3.dbcls.jp/>

[wiki](#)[Timeline](#)[Re](#)

DBCLS BioHackathon 2010

The 3rd DBCLS BioHackathon for interpreting biological knowledge with Semantic Web technologies will be held during 2010/2/ 8-12 in Japan.

- [Participants](#)
- [Schedule](#)
- [Symposium](#)
- [PosterSession](#)
- [OpenSpace](#)
- [Hackathon](#)
- [MeetingReport](#)

Objectives

DBCLS is working on the integration of biological resources. To achieve this goal, we have been organizing BioHackathons since 2008 to survey environments with open source software and public services. Themes of the hackathons evolved year by year, and we are continuously providing a platform for gathering and utilizing state of the art technologies to emerging demands in life sciences.

DBCLS BioHackathon for interpreting biological knowledge with Semantic Web technologies is working on the integration of biological resources.

To achieve this goal, we have been organizing BioHackathons since 2008 to survey existing efforts and develop integrated environments with open source software and public services.

Semantic Web technologies

<http://rewerse.net/A2/Overview.htm>



Rewerse Working group A2: Adding Semantics to the Bioinformatics Web

Sections

[Overview](#)

[Demos](#)

[Deliverables](#)

[Participants](#)

[Publications](#)

Overview

Objectives

The objective of the WG is to create the core of a Bioinformatics Semantic Web populated by a number of sample data sources and applications representative of the use of the Web in Bioinformatics and to demonstrate novel, reasoning-based solutions dealing with the following problems:

- Rules for mediation and to formulate complex queries
- Consistent integration of Bioinformatics data
- Adaptive portals for molecular biologists

Bioinformatics is an ideal field for testing Semantic Web technologies for three reasons: First, Web-based systems and Web databases have been applied very early in the field; second, the dramatic increase of data produced in the field calls for novel processing methods; third, the high heterogeneity of Bioinformatics data require semantic-based integration.

Consider the following scenario: a biologist obtains a novel DNA sequence nothing is known about. He or she wants to run an alignment, but has specific requirements. These requirements are captured as rules and constraints, which are taken into account by the online accessible semantic Web enabled sequence comparison service.

A2: Bioinformatics
rewerse.net

The objective of the WG is to create the core of a Bioinformatics Semantic Web populated by a number of sample data sources and applications representative of the use of the Web in Bioinformatics and to demonstrate novel, reasoning-based solutions dealing with the following problems:

- Rules for mediation and to formulate complex queries
- Consistent integration of Bioinformatics data
- Adaptive portals for molecular biologists

<http://www.ida.liu.se/~iislab/projects/>

SAMBO

(System for Aligning and Merging Biomedical Ontologies)



KitAMO

(a ToolKit for Aligning and Merging Ontologies)



Gene ontologies

<http://bioportal.bioontology.org/>



Welcome to the NCBO Bioportal

Use BioPortal to access and share ontologies that are actively used in biomedical communities. You can search for terms in ontologies (try typing "Melanoma" in the "Search all ontologies" column), browse a list of ontologies in BioPortal (type "NCI Thesaurus" in the "Find an ontology" box in the middle column), search biomedical resources that we automatically annotated with (try typing "Melanoma" in the "Search resources" box in the right column). You can [create ontology-based annotations for your own text](#), [link your own project that uses ontologies to the existing ontologies](#), [find and create relations between terms in different ontologies](#), review and comment on ontologies and their components as you [browse](#) them. [Sign in to BioPortal](#) to submit a new ontology-based project, provide comments on ontologies or add ontology mappings.

Search all ontologies

[Advanced Search](#)

Find an ontology

[Browse Ontologies >](#)

Search resources

[Advanced Resource Search](#)

Most Viewed Ontologies (August, 2010)

Ontology	Views
SNOMED Clinical Terms	2044
NCI Thesaurus	1294
NCBI organismal classification	1177
RadLex	738
MedDRA	639

Latest Notes

[Order Chiroptera has been included \(Malaria Ontology\)](#) 15 days ago by topalis
Bats are now correctly appearing as members of Chiroptera.

[New Relationship Proposal: is_a \(Malaria Ontology\)](#) 24 days ago by slozano

[New Term Proposal: ventral fin lepidotrichium](#)

Latest Mappings

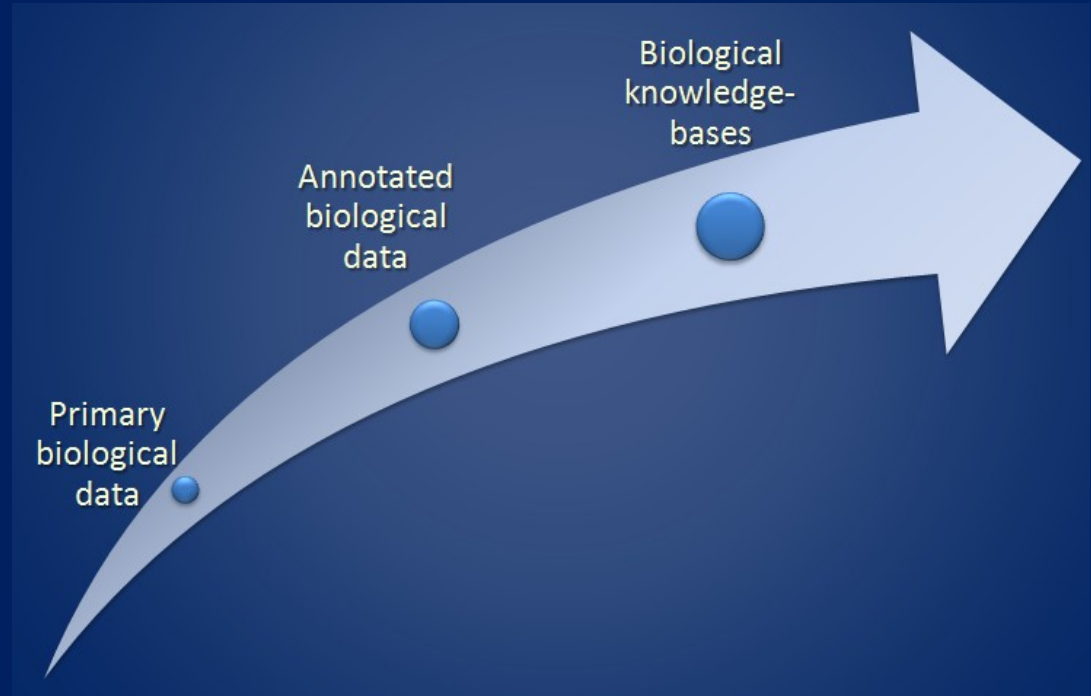
[Patient \(Health Level Seven\) => Interventions \(SNOMED Clinical Terms\)](#)
08/06/10 nigam

[Melanoma \(NCI Thesaurus\) => Malignant melanoma \(SNOMED Clinical Terms\)](#)
08/02/10 nigam

[Malignant melanoma \(SNOMED Clinical Terms\) => Melanoma \(NCI Thesaurus\)](#)
08/02/10 nigam

New IT Solutions for bioinformatics

- Data compression
- Web semantics
- Web services
- Gridification
- Biocomputing



Recent publications on Web services for bioinformatics

- ❑ Kalas M, Puntervoll P, Joseph A et al. BioXSD: the common data-exchange format for everyday bioinformatics web services// Bioinformatics. 2010 Sep 15;26(18):i540-6.
- ❑ Zappa A, Miele M, Romano P. IBWS: IST Bioinformatics Web Services// Nucleic Acids Res. 2010 Jul 1;38 Suppl:W712-8
- ❑ Katayama T, Nakao M, Takagi T. TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services// Nucleic Acids Res. 2010 Jul 1;38 Suppl:W706-11
- ❑ Smedley D, Schofield P, Chen CK et al Finding and sharing: new approaches to registries of databases and services for the biomedical sciences // Database (Oxford). 2010 Jul 6;2010: baq014.
- ❑ Ramírez S, Muñoz-Mérida A, Karlsson J et al. MOWServ: a web client for integration of bioinformatic resources// Nucleic Acids Res. 2010 Jul 1;38 Suppl:W671-6.

http://www.biocatalogue.org/

The screenshot shows the BioCatalogue website interface. At the top, the logo "BioCatalogue beta" is displayed with the tagline "The Life Science Web Service Registry". A navigation bar includes a search box, a "Go!" button, and links for "Home", "Services", "Register a Service", and "Providers". Below the navigation bar, a banner reads "The BioCatalogue: providing a curated catalogue of Life Science Web Services". A green box states: "The BioCatalogue currently has **1719 services**, **127 service providers** and **420 members**".

Latest Activity

Last 7 days

- Carsten Schnober **joined** the BioCatalogue
- Franck Tanoh **added** a description annotation to Service Provider: [Protein Data Bank, USA \(RCSB PDB\)](#)
- Franck Tanoh **added** a description annotation to Service Provider: [Protein Data Bank Japan \(PDBJ\)](#)
- Franck Tanoh **added** a description annotation to Service Provider: [UniProt](#)
- Franck Tanoh **added** a category annotation to Service: [PDB](#)
- Franck Tanoh **added** a category annotation to Service: [PDB](#)

DISCOVER

"Web Services are hard to find"

- Find the right Web Service
- Powerful search and filtering
- Information from providers and community

[More info](#)

REGISTER

"My Web Services are not visible"

- Easily register Web Services
- Instantly available to everyone
- Providers can advertise, describe and monitor their Services

[More info](#)

ANNOTATE

"Web Services are poorly described"

- Anyone can describe and annotate
- Ongoing expert curation
- Social curation by the community

[More info](#)

MONITOR

"Web Services are volatile"

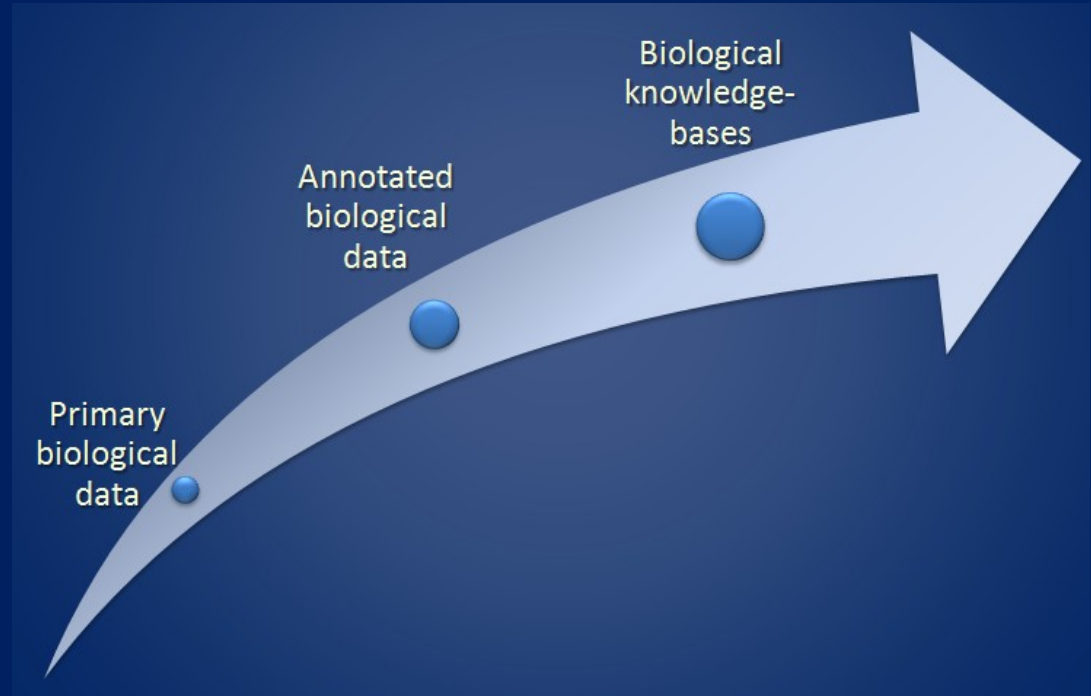
- Services change and get outdated
- BioCatalogue monitors Services
- Monitors availability and reliability

[More info](#)


BioCatalogue: a universal catalogue of web services for the life sciences// NAR, Web servers issue, 2010

New IT Solutions for bioinformatics

- Data compression
- Web semantics
- Web services
- **Gridification**
- Biocomputing



<http://www.embracegrid.info/>



EMBRACE
Grid
Network of excellence

*A EUROPEAN MODEL FOR BIOINFORMATICS
RESEARCH AND COMMUNITY EDUCATION*

HOME CONTACT PARTNERS TRAINING WORK PACKAGES PRODUCTS

Information for:

- » [The general public](#)
- » [Bioinformaticians](#)
- » [Industry](#)
- » [Biomedical scientists](#)
- » [Students](#)
- » [Journalists](#)
- » [Jobs](#)


Other information:

- » [News](#)
- » [Project abstract](#)
- » [Description of work](#)
- » [Technology watch](#)

EMBRACE Network of Excellence

A European Model for Bioinformatics Research and Community Education


The objective of EMBRACE is to draw together technology in the biomolecular sciences. The network will work to integrate the major service technologies. The integration effort will involve service providers and end-user biologists, their own local or proprietary data and tools



BioSapiens Network

A European Virtual Institute for Genome Annotation

Home DAS Portal Partners Training Work Packages Meetings News Restricted Area




» [About BioSapiens](#)

» [Management Structure](#)


» [Vacancies](#)

» [BioSapiens Book](#)


» [Publications](#)



Funded by



Coordinated by



BioSapiens Network of Excellence

A European Virtual Institute for Genome Annotation

BioSapiens is a Network of Excellence, funded by the European Union's 6th Framework Programme, and made up of bioinformatics researchers from 25 institutions based in 14 countries throughout Europe.

The objective of the BioSapiens is to provide a large scale, coordinated effort to annotate genome data by laboratories around Europe, using both informatics tools and input from experimentalists.

Links

- » [EMBRACE](#)
- » [Enfin](#)

User Login

Username:

Password:

[Send me my password](#)

<http://www.biosapiens.info/>

<https://cabig.nci.nih.gov/>



National Cancer Institute



caBIG®

Cancer Biomedical
Informatics Grid®

Home
Getting Connected
Setting the Stage
Finding caBIG® Tools
caBIG® Compatibility
caCORE
caGrid
Data Sharing
Getting Support
Knowledge Centers
Service Providers
Training Portal
About caBIG®
How caBIG® Operates
Communications Resources
Scientific Articles
News Articles
Contact Us

главная » [concepts](#) » cancer common ontologic representation environment (cacore)

Cancer Common Ontologic Representation Environment (caCORE)

🔴 **caCORE 3.x APIs and Grid Services Decommissioned August 7, 2009.** [Announcement posted](#) 📄.

Key to achieving interoperability and compatibility, caCORE tools and APIs are developed by the National Cancer Institute Center for Bioinformatics and Information Technology (NCI CBIIT) to provide the building blocks for development of interoperable information management systems. This ultimately enables interoperability and data sharing from the scientific bench to the clinical bedside and back.

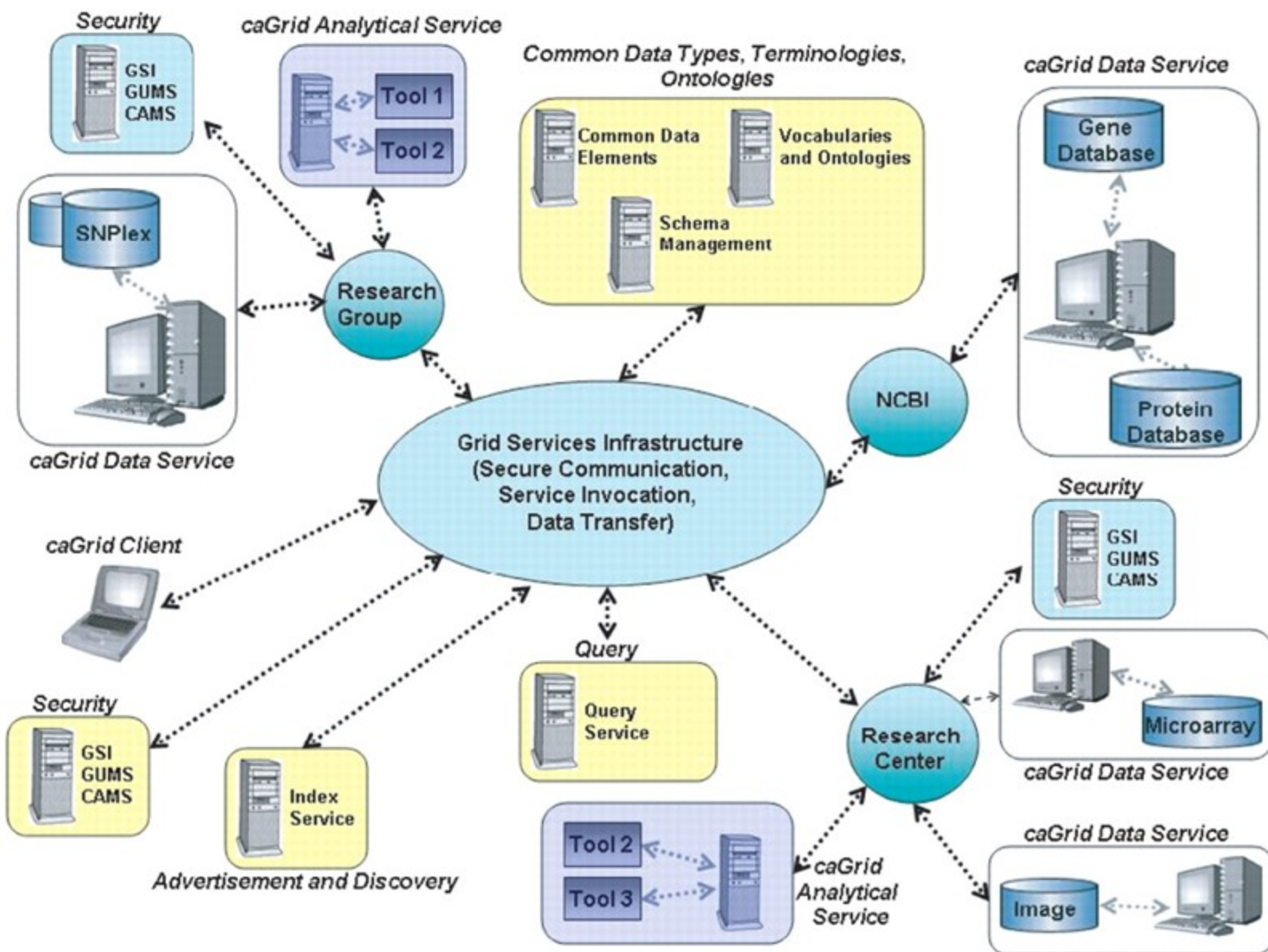
caCORE includes the following key components:

- [EVS](#) (Enterprise Vocabulary Services) for hosting and managing vocabulary.
- [caDSR](#) (Cancer Data Standards Registry and Repository) for hosting and managing metadata.
- [caCORE SDK](#), the GUI-based [caCORE Workbench](#), and associated tools for model-driven software engineering of systems which can be easily integrated with [caGrid](#).

[EVS](#) and the [caDSR database and tools](#) are the basis of the semantic foundation for interoperable data and analytical services.

Developers use caCORE components to create "caCORE-like" systems. By definition these systems have object-oriented information models registered in caDSR whose meaning is linked to EVS vocabularies, and have a REST-based interface for accessing the data. The [caBIO data service](#) 📄 is an example of a

caBIG (Cancer Biomedical Informatics Grid)



<http://dockinggrid.gforge.inria.fr/>

ANR DOCK *Financé par*
Molecular Docking on Grids ANR

Objectives and Project Outlines
Project Coordinators and Partners
Project
Project Organization
Softwares
Download
Meetings
Publications
Related links
Deliverables

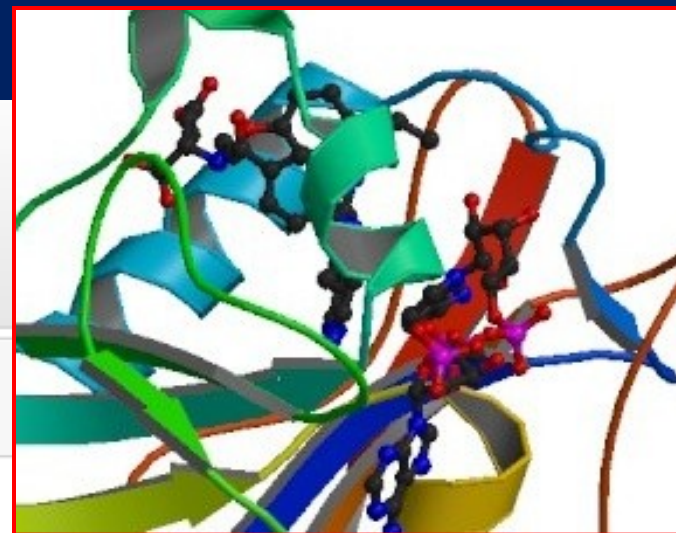
Objectives and Project Outlines

Acronym of the project: Docking@Grid

Title of the project: Conformational Sampling and Docking on Grids

Goals and Summary

Molecular modelling – and notably the conformational sampling and docking procedures for understanding the interaction mechanisms between (macro)molecules involved in pl obvious interest in computationally predicting the binding modes of partners involved in participating in regulatory processes within the living cell, such approach may equally be means to interfere with the normal or pathological process (rational drug design). Howe combinatorial complexity (molecule size, number of degrees of freedom) that represent currently available computing power, hence the three imperative research directions in the search for mathematical models of maximum simplicity that nevertheless provide a



<http://www.bioinfoGRID.eu/>



BioinfoGRID

Bioinformatics Grid Application for life science

[Home](#)[The Project](#)[Partners](#)[Documentation](#)[Grid access](#)[Project Events](#)[Events](#)[Links](#)

You are here: Home


Navigation


- ▶ [The Project](#)
- ▶ [Partners](#)
- ▶ [Documentation](#)
- ▶ [Grid access](#)
- ▶ [Project Events](#)
- ▶ [Events](#)
- ▶ [Links](#)

Log In

Login Name

Password

 Log in

 [Forgot your pwd?](#)

The BioinfoGRID Project



The European Commission promotes the Bioinformatics Grid Application for life science (Bioinformatics Grid) in order to carry out Bioinformatics research and to develop new applications in the technology that represents the natural evolution of the Web.

Grid networking promises to be a very important step forward in the Information Technology of thousands of interconnected computers possible, allowing the shared use of calculating power, data storage and communication between computers and aims instead to transform the global network of computers into a vast network.

The BioinfoGRID White Paper



The BioinfoGRID White paper outlines guidelines and recommendations for the scientific community to use the BioinfoGRID project.

The BioinfoGRID European project aims to promote the Bioinformatics applications for life science technology. More specifically the BioinfoGRID project evaluates applications in the fields of data calculation times by distributing the calculation on thousands of computers using the Framework Program). The massive potential of Grid technology is indispensable when dealing with large data, for example, in searching the human genome or when carrying out docking simulation. Organisations related to the BioinfoGRID project are able to run Bioinformatics challenges of biology.

The BioinfoGRID white paper provides guidelines and recommendations that have been drawn from our experience. In our findings we give advice to different user communities on how to take advantage of the power of the GRID.



[Download the BioinfoGRID White Paper in pdf format \(3.2MB\)](#)

Украинский Академический Грид

[Главная](#)[Карта сайта](#)[Новости](#)[Контакты](#)[Мониторинг УАГ](#)

О Гриде

- » Что такое Грид?
- » Проекты Грид
- » Грид в Украине
- » Литература
- » Ресурсы

Проекты EGEE и EGI

- » Что такое EGEE?
- » EGEE Activities
- » Партнеры EGEE
- » События EGEE
- » Грид-приложения
- » Что такое EGI?

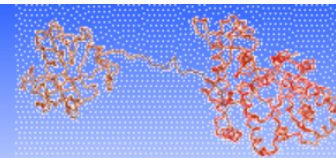
Полезные ссылки

Добро пожаловать в Украинский Академический Грид!



<http://uag.bitp.kiev.ua/>

MolDynGrid Virtual Lab

[Навчання](#) [Розрах](#)

Ресурси

- Форум ВЛ MolDynGrid
- Український академічний Грид
- Грид-інфраструктура для наукових та освітніх установ України
- Обчислювальні кластери
- Грид-монітори

Корисні посилання

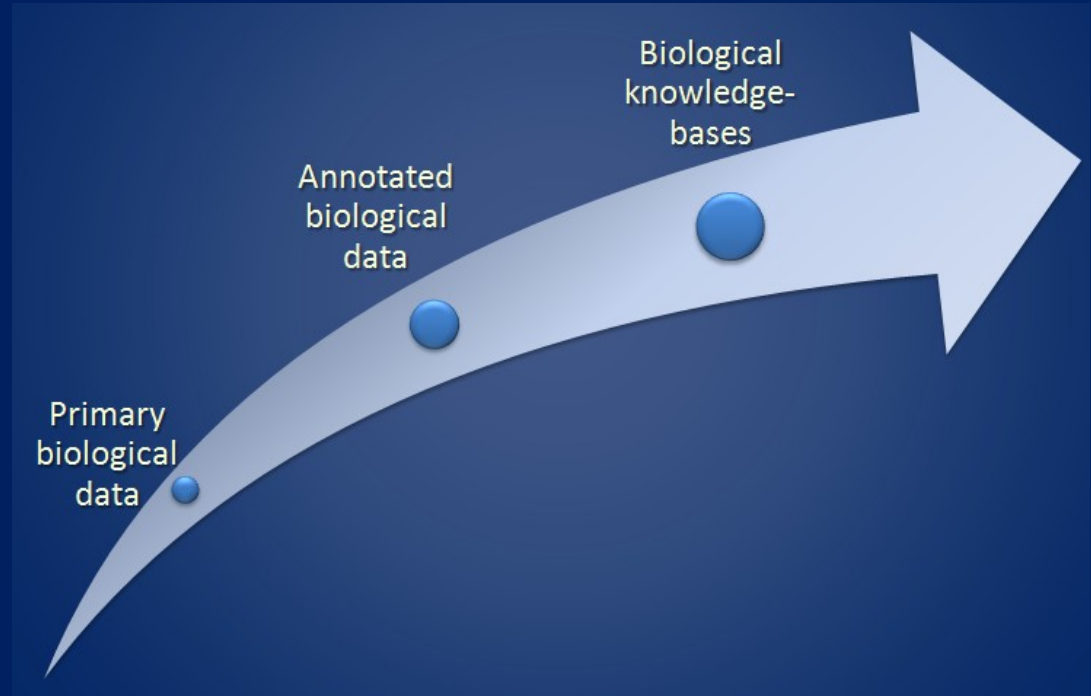
Віртуальна лабораторія MolDynGrid створена у вересні 2008 р. для вирішення задач в галузях структурної біології і біоінформатики, які потребують значних витрат машинного часу та оперують великими об'ємами інформації. Мета створення ВО полягає в розробці ефективної інфраструктури для проведення in silico розрахунків молекулярної динаміки (МД) біологічних макромолекул (білків, нуклеїнових кислот та їхніх комплексів) у водно-іонному оточенні в часовому інтервалі до 100 нс. MolDynGrid є частиною проекту розвитку Грид-сегменту Національної академії наук України з обчислювальними молекулярної біофізики (ІТФ) та інших наукових центрів, зареєстрованих членів за умови дотримання правил користування.

<https://moldyngrid.org/>

<http://uag.bitp.kiev.ua/file/applications/biomedical-applications-ru.pdf>

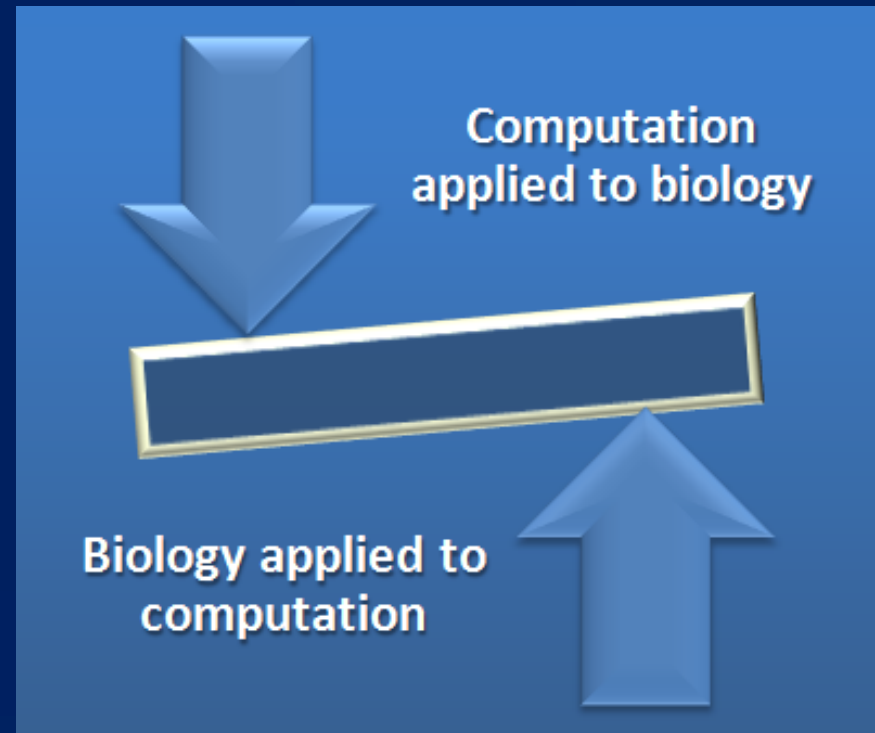
New IT Solutions for bioinformatics

- Data compression
- Web semantics
- Web services
- Gridification
- Biocomputing



Two Kinds of Computation Motivated by Biology

- **Computation applied to biology**
 - Bioinformatics
 - computational biology
 - modeling DNA, cells, organs, populations, etc.
- **Biology applied to computation**
 - biologically-inspired computation
 - neural networks
 - artificial life, etc.





Bioinformatics vs biocomputing

Bioinformatics: the application of computer technology to the management of biological information. Is associated with knowledge extraction and interpretation of data

Biomolecular computing: the use of biological and chemical processes to perform computations

Bio-inspired computing: the use of biological paradigms (e.g., neural nets, genetic algorithms) in the design of computational algorithms. Algorithms may be implemented in *any* appropriate technology.

Useful links

<http://www.biochem.oulu.fi/Biocomputing/juffer/Teaching/Biocomputing/>

<http://www.cs.utk.edu/~mclennan/Courses/420>

IT solutions from biology data

BIO-INSPIRED COMPUTING (BIOCOMPUTING)

- Genetic Programming
- Evolutionary algorithms
- Swarm Intelligence
- Cellular Automata
- Neural Computing / Pattern recognition using *neural networks*
(the most widely used form of BIC in industry and science)
- Artificial Immune System methods
- etc

<http://www.macs.hw.ac.uk/~dwcorne/Teaching/bic.html>

IT solutions from biology data

DATA COMPRESSION-BASED APPROACHES TO ANALYSIS OF BIOLOGICAL NETWORKS

Tatsuya Akutsu (Kyoto University)

From Plenary talk on **The Fourth International Conference on Computational Systems Biology (ISB2010)**

Suzhou, China, September 9-11, 2010

The human genome consists of around 3 billion base pairs whereas the number of cells in the human body is estimated to be 60 trillion.

Therefore, it is considered that information on the human body consisting of 60 trillion cells is compressed into 3 billion base pairs.

Deciphering this data compression mechanism is one of major goals of systems biology.

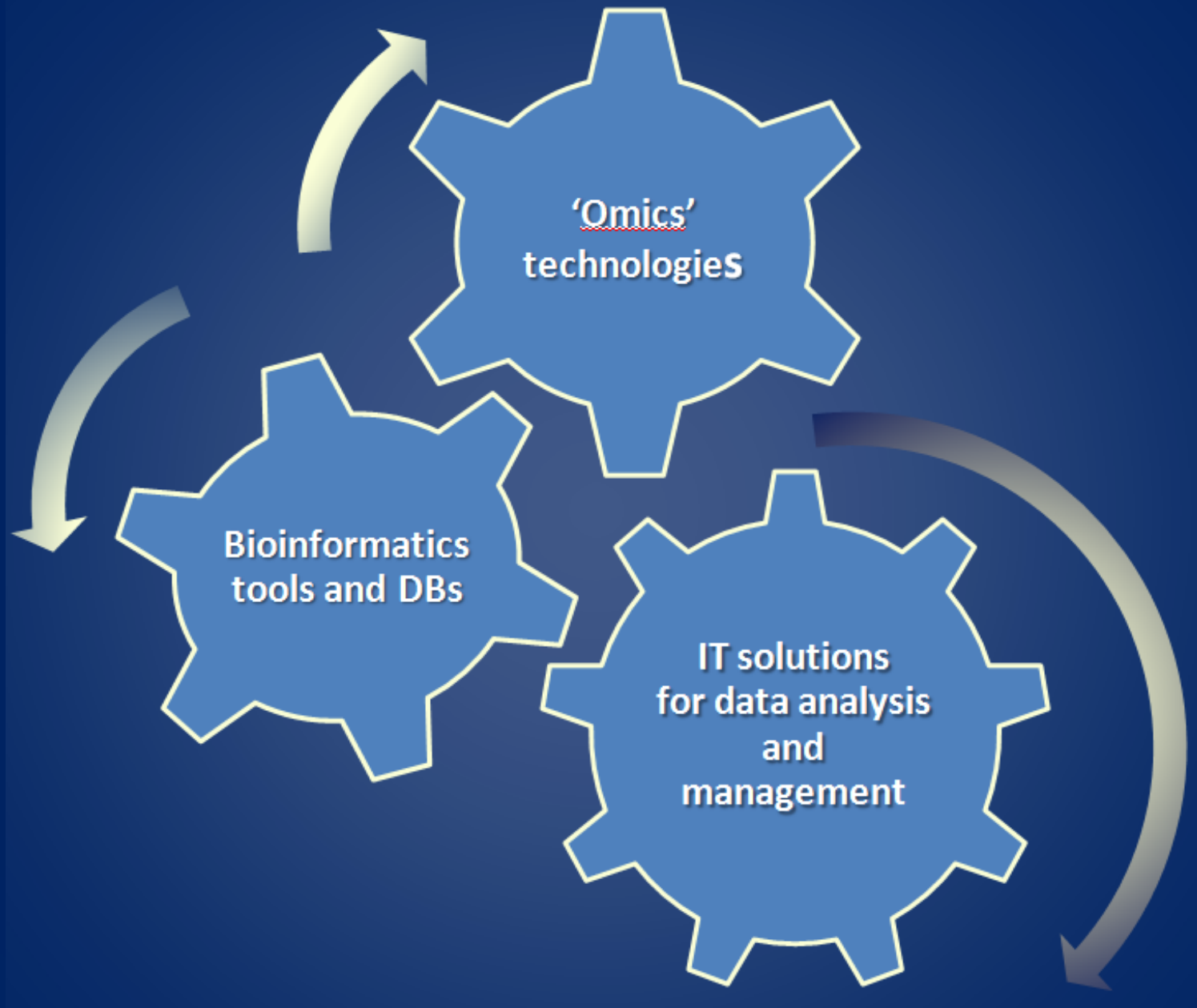
Lecture Notes in Operations Research 13

Series Editors: Ding-Zhu Du and Xiang-Sun Zhang

Computational Systems Biology

<http://www.aporc.org/LNOR/13/>

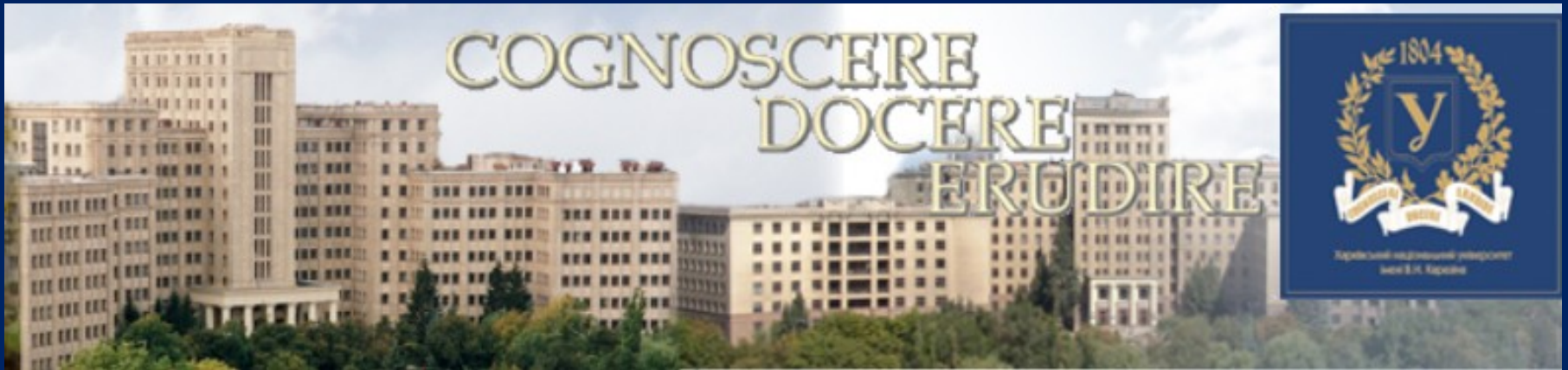
Resume



Acknowledgements to:

Deputy Dean
on science affairs
Professor
Dmytro B. Buy

for invitation to take part in the
Conference with plenary talk



tbarannik@univer.kharkov.ua