

Genetics and Bioinformatics

Kristel Van Steen, PhD²

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

Lecture 1: Setting the pace

1 Bioinformatics – what's in a name ?

Genetics

Molecular biology

Bioinformatics

2 Evolving trends in bioinformatics

Challenges

Topics in bioinformatics from a journal's perspective

3 Bioinformatics software

R and Bioconductor

1 Bioinformatics – what’s in a name?

Genetics

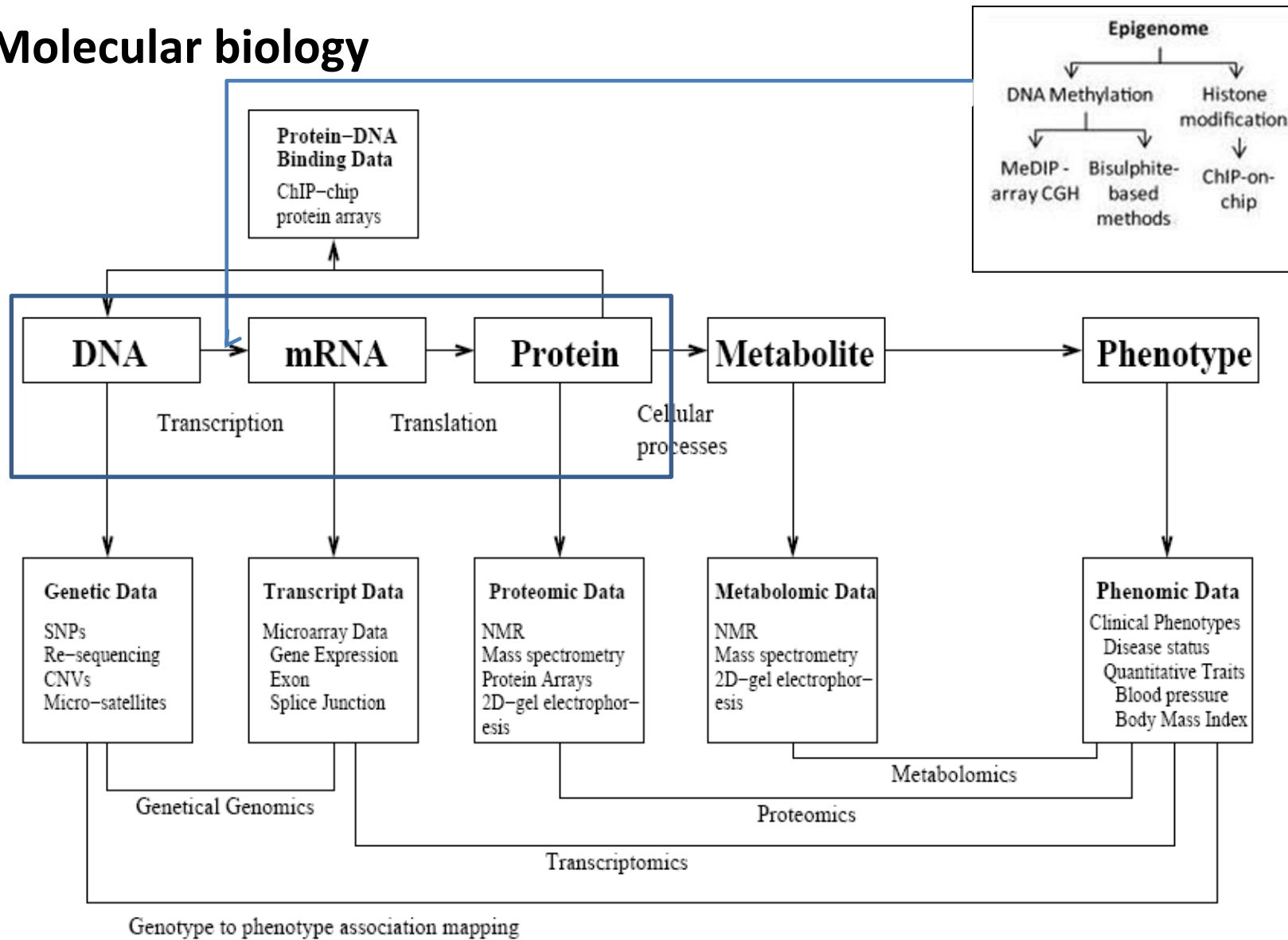
- Genetics is the study of how traits such as hair color, eye color, and risk for disease are passed (“inherited”) from parents to their children. Genetics influence how these inherited **traits** can be different from person to person.
- Your genetic information is called your genetic code or **genome**.

Your genome is made up of a chemical called **deoxyribonucleic acid (DNA)** and is stored in almost every cell in your body.



Genomes in a human: 1
Genes in a human genome: 20,000
Cells in a human body: 75-100 trillion
Chromosomes in a human cell: 46

Molecular biology



(adapted from: Davies et al 2009, Integrative genomics and functional explanation)

Epigenetics – could the central dogma be in danger?

If parents are able to pass environmental information, in the form of epigenetic modifications, on to their offspring as well as their genetic code, epigenetic inheritance adds a whole new dimension to the modern picture of evolution. For over a hundred years we have accepted that the genetic code changes slowly, through the processes of random mutation and natural selection. Epigenetics creates the possibility for a much more rapid response to signals from the environment. It requires a completely different concept of information transfer – experiences had generations ago, such as a famine during your grandmothers time, could influence the way that your body develops, even in today's more plentiful western world.

(<http://blogs.mcgill.ca/osscontributors/2014/01/07/>)

Computational biology

- Biology = noun
- Computational = adjective

When I use my method (or those of others) to answer a biological question, I am doing science. I am learning new biology. The criteria for success has little to do with the computational tools that I use, and is all about whether the new biology is true and has been validated appropriately and to the standards of evidence expected among the biological community. The papers that result report new biological knowledge and are science papers. This is computational biology. (https://rbaltman.wordpress.com)

Computational biology = the study of biology using computational techniques. The goal is to learn new biology, knowledge about living systems. It is about science.

Bioinformatics

- Bio(logy) + Informatics (2 nouns)

When I build a method (usually as software, and with my staff, students, post-docs—I never unfortunately do it myself anymore), I am engaging in an engineering activity: I design it to have certain performance characteristics, I build it using best engineering practices, I validate that it performs as I intended, and I create it to solve not just a single problem, but a class of similar problems that all should be solvable with the software. I then write papers about the method, and these are engineering papers. This is bioinformatics.

(<https://rbaltman.wordpress.com>)

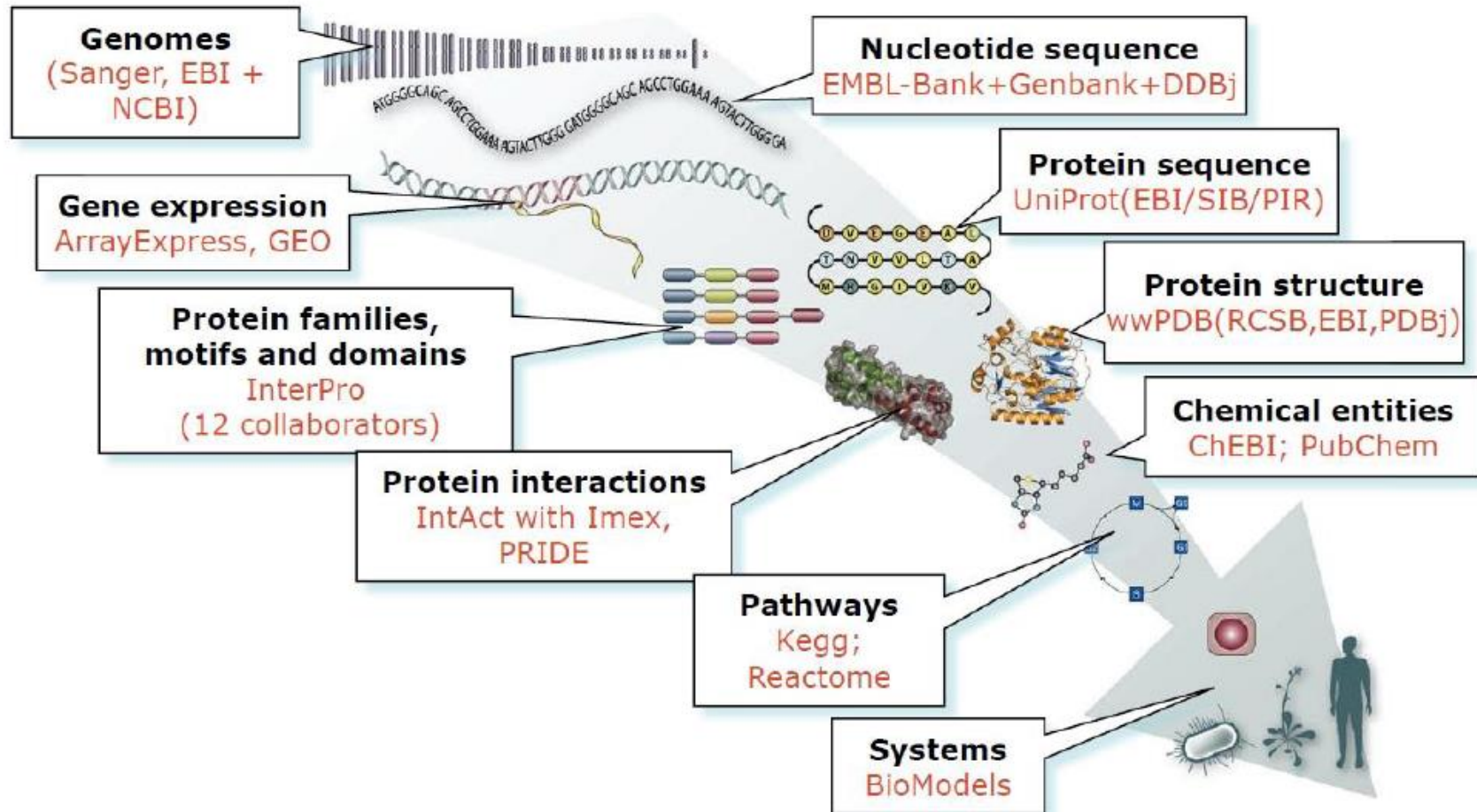
Bioinformatics = the creation of tools (algorithms, databases) that solve problems. The goal is to build useful tools that work on biological data. It is about engineering.

Genetic epidemiology

- Genetic epidemiology is a particular sub-discipline of epidemiology which considers genetic influences on human traits.
- As with any other epidemiologic studies, genetic epidemiology studies aim
 - to assess the public health importance of diseases,
 - to identify the populations at risk,
 - to identify the causes of the disease,
 - and to evaluate potential treatment or prevention strategies based on those findings
- Strategies of analysis include population studies and family studies.
 - Huge challenge is to combine big data repositories from population or family-studies (e.g., genomics, transcriptomics, metagenomics, metabolomics, epigenomics, ...) with clinical and demographic data:

BIOINFORMATICS

Integration with high-throughput omics (data-bases)



(Janet Thornton, EBI)

2 Evolving trends in bioinformatics

Challenges in bioinformatics

- Data deluge (availability, what to archive and what not?, ...)
- Knowledge management (accessibility, usability, ...)
- Predicting, not just explaining (what comes first: hypothesis generation, data collection? ...)
- Precision medicine (alias: personalized medicine; holistic approach – correlating different causal associations – versus a reductionist approach – targeting very specific biomarkers, negative gold standards ... negative controls)
- Speciation (loss in biodiversity, evolutionary units, “integrative taxonomy”: molecular, morphological, ecological and environmental information)
- Inferring the tree of life (unresolved orthology assignment, gene sampling pyramid)

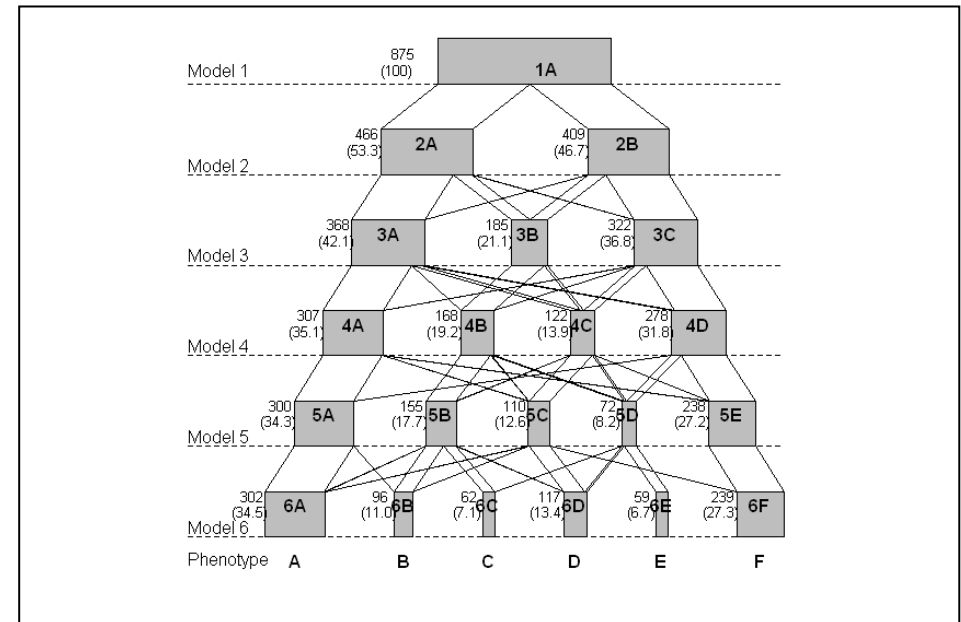
Topics in bioinformatics from a journal's perspective

(source: Scope of the journal "Bioinformatics")

Data and Text Mining

This category includes: New methods and tools for extracting biological information from text, databases and other sources of information. Description of tools to organize, distribute and represent this information. New methods for inferring and predicting biological features based on the extracted

information. The submission of databases and repositories of annotated text, computational tools and general methodology for the work in this area are encouraged, provided that they have been previously tested.



The journal

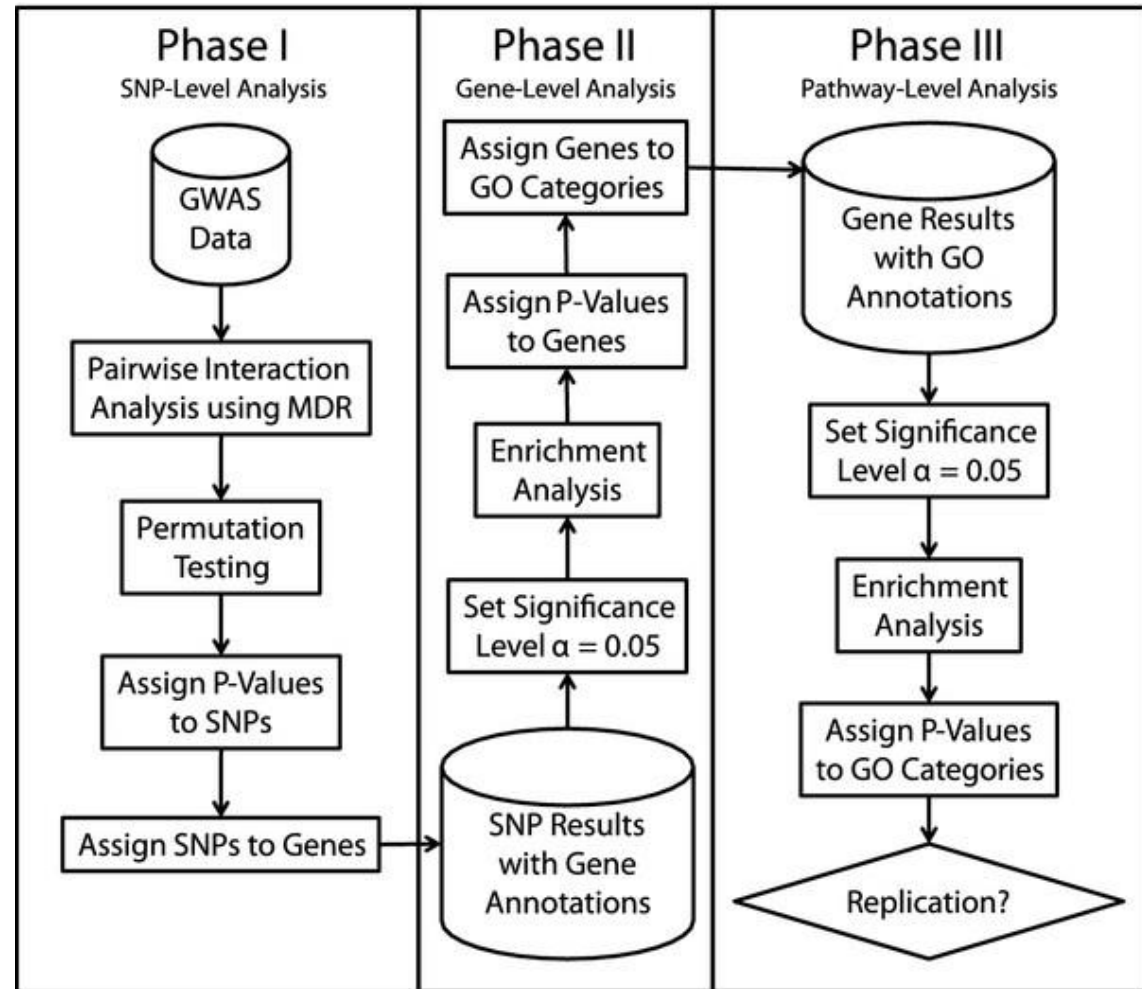
BioData Mining is an open access, peer reviewed, online journal encompassing research on all aspects of data mining applied to high-dimensional biological and biomedical data, focusing on computational aspects of knowledge discovery from large-scale genetic, transcriptomic, genomic, proteomic, and metabolomic data.

Topical areas include, but are not limited to:

- Development, evaluation, and application of novel data mining and machine learning algorithms.
- Adaptation, evaluation, and application of traditional data mining and machine learning algorithms.
- Open-source software for the application of data mining and machine learning algorithms.
- Design, development and integration of databases, software and web services for the storage, management, retrieval, and analysis of data from large scale studies.
- Pre-processing, post-processing, modeling, and interpretation of data mining and machine learning results for biological interpretation and knowledge discovery.

Databases and Ontologies

This category includes: Curated biological databases, data warehouses, eScience, web services, database integration, biologically-relevant ontologies.



(Kim et al. 2012)

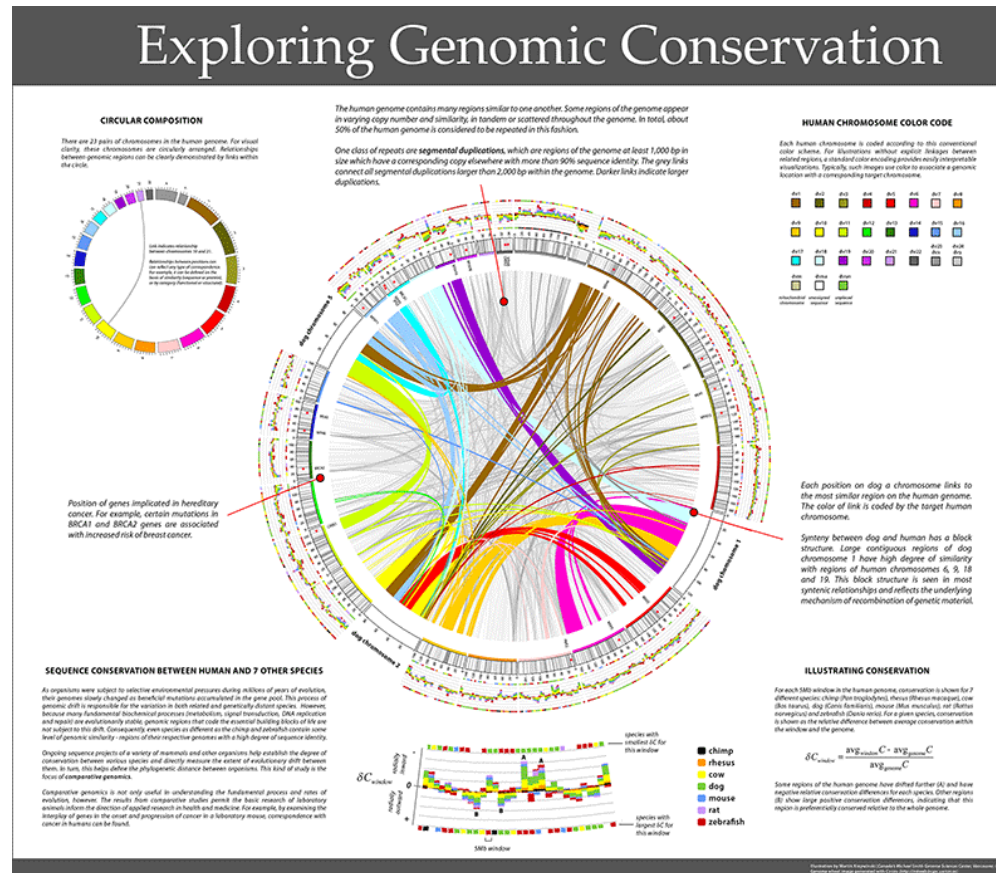
Bioimage Informatics

This category includes novel methods for the acquisition, analysis and modeling of images produced by modern microscopy, with an emphasis on the application of innovative computational methods to solve challenging and significant biological problems at the molecular, sub-cellular, cellular, and tissue levels.

This category also encourages large-scale image informatics methods/applications/software, joint analysis of multiple heterogeneous datasets that include images as a component, and development of bioimage-related ontologies and image retrieval methods.

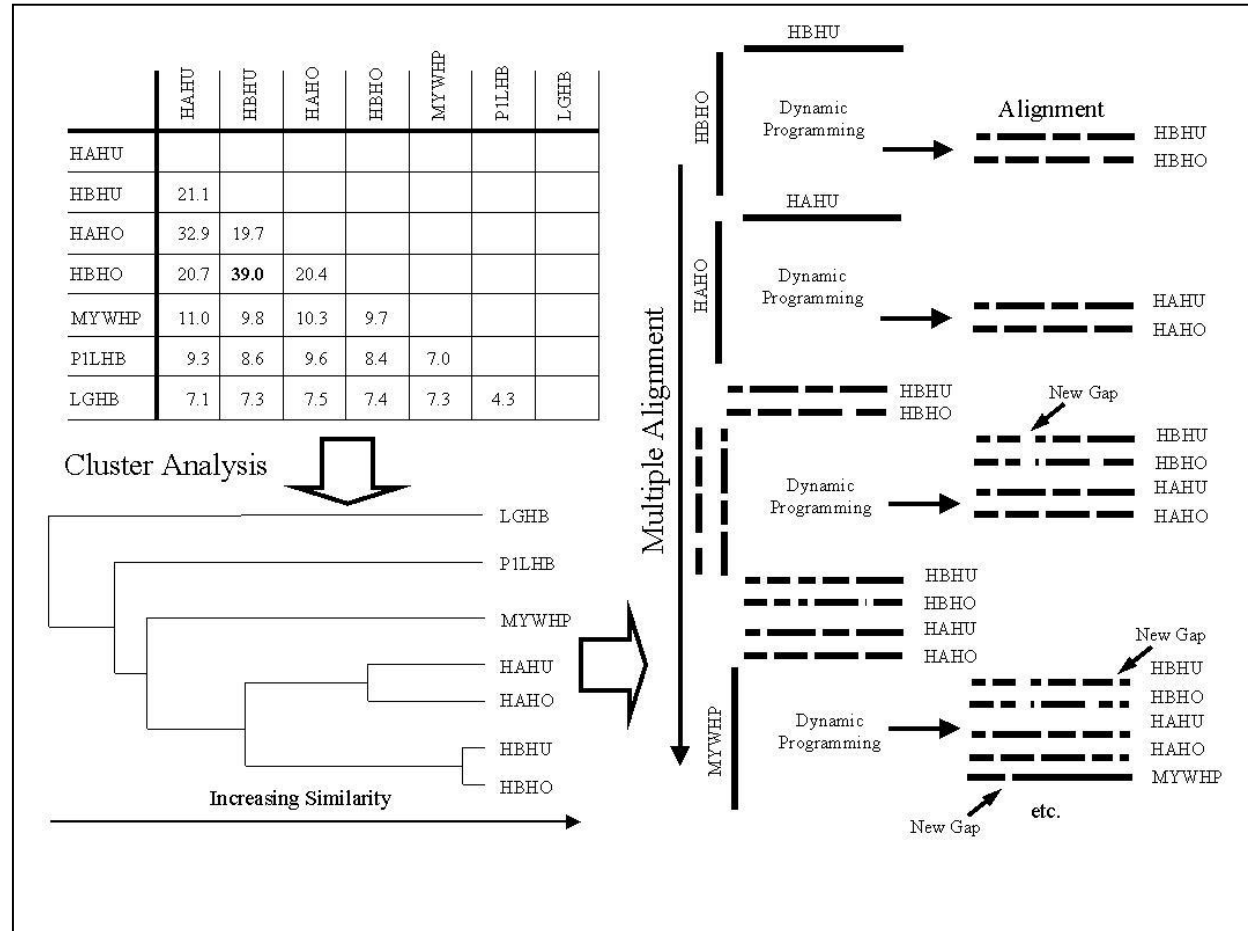
Genome analysis

This category includes: Comparative genomics, genome assembly, genome and chromosome annotation, identification of genomic features such as genes, splice sites and promoters.



Sequence analysis

This category includes: Multiple sequence alignment, sequence searches and clustering; prediction of function and localisation; novel domains and motifs; prediction of protein, RNA and DNA functional sites and other sequence features.

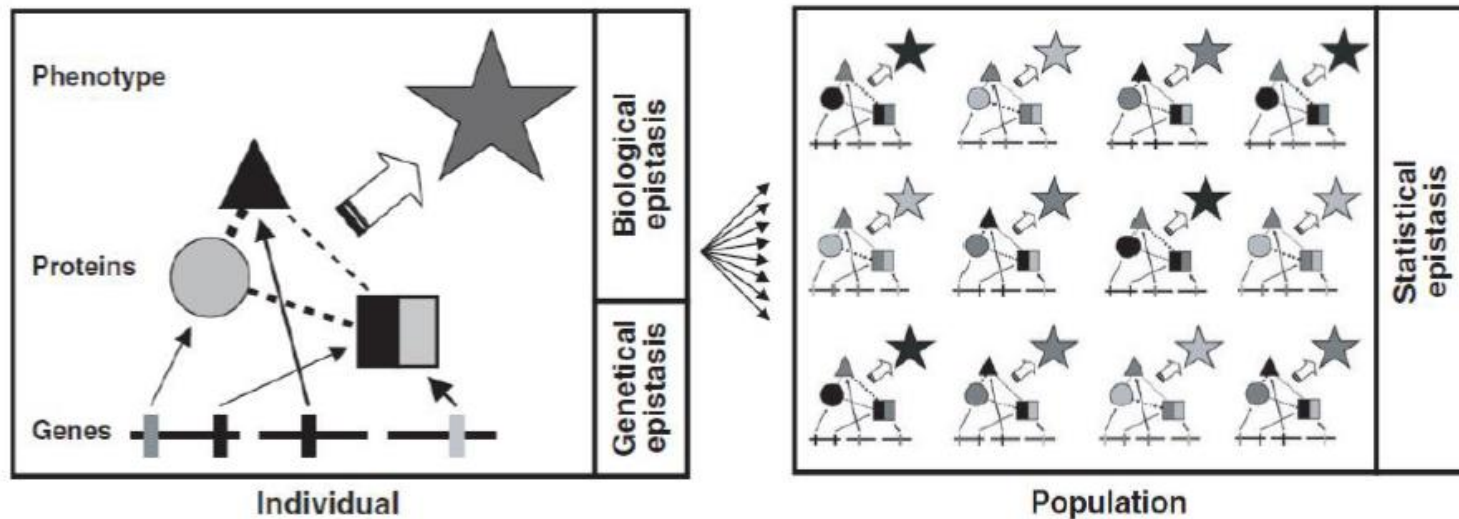


Phylogenetics

This category includes: novel phylogeny estimation procedures for molecular data including nucleotide sequence data, amino acid data, whole genomes, SNPs, etc., simultaneous multiple sequence alignment and phylogeny estimation, phylogenetic approaches for any aspect of molecular sequence analysis (see Sequence Analysis scope), models of molecular evolution, assessments of statistical support of resulting phylogenetic estimates, comparative methods, coalescent theory, approaches for comparing phylogenetic trees, methods for testing and/or mapping character change along a phylogeny.

Gene Expression

This category includes a wide range of applications relevant to the high-throughput analysis of expression of biological quantities, including microarrays (nucleic acid, protein, array CGH, genome tiling, and other arrays), RNA-seq, proteomics and mass spectrometry. Approaches to data analysis to be considered include statistical analysis of differential gene expression; expression-based classifiers; methods to determine or describe regulatory networks; pathway analysis; ...



(Moore 2005)

Systems Biology

This category includes whole cell approaches to molecular biology. Any combination of experimentally collected whole cell systems, pathways or signaling cascades on RNA, proteins, genomes or metabolites that advances the understanding of molecular biology or molecular medicine will be considered. Interactions and binding within or between any of the categories will be considered including protein interaction networks, regulatory networks, metabolic and signaling pathways. Detailed analysis of the biological properties of the systems are of particular interest.

A sample of literature-based bioinformatics resources

BioData Mining

Bioinformatics

BMC Bioinformatics

Briefings in Bioinformatics

Genome Biology

Genome Medicine

Journal of Integrative Bioinformatics

(<http://www.bioinformatics.org/wiki/Journals>)

3 Bioinformatics software

Data access and analysis – data mining

NCBI

Entrez, The Life Sciences Search Engine

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

Search across databases GO Clear Help

Welcome to the Entrez cross-database search page

PubMed: biomedical literature citations and abstracts	Books: online books
PubMed Central: free, full text journal articles	OMIM: online Mendelian Inheritance in Man
Site Search: NCBI web and FTP sites	
Nucleotide: Core subset of nucleotide sequence records	dbGaP: genotype and phenotype
EST: Expressed Sequence Tag records	UniGene: gene-oriented clusters of transcript sequences
GSS: Genome Survey Sequence records	CDD: conserved protein domain database
Protein: sequence database	UniSTS: markers and mapping data
Genome: whole genome sequences	PopSet: population study data sets
Structure: three-dimensional macromolecular structures	GEO Profiles: expression and molecular abundance profiles

Principal internet resources for genome browsers and databases

Resource	Web address	Description	Sponsoring organizations
Open Helix	http://www.openhelix.com/tutorials.shtml	On-line tutorial material for all of the genome databases.	OpenHelix, LLC
UCSC Genome Browser	http://genome.ucsc.edu	Comprehensive, multi-species genome database providing genome browsing and batch querying.	Genome Bioinformatics Group, University of California, Santa Cruz
Ensembl Browser	http://www.ensembl.org	Comprehensive, multi-species genome database providing genome browsing and batch querying.	European Bioinformatics Institute (EBI) and the Sanger Center
NCBI MapViewer	http://www.ncbi.nlm.nih.gov/mapview	Multi-species genome browser focusing especially on genome mapping applications.	National Center for Biotechnology Information (NCBI)

Biomart	http://www.biomart.org/	Genome-database, batch-querying interface used by Ensembl and several single-genome databases.	Ontario Institute for Cancer Research and European Bioinformatics Institute
Galaxy	http://main.g2.bx.psu.edu	Integrated toolset for analyzing genome batch-querying data.	Center for Comparative Genomics and Bioinformatics. Penn State University
Taverna	http://taverna.sourceforge.net	Toolset for creating pipelines of bioinformatics analyses implemented via the Web services protocol.	Open Middleware Infrastructure Institute, University of Southampton (OMII-UK)
GMOD	http://www.gmod.org	Repository of software tools for developing generic genome databases.	A consortium of organizations operating as the Generic Model Organism Database project

(Schattner et al. 2009)

Bioinformatics tools

The screenshot shows the website for the Department of Computational Medicine & Bioinformatics. The header includes the department name and navigation links: HOME, GRADUATE PROGRAM, CCMB, SEMINARS, BIOINFORMATICS CORE, and TRANSMART. A search bar is visible in the top right. The main content area features a dropdown menu for 'Bioinformatics Core' tools, listing various categories such as Analysis package, Database search, Gene expression, and Proteomics. A table with columns 'Tool' and 'Description' is partially visible at the bottom.

Department of Computational Medicine & Bioinformatics

CONTACT US | SITE

HOME GRADUATE PROGRAM CCMB SEMINARS **BIOINFORMATICS CORE** TRANSMART

SEARCH

Home » Bioinformatics Core

Bioinformatics Core

- About Us
- Bioinformatics Tools
- People
- Services/Costs

- Any -

- Analysis package with multiple programs/applications
- Analysis Tools
- Database search (Similarity & Homology)
- Gene expression and microarray
- Gene set enrichment testing
- Genome browser and comparative analysis
- Information retrieval, literature mining and NLP
- Molecular data
- Next-generation sequencing read alignment and assembly programs
- Ontologies/Standards
- Others
- Pathways analysis
- Protein-protein interactions
- Proteomics
- Sequence alignments
- ToolKits/Programming
- Transcription factor binding motif search tools
- Visualization tools
- Workflow tools

- Any -

Apply

Tool	Description

(<http://www.ccmb.med.umich.edu/bioinf-core/tools>)

Bioconductor (TA- sessions)

Open Access

Method

Bioconductor: open software development for computational biology and bioinformatics

Robert C Gentleman¹, Vincent J Carey², Douglas M Bates³, Ben Bolstad⁴, Marcel Dettling⁵, Sandrine Dudoit⁴, Byron Ellis⁶, Laurent Gautier⁷, Yongchao Ge⁸, Jeff Gentry¹, Kurt Hornik⁹, Torsten Hothorn¹⁰, Wolfgang Huber¹¹, Stefano Iacus¹², Rafael Irizarry¹³, Friedrich Leisch⁹, Cheng Li¹, Martin Maechler⁵, Anthony J Rossini¹⁴, Gunther Sawitzki¹⁵, Colin Smith¹⁶, Gordon Smyth¹⁷, Luke Tierney¹⁸, Jean YH Yang¹⁹ and Jianhua Zhang¹

Published: 15 September 2004

Genome Biology 2004, 5:R80

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2004/5/10/R80>

Received: 19 April 2004

Revised: 1 July 2004

Accepted: 3 August 2004

© 2004 Gentleman et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[Home](#)[Install](#)[Help](#)[Developers](#)[About](#)Search:

About *Bioconductor*

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, [1024 software packages](#), and an active user community. Bioconductor is also available as an [AMI](#) (Amazon Machine Image) and a series of [Docker](#) images.

News

- Bioconductor [F1000 Research Channel](#) launched.
- Bioconductor [3.1](#) is available.
- Orchestrating high-throughput genomic analysis with *Bioconductor* ([abstract](#)) and other [recent literature](#)

Install »

Get started with *Bioconductor*

- [Install Bioconductor](#)
- [Explore packages](#)
- [Get support](#)
- [Latest newsletter](#)
- [Follow us on twitter](#)
- [Install R](#)

Learn »

Master *Bioconductor* tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)
- [Community resources](#)
- [Videos](#)

Use »

Create bioinformatic solutions with *Bioconductor*

- [Software](#), [Annotation](#), and [Experiment](#) packages
- [Amazon Machine Image](#)
- [Latest release announcement](#)

Develop »

Contribute to *Bioconductor*

- [Use Bioc 'devel'](#)
- 'Devel' [Software](#), [Annotation](#) and [Experiment](#) packages
- [Package guidelines](#)
- [New package submission](#)

(<http://www.bioconductor.org/>)

R (TA- sessions)

- R is a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc.
 - Consult the R project homepage for further information.
 - The “R-community” is very responsive in addressing practical questions with the software (but consult the FAQ pages first!)
- CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R.



[\[Home\]](#)

Download

[CRAN](#)

R Project

[About R](#)

[Contributors](#)

[What's New?](#)

[Mailing Lists](#)

[Bug Tracking](#)

[Conferences](#)

[Search](#)

R Foundation

[Foundation](#)

[Board](#)

[Members](#)

[Donors](#)

[Donate](#)

Documentation

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

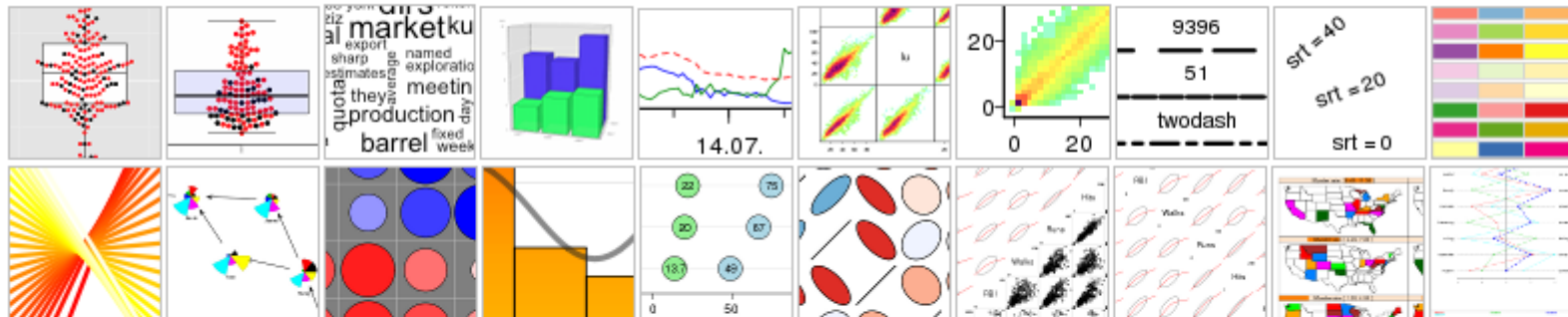
News

- [R version 3.2.2 \(Fire Safety\)](#) has been released on 2015-08-14.
- [The R Journal Volume 7/1](#) is available.
- [R version 3.1.3 \(Smooth Sidewalk\)](#) has been released on 2015-03-09.
- [useR! 2015](#), will take place at the University of Aalborg, Denmark, June 30 - July 3, 2015.
- [useR! 2014](#), took place at the University of California, Los Angeles, USA June 30 - July 3, 2014.

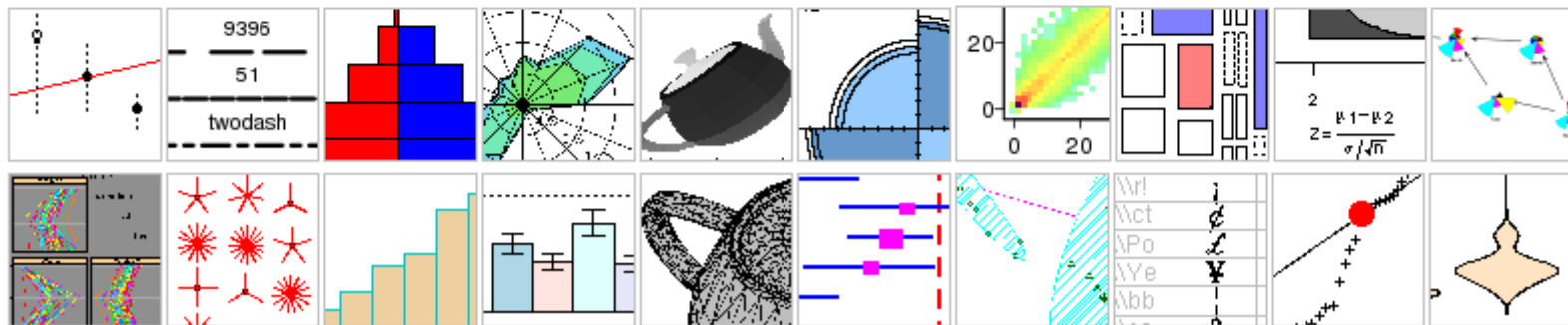
(<https://www.r-project.org/>)

The R graph gallery

» Last entries ...



» Random entries



- One of R's strengths is the ease with which well-designed publication-quality plots can be produced ...