

# B I O I N F O R M A T I C S

**Kristel Van Steen, PhD<sup>2</sup>**

**Montefiore Institute - Systems and Modeling**

**GIGA - Bioinformatics**

**ULg**

**[kristel.vansteen@ulg.ac.be](mailto:kristel.vansteen@ulg.ac.be)**

## **CHAPTER 1: BIOINFORMATICS IN A NUTSHELL**

### **1 Bioinformatics: a new field in engineering**

#### **1 a Introduction**

#### **1 b The origins of computational biology**

#### **1 c The origins of bioinformatics**

### **2 Definition of bioinformatics**

#### **2 a What is bioinformatics**

#### **2 b The importance of bioinformatics**

## **3 Topics in bioinformatics**

## **4 Evolving research trends in bioinformatics**

### **4 a Introduction**

### **4 b Bioinformatics timeline**

### **4 c Careers in bioinformatics**

## **5 Bioinformatics software**

### **5 a Introduction**

### **5 b R and Bioconductor**

# 1 Bioinformatics: a new field in engineering

## 1 a Introduction

- Bioinformatics can be broadly defined as the application of computer techniques to biological data.
- This field has arisen in parallel with the development of automated high-throughput methods of biological and biochemical discovery that yield a variety of forms of experimental data, such as DNA sequences, gene expression patterns, and three-dimensional models of macromolecular structure.
- The field's rapid growth is spurred by the vast potential for new understanding that can lead to new treatments, new drugs, new crops, and the general expansion of knowledge.

([http://findarticles.com/p/articles/mi\\_qa3886/is\\_200301/ai\\_n9182276/](http://findarticles.com/p/articles/mi_qa3886/is_200301/ai_n9182276/))

## Introduction

- Bioinformatics encompasses everything from data storage and retrieval to computational testing of biological hypotheses.
- The data and the techniques can be quite diverse, including such tasks as finding genes in DNA sequences, finding similarities between sequences, predicting structure of proteins, correlating sequence variation with clinical data, and discovering regulatory elements and regulatory networks.
- Bioinformatics systems include multi-layered software, hardware, and experimental solutions that bring together a variety of tools and methods to analyze immense quantities of noisy data.

([http://findarticles.com/p/articles/mi\\_qa3886/is\\_200301/ai\\_n9182276/](http://findarticles.com/p/articles/mi_qa3886/is_200301/ai_n9182276/))

## 1 b The origins of computational biology

### Computational biology

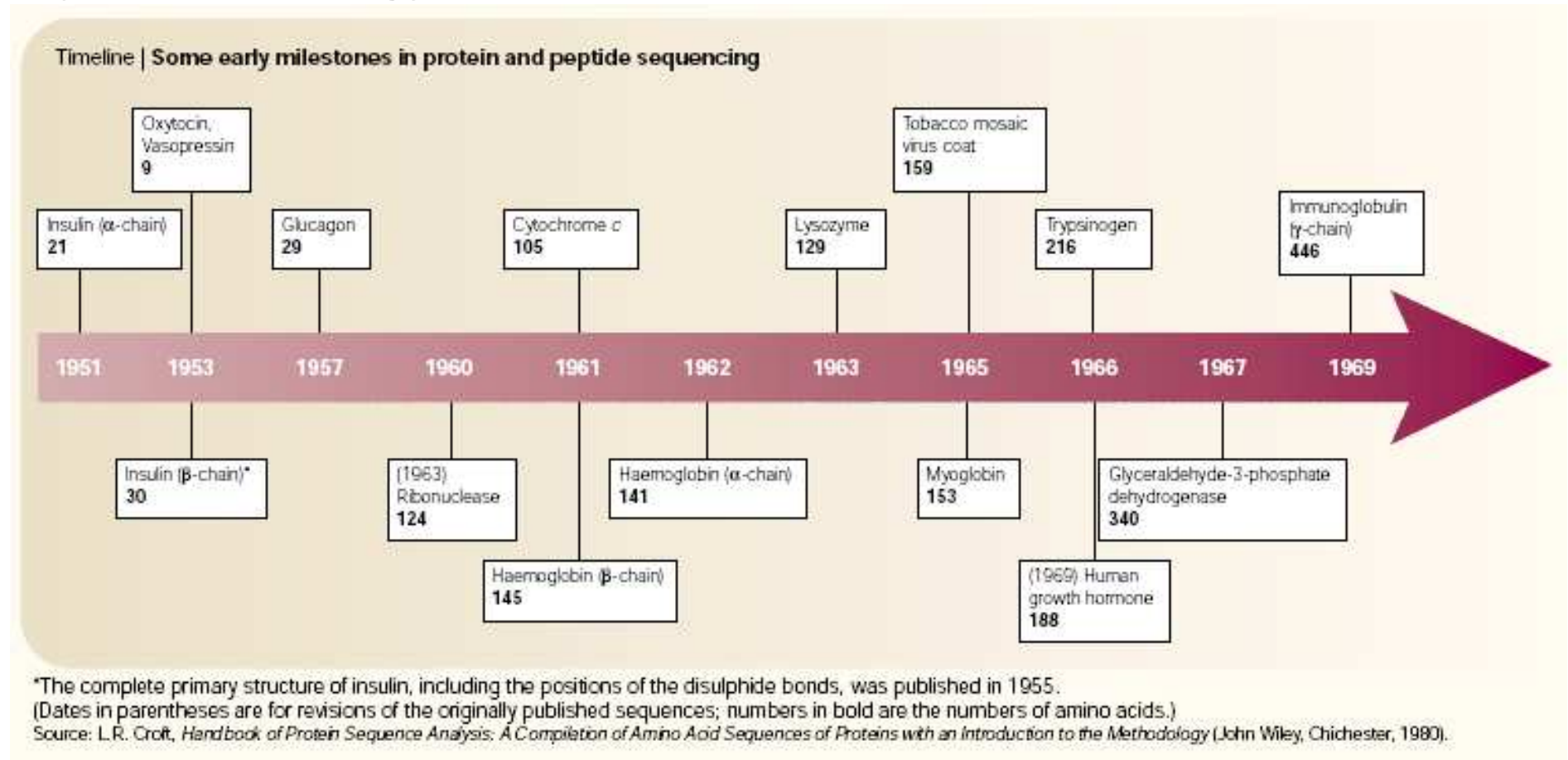
- Three important factors facilitated the emergence of computational biology during the early 1960s.
  - First, an expanding collection of amino-acid sequences provided both a source of data and a set of interesting problems that were infeasible to solve without the number-crunching power of computers.
  - Second, the idea that macromolecules carry information became a central part of the conceptual framework of molecular biology.
    - Thinking in terms of macromolecular information provided an important conceptual link between molecular biology and the computer science from which formal information theory had arisen.

## Computational biology

- Third, high-speed digital computers, which had been developed from weapons research programmes during the Second World War, finally became widely available to academic biologists.
- The idea that proteins carry information encoded in linear sequences of amino acids is commonplace today, but it has a relatively short history.
- This idea first emerged during the decades following the Second World War, a time that one main participant, Emil Smith, later described as a “heroic period” in protein biochemistry.
- The watershed event of this period was the first successful sequencing of a complete protein, INSULIN, by Frederick Sanger and his colleagues<sup>8–10</sup> at Cambridge University during the decade 1945–1955.

(Hagen 2000)

## Computational biology





## Macromolecular information

- Once the polypeptide theory became firmly established and methods for sequencing proteins were readily available, the idea of proteins as information-carrying macromolecules became widespread.
- This general idea developed within three broadly overlapping contexts:
  - the genetic code,
  - the three-dimensional structure of a protein in relation to its function, and
  - protein evolution.

(Hagen 2000)

## The genetic code

- Concurrent developments in the molecular biology of the gene provided a compelling theoretical context for discussing how genetic information was transferred from a sequence of nucleotides to a sequence of amino acids.
- However, the sequencing of DNA and RNA presented formidable technical hurdles that were not fully overcome until the early 1970s.
- So, although molecular biologists learned a great deal about the genetic code, the actual nucleotide sequences of genes remained largely unknown during the 1960s.
- With a growing collection of amino-acid sequences, the idea of molecular information could therefore be explored with proteins in ways not applicable to nucleic acids.

(Hagen 2000)

## Protein structure

- Experiments carried out by Christian Anfinsen and his colleagues at the National Institutes of Health in the late 1950s indicated that the sequence of amino acids completely specified the three-dimensional structure of the protein.
- In practical terms, knowing the sequence did not necessarily allow biochemists to correctly predict the complete structure of the protein.
- Nevertheless, sequence data played a key role in interpreting the X-ray diffraction images used by John Kendrew and Max Perutz to determine the three-dimensional structures of MYOGLOBIN and HAEMOGLOBIN
- Hence, combining the best of all worlds?

(Hagen 2000)

## Protein evolution

- The idea that linear information could determine the structure and function of proteins fits squarely within a dominant tradition in twentieth-century biochemistry
- During the 1960s, biochemists and molecular biologists were increasingly drawn to evolutionary questions.
- For example, Emile Zuckerkandl and Linus Pauling referred to proteins and nucleic acids as “semantides”, whose information-carrying sequences of subunits could be used to document evolutionary history.
- Derived from ‘semanteme’, the fundamental unit of meaning used by linguists to study human speech, semantides were to be the analogous biochemical units (hence the chemical suffix — ide) for evolutionary studies.

(Hagen 2000)

## Protein evolution

- Comparisons of similar proteins, such as myoglobin and haemoglobin, provided evidence for molecular evolution by gene duplication.
- Comparison of homologous proteins drawn from various species could be used to trace phylogenetic relationships among both the proteins themselves and the species that carried them.
- In some cases, such comparisons could also be used to recreate the ancestral proteins from which present-day molecules evolved.
- Assuming that amino-acid substitution rates were relatively constant within a given protein, paleogeneticists (later more commonly known as molecular evolutionists) had a 'molecular clock' by which evolutionary events might be reliably dated.

(Hagen 2000)

## 1 c The origins of bioinformatics

### The emergence of computational biology

- By the early 1960s, computers were becoming widely available to academic researchers.
- According to surveys conducted at the beginning of the decade, 15% of colleges and universities in the United States had at least one computer on campus, and most principal research universities were purchasing so-called 'second generation' computers, based on transistors, to replace the older vacuum-tube models.
- The first high-level programming language FORTRAN (formula translation), was introduced by the International Business Machines (IBM) corporation in 1957.
- It was particularly well suited to scientific applications, and compared with the earlier machine languages, it was relatively easy to learn (Hagen 2000)

## The emergence of bioinformatics

- By 1970, computational biologists had developed a diverse set of techniques for analyzing molecular structure, function and evolution.
- Although originally designed for studying proteins, many of these computing techniques could be adapted for studying nucleic acids.
- Some of these techniques survive today or have lineal descendants that are used in bioinformatics.
- In other cases, they stimulated the development of more refined techniques to correct deficiencies in the original methods.
- Although the nascent field was later revolutionized by the advent of genome projects, large-scale computer networks, immense databases, supercomputers and powerful desktop computers, today's bioinformatics also rests on the important intellectual and technical foundations laid by scientists at an earlier period in the computer era. (Hagen 2000)

## 2 Definitions for bioinformatics

### 2 a What is bioinformatics?

- Some say that bioinformatics is NOT a new discipline, but that it is the same as computational biology and thus actually pretty old ...
- Bioinformatics tries to solve problems in (molecular) biology by the use of computers, and for this reason Bioinformatics and Computational Biology are often considered synonyms



## What is bioinformatics?

- Some people however prefer to clearly differentiate between bioinformatics and computational biology

<b>Bioinformatics</b>	<b>Computational biology</b>
Research, development or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, analyze, or visualize such data	Development and application of data-analytical, theoretical methods, mathematical modeling and computational simulation to the study of biological, behavioral, and social systems.

(BISTIC Definition Committee, NIH, 2000)

## What is bioinformatics?

- In general, bioinformatics is a field of science in which biology, computer science, and information technology merge into a single discipline.
- The ultimate goal of this interdisciplinary field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.

(Y vd Peer 2008)

## 2 b The importance of bioinformatics

- The main challenge facing the bioinformatics community today is the intelligent and efficient storage of this mass of data.
- It is then their responsibility to provide easy and reliable access to this data.
- The data itself are meaningless before analysis and the sheer volume present makes it impossible for even a trained biologist to begin to interpret it manually.
- Therefore, incisive computer tools must be developed to allow the extraction of meaningful biological information.
- There are three central biological processes around which bioinformatics tools must be developed:
  - DNA sequence determines protein sequence
  - Protein sequence determines protein structure
  - Protein structure determines protein function

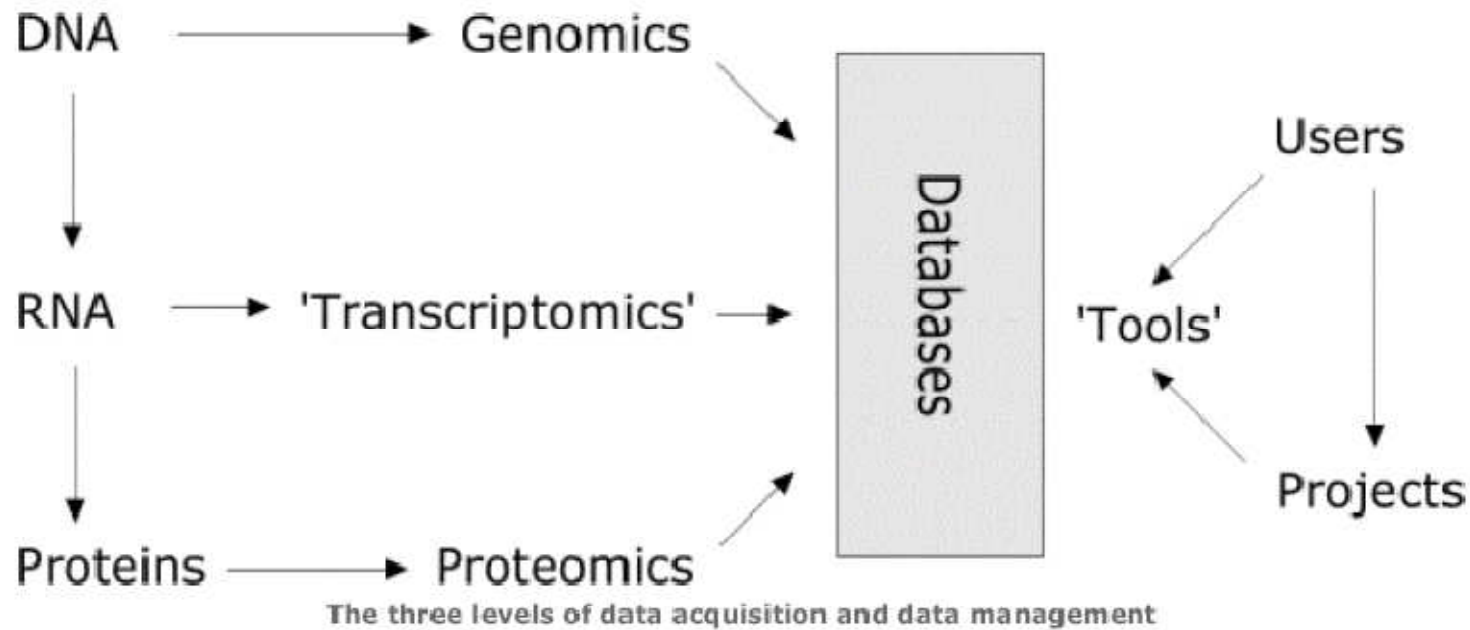
(Y vd Peer 2008)

## The importance of bioinformatics

- Basically, bioinformatics can be said to have 3 major sub-disciplines:
  - the development of new algorithms and statistics (with which to assess relationships among members of large data sets)
  - the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures
  - the development and implementation of tools that enable efficient access and management of different types of information (eg. database development).

(Y vd Peer 2008)

## The importance of bioinformatics



(Y vd Peer 2008)

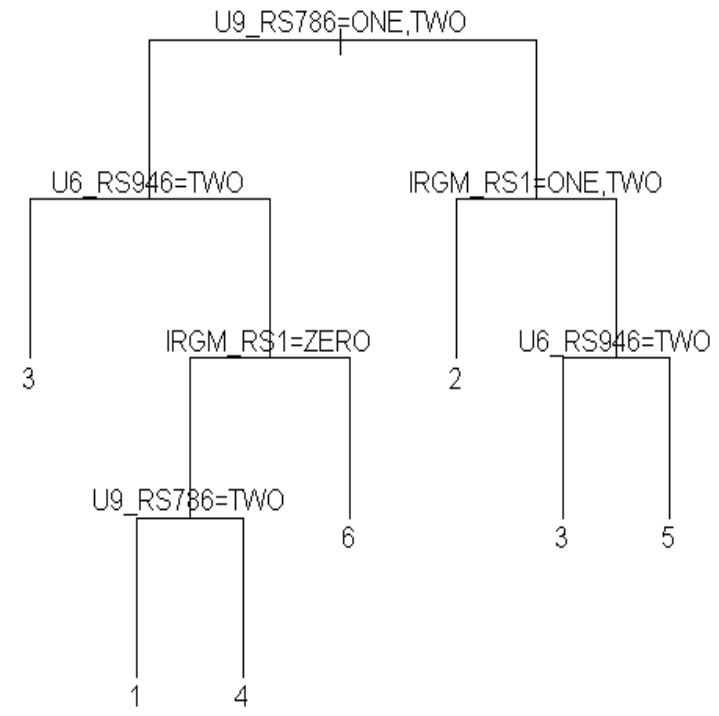
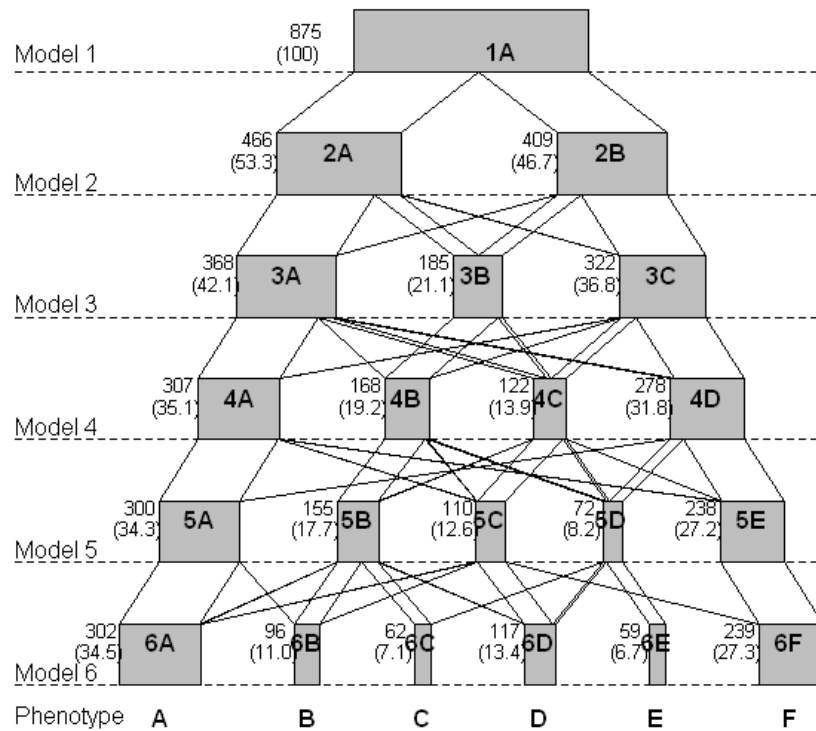
## 3 Topics in bioinformatics

(source: Scope Guidelines 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.
  - Methods for inferring and predicting biological features based on the extracted information.

# Data mining and clustering



## Databases and Ontologies

- This category includes:
  - Curated biological databases
  - Data warehouses
  - eScience
  - Web services
  - Database integration
  - Biologically-relevant ontologies



## Data bases and ontologies

- Collect, organize and classify data
- Query the data
- Retrieve entries based on keyword searches



## 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

## Sequence alignment

- After collection of a set of related sequences, how can we compare them as a set?
- How should we line up the sequences so that the most similar portions are together?
- What do we do with sequences of different length?

```

                2430          2440          2450          2460          2470
HSA128 CACTTCCCCTAT---GCAGGTGTCCAACGGATGTGTGAGTAAAATTCTGGGCAGGTATTA
      ::  :::::  ::  ::::::::::::::::::::::::::::::::::::::::::::::::::::
pax6  CATTTCCCGAATTCTGCAGGTGTCCAACGGATGTGTGAGTAAAATTCTGGGCAGGTATTA
      540      550      560      570      580      590

      2480      2490      2500      2510      2520      2530
HSA128 CGAGACTGGCTCCATCAGACCCAGGGCAATCGGTGGTAGTAAACCGAGAGTAGCGACTCC
      ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
pax6  CGAGACTGGCTCCATCAGACCCAGGGCAATCGGTGGTAGTAAACCGAGAGTAGCGACTCC
      600      610      620      630      640      650

      2540      2550      2560      2570      2580      2590
HSA128 AGAAGTTGTAAGCAAAATAGCCAGTATAAGCGGGAGTGCCCGTCCATCTTTGCTTGGGA

```

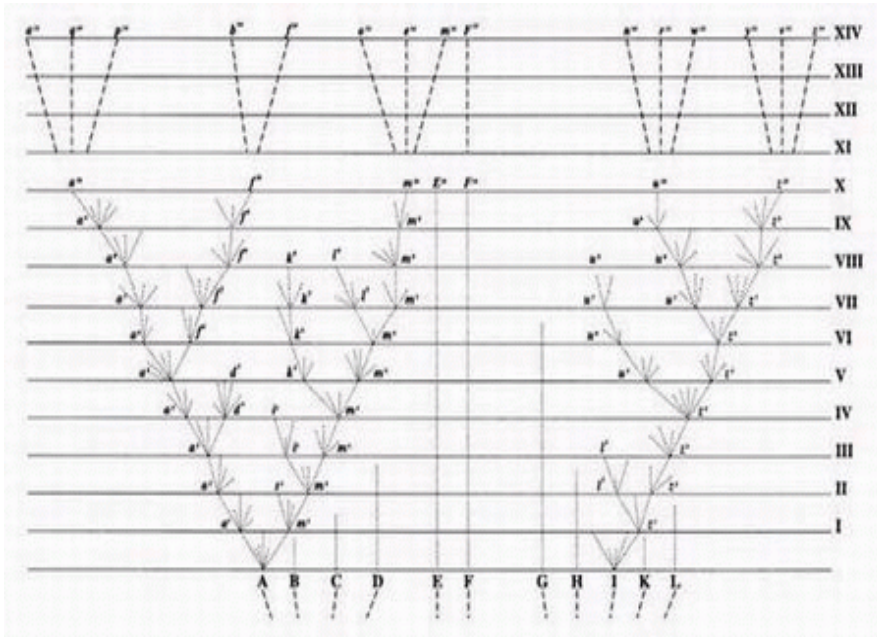
## Genome analysis

- This category includes:
  - Genome assembly
  - Genome and chromosome annotation
  - Gene finding
  - Alternative splicing
  - EST analysis
  - ***Comparative genomics***

## Phylogenetics

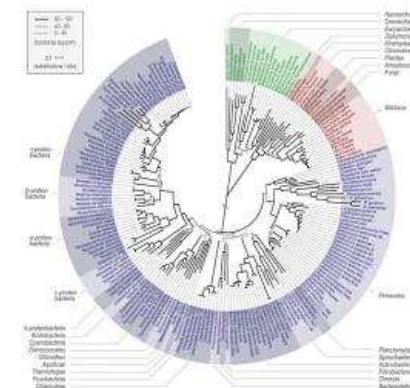
- This category includes:
  - novel phylogeny estimation procedures for molecular data including nucleotide sequence data, amino acid data, SNPs, etc.,
  - simultaneous multiple sequence alignment and
  - phylogeny estimation, using phylogenetic approaches for any aspect of molecular sequence analysis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

## Darwin's tree of life

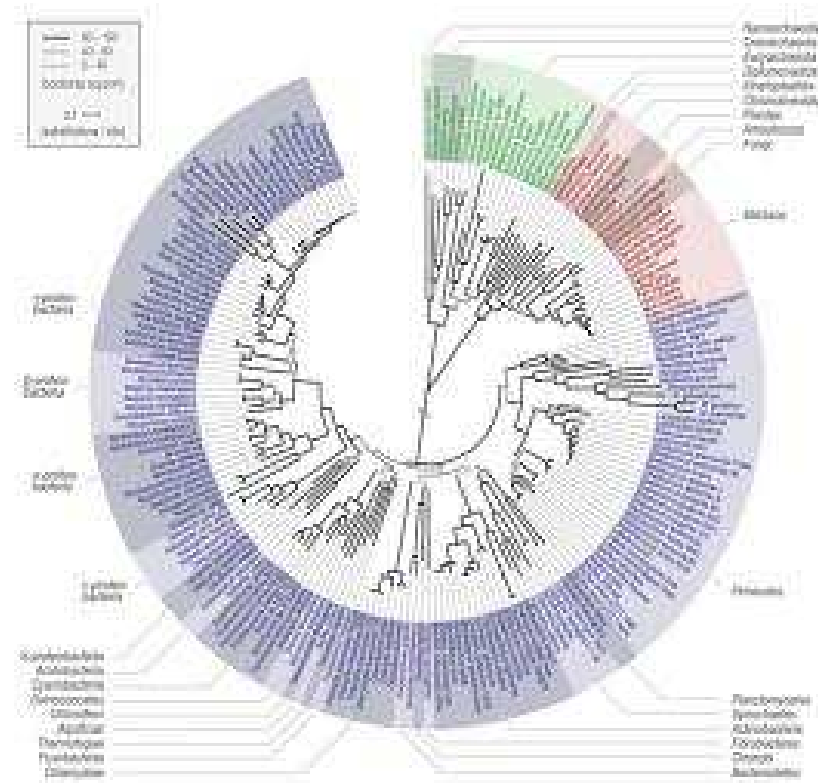


The Tree of Life image that appeared in Darwin's *On the Origin of Species by Natural Selection*, 1859. It was the book's only illustration

A group at the European Molecular Biology Laboratory (EMBL) in Heidelberg has developed a computational method that resolves many of the remaining open questions about evolution and has produced what is likely the most accurate tree of life ever:

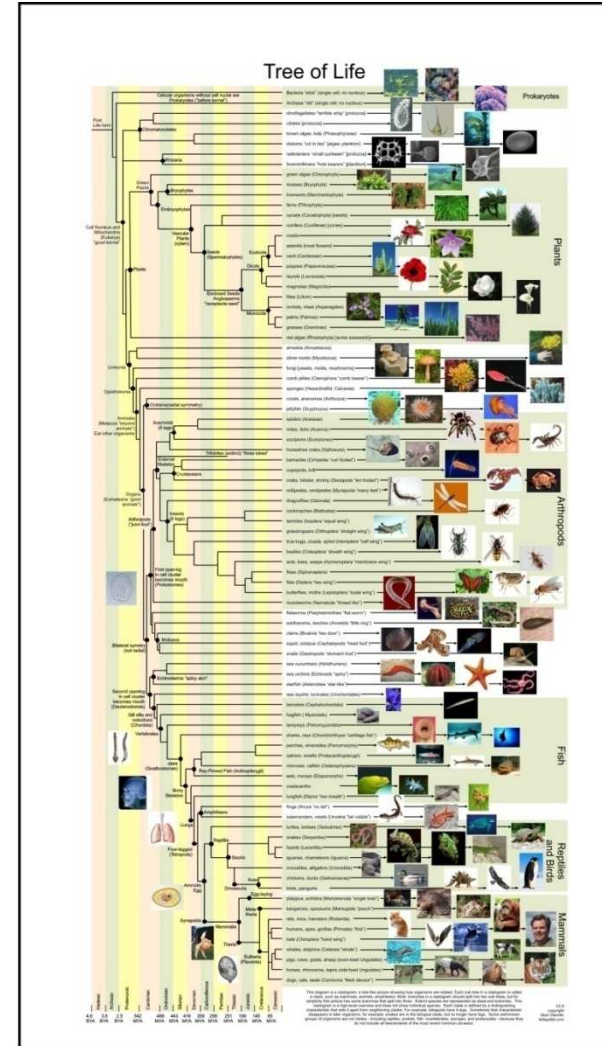


# Modern tree of life



# Modern tree of life

- A modern phylogenetic tree.  
Species are divided into *bacteria*, *archaea*, which are similar to bacteria but evolved differently, and *eucarya*, characterised by a complex cell structure
- A beautiful presentation can be downloaded from  
[http://tallapallet.com/tree\\_of\\_life.htm](http://tallapallet.com/tree_of_life.htm)





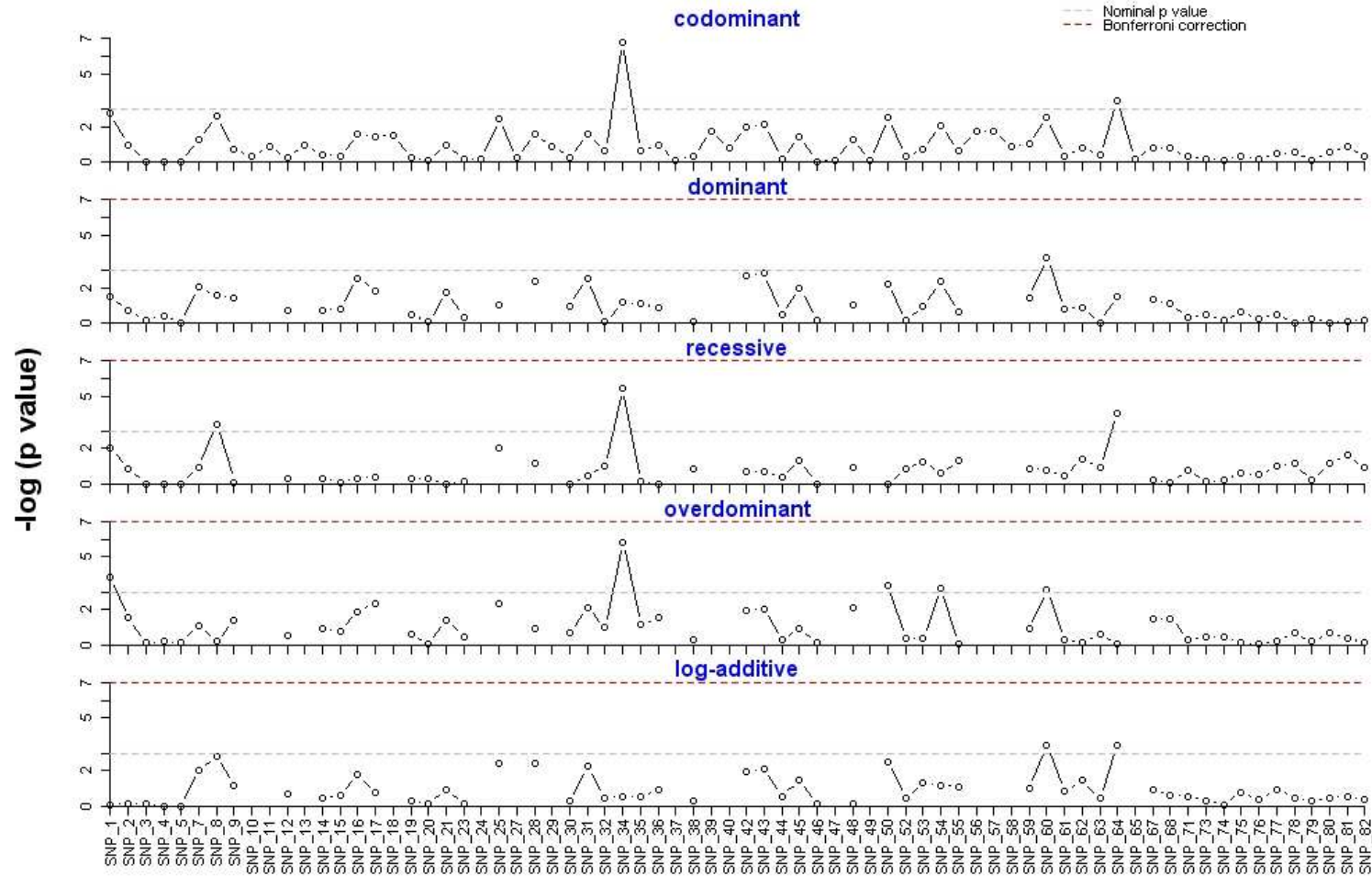
## Structural Bioinformatics

- This category includes:
  - New methods and tools for structure prediction, analysis and comparison;
  - new methods and tools for model validation and assessment;
  - new methods and tools for docking;
  - models of proteins of biomedical interest;
  - protein design;
  - structure based function prediction.

## Genetics and Population Analysis

- This category includes:
  - Segregation analysis,
  - linkage analysis,
  - ***association analysis***,
  - map construction,
  - population simulation,
  - haplotyping,
  - linkage disequilibrium,
  - pedigree drawing,
  - marker discovery,
  - power calculation,
  - genotype calling.

# Genome wide genetic association analysis

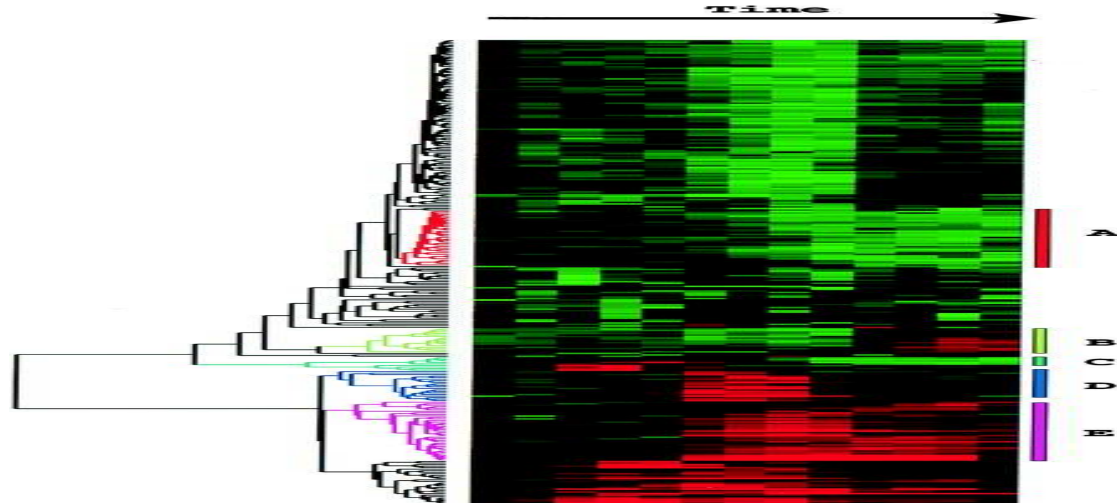


## Gene Expression

- 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), EST, SAGE, MPSS, and related technologies, proteomics and mass spectrometry.
  - Approaches to data analysis in this area include statistical analysis of differential gene expression; expression-based classifiers; methods to determine or describe regulatory networks; pathway analysis; integration of expression data; expression-based annotation (e.g., Gene Ontology) of genes and gene sets, and other approaches to meta-analysis.

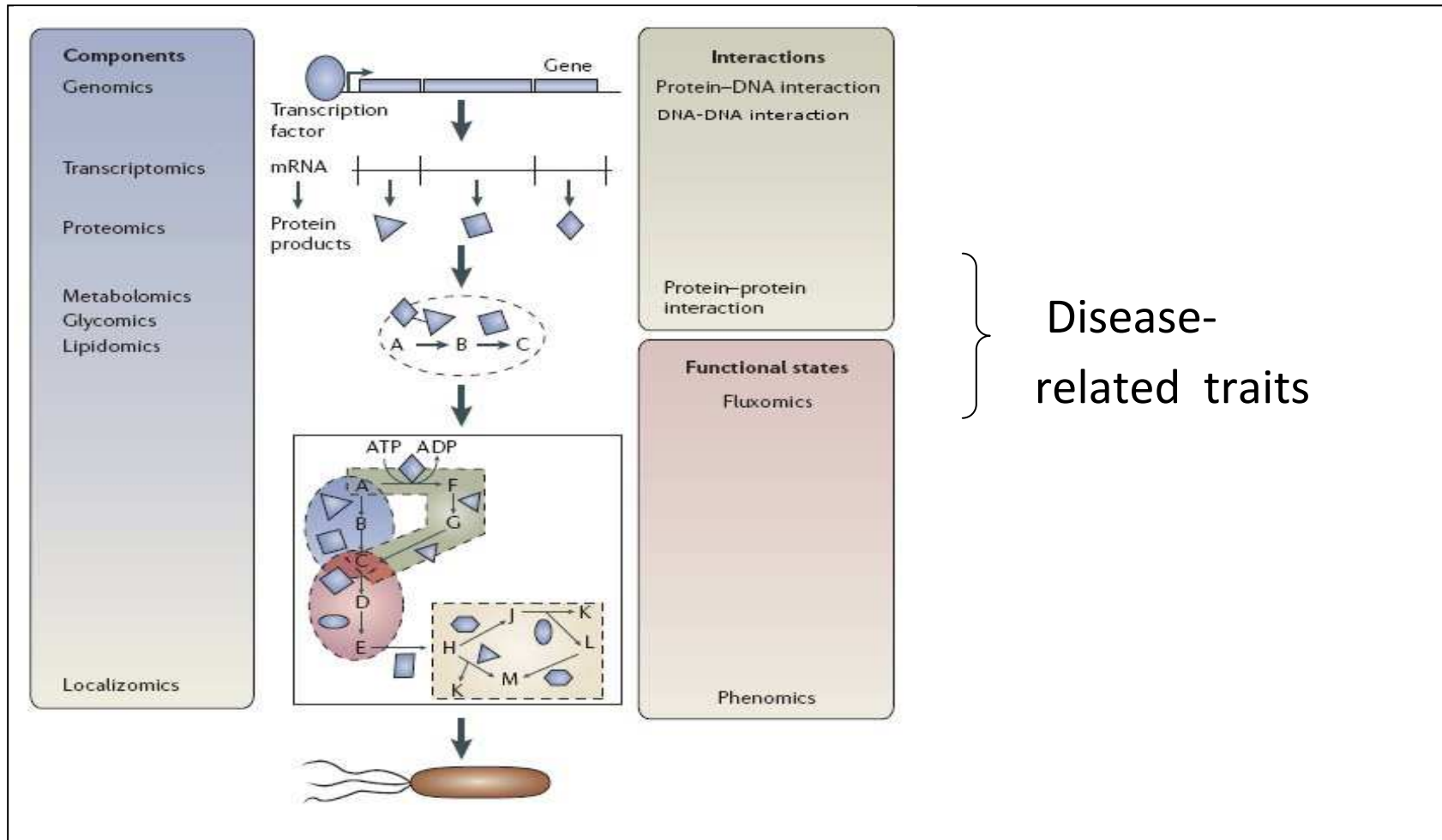
## Analysis of gene expression studies

- Technologies have now been designed to measure the relative number of copies of a genetic message (levels of gene expression) at different stages in development or disease or in different tissues. Such technologies, such as DNA microarrays are growing in importance.



## 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 fall under systems biology;
  - interactions and binding within or between any of the categories including protein interaction networks, regulatory networks, metabolic and signaling pathways.



(Joyce et al 2005 - the model organism as a system: integrating omics data sets)

## 4 Evolving research trends in bioinformatics

### 4 a Introduction

- The questions asked and answered during the early days of bioinformatics were quite different than those that are relevant nowadays.
- At the beginning of the "genomic revolution", a bioinformatics concern was the creation and maintenance of a database to store biological information, such as nucleotide and amino acid sequences.
- Development of this type of database involved not only design issues but the development of complex interfaces whereby researchers could both access existing data as well as submit new or revised data



## Introduction

- To learn more about “early bioinformatics”, please refer to the background reading section (Ouzounis et al 2003)

*BIOINFORMATICS*

**REVIEW**

Vol. 19 no. 17 2003, pages 2176–2190  
DOI: 10.1093/bioinformatics/btg309

---



### ***Early bioinformatics: the birth of a discipline— a personal view***

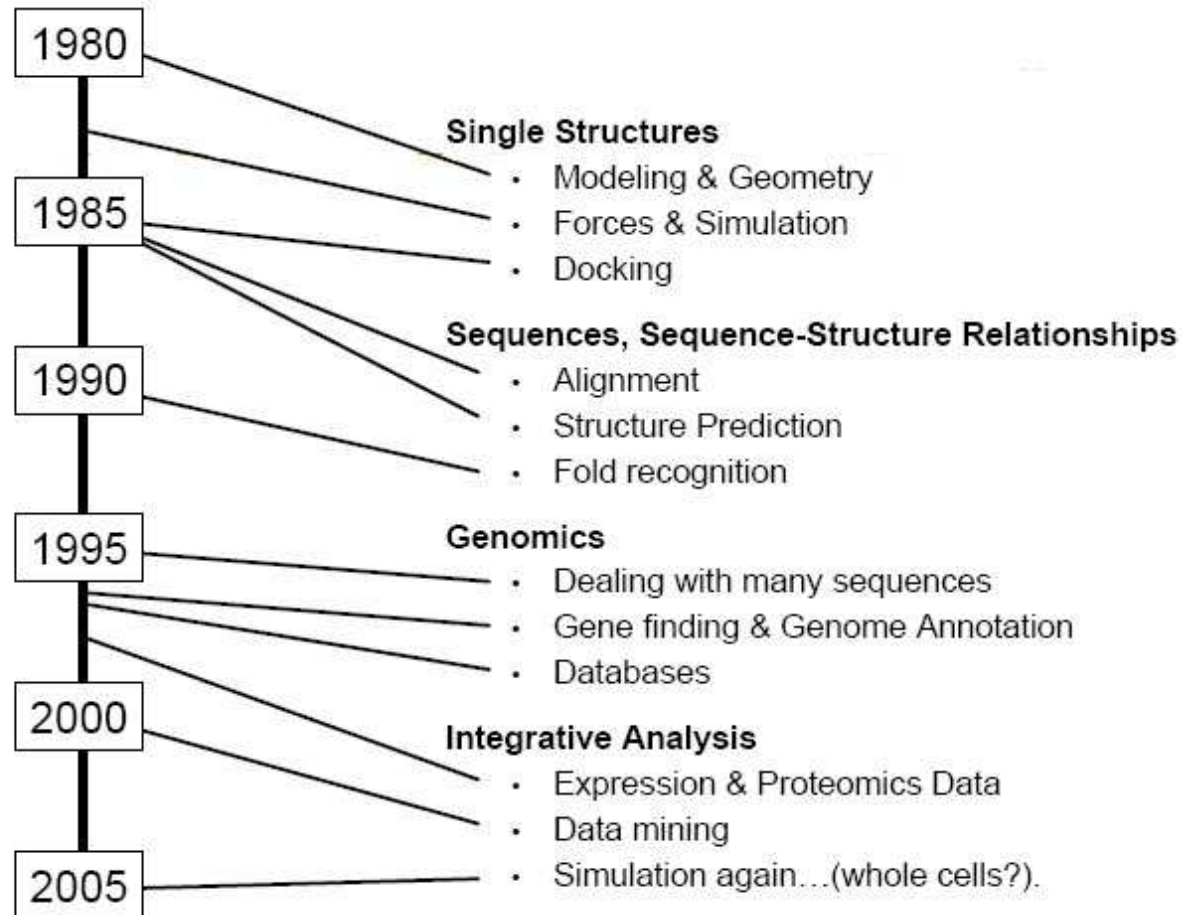
*Christos A. Ouzounis<sup>1,\*</sup> and Alfonso Valencia<sup>2</sup>*

*<sup>1</sup>Computational Genomics Group, The European Bioinformatics Institute, EMBL  
Cambridge Outstation, Cambridge CB10 1SD, UK, <sup>2</sup>Protein Design Group, National  
Center for Biotechnology, CNB-CSIC Campus U. Autonoma Cantoblanco, Madrid  
28049, Spain*

Received on December 13, 2002; revised on May 25, 2003; accepted on March 28, 2003

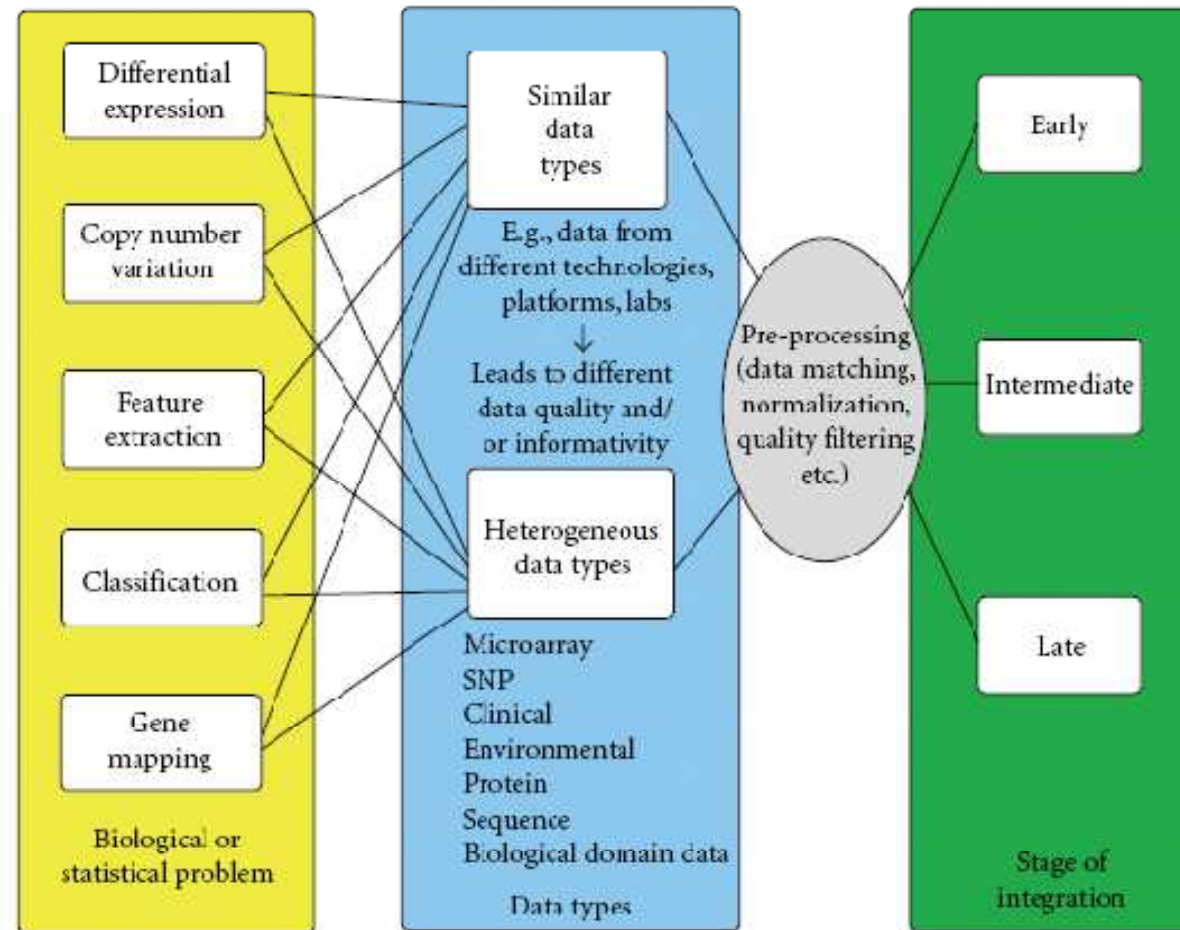
---

## 4 b Bioinformatics timeline



(S-Star presentation; Choo)

## Conceptual framework for data integration in genetics and genomics



(Hamid et al 2009 – data integration in genetics and genomics: methods and challenges)

## 4 c Careers in bioinformatics

- Job sectors include:
  - Academia
  - Research institutes
  - Biotechnology
  - Bioinformatics
  - Pharmaceutical
  - Agriculture
  - Biodiversity

## Careers in bioinformatics

# BioinformaticsBlog.org

bioinformatics in academia and industry; tricks and techniques and life as a bioinformatician

### About the BioinformaticsBlog

The BioinformaticsBlog is a blog dedicated to describing experiences and opinions with bioinformatics software, philosophy and infrastructure. This has been a work in progress for the last 5 years, but as a New Years Resolution for 2009 I am hopeful that it might spring to the forefront of our awareness and be of benefit to a few in the community!

As bioinformaticians we have dedicated much of our working lives to facing the chaos that is the interface between biological data, systems biology and information technology. Following my own roller-coaster ride through academia and industry, having worked with fascinating and talented bioinformaticians in three different countries I have my own views of the subject. I have interests in open-source software, high-performance and distributed bio-computing, high-throughput biotechnologies and meta aggregation of biological data. Hopefully this

### Pages

- » [About the BioinformaticsBlog](#)
- » [Links, pit-stops and destinations](#)

### Archives

- » [August 2009](#)
- » [June 2009](#)
- » [April 2009](#)
- » [March 2009](#)
- » [February 2009](#)
- » [January 2009](#)

### Categories

- » [absolutely nothing at all to do with bioinformatics](#) (13)
- » [best working practices](#) (8)

## Bioinformatics crosses many disciplines



**Statistical Genetics Research Club ([www.statgen.be](http://www.statgen.be))**

## 5 Bioinformatics Software

### 5 a Introduction

- Go commercial or not?
  - The advantage of commercial packages is the support given, and the fact that the programs that are part of the same package are mutually compatible. The latter is not always the case with freeware or shareware
  - The disadvantage is that some of these commercially available software packages are rather expensive ...
- One of the best known commercial software packages in bioinformatics is the GCG (Genetics Computer Group) package
- One of the best known non-commercial software environments is R with BioConductor

## 5 b R and Bioconductor

- 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!)
- Bioconductor is an open source and open development software project to provide tools for the analysis and comprehension of genomic data, primarily based on the R programming language, but containing contributions in other programming languages as well.
- 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.



# The R environment

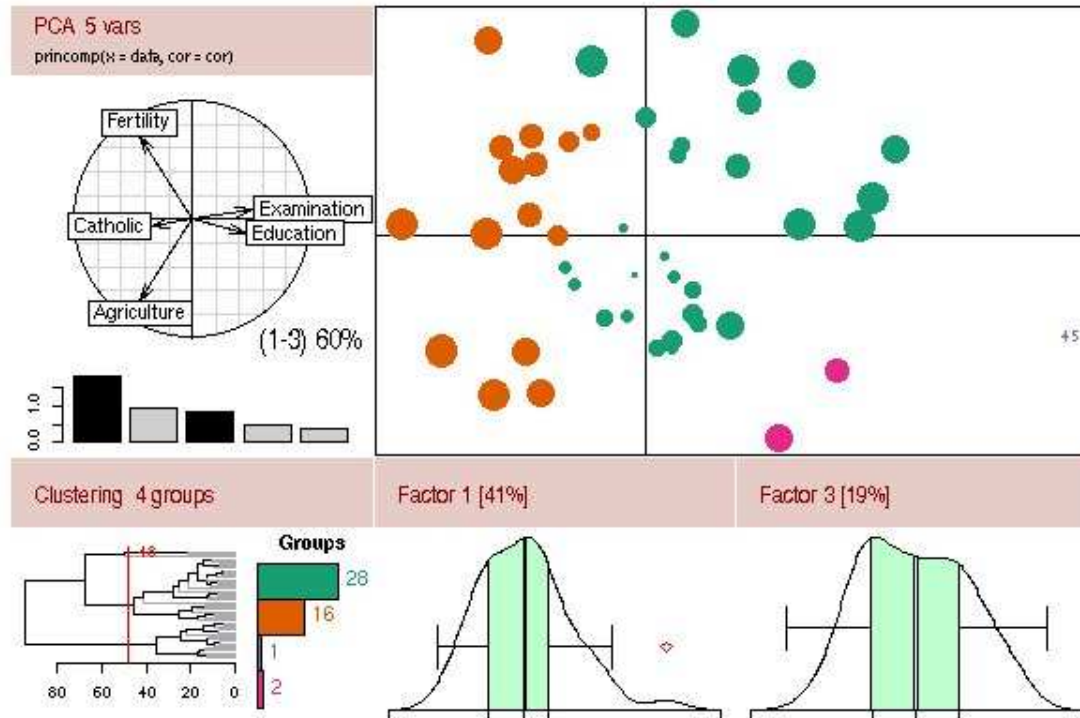


## The R Project for Statistical Computing

About R  
[What is R?](#)  
[Contributors](#)  
[Screenshots](#)  
[What's new?](#)

Download,  
Packages  
[CRAN](#)

R Project  
[Foundation](#)  
[Members & Donors](#)  
[Mailing Lists](#)  
[Bug Tracking](#)  
[Developer Page](#)  
[Conferences](#)  
[Search](#)



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

# Bioconductor

**BIOCONDUCTOR**  
open source software for bioinformatics

Bioconductor is an open source and open development software project for the analysis and comprehension of genomic data.

home getting started overview downloads documentation publications workshops cabig

**project news**

- ▶ 2009-01-07  
R, the open source platform used by Bioconductor, featured in a series of articles in the New York Times.  
[More...](#)

**QUICK LINKS**

- ▶ Getting Started
- ▶ Installation
- ▶ Downloads
- ▶ Software
- ▶ Workshops

**BioC2009 Conference**  
Seattle, WA, 27-28 July 2009. [Conference Material](#)

**Gene expression based on sequencing technologies**  
Copenhagen, Denmark, 24-25 August 2009. [Details and Registration](#)

**Bioconductor 2.4 released – April 21, 2009**  
Following the usual 6-month cycle, the Bioconductor community has released Bioconductor 2.4

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

The screenshot shows the Bioconductor website's installation instructions page. On the left is a navigation menu with links for Getting Started, Overview, Downloads, Documentation, Workflows, Installation, FAQ, Package Slides, Annual Reports, Monograph, Publications, Workshops, Developers, and News. The main content area is titled 'Installation Instructions' and 'Install R'. It contains a three-step list: 1. Download R from CRAN, 2. Start the R program, and 3. Start the R help browser. Below this is a section for installing standard Bioconductor packages, which includes a code block for the `biocLite.R` script. A right-hand sidebar contains a search box and a 'News' section with a link to a 2009-01-07 article.

[Getting Started](#)  
[Overview](#)  
[Downloads](#)  
[Documentation](#)  
[Workflows](#)  
[Installation](#)  
[FAQ](#)  
[Package Slides](#)  
[Annual Reports](#)  
[Monograph](#)  
[Publications](#)  
[Workshops](#)  
[Developers](#)  
[News](#)

## Installation Instructions

### Install R

1. Download the most recent version of R from [The Comprehensive R Archive Network \(CRAN\)](#). The [R FAQ](#) and the [R Installation and Administration Manual](#) contain detailed instructions for installing R on various platforms (Linux, OS X, and Windows being the main ones).
2. Start the R program; on Windows and OS X, this will usually mean double-clicking on the R application, on UNIX-like systems, type "R" at a shell prompt.
3. As a first step with R, start the R help browser by typing "help.start()" in the R command window. For help on any function, e.g. the "mean" function, type "? mean".

### Install standard Bioconductor packages

Install BioConductor packages using the `biocLite.R` installation script. In an R command window, type the following:

```
source("http://bioconductor.org/biocLite.R")
biocLite()
```

This installs the following packages: `affy`, `affydata`, `affyPLM`, `annaffy`, `annotate`, `Biobase`, `Biostrings`, `DynDoc`, `gcrma`, `genefilter`, `genefilter`, `genefilter`, `genefilter`, `hgu95av2.db`, `limma`, `marray`, `matchprobes`, `multtest`, `ROC`, `vsn`, `xtable`, `affyQCReport`. After downloading and installing these packages, the script prints

In this site  search

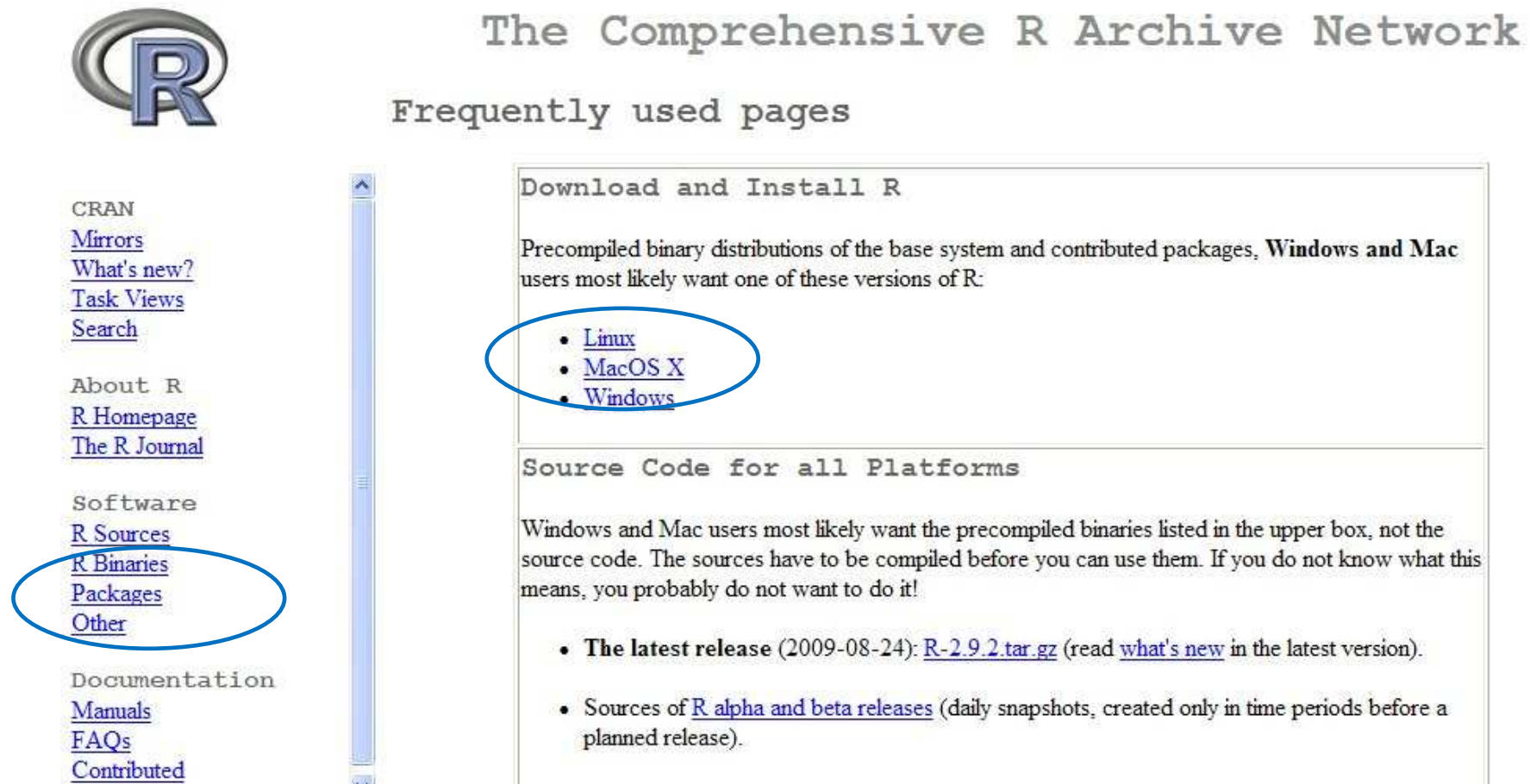
News

2009-01-07  
R, the open source platform used by Bioconductor, featured in a series of articles in the New York Times.  
[More...](#)

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

## R comprehensive network

- Use the CRAN mirror nearest to you to minimize network load.



The screenshot shows the CRAN website interface. On the left is a navigation menu with links for CRAN, Mirrors, What's new?, Task Views, Search, About R, R Homepage, The R Journal, Software, R Sources, R Binaries, Packages, Other, Documentation, Manuals, FAQs, and Contributed. The 'R Binaries' link is circled in blue. The main content area is titled 'The Comprehensive R Archive Network' and 'Frequently used pages'. It features two sections: 'Download and Install R' and 'Source Code for all Platforms'. The 'Download and Install R' section contains a list of links for Linux, MacOS X, and Windows, which are also circled in blue. The 'Source Code for all Platforms' section provides information about precompiled binaries and source code, with a list of links for the latest release and alpha/beta releases.

**The Comprehensive R Archive Network**

Frequently used pages

**Download and Install R**

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Linux](#)
- [MacOS X](#)
- [Windows](#)

**Source Code for all Platforms**

Windows and Mac users most likely want the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- **The latest release** (2009-08-24): [R-2.9.2.tar.gz](#) (read [what's new](#) in the latest version).
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).

## R packages



CRAN

[Mirrors](#)

[What's new?](#)

[Task Views](#)

[Search](#)

About R

[R Homepage](#)

[The R Journal](#)

Software

[R Sources](#)

[R Binaries](#)

[Packages](#)

[Other](#)

Documentation

[Manuals](#)

[FAQs](#)

[Contributed](#)

### Contributed Packages

#### Installation of Packages

Please type `help("INSTALL")` or `help("install.packages")` in R for information on how to install packages from this directory. The manual [R Installation and Administration](#) (also contained in the R base sources) explains the process in detail.

[CRAN Task Views](#) allow you to browse packages by topic and provide tools to automatically install all packages for special areas of interest. Currently, 24 views are available.

#### Daily Package Check Results

All packages are tested regularly on machines running [Debian GNU/Linux](#). Packages are also checked under MacOS X and Windows, but only at the day the package appears on CRAN.

The results are summarized in the [check summary](#) (some [timings](#) are also available). Additional details for Windows checking and building can be found in the [Windows check summary](#).

#### Writing Your Own Packages

The manual [Writing R Extensions](#) (also contained in the R base sources) explains how to write new packages and how to contribute them to CRAN.

---

#### Available Bundles and Packages

## R packages

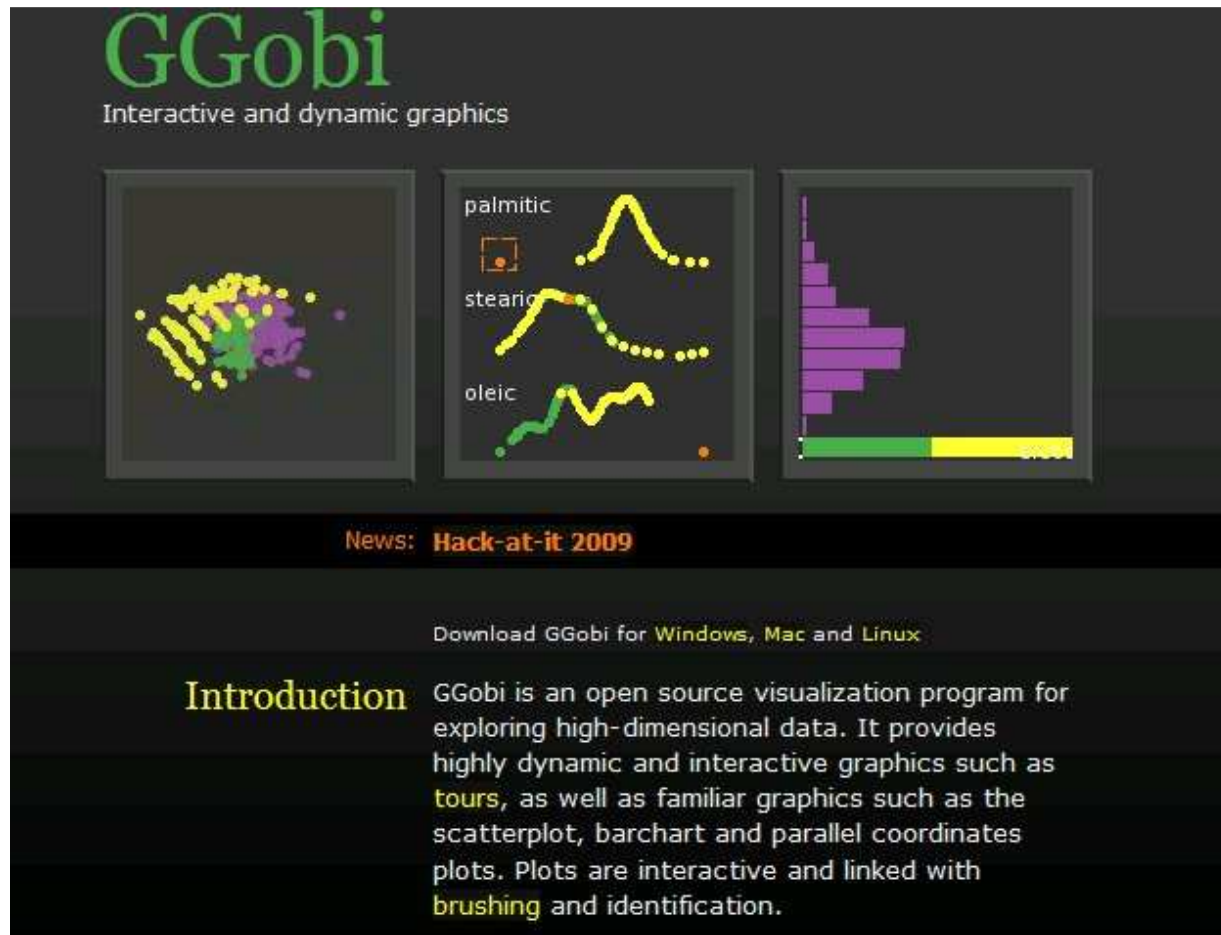
- Go to <http://cran.r-project.org/doc/manuals/R-admin.html> for details on how to install the packages
- Having Bioconductor libraries and packages already installed on your laptop, and also the "ALL" dataset, installed on your laptop prior the lab is a good idea.
- A comprehensive R & BioConductor manual can be obtained via [http://faculty.ucr.edu/~tgirke/Documents/R\\_BioCond/R\\_BioCondManual.html](http://faculty.ucr.edu/~tgirke/Documents/R_BioCond/R_BioCondManual.html)
- Some other useful supporting documents to get started in R are provided on the course website

## Exploratory analysis of omics data

- exploRase leverages the synergy of the statistical analysis platform R with GGobi, a tool for interactive multivariate visualization.
- R provides a wide array of analysis functionality, including Bioconductor.
- Unfortunately, biologists are often discouraged from using the script-driven R as it requires some programming skill.
- Similarly, the usefulness of GGobi is not obvious to those unfamiliar with interactive graphics and exploratory data analysis.
- exploRase attempts to solve this problem by providing access to R analysis and GGobi graphics through a simplified GUI designed for use in Systems Biology research.
- It provides a framework for convenient loading and integrated analysis and visualization of transcriptomic, proteomic, and metabolomic data.

(<https://secure.bioconductor.org/BioC2009/>)

## GGobi



The screenshot displays the GGobi software interface. At the top left, the logo "GGobi" is shown in green, with the tagline "Interactive and dynamic graphics" below it. The main area contains three panels: a scatter plot on the left with yellow, green, and purple points; a central panel with three line plots labeled "palmitic", "stearic", and "oleic" in yellow, each showing a different data trend; and a bar chart on the right with purple bars. Below the plots, there is a "News: Hack-at-it 2009" section. Further down, a link says "Download GGobi for Windows, Mac and Linux". At the bottom, an "Introduction" section describes GGobi as an open source visualization program for high-dimensional data, mentioning features like "tours", "brushing", and "identification".

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



# exploRase

**exploRase**

Overview Download Getting Started Documentation

**Metabolic Network Exchange**

Workshops  
Data Export  
MetNet Web

Functional Genomics Tools  
MetNet Plugins for Cytoscape  
exploRase  
PathBinder  
AtGeneSearch  
MetaOmGraph  
PubMed Assistant  
MetNetDB  
BirdsEyeView

exploRase 0.10  
bio1.yes.mean  
bio1.no.mean  
d# bio1.no.mean.bio1.yes.mean

ExploRase with GGobi graphics

([http://metnet.vrac.iastate.edu/MetNet\\_exploRase.htm](http://metnet.vrac.iastate.edu/MetNet_exploRase.htm))

- Installing is ease: open R and type  
`source("http://www.metnetdb.org/exploRase/files/installer.R")`

## Data mining

- A comprehensive analysis of high-throughput biological experiments involves integration and visualization of a variety of data sources.
- Much of this (meta) data is stored in publicly available databases, accessible through well-defined web interfaces.
  - One simple example is the annotation of a set of features that are found differentially expressed in a microarray experiment with corresponding gene symbols and genomic locations.
- BioMart is a generic, query oriented data management system, capable of integrating distributed data resources.
- It is developed at the European Bioinformatics Institute (EBI) and Cold Spring Harbour Laboratory (CSHL).

(<https://secure.bioconductor.org/BioC2009/>)

## Data mining

- Extremely useful is biomaRt, which is a software package aimed at integrating data from BioMart systems into R, providing efficient access to a wealth of biological data from within a data analysis environment and enabling biological database mining.
- In addition to the retrieval of annotation, one is interested in making customized graphics displaying both the annotation along with experimental data.
- Moreover, the Bioconductor package GenomeGraphs provides a unified framework for plotting data along the chromosome.

(<https://secure.bioconductor.org/BioC2009/>)

# BioMart



- HOME
- MARTVIEW
- MARTSERVICE
- DOCS
- CONTACT
- NEWS
- CREDITS

## BioMart Project

BioMart is a query-oriented data management system developed jointly by the [Ontario Institute for Cancer Research \(OICR\)](#) and the [European Bioinformatics Institute \(EBI\)](#).

The system can be used with any type of data and is particularly suited for providing 'data mining' like searches of complex descriptive data. BioMart comes with an 'out of the box' website that can be installed, configured and customised according to user requirements. Further access is provided by graphical and text based applications or programmatically using web services or API written in Perl and Java. BioMart has built-in support for query optimisation and data federation and in addition can be configured to work as a DAS 1.5 Annotation server. The process of converting a data source into BioMart format is fully automated by the tools included in the package. Currently supported RDBMS platforms are MySQL, Oracle and Postgres.

BioMart is completely Open Source, licensed under the LGPL, and freely available to anyone without restrictions.

### Powered by BioMart software:

- [BioMart Central Portal](#)
- [Ensembl](#)
- [Ensembl Bacteria](#)
- [Ensembl Metazoa](#)
- [Ensembl Protists](#)
- [Dictybase](#)
- [Wormbase](#)
- [Gramene](#)
- [Europhenome](#)
- [UniProt](#)
- [InterPro](#)
- [HGNC](#)
- [Rat Genome Database](#)
- [DroSpeGe](#)
- [ArrayExpress DW](#)
- [Eurexpress](#)
- [HapMap](#)
- [GermOnLine](#)
- [PRIDE](#)
- [PepSeeker](#)
- [VectorBase](#)
- [HTGT](#)
- [Pancreatic Expression Database](#)
- [Reactome](#)
- [EU Rat Mart](#)
- [Paramecium DB](#)
- [International Potato Center \(CIP\)](#)

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

# biomaRt

**biomaRt**

**Interface to BioMart databases (e.g. Ensembl, Wormbase and Gramene)**

In recent years a wealth of biological data has become available in public data repositories. Easy access to these valuable data resources and firm integration with data analysis is needed for comprehensive bioinformatics data analysis. biomaRt provides an interface to a growing collection of databases implementing the BioMart software suite (<http://www.biomaRt.org>). The package enables retrieval of large amounts of data in a uniform way without the need to know the underlying database schemas or write complex SQL queries. Examples of BioMart databases are Ensembl, Uniprot, Gramene, Wormbase and HapMap. These major databases give biomaRt users direct access to a diverse set of data and enable a wide range of powerful online queries from gene annotation to database mining.

Author      Steffen Durinck , Wolfgang Huber , Sean Davis  
 Maintainer   Steffen Durinck

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("biomaRt")
```

**Documentation**

The biomaRt users guide [PDF](#) [R Script](#)  
[Reference Manual](#)

**Details**

biocViews	<a href="#">Annotation</a>
Depends	methods
Imports	<a href="#">XML</a> , <a href="#">RCurl</a>
Suggests	<a href="#">annotate</a>
System Requirements	

(<http://www.bioconductor.org/packages/devel/bioc/html/biomaRt.html>)

## biomaRt

### 4.2 Task 2: Annotate a set of EntrezGene identifiers with GO annotation

In this task we start out with a list of EntrezGene identifiers and we want to retrieve GO identifiers related to biological processes that are associated with these entrezgene identifiers. Again we look at the output of `listAttributes` and `listFilters` to find the filter and attributes we need. Then we construct the following query:

```
> entrez = c("673", "837")
> getBM(attributes = c("entrezgene", "go_biological_process_id"), filters = "entrezgene", values = entrez,
+       mart = ensembl)
```

	entrezgene	go_biological_process_id
1	673	GO:0051291
2	673	GO:0006916
3	673	GO:0009887
4	673	GO:0007264
5	673	GO:0006468
6	673	GO:0000165
7	673	GO:0007242
8	673	GO:0007165
9	837	GO:0006917
10	837	GO:0006508
11	837	GO:0042981
12	837	GO:0006915

(<http://www.bioconductor.org/packages/devel/bioc/vignettes/biomaRt/inst/doc/biomaRt.pdf>)

# GenomeGraphs

## GenomeGraphs

### Plotting genomic information from Ensembl

Genomic data analyses requires integrated visualization of known genomic information and new experimental data. GenomeGraphs uses the biomaRt package to perform live annotation queries to Ensembl and translates this to e.g. gene/transcript structures in viewports of the grid graphics package. This results in genomic information plotted together with your data. Another strength of GenomeGraphs is to plot different data types such as array CGH, gene expression, sequencing and other data, together in one plot using the same genome coordinate system.

Author: Steffen Durinck, James Bullard

Maintainer: Steffen Durinck

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("GenomeGraphs")
```

#### Vignettes (Documentation)

[GenomeGraphs.pdf](#)

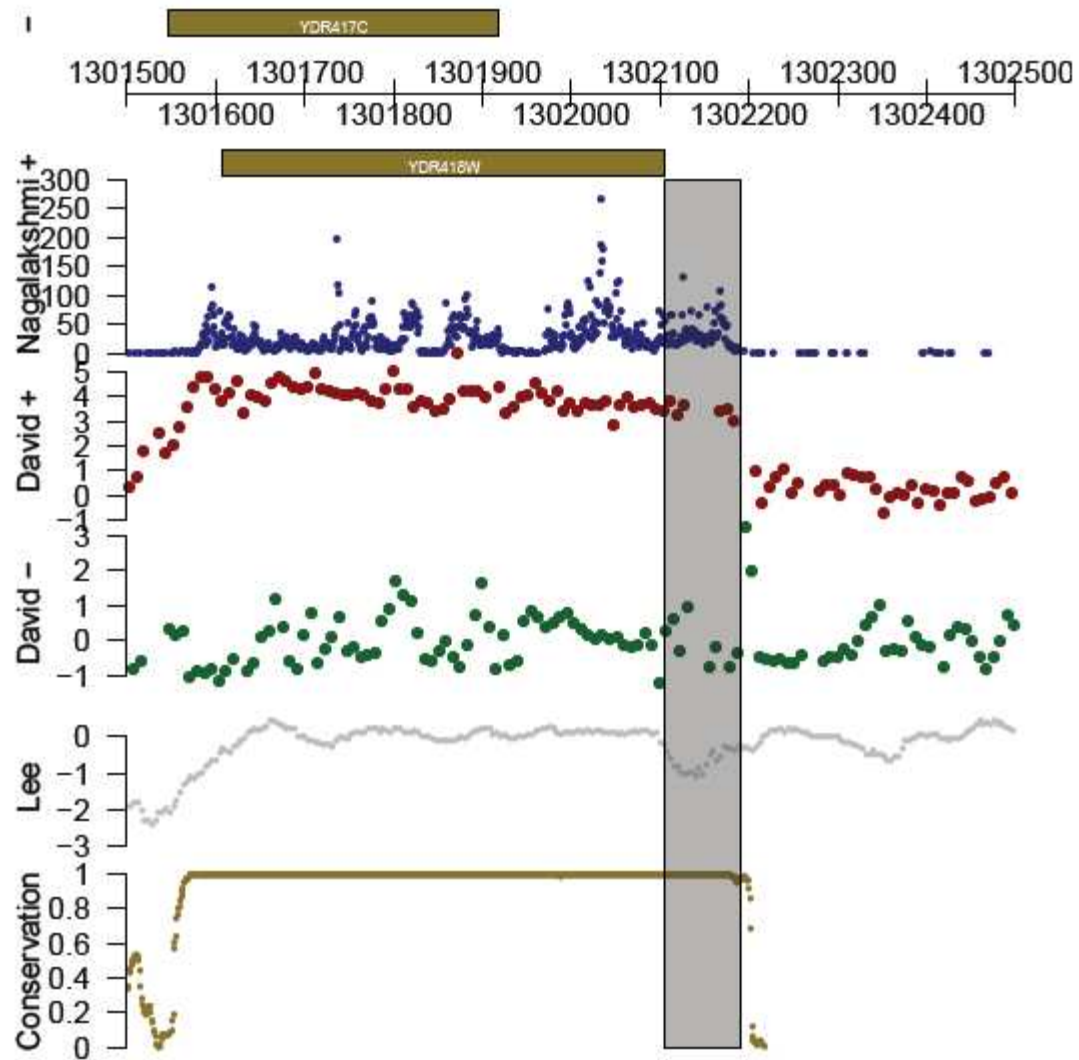
#### Package Downloads

Source	<a href="#">GenomeGraphs_1.0.1.tar.gz</a>
Windows binary	<a href="#">GenomeGraphs_1.0.1.zip</a>
OS X binary	<a href="#">GenomeGraphs_1.0.1.tgz</a>

#### Details

biocViews	<a href="#">Visualization</a> , <a href="#">Microarray</a>
Depends	methods, biomaRt, grid
Suggests	
Imports	

(<http://www.bioconductor.org/packages/2.2/bioc/html/GenomeGraphs.html>)

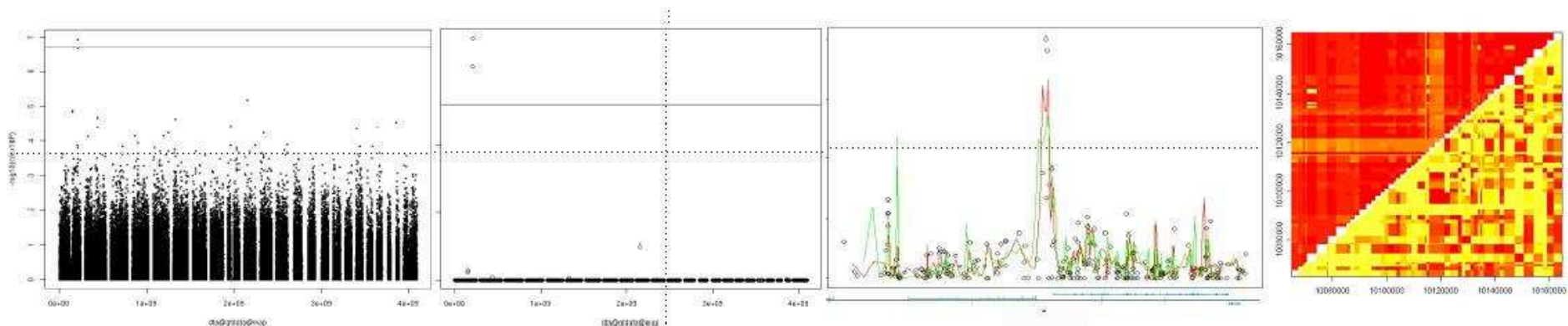




## Genome wide analysis

- With the recent explosion in availability of genome-wide data, handling large-scale datasets efficiently has become a common problem.
- In both cleaning and analyzing such datasets, the computational tasks involved are typically straightforward, but must be implemented millions of times.
- R can be used to tackle these problems, in a powerful and flexible way.

(<https://secure.bioconductor.org/BioC2009/>)



(<http://mga.bionet.nsc.ru/~yurii/ABEL/GenABEL/>)

## Biostrings

- The Biostrings package provides the infrastructure for representing and manipulating large nucleotide sequences (up to hundreds of millions of letters) in Bioconductor as well as fast pattern matching functions for finding all the occurrences of millions of short motifs in these large sequences.
- This is achieved by providing string containers that were designed to be memory efficient and easy to manipulate.

(<https://secure.bioconductor.org/BioC2008/>)

(<https://secure.bioconductor.org/BioC2009/>)

# Biostrings

## Biostrings

### String objects representing biological sequences, and matching algorithms

Memory efficient string containers, string matching algorithms, and other utilities, for fast manipulation of large biological sequences or set of sequences.

Author H. Pages, R. Gentleman, P. Aboyoun and S. DebRoy

Maintainer H. Pages

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("Biostrings")
```

#### Vignettes (Documentation)

[Alignments.pdf](#)

[Biostrings2Classes.pdf](#)

[DNAStringVectorization.pdf](#)

[GenomeSearching.pdf](#)

#### Package Downloads

Source [Biostrings\\_2.8.18.tar.gz](#)

Windows binary [Biostrings\\_2.8.18.zip](#)

OS X binary [Biostrings\\_2.8.18.tgz](#)

#### Details

biocViews	<a href="#">SequenceMatching</a> , <a href="#">Genetics</a> , <a href="#">Infrastructure</a>
Depends	R, methods, stats
Suggests	BSgenome, BSgenome.Celegans.UCSC.ce2, BSgenome.Dmelanogaster.UCSC.dm3, drosophila2probe, hgu95av2probe, RUnit

(<http://www.bioconductor.org/packages/2.2/bioc/html/Biostrings.html>)

## Pairwise sequence alignment using Biostrings

- Pairwise sequence alignment is a technique for finding regions of similarity between two sequences of DNA, RNA, or protein.
- It has been employed for decades in genomic analysis to answer questions on functional, structural, or evolutionary relationships between the two sequences as well as to assess the quality of data from sequencing technologies.
- The `pairwiseAlignment()` function from the Biostrings package in the development version of Bioconductor can be used to solve the (Needleman-Wunsch) global alignment, (Smith-Waterman) local alignment, and (ends-free) overlap alignment problems with or without affine gaps using either a constant or quality-based substitution scoring scheme.

(<https://secure.bioconductor.org/BioC2008/>)

## Biostrings

Note that some of the ORF sequences are represented in reverse complement form.

### 3 Optimal Pairwise Alignments

The function `pairwiseAlignment` solves the (Needleman-Wunsch) global, the (Smith-Waterman) local, and the overlap optimal pairwise alignment problems. The solution to each of these problems is dependent on the specified substitution scores and the gap penalties:

- **Substitution Scores:** The substitution scores can either be fixed for each pairing of letters within the two strings or be dependent on the qualities associated with those letters. When the scores are fixed by pairing, the `substitutionMatrix` argument takes a matrix with the appropriate alphabets as dimension names. When the scores are quality-based, the `patternQuality` and `subjectQuality` arguments accept the equivalent of [0-99] numeric quality values for the respective strings.
- **Gap Penalties:** Gaps have the potential to incur a cost when they are introduced and when they are extended in an optimal pairwise alignment. The former is regulated by the `gapOpening` argument and the latter by the `gapExtension` argument.

The `pairwiseAlignment` function uses memory and computation time proportional to the product of the two string lengths.

The BLOSUM50 matrix is available in this package as a matrix:

```
> data(BLOSUM50)
> BLOSUM50[1:4, 1:4]

  A  R  N  D
A  5 -2 -1 -2
R -2  7 -1 -2
N -1 -1  7  2
```

(<http://www.bioconductor.org/packages/2.2/bioc/vignettes/Biostrings/inst/doc/Alignments.pdf>)

## Efficient string manipulation and genome-wide motif searching with Biostrings and the BSgenome data packages

- The Bioconductor project also provides a collection of "BSgenome data packages".
- These packages contain the full genomic sequence for a number of commonly studied organisms.
- The Biostrings package together with the BSgenome data packages provide an efficient and convenient framework for genome-wide sequence analysis.
- Noteworthy are the built-in masks in the BSgenome data packages; the ability to inject SNPs from a SNPlocs package into the chromosome sequences of a given species (only Human supported for now); and the matchPDict() function for efficiently finding all the occurrences in a genome of a big dictionary of short motifs (like one typically gets from an ultra-high throughput sequencing experiment).

(<https://secure.bioconductor.org/BioC2008/>)

**Bookmarks** Options ▾

- The Biostrings-based genome data packages
- Finding an arbitrary nucleotide pattern in a chromosome
- Finding an arbitrary nucleotide pattern in an entire genome
- Some precautions when using matchPattern
- Masking the chromosome sequences
- Hard masking
- Injecting known SNPs in the chromosome sequences
- Finding all the patterns of a constant

Efficient genome searching with Biostrings and the BSgenome data packages

Harv  Pagh  
July 6, 2009

Contents

1 The Biostrings-based genome data packages	1
2 Finding an arbitrary nucleotide pattern in a chromosome	2
3 Finding an arbitrary nucleotide pattern in an entire genome	5
4 Some precautions when using matchPattern	9
5 Masking the chromosome sequences	10
6 Hard masking	15
7 Injecting known SNPs in the chromosome sequences	15
8 Finding all the patterns of a constant with a dictionary in an entire genome	15
9 Session info	17

1 The Biostrings-based genome data packages

The Bioconductor project provides data packages that contain the full genome sequences of a given organism. These packages are called *Biostrings-based genome data packages* because the sequences they contain are stored in some of the basic containers defined in the Biostrings package, like the `DNAString`, the `DNAStringSet` or the `MaskedDNAString` containers. Regardless of the particular sequence data that they contain, all the Biostrings-based genome data packages are very similar and can be manipulated in a consistent and easy way. They all require the BSgenome package in order to work properly. This package, unlike the Biostrings-based genome data packages, is a software package that provides the infrastructure needed to support them (this is why the Biostrings-based genome data packages are also called *BSgenome data packages*). The BSgenome package itself requires the Biostrings package.

See the main page for the `available.genomes` function (`available.genomes`) for more information about how to get the list of all the BSgenome data packages currently available in your version of Bioconductor (you need an internet connection so that `available.genomes` can query the Bioconductor package repositories).

More genomes can be added if necessary. Note that the process of making a BSgenome data package is not yet documented but you are welcome to ask for help on the `bio-devel` mailing list (<http://bioconductor.org/lists/mail-listen.html>) if you need a genome that is not yet available.

(<http://www.bioconductor.org/packages/bioc/vignettes/BSgenome/inst/doc/GenomeSearching.pdf>)

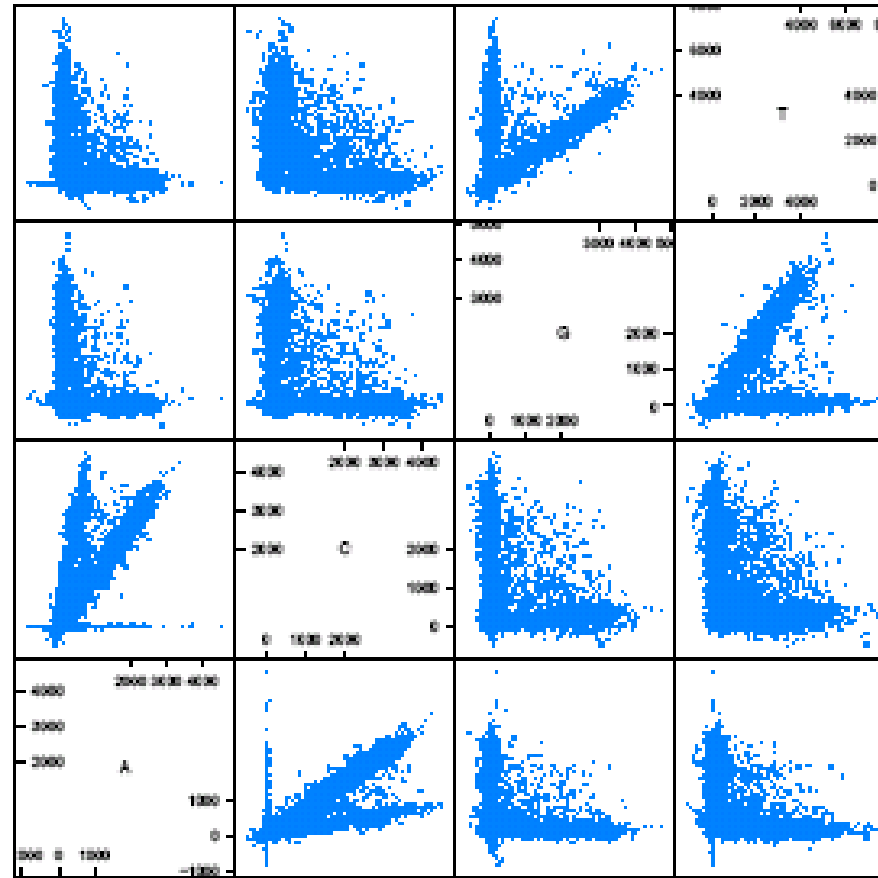
## **ShortRead: tools for input and quality assessment of high-throughput sequence data**

- Short reads are DNA sequences derived from ultra-high throughput sequencing technologies.
- Data typically consists of hundreds of thousands to tens of millions of reads, ranging from 10's to 100's of bases each. The ShortRead package is another R package that is available in the development version of Bioconductor.
- ShortRead provides methods for importing short reads into R data structures such as those used in the Biostrings package.
- ShortRead provides quality assessment tools for some specific technologies, and provides simple building blocks allowing creative and fast exploration and visualization of data.

(<https://secure.bioconductor.org/BioC2008/>)



## ShortRead for quality control



Scatter Plot Matrix

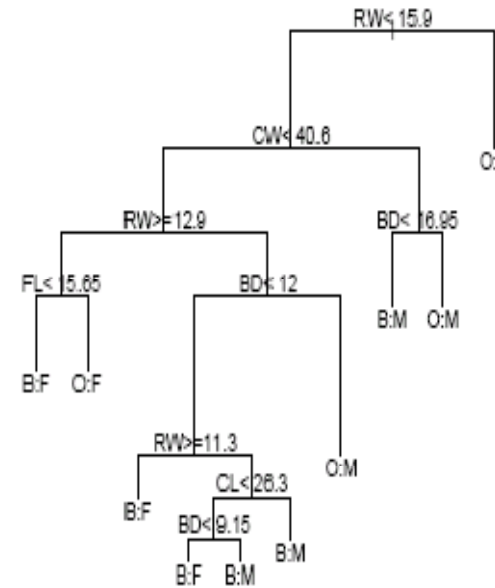
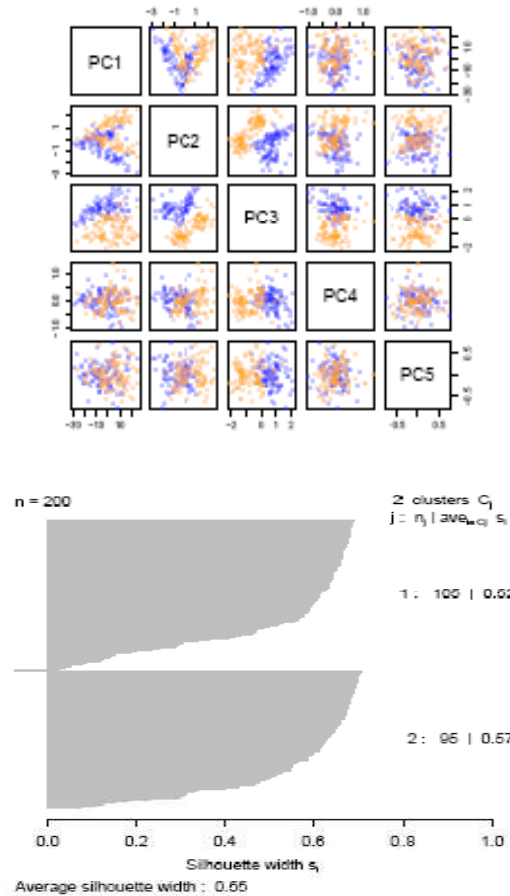
(<http://www.bioconductor.org/workshops/2009/SSCMay09/ShortRead/IOQA.pdf>)

## Machine learning with Bioconductor

- The facilities of the MLInterfaces package are numerous.
- MLInterfaces facilitates answering questions like:
  - Given an ExpressionSet, how can we reason about clustering and opportunities for dimensionality reduction using unsupervised learning techniques?
  - For an ExpressionSet with labeled samples, how can we build and evaluate classifiers from various families of prediction algorithms?
  - How do we specify feature-selection and cross-validation processes for machine learning in MLInterfaces?

(<https://secure.bioconductor.org/BioC2008/>)

# MLInterfaces, towards a uniform interface for machine learning applications



- Looking for the tree in the forest?

## Random Jungle

**Random Jungle is a fast implementation of RandomForest(TM) for high dimensional data\***

**Welcome to RandomJungle.com!**

Random Jungle provides a free random forest implementation for high dimensional data. It is intended to be widely useful, and usable across a broad spectrum of applications.

### **News**

**Latest version: 0.8.3**



(<http://randomjungle.com/>)

## Bioconductor Task View: Clustering

### Subview of

- [Statistics](#)

### Packages in view

Package	Maintainer	Title
<a href="#">adSplit</a>	Claudio Lottaz	Annotation-Driven Clustering
<a href="#">clusterStab</a>	James W. MacDonald	Compute cluster stability scores for microarray data
<a href="#">CORREP</a>	Dongxiao Zhu	Multivariate Correlation Estimator and Statistical Inference Procedures.
<a href="#">ctc</a>	Antoine Lucas	Cluster and Tree Conversion.
<a href="#">flowClust</a>	Raphael Gottardo	Clustering for Flow Cytometry
<a href="#">geneRecommender</a>	Greg Hather	A gene recommender algorithm to identify genes coexpressed with a query set of genes
<a href="#">hopach</a>	Katherine S. Pollard	Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH)
<a href="#">maanova</a>	Hyuna Yang	Tools for analyzing Micro Array experiments
<a href="#">made4</a>	Aedin Culhane	Multivariate analysis of microarray data using ADE4
<a href="#">maigesPack</a>	Gustavo H. Esteves	Functions to handle cDNA microarray data, including several methods of data analysis
<a href="#">MantelCorr</a>	Brian Steinmeyer	Compute Mantel Cluster Correlations
<a href="#">Mfuzz</a>	Matthias Futschik	Soft clustering of time series gene expression data
<a href="#">MLInterfaces</a>	V. Carey	Uniform interfaces to R machine learning procedures for data in Bioconductor containers
<a href="#">puma</a>	Richard Pearson	Propagating Uncertainty in Microarray Analysis
<a href="#">SAGx</a>	Per Broberg.	Statistical Analysis of the GeneChip

## Gene set enrichment analysis with R

- Gene Set Enrichment Analysis (GSEA) - the identification of expression patterns by groups of genes rather than by individual genes - is fast becoming a regular part of microarray data analysis.
- GSEA is a dynamically evolving field, with a variety of approaches on offer and with a clear standard yet to emerge.
- Similarly, R/Bioconductor offers a variety of packages and tools for GSEA, including the packages "Category" and "GSEAlm", and libraries such as "GSEABase" and "GOstats".

(<https://secure.bioconductor.org/BioC2008/>)

## Navigating protein interactions with R and BioC

- BioConductor offers tools for performing a protein interaction analysis using Bioconductor packages including RpsiXML, ppiStats, graph, RBGL, and apComplex.
- Such an analysis may involve
  - compiling from different molecular interaction repositories and
  - converting these files into R graph objects,
  - conducting statistical tests to assess sampling, coverage, as well as systematic and stochastic errors,
  - using specific algorithms to search for features such as clustering coefficient and degree distribution,
  - estimating features from different data types: physical interactions, co-complexed interactions, genetic interactions, etc.

(<https://secure.bioconductor.org/BioC2008/>)

## Microarray analysis

- One of the most common tasks when analyzing microarrays is to make comparisons between sample types, and the limma package in R is one of the more popular packages for this task.
- The limma package is quite powerful and allows users to make relatively complex comparisons.
- However, this power comes with a cost in complexity.

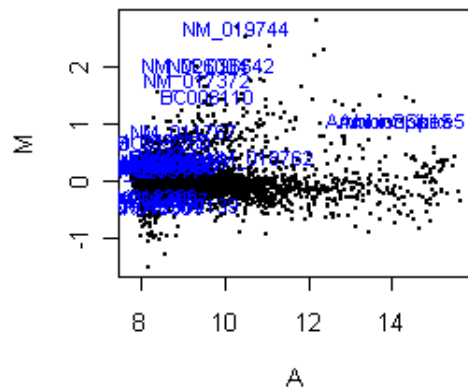
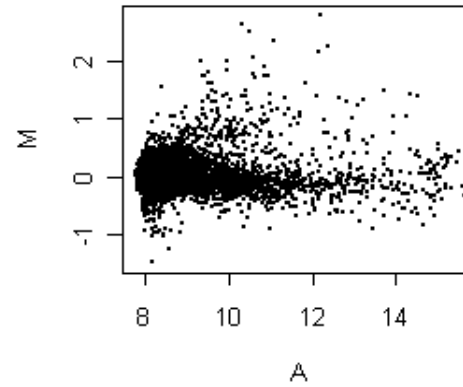
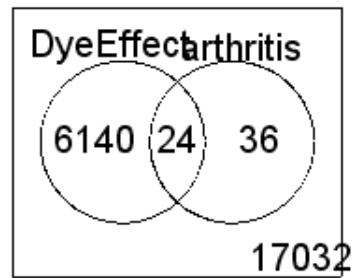
(<https://secure.bioconductor.org/BioC2008/>)

- Furthermore, GGTools can be used for investigating relationships between DNA polymorphisms and gene expression variation
- It provides facilities to for importing genotype and expression data from several platforms.

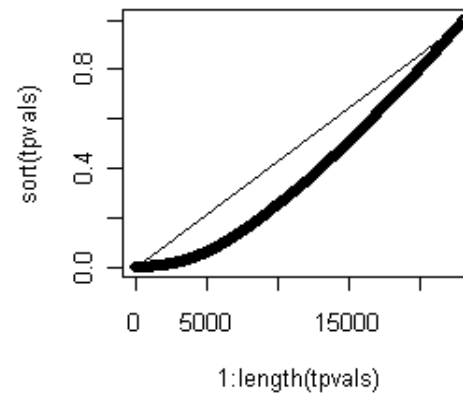
(<https://secure.bioconductor.org/BioC2008/>)



# Limma

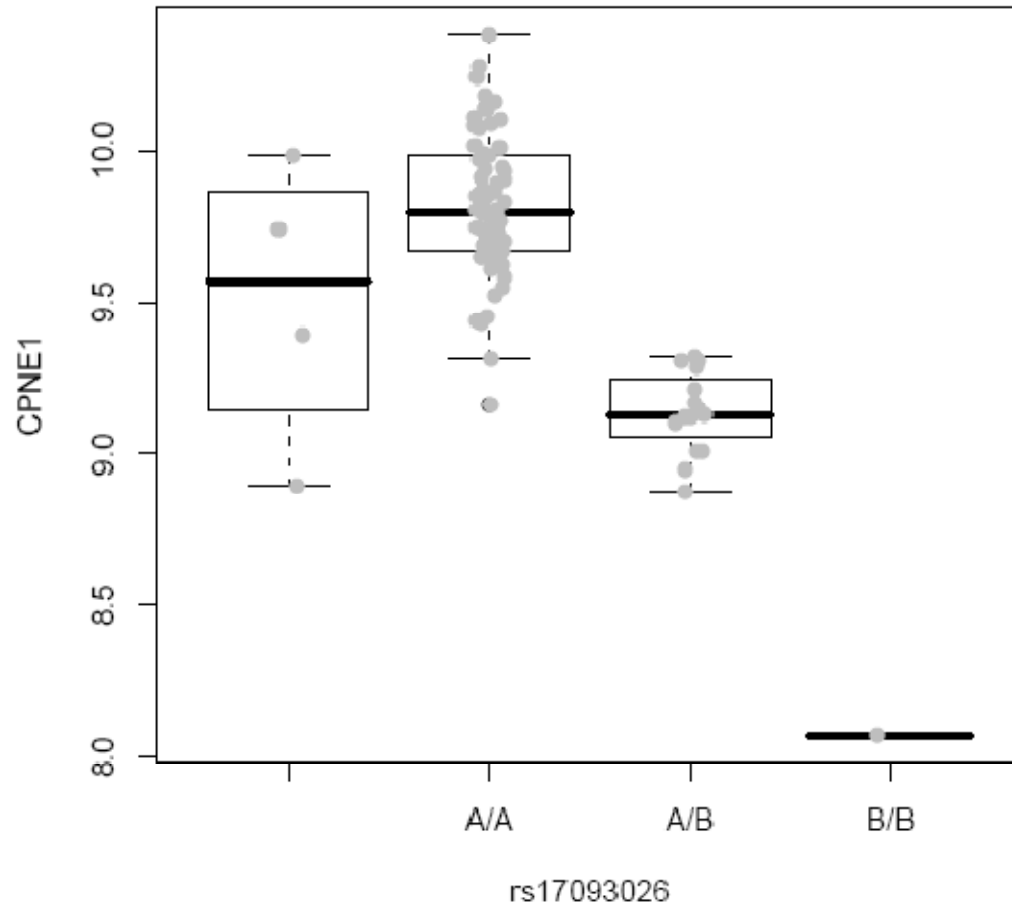


Student-t p-values



(Boer 2005)

## GGtools



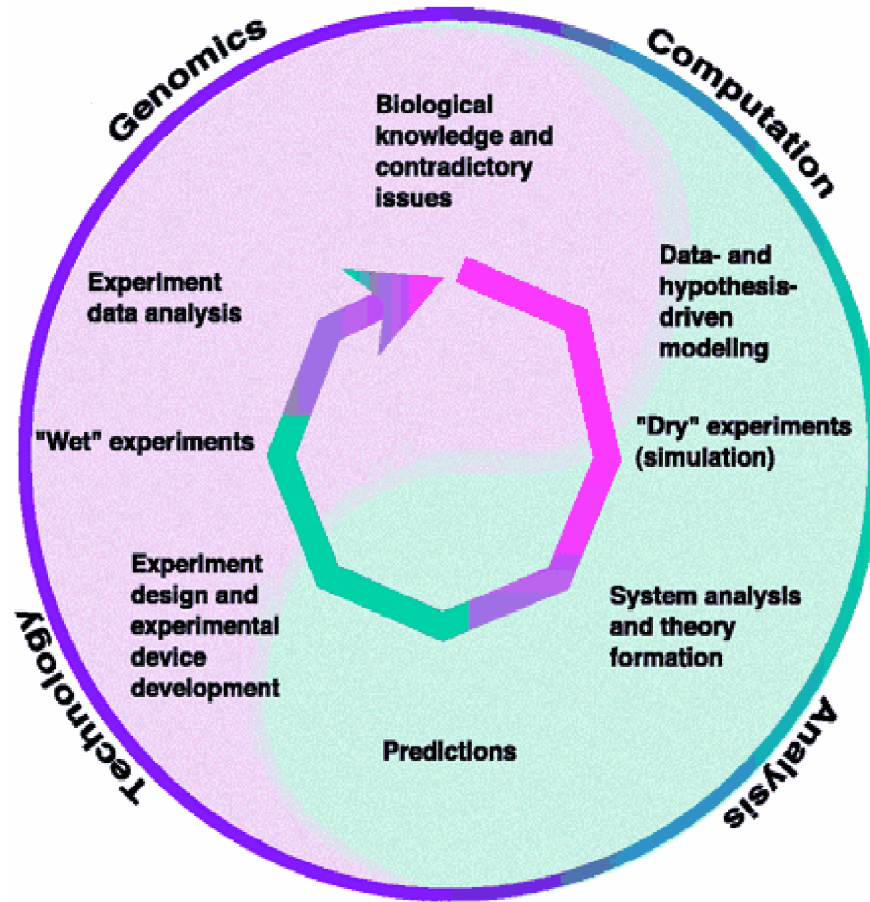
(<http://www.bioconductor.org/packages/2.2/bioc/vignettes/GGtools/inst/doc/GGoverview2008.pdf>)

## Copy number data analysis

- TCGA (The Cancer Genome Atlas) is a comprehensive cancer molecular characterization data repository supported by NIH.
- Its data portal currently contains genomic copy number, expression (exon, mRNA, miRNA), SNP, DNA methylation, and sequencing data of brain and ovarian tumors. More cancer types will be included in the years to come.
- With its large collection of samples (aimed at 500 samples for each tumor type), TCGA data will be extremely useful to cancer researchers.
- Several Bioconductor's packages can be used to process the raw arrayCGH data, identify DNA copy number alterations within samples, and find genomic regions of interest across samples, or to carry out classification and significance testing based on copy number data.

(<https://secure.bioconductor.org/BioC2009/>)

## The importance of bioinformatics software



(Kitano 2002)

## References:

- Hagen 2000. The origins of bioinformatics. Nature Reviews Genetics (Perspectives)
- Hughey et al 2003. Bioinformatics: a new field in engineering education. Journal of Engineering Education
- Perez-Iratxeta et al 2006. Evolving research trends in bioinformatics. Briefings in bioinformatics
- URL: [www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html](http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html)
- URL: <http://www.ebi.ac.uk/2can/bioinformatics/>

## Background reading:

- [http://faculty.ucr.edu/~tgirke/Documents/R\\_BioCond/R\\_BioCondManual.html](http://faculty.ucr.edu/~tgirke/Documents/R_BioCond/R_BioCondManual.html)
- Ouzounis et al. 2003. Early bioinformatics: the birth of a discipline – a personal view. Bioinformatics (Review)

## In-class discussion document

Elkin P (2003). Primer on medical genomics. Part V: Bioinformatics. *Mayo Clin Proc*, 78: 57-64

Questions: In class reading\_1.pdf

### Preparatory reading:

- Tefferi et al 2002. Primer on medical genomics. Part II: Background principles and methods in molecular genetics. *Mayo Clin Proc*, 77:785-808.