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

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

CHAPTER 1: WHAT IT MEANS AND DOES NOT MEAN

1 Bioinformatics: a “new” field in engineering

1.1 A gentle introduction

1.2 Bioinformatics – what’s in a name?

1.3 The origins of bioinformatics

2 Definition of bioinformatics

2.1 A “clear” definition for bioinformatics

2.2 Topics in bioinformatics from a journal’s perspective

3 Evolving research trends in bioinformatics

3.1 Introduction

3.2 Bioinformatics timeline

3.3 Careers in bioinformatics

4 Bioinformatics software

4.1 Introduction

4.2 R and Bioconductor

4.3 Example R packages

1 Bioinformatics: a “new” field in engineering

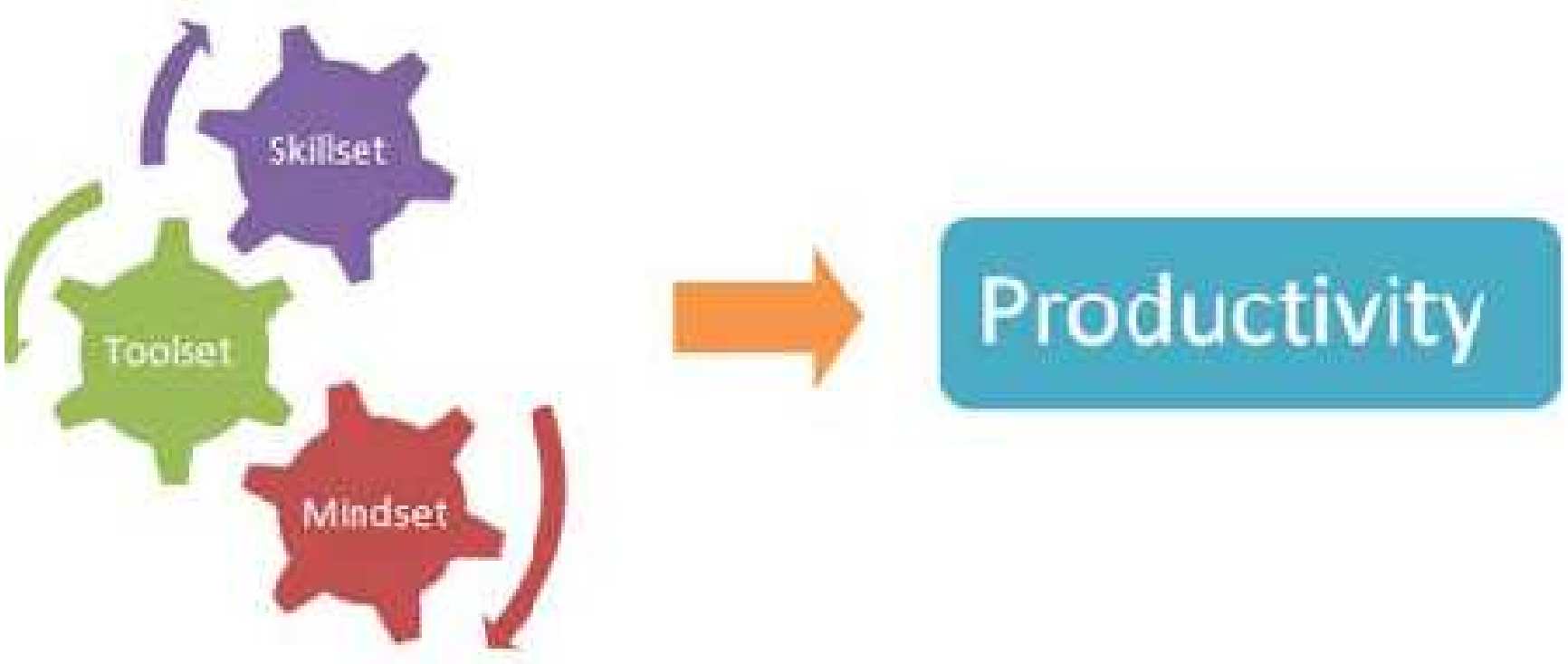
1.1 A gentle introduction

- You know who I am and how the bioinformatics course will be organized
- But who are you?

- **http://www.youtube.com/watch?v=MULMbqQ9LJ8**

(Ref: ”Dammit Jim, I’m a doctor, not a bioinformatician” – Golden Helix”)

- It takes more than just brains to make advances in genetics:



Skillset

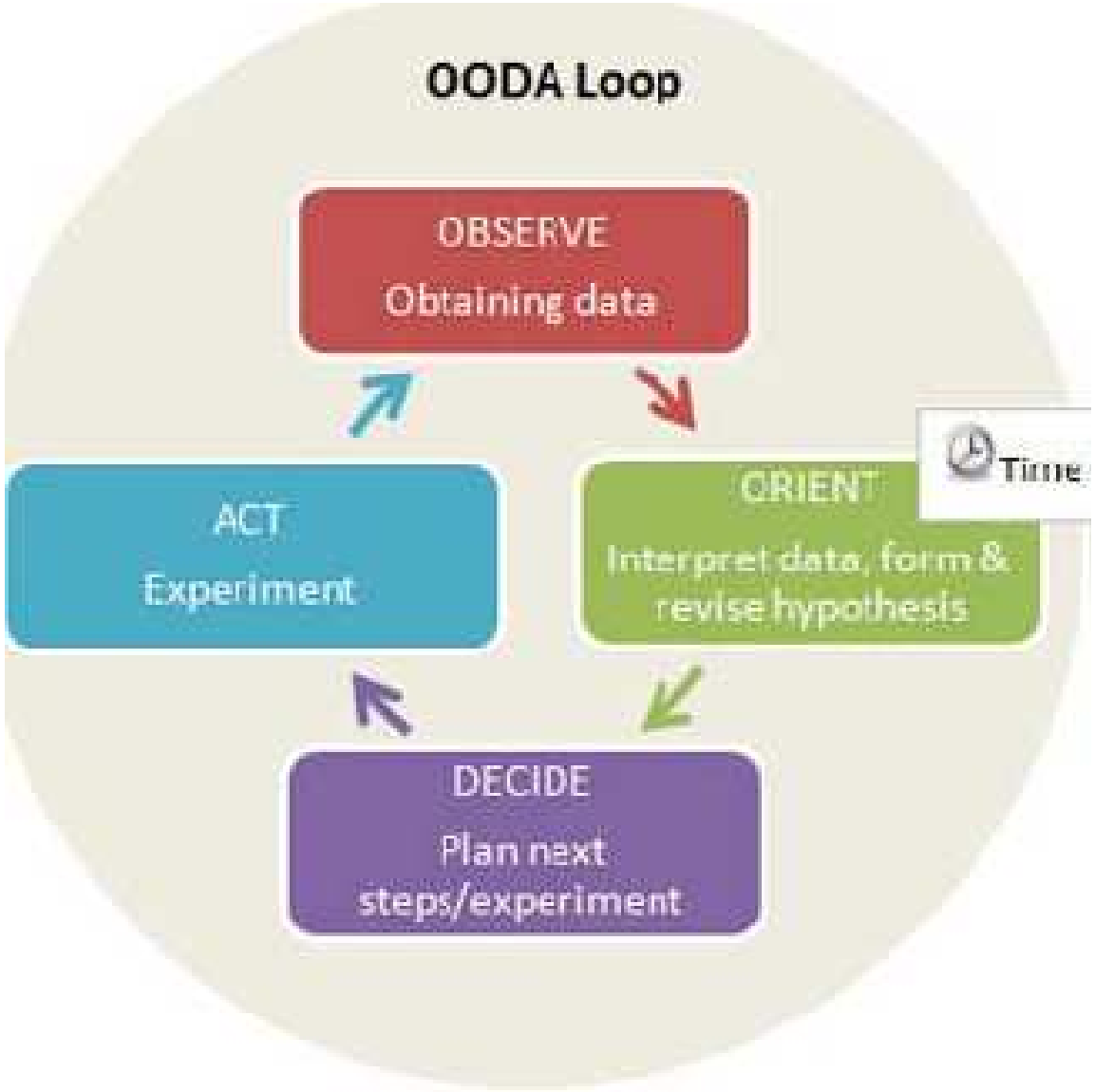
- The free software tools used today require highly skilled bioinformatics professionals, which are often in short reply …
- One must have competences in several disciplines: computer science, statistics and genetics.
- Why does someone virtually have to be a computer programmer in order to perform genetics research?

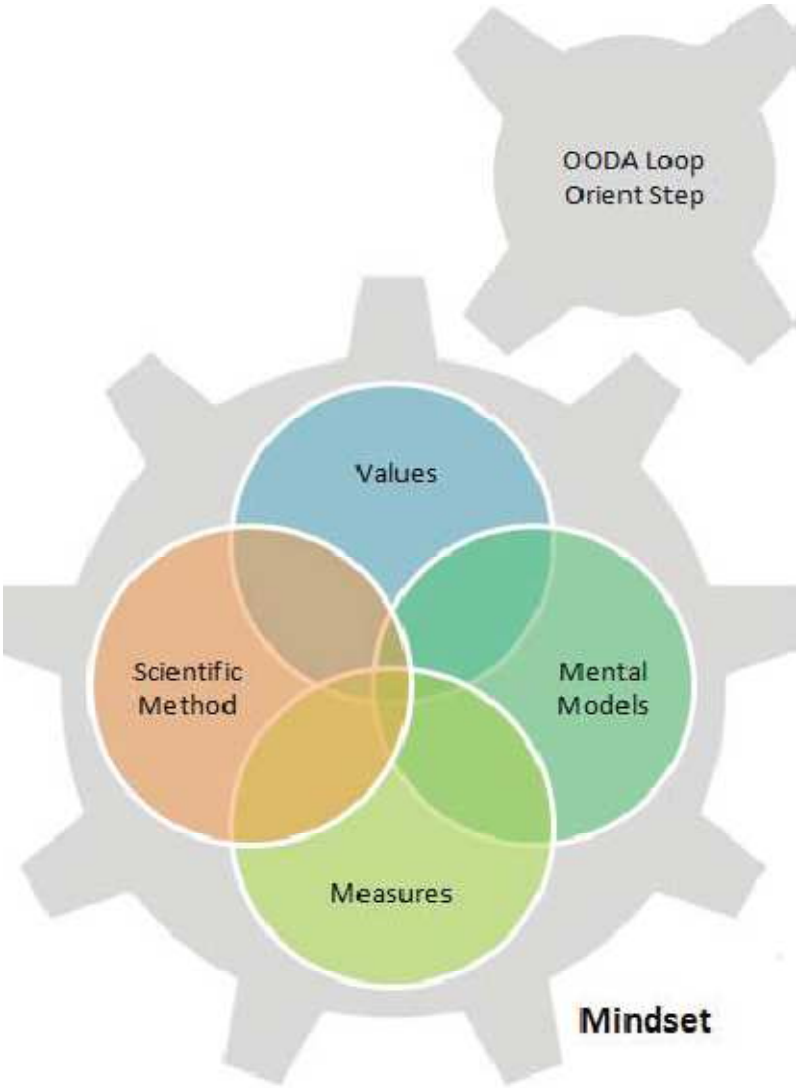
Toolset

- There are pressing needs in software tools and infrastructure for high-throughput sequence research:
  - Robust, well-documented, and well-supported; graphical user interface
  - Most of the "in-house" informatics tools developed so far are optimized only for local applications
  - It may only run on large, local computational clusters
  - It may require a dedicated group of local bioinformatics experts to maintain or update
- Foundational to this problem is the fact that academia is the birthplace of most new statistical and computational methods in genetic research.
- Variety of data formats → need for standardization and optimized transparent work flow systems
- Why is keeping software updated and "advertising" it that hard?

Mindset

- "Publish or perish": refers to the pressure to publish work constantly to further or sustain a career in academia. The competition for tenure-track faculty positions in academia puts increasing pressure on scholars to publish new work frequently
- Publications are a way to build up reputation, not the software tools they develop to bring the work into practice and increase a collective productivity
- There is a need for bioinformaticians that are able to make sense of available software, and apply it to large data sets. This involves project-oriented work ←→ new developmental research
- Observe – Orient – Decide – Act


OODA Loop
OBSERVE
Obtaining data
ORIENT
Interpret data, form &
revise hypothesis
DECIDE
Plan next
steps/experiment
ACT
Experiment
Time


OODA Loop
Orient Step
Values
Scientific
Method
Mental
Models
Measures
Mindset

- If productivity in our field is measured not only by volume of publications, but also by the quality of the causal theoretical models for biological processes, we have a number of systemic and interrelated obstacles to productivity in our field:
  - Bioinformatics has become the constrained resource limiting the pace of genetic research—there is a skillset deficit in the field as a whole.
  - The software toolset for genetic research, produced and broadly used in academia, has serious shortcomings for productivity. For the most part, it can only be operated well by the constrained resource.
  - The mindset embodied in reputation as the prime metric of academia reinforces the toolset deficit.
  - The toolset and mindset inhibits the reproducibility of research, a cornerstone to the scientific method and the productivity that method provides us.

"Almost any bioinformatician started off lacking skills in statistics, computer science, or biology and had to learn a domain-appropriate subset of the rest generally through experience and, perhaps, being paired with a capable mentor."

"… And that's my two SNPs"

1.2 Bioinformatics – what's in a name?

Towards a definition

- 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/)

- 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/)

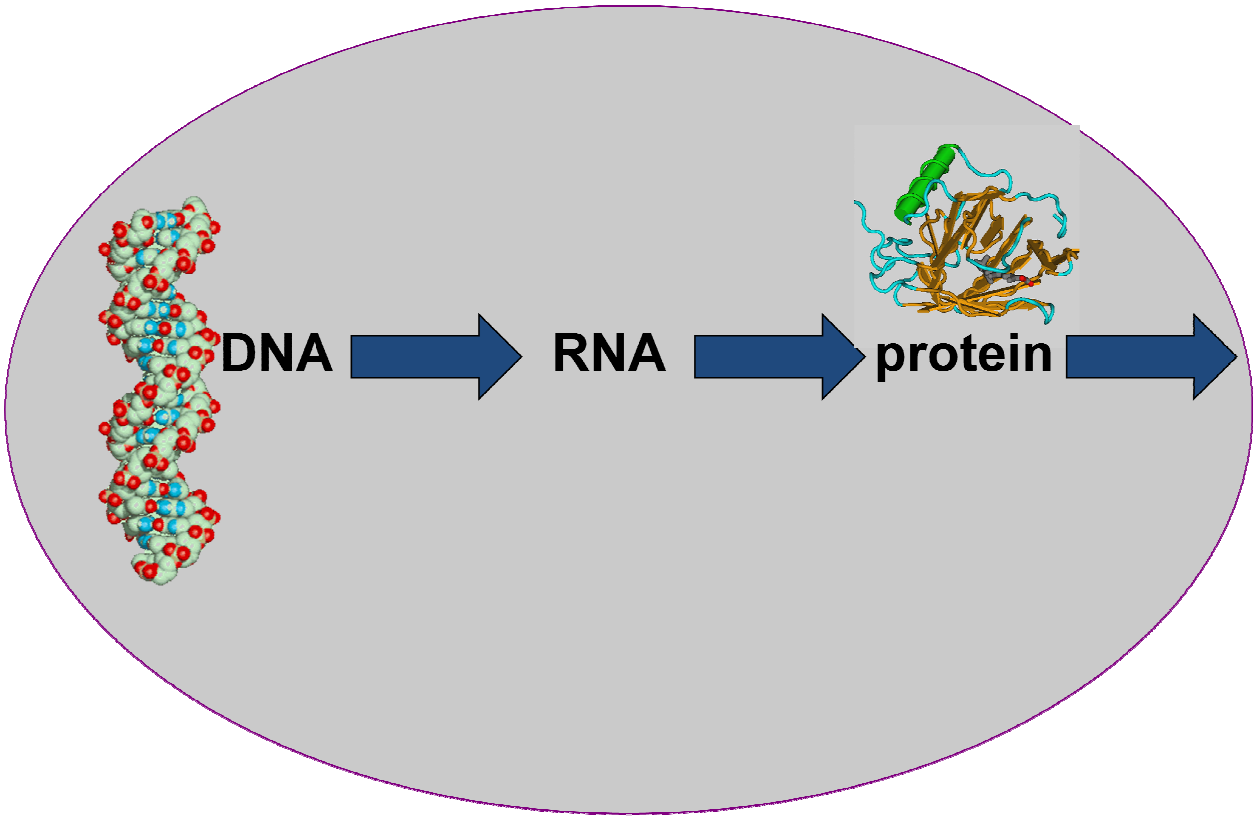
Biosciences

- What is the goal of biosciences?
- Ultimately, the complete understanding of life phenomena
  - Complex organization
  - Regulatory mechanism (homeostasis)
  - Growth & development
  - Energy utilization
  - Response to the environmental stimuli
  - Reproduction (DNA guaranties exact replication)
  - Evolution (capacity of species to change over time)

Biosciences

- It clearly goes beyond human biology / genetics (although we will put emphasis on human genetics data analyses)
  - Life's diversity results from the variety of molecules in cells
  - A spider's web-building skill depends on its DNA molecules
  - DNA also determines the structure of silk proteins
  - These make a spiderweb strong and resilient

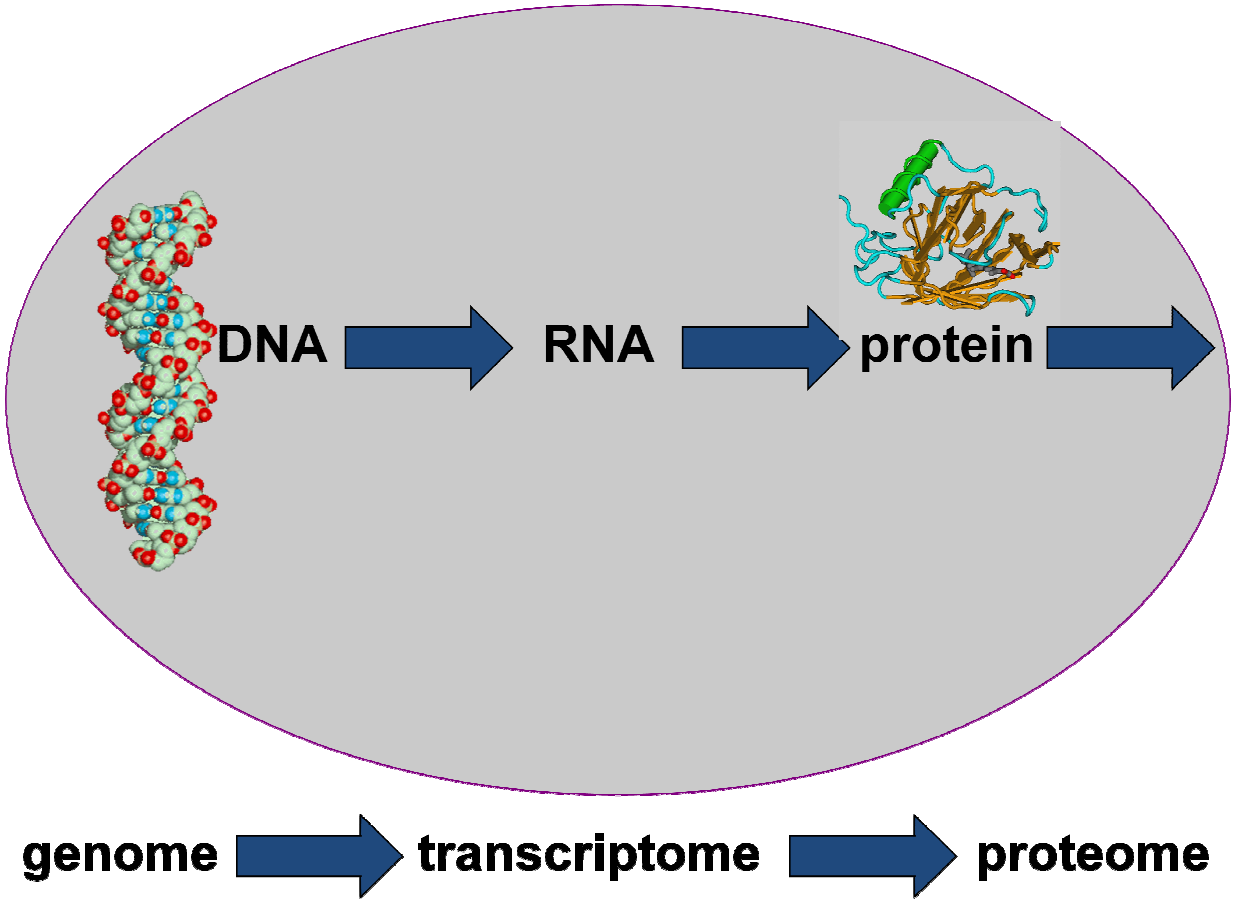
- We will talk about molecular genetics, to set the pace (Chapter 2) and discuss the “central dogma of molecular biology”



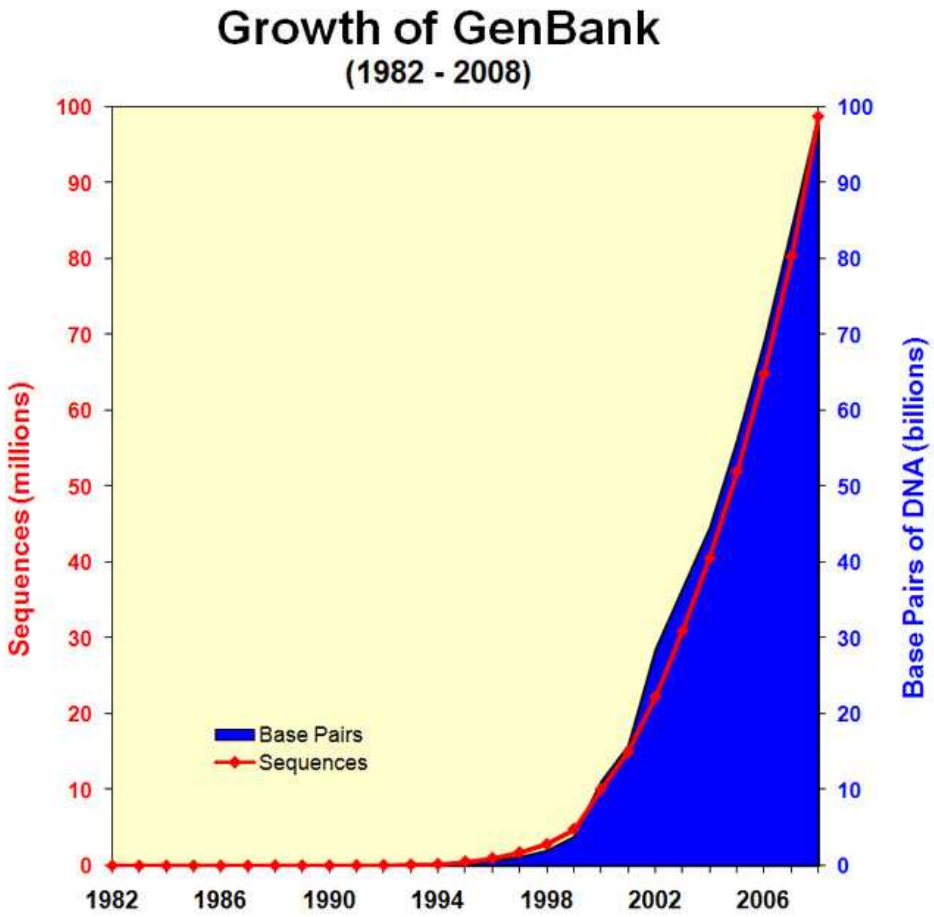
Paradigm shift in biosciences

- So far, biologists have focused certain phenotypes and hunted the genes responsible, one at a time
- New trend is:
  - Catalog all the parts: genes and proteins → **genomics and proteomics**
  - Understand how each part works → **functional genomics**
  - Model & simulate the collective behavior of the parts → **systems biology**

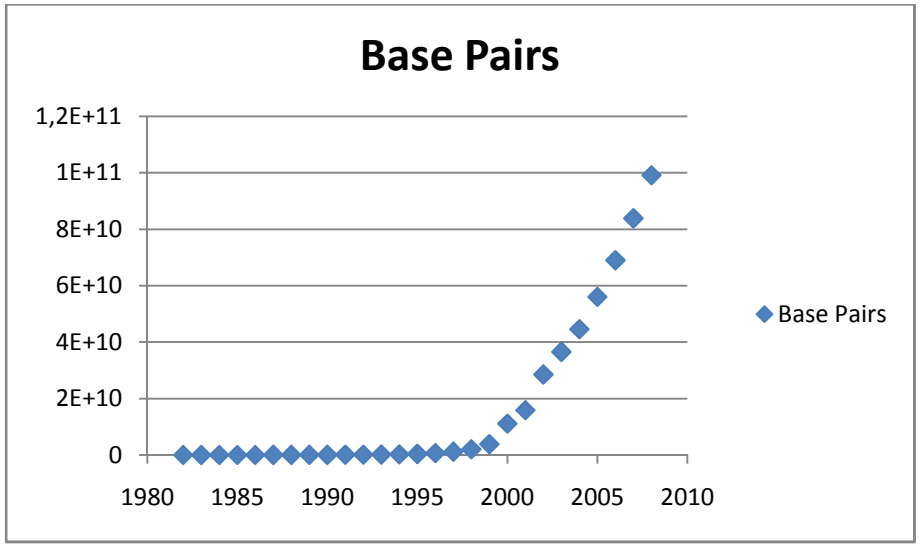
Central dogma of molecular biology

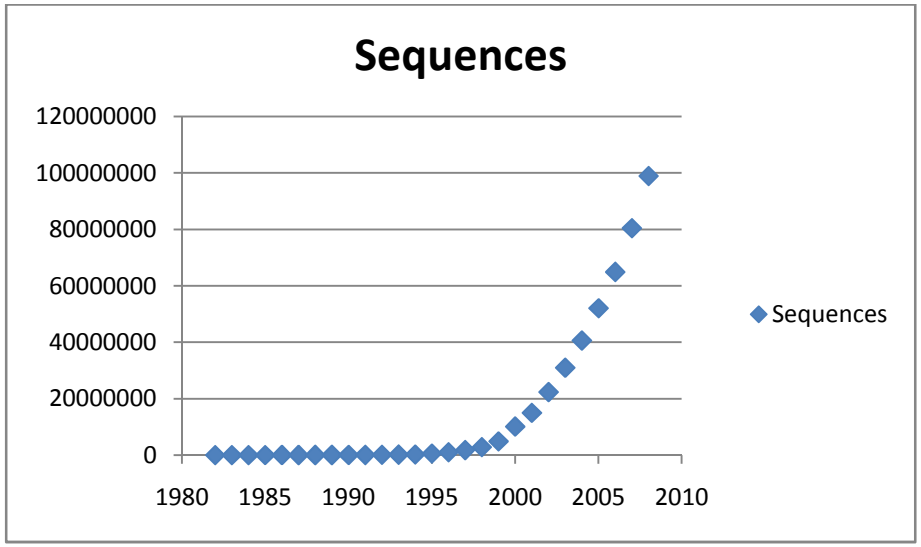


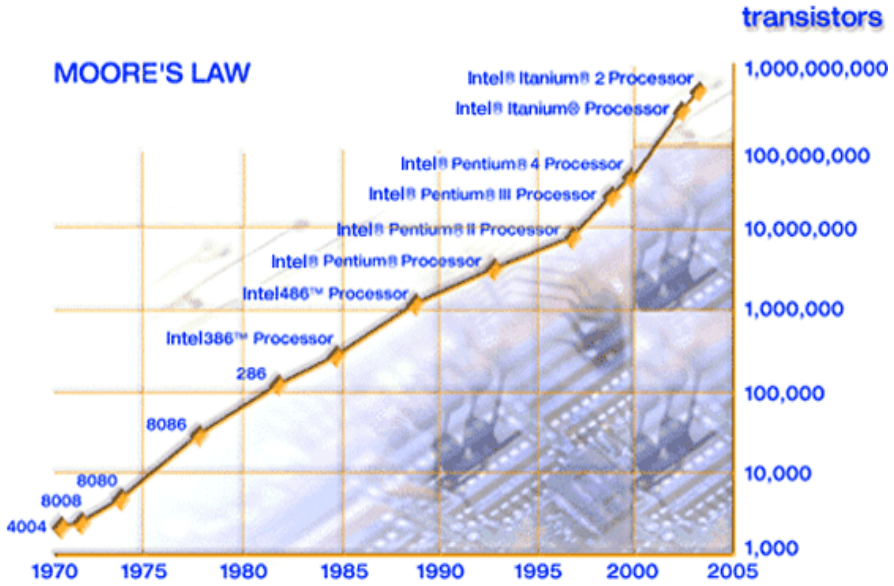
Central dogma of bioinformatics and genomics



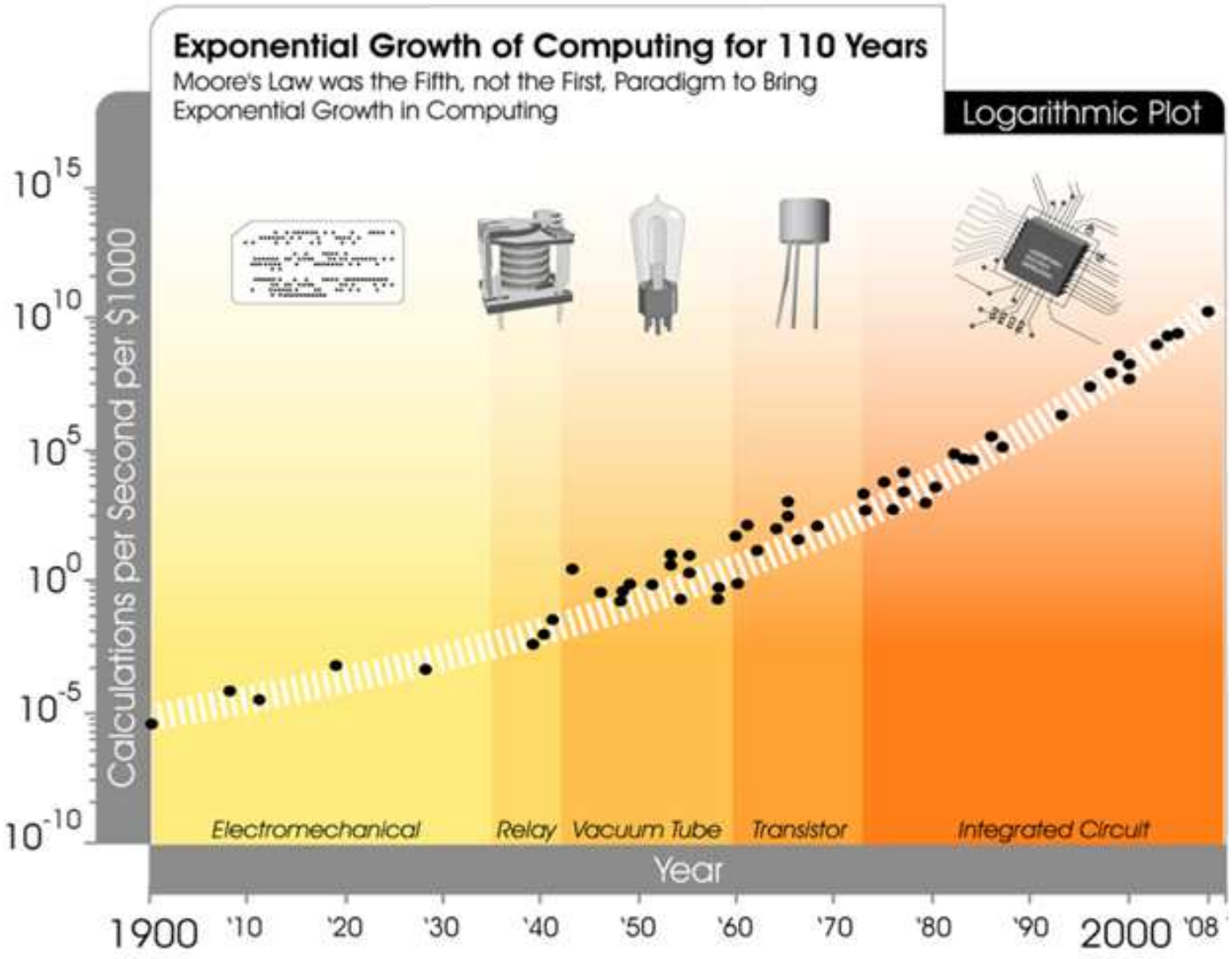
(http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html)



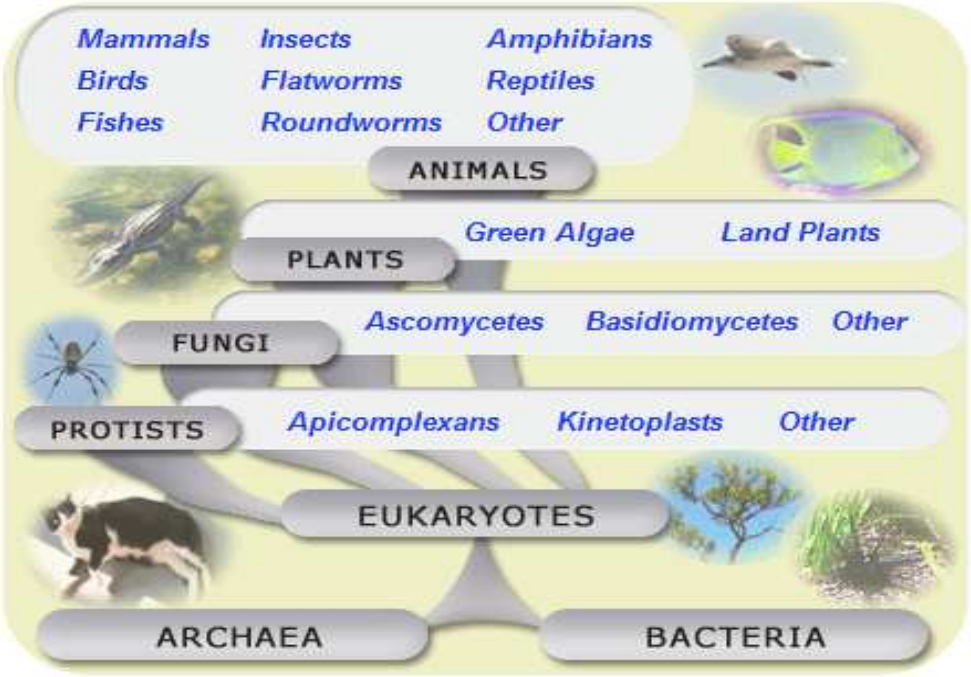




- With $1,000 genome sequencing technologies in 10 years coupled with functional data, we need better IT solutions!


Exponential Growth of Computing for 110 Years
Moore's Law was the Fifth, not the First, Paradigm to Bring Exponential Growth in Computing
Logarithmic Plot
Calculations per Second per $1000
10^{15}
10^{10}
10^{5}
10^{0}
10^{-5}
10^{-10}
Electromechanical
Relay
Vacuum Tube
Transistor
Integrated Circuit
Year
1900
'10
'20
'30
'40
'50
'60
'70
'80
'90
2000
'08

Explosion of data: multiple genomes (finished)



- Human genes: 25,000
- Human genome: $3x10^9$ bp
- DNA-protein or protein-protein interactions

Where to look for additional info? - http://www.ncbi.nlm.nih.gov/sites/gquery

BMB
reports

Mini Review

Genome data mining for everyone

*Gir Won Lee & Sangsoo Kim**

Department of Bioinformatics, Soongsil University, Seoul 156-743, Korea

The genomic sequences of a huge number of species have been determined. Typically, these genome sequences and the associated annotation data are accessed through Internet-based genome browsers that offer a user-friendly interface. Intelligent use of the data should expedite biological knowledge discovery. Such activity is collectively called data mining and involves queries that can be simple, complex, and even combinational. Various tools have been developed to make genome data mining available to computational and experimental biologists alike. In this mini-review, some tools that have proven successful will be introduced along with examples taken from published reports. [BMB reports 2008; 41(11): 757-764]

Feature[a]	UCSC Genome Browser[b]	Ensembl Genome Browser[c]
Number of organisms hosted by the browser	47 eukaryotes • 14 mammals • 10 other vertebrates • 3 deuterostomes • 13 insects • 6 nematodes • 1 fungus	39 eukaryotes • 25 mammals • 7 other vertebrates • 2 chodates • 3 insects • 1 nematode • 1 fungus
Genome-wide comparisons between species	28-way genome alignments[d]	multi-genome alignments, synteny blocks
Gene-by-gene orthologs/paralogs	orthology over 6 model organisms	orthology/paralogy over all the organisms in the project based on TreeBeST[e]
Functional data types that can be viewed alongside genome sequence	gene expression, protein motifs, ENCODE data[f]	gene expression, protein motifs, regulatory elements via DAS[g]
Methods for mining and bulk sequence downloading	Gene Sorter and Table Browser	BioMart

Top 10 challenges for bioinformaticiancs

- Having biosciences in mind:
  - Precise models of where and when transcription will occur in a genome (initiation and termination)
  - Precise, predictive models of alternative RNA splicing
  - Precise models of signal transduction pathways; ability to predict cellular responses to external stimuli
  - Determining protein:DNA, protein:RNA, protein:protein recognition codes
  - Accurate protein structure prediction

Top 10 challenges for bioinformaticiancs (continued)

- Rational design of small molecule inhibitors of proteins
- Mechanistic understanding of protein evolution
- Mechanistic understanding of speciation
- Development of effective gene ontologies: systematic ways to describe gene and protein function
- Education: development of bioinformatics curricula

- These are from an academic point of view …

:

How will we address these challenges in this course?

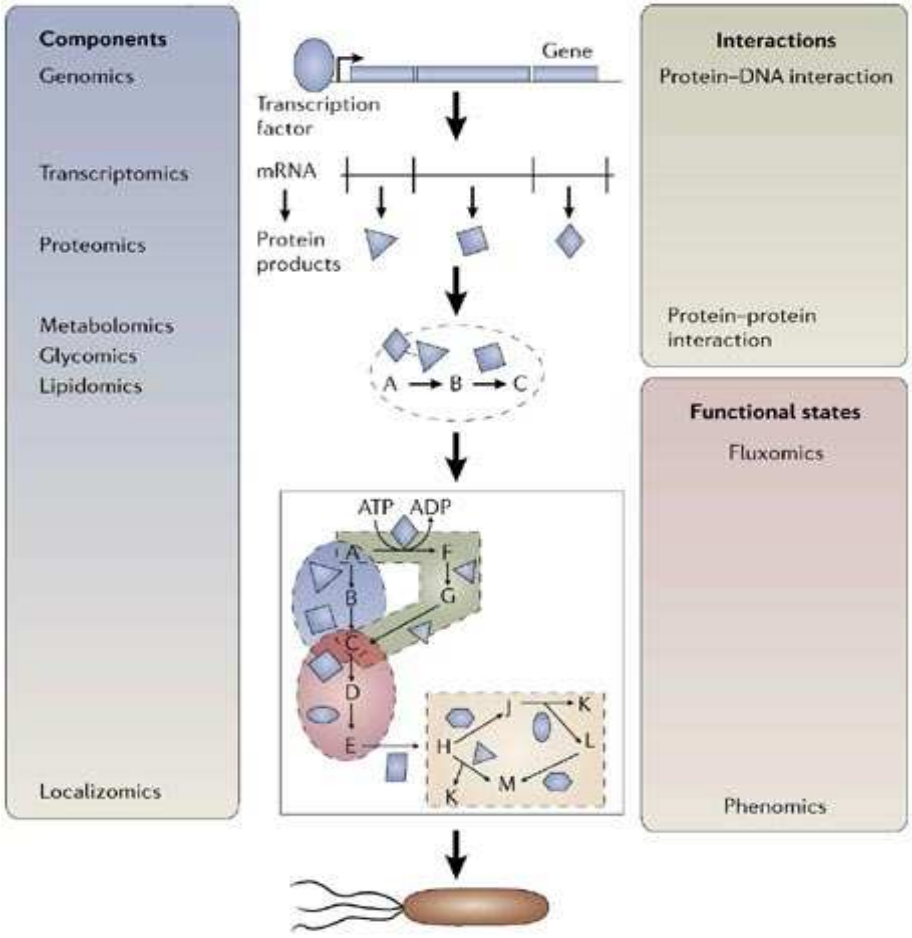
- We will revise some molecular biology concepts (CH2)
- We will introduce some historically important ways to find patterns in sequences (CH3)
- We will give a primer on how to compare sequences, with indications about its relevance to phylogenetic analysis (CH4)
- We will then focus on the human genome and address:
    - ‘Statistical’ aspects in the genomewide association analysis of Single Nucleotide Polymorphisms (SNPs) (CH5)
    - Add additional levels of complexity:
        - Gene-gene interactions (CH6)
        - Gene-environment interactions: integrating the genome with the exposome (CH6)
        - Family-structure (CH7)

How will we address these challenges in this course?

- In all of the above, we will set pointers towards:
  - mathematical modeling / algorithm developments
  - simulation of biological processes
  - graphical visualization
- The last class will be a surprise GUEST lecture, with some “field workers” from different backgrounds, using bioinformatics tools on a case study
- Such a case study will serve multiple purposes:
  - Be aware that this is an INTRODUCTION course in bioinformatics
  - Get you WARMED UP for future work in this field …..
  - When interested, do not hesitate to CONTACT ME!

Statistical Genetics Research Club (www.statgen.be)

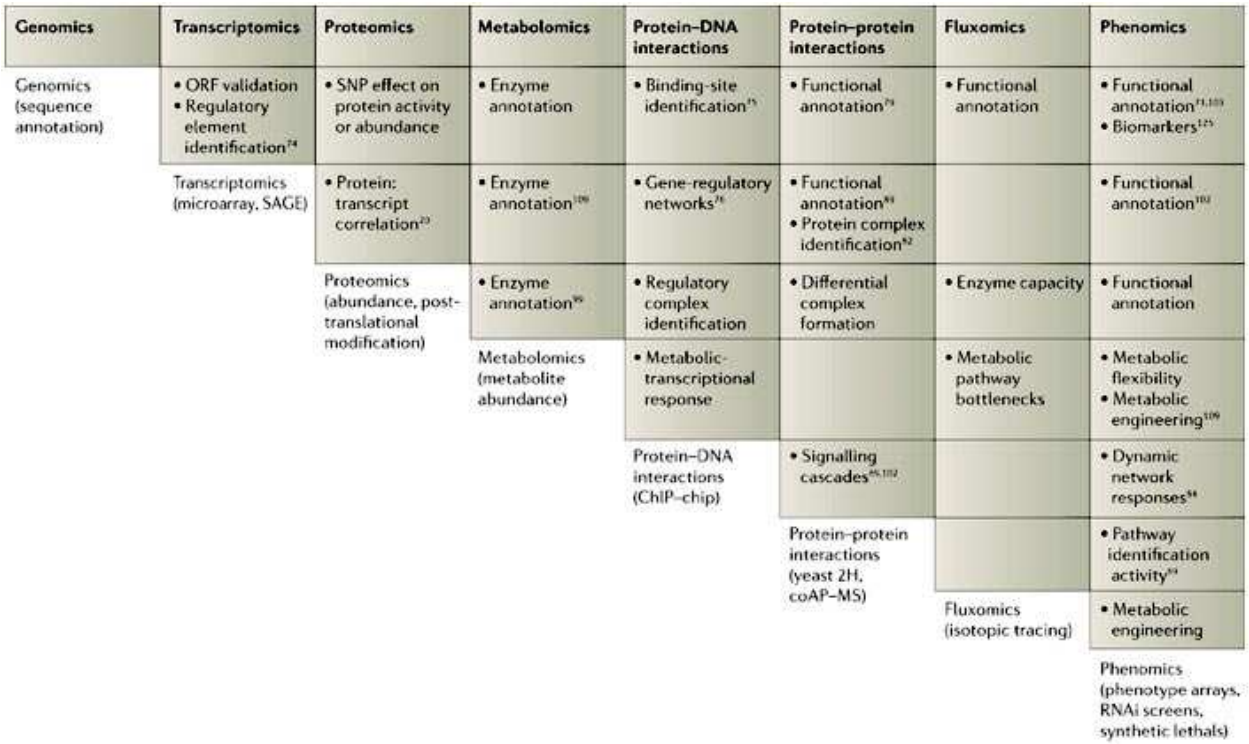
An integrated view (1) - (Joyce *et al. Nature Reviews Molecular Cell Biology* 2006)



Omics data

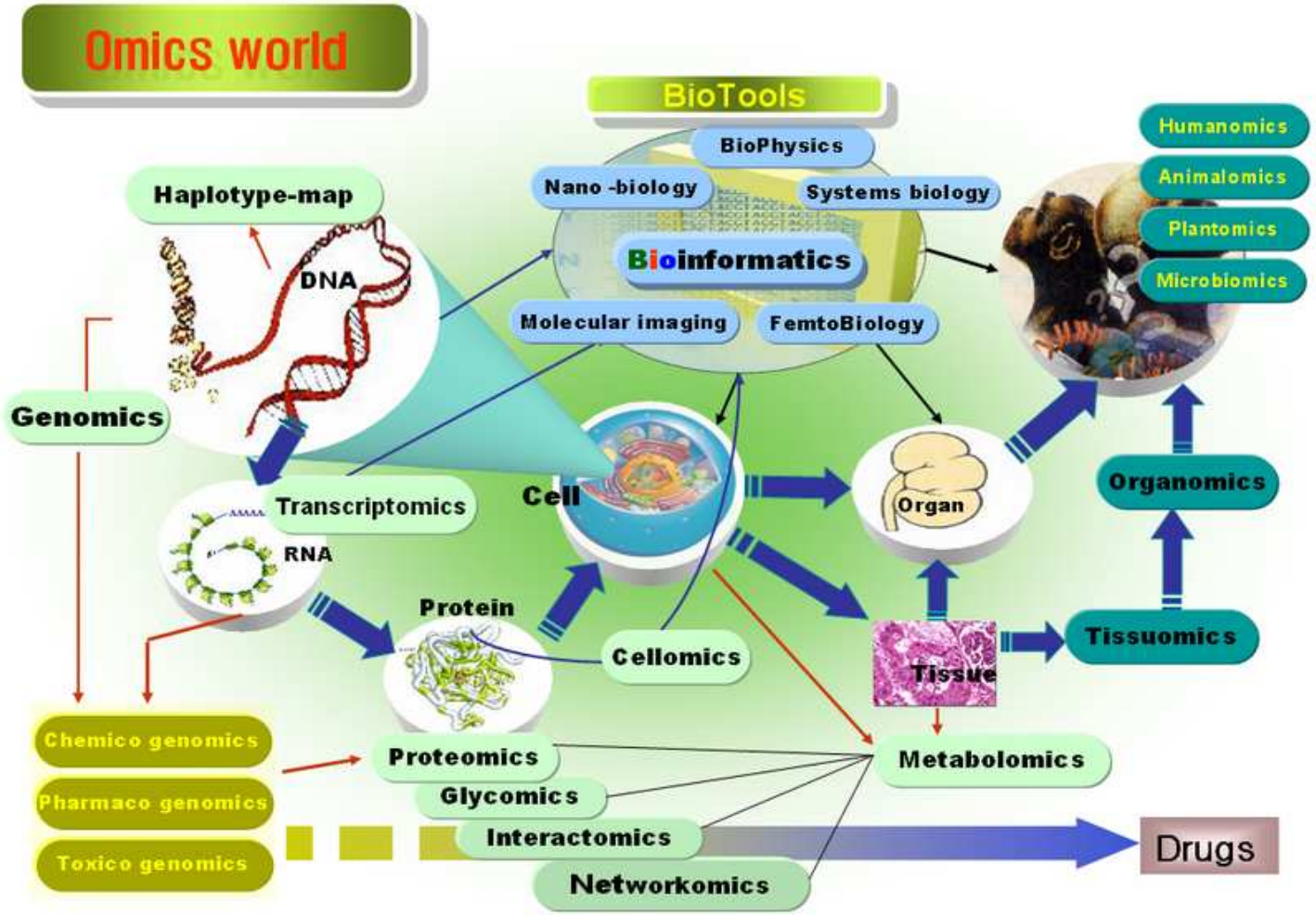
- In the Omics era, we see proliferation of genome/proteome-wide high throughput data that are available in public archives
    - Comparative genome sequences
    - Sequence variation & phenotypes
    - Epigenetics & chromatin structure
    - Regulatory elements & gene expression
    - Protein expression, modification & localization
    - Protein domain, structure, interaction
    - Metabolic, signal, regulatory pathways
    - Drug, toxicogenomics, toxicoproteomics

An integrated view: multi-omics



Genomics	Transcriptomics	Proteomics	Metabolomics	Protein–DNA interactions	Protein–protein interactions	Fluxomics	Phenomics
Genomics (sequence annotation)	• ORF validation • Regulatory element identification	• SNP effect on protein activity or abundance	• Enzyme annotation	• Binding-site identification	• Functional annotation	• Functional annotation	• Functional annotation • Biomarkers
	Transcriptomics (microarray, SAGE)	• Protein: transcript correlation	• Enzyme annotation	• Gene-regulatory networks	• Functional annotation • Protein complex identification		• Functional annotation
		Proteomics (abundance, post-translational modification)	• Enzyme annotation	• Regulatory complex identification	• Differential complex formation	• Enzyme capacity	• Functional annotation
			Metabolomics (metabolite abundance)	• Metabolic-transcriptional response		• Metabolic pathway bottlenecks	• Metabolic flexibility • Metabolic engineering
				Protein–DNA interactions (ChIP–chip)	• Signalling cascades		• Dynamic network responses
					Protein–protein interactions (yeast 2H, coAP–MS)		• Pathway identification activity
						Fluxomics (isotopic tracing)	• Metabolic engineering
							Phenomics (phenotype arrays, RNAi screens, synthetic lethals)

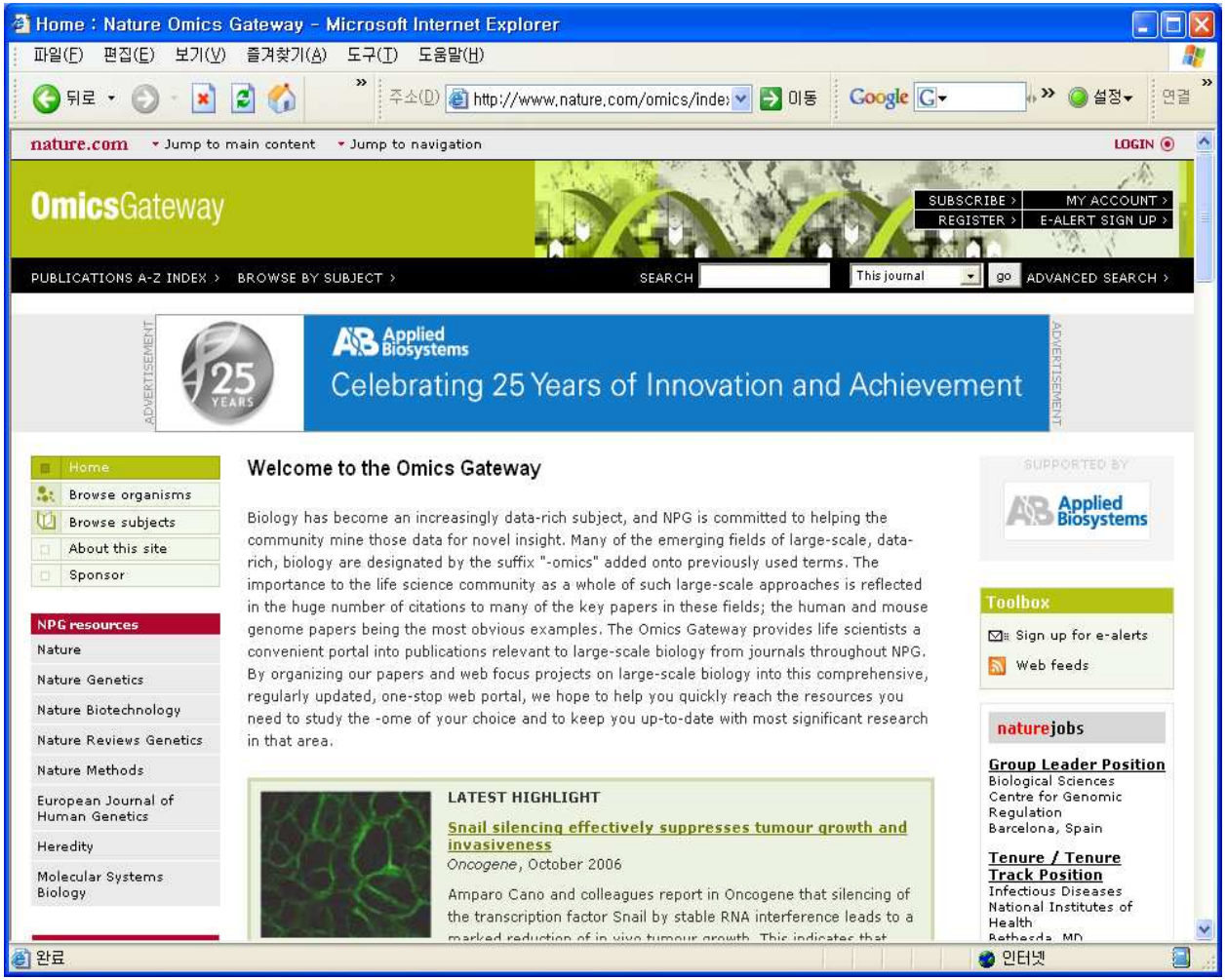
An integrated view (2): from multi-omics to multi-data types



No need to restrict to a single species

	human	mammal	vertebrate	animal	eukaryote
Genome sequence					
Chromatin structure					
Transcription & regulation					
Protein expression					
PTM & localization					
Structure & interaction					

Where to look for additional info? - http://www.nature.com/omics/index.html



1.3 The origins of bioinformatics

Bioinformatics is often *confused with* computational biology

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

Computational biology

- *"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."*

(http://rbaltman.wordpress.com/2009/02/18/bioinformatics-computational-biology-same-no/)

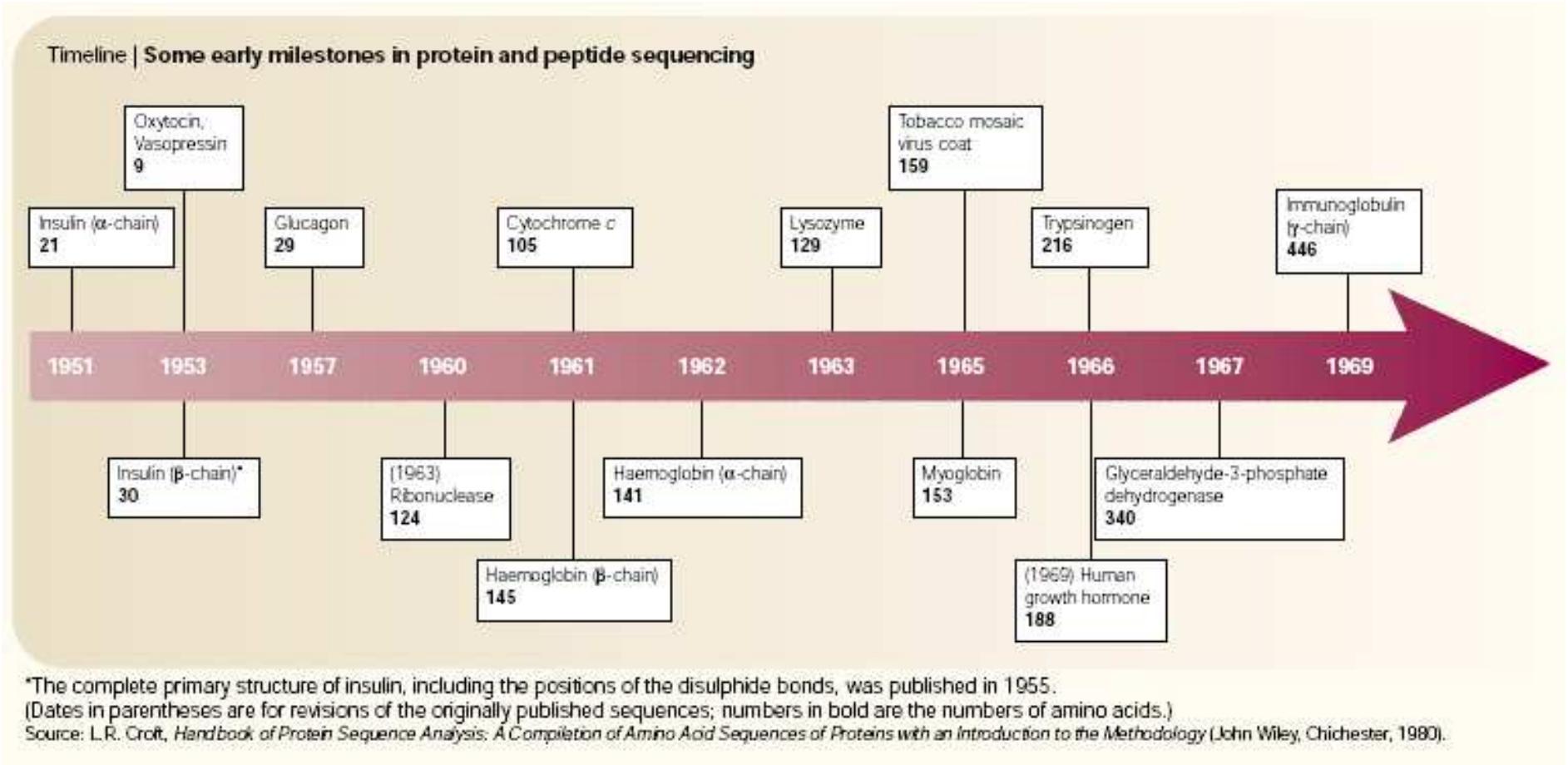
- Three important factors facilitated the emergence of computational biology during the early 1960s.
  - First, an expanding collection of amino-acid sequences provided both a source of data and a set of interesting problems that were infeasible to solve without the number-crunching power of computers.
  - Second, the idea that macromolecules (proteins carry information encoded in linear sequences of amino acids) carry information became a central part of the conceptual framework of molecular biology.
  - Third, high-speed digital computers, which had been developed from weapons research programs during the Second World War, finally became widely available to academic biologists.

(Hagen 2000)

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 computational biology



*The complete primary structure of insulin, including the positions of the disulphide bonds, was published in 1955.
(Dates in parentheses are for revisions of the originally published sequences; numbers in bold are the numbers of amino acids.)
Source: L.R. Croft, *Handbook of Protein Sequence Analysis: A Compilation of Amino Acid Sequences of Proteins with an Introduction to the Methodology* (John Wiley, Chichester, 1980).

The emergence of computational biology

- By 1970, computational biologists had developed a diverse set of techniques for analyzing molecular structure, function and evolution.
- The idea of proteins acting as information-carrying macromolecules consecutively lead to developments in 3 broadly overlapping contexts
- These contexts are:

    - the genetic code,
    - the three-dimensional structure of a protein in relation to its function, and
    - the protein evolution

(Hagen 2000)

The emergence of bioinformatics

- Some of these techniques, initially developed by computational biologists, 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.
- The field later became 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)

Bioinformatics

- *"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."*

(http://rbaltman.wordpress.com/2009/02/18/bioinformatics-computational-biology-same-no/)

2 Definitions for bioinformatics

2.1 A “clear” definition for bioinformatics

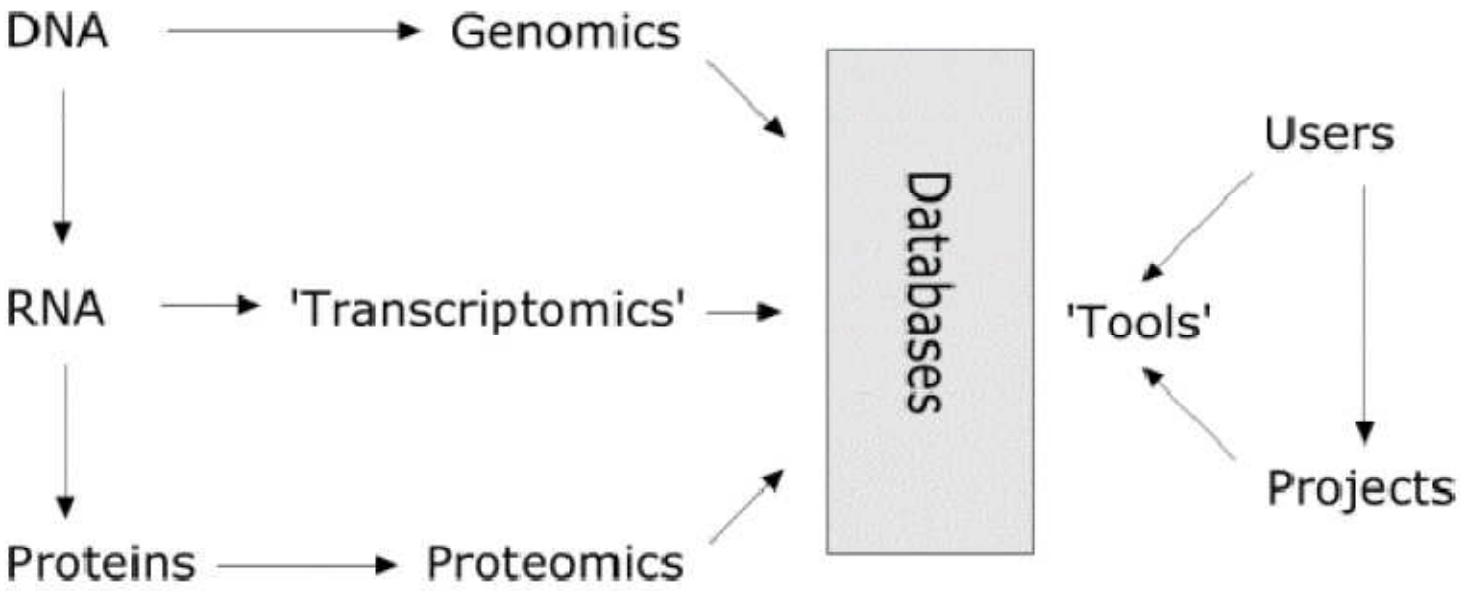
Bioinformatics	**Computational biology**
Research, development or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, analyze, or visualize such data	Development and application of data-analytical,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

(BISTIC Definition Committee, NIH, 2000)

Bioinformaticians are jack-of-all-trades

- 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 three levels of data acquisition and data management

(Y vd Peer 2008)

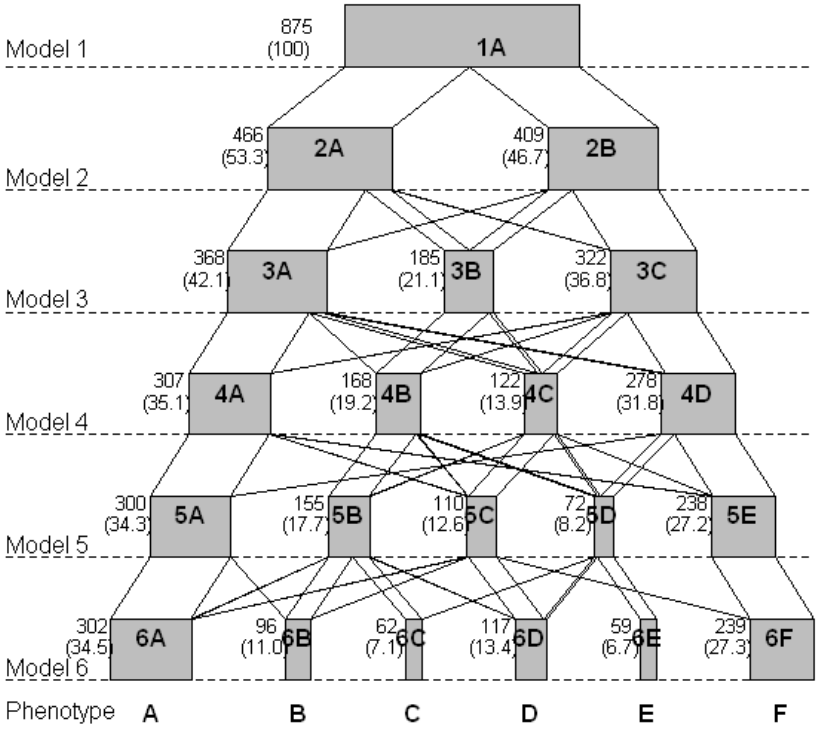
2.2 Topics in bioinformatics from a journal's perspective

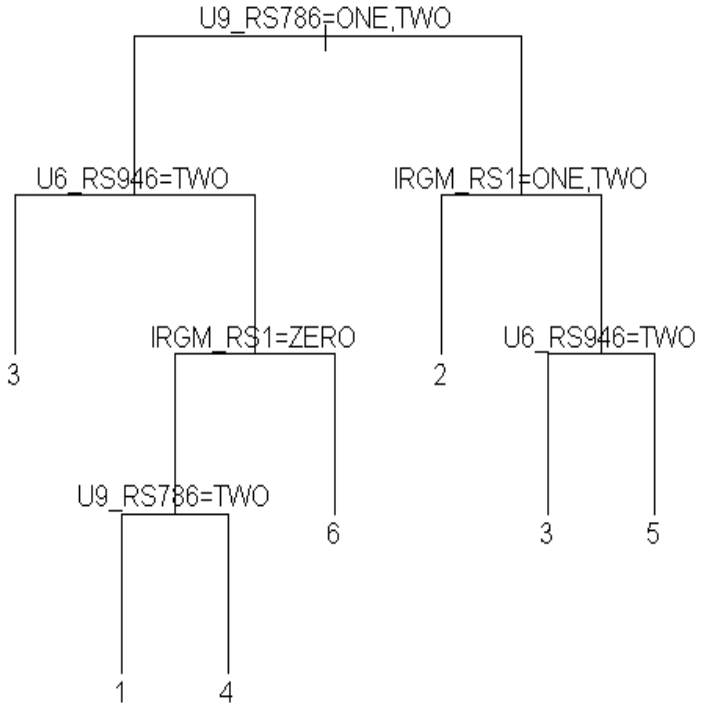
(source: Scope Guidelines of the journal "Bioinformatics")

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?

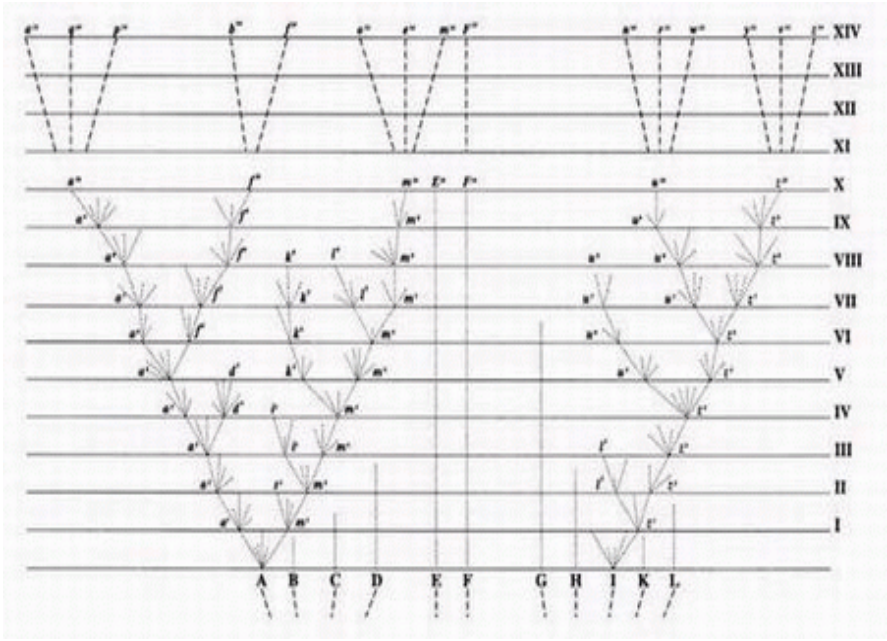
Genome analysis

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

Phylogenetics

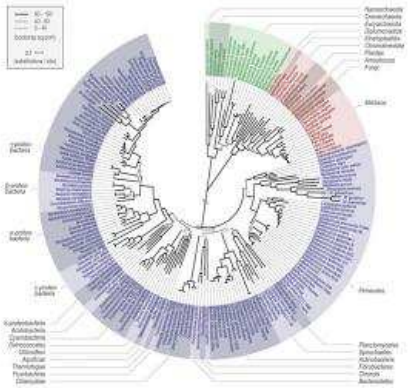
- This category includes:
  - novel phylogeny estimation procedures for molecular data including nucleotide sequence data, amino acid data, SNPs, etc.,
  - simultaneous multiple sequence alignment and
  - phylogeny estimation, using phylogenetic approaches for any aspect of molecular sequence analysis (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

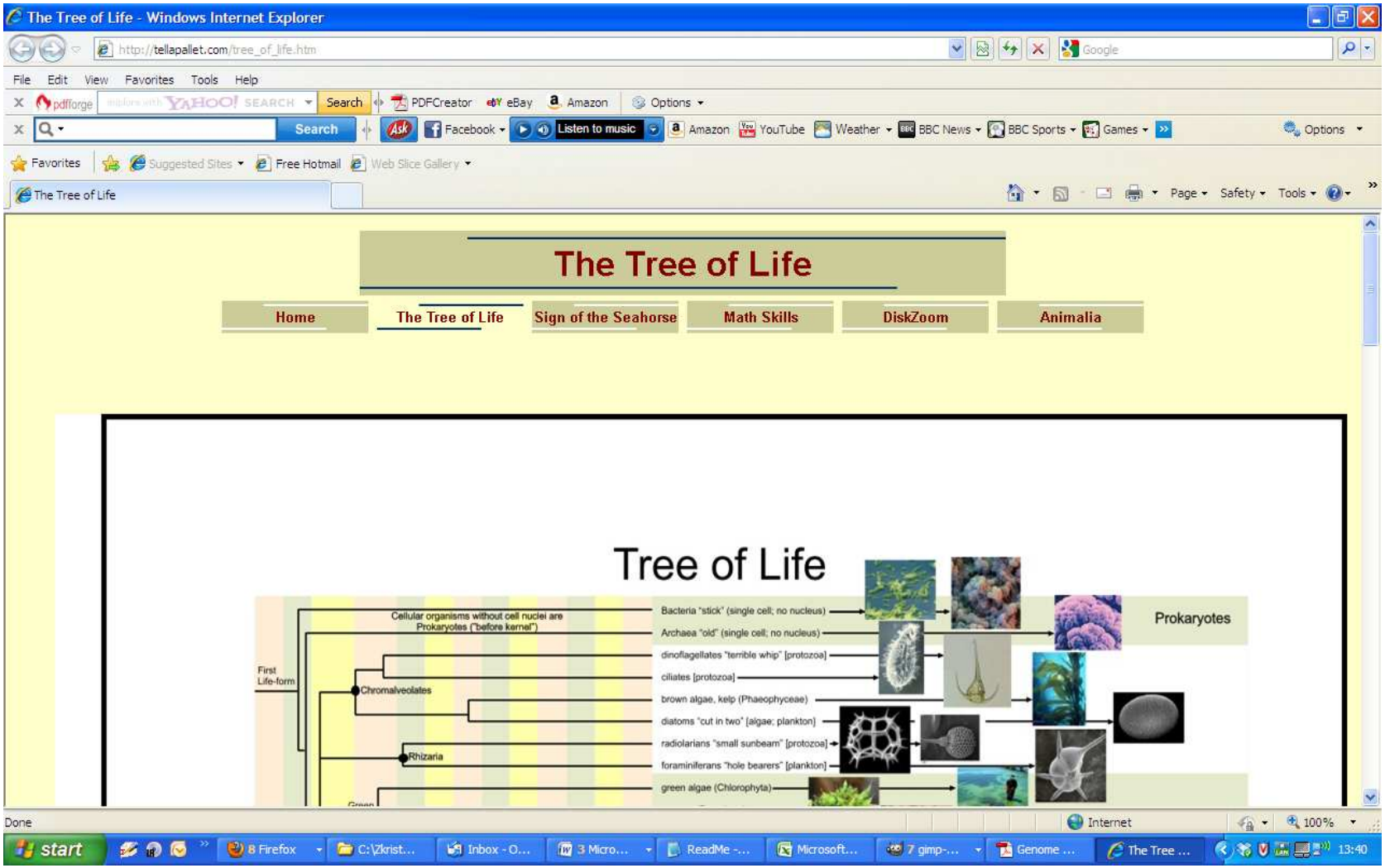


The Tree of Life image that appeared in Darwin's On the Origin of Species by Natural Selection, 1859. It was the book's only illustration

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 trees of life http://tellapallet.com/tree_of_life.htm



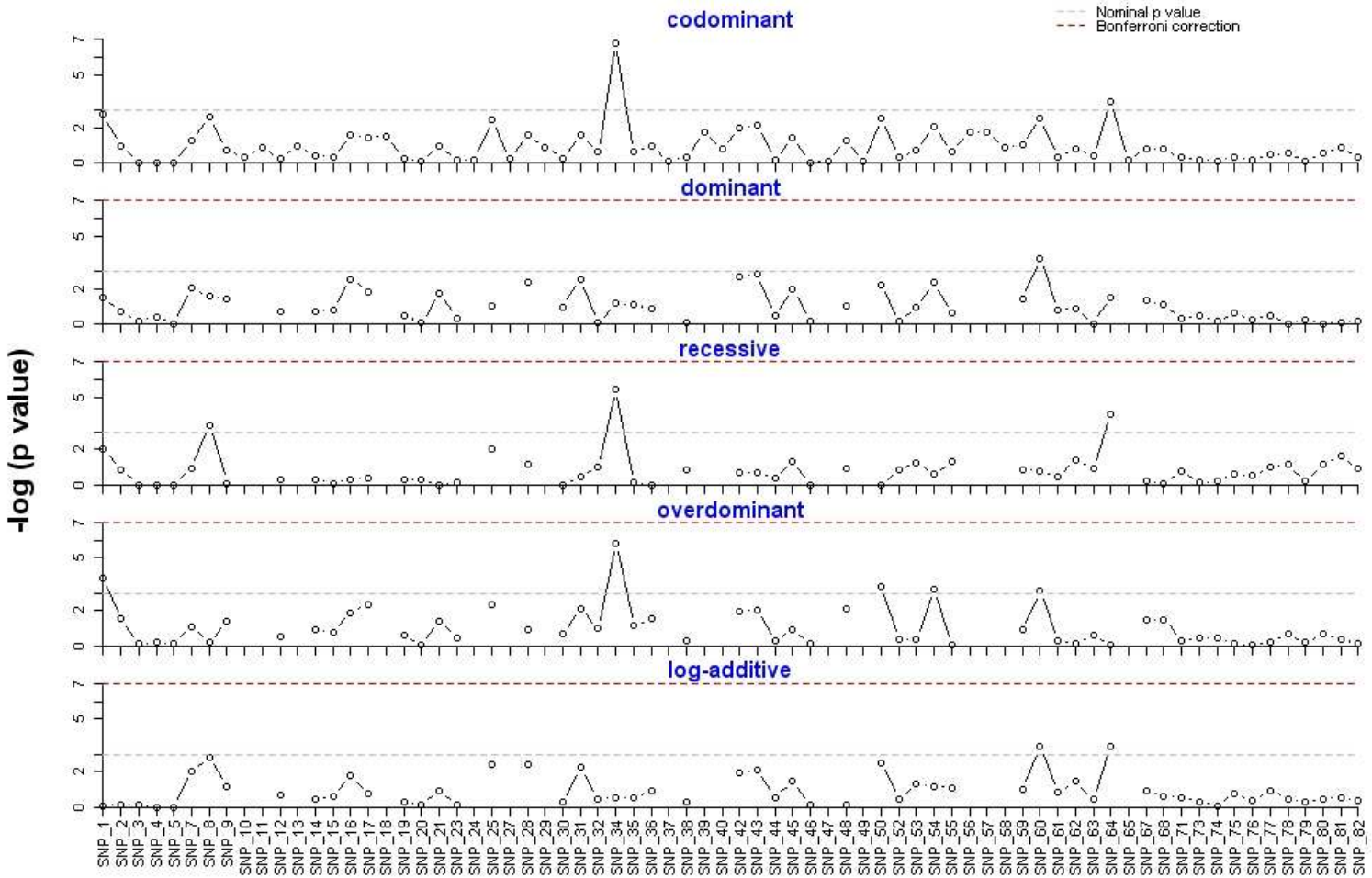
Structural Bioinformatics

- 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

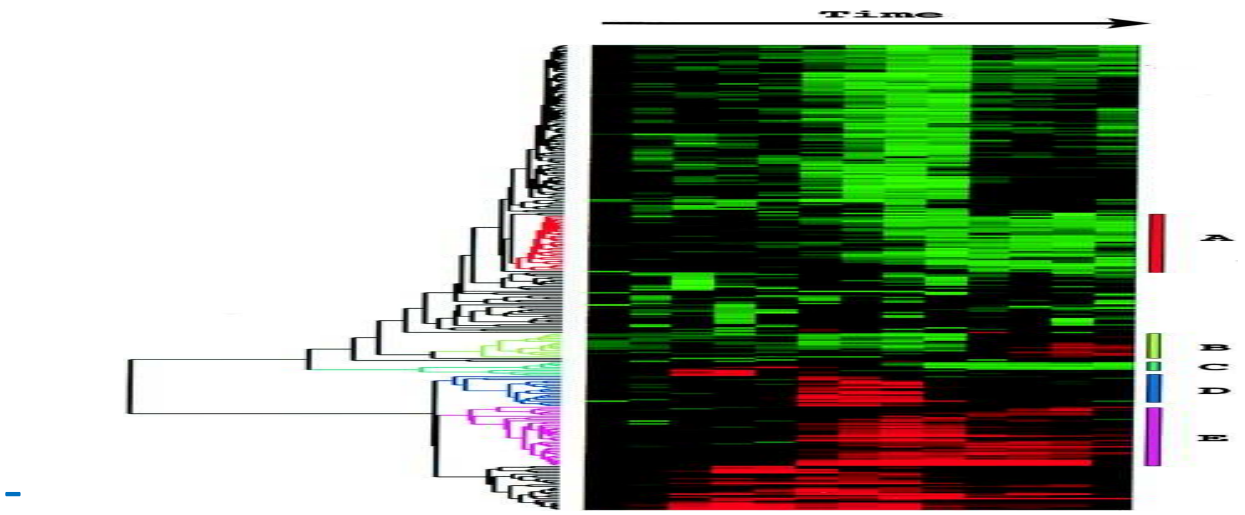


Gene Expression

- 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.

3 Evolving research trends in bioinformatics

3.1 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

3.2 “Early bioinformatics”

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

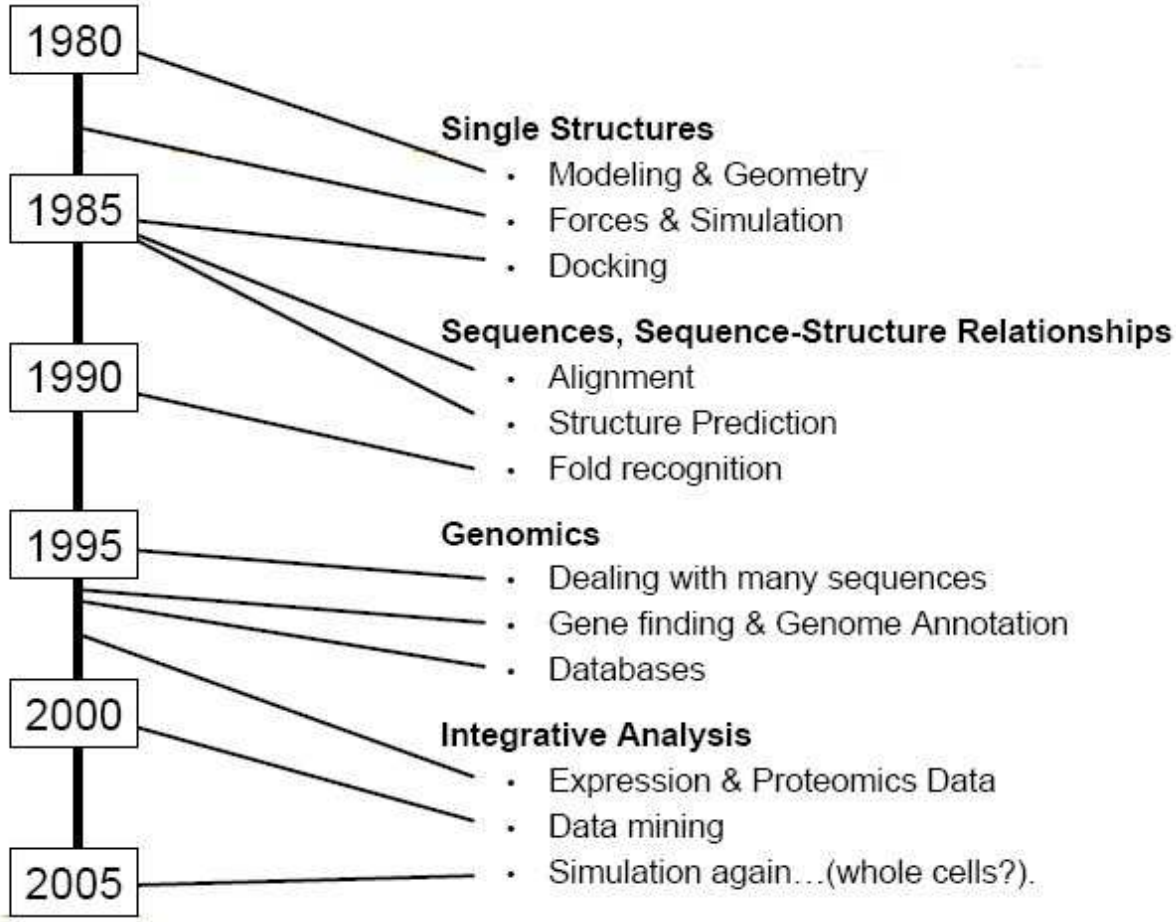
Christos A. Ouzounis[1,*] and Alfonso Valencia[2]

[1]Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK, [2]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

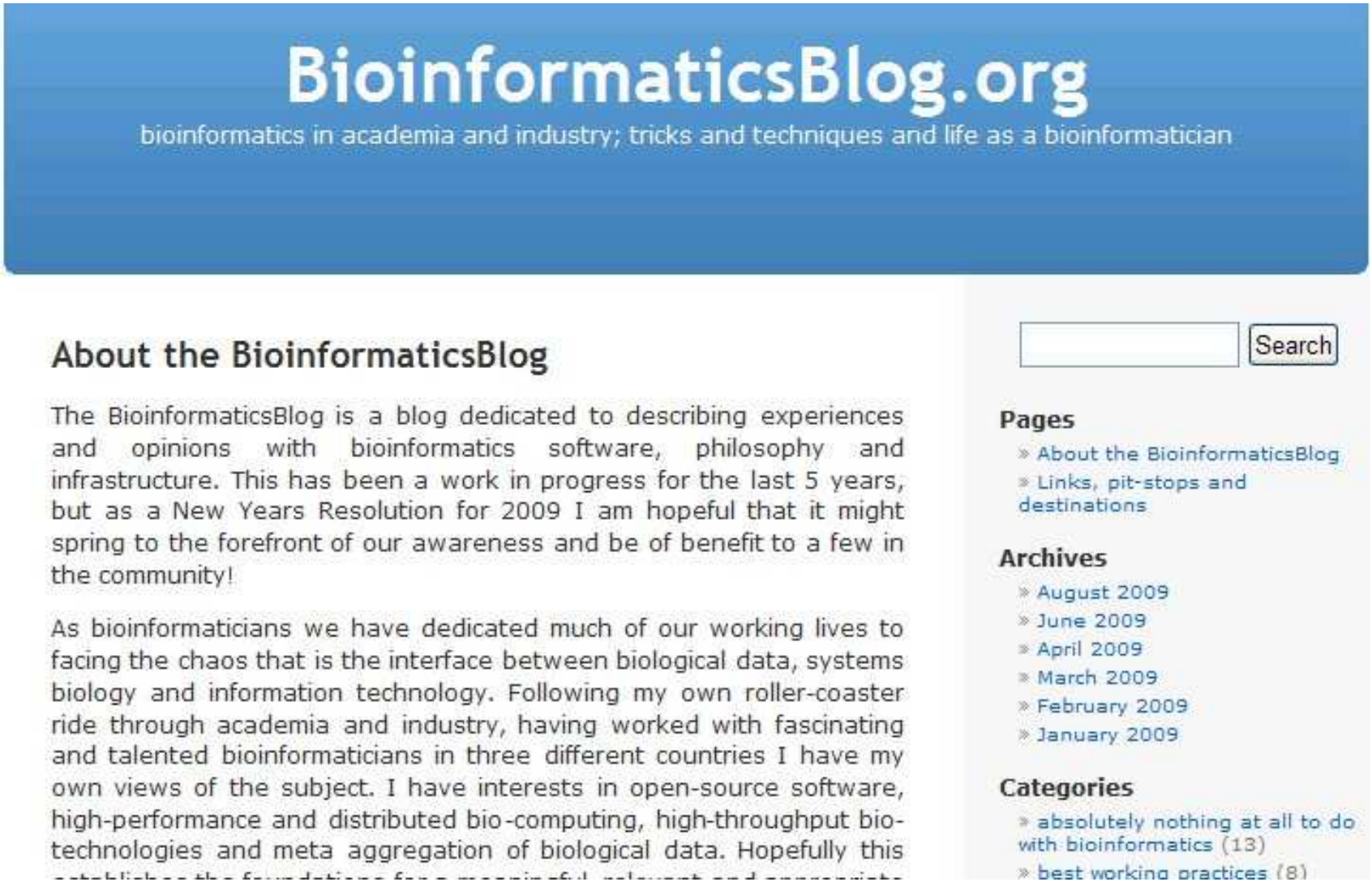
(Ouzounis et al 2003)

3.3 “Later bioinformatics”



(S-Star presentation; Choo)

3.4 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 bio-technologies and meta aggregation of biological data. Hopefully this

Search

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)

4 Bioinformatics Software

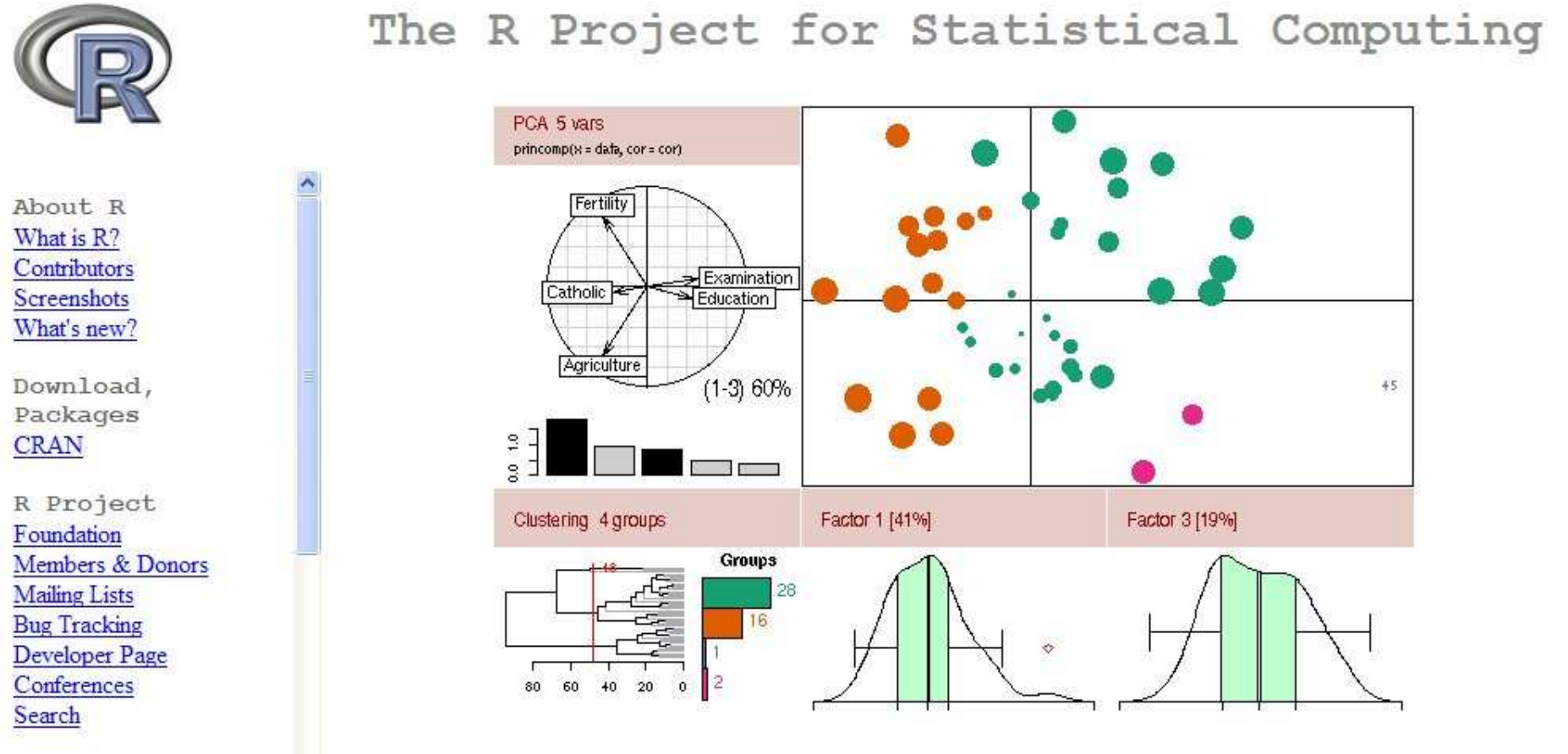
4.1 Introduction

- 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

4.2 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.

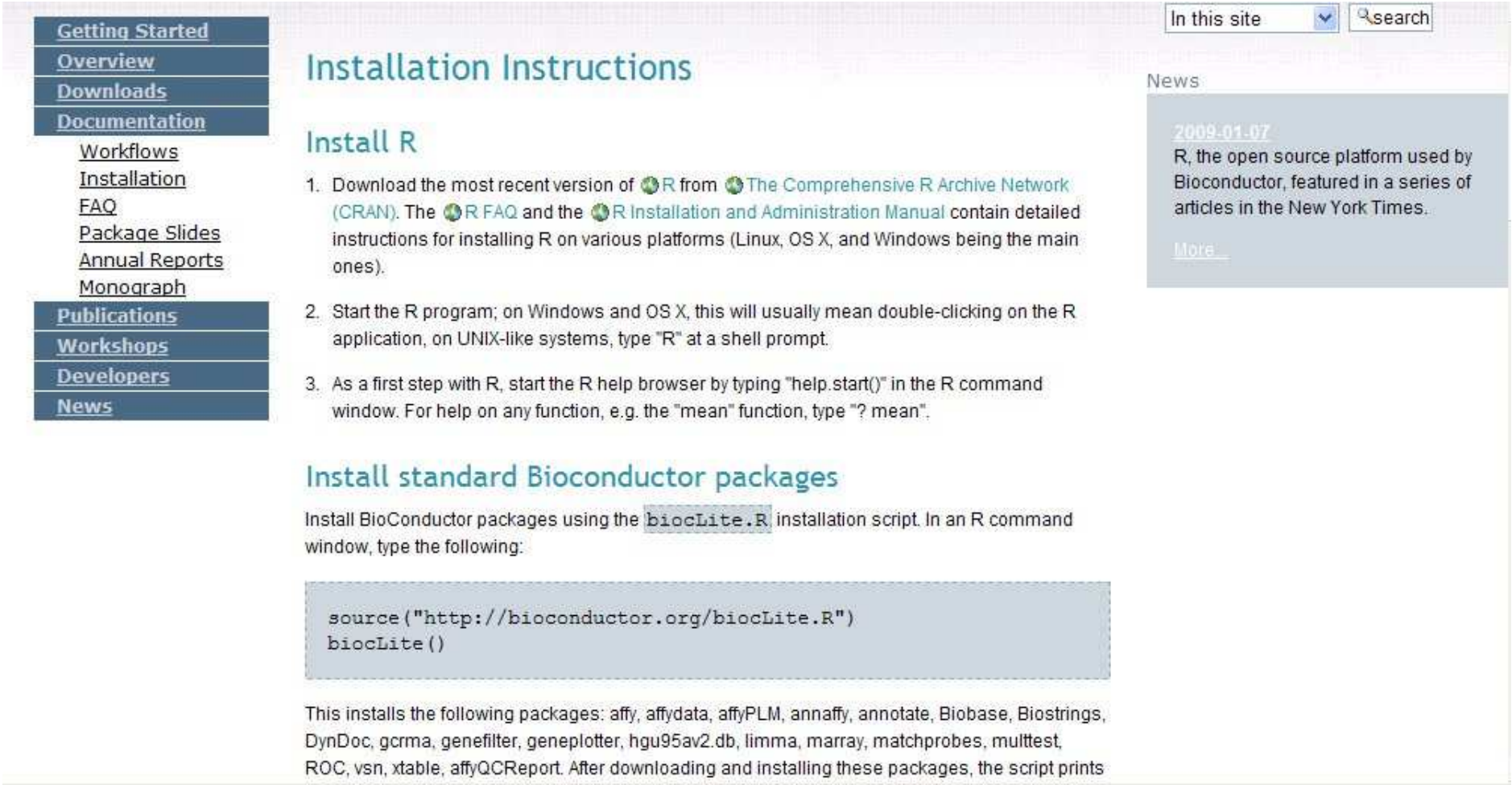
The R environment



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

Bioconductor

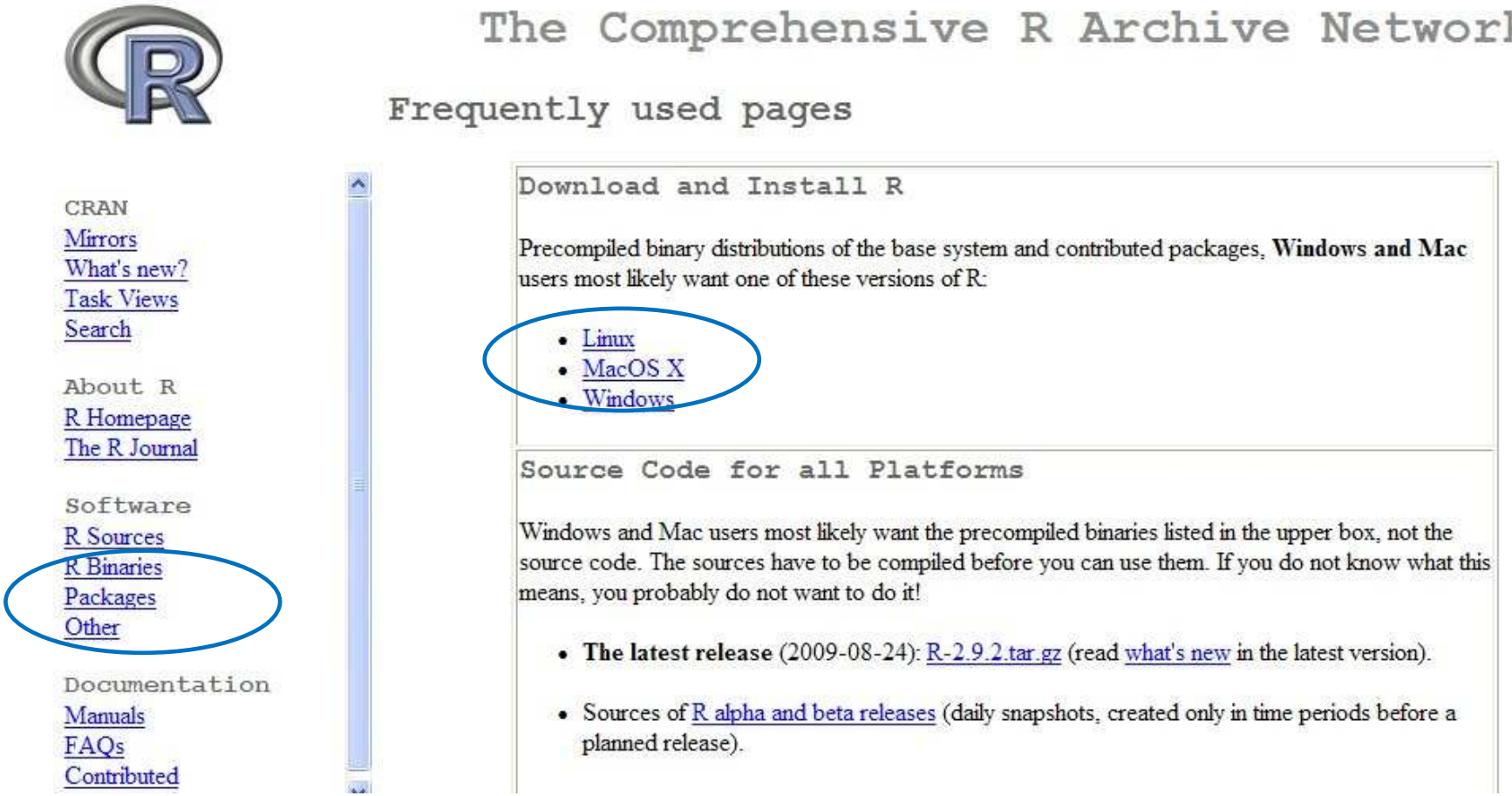
(http://www.bioconductor.org/)



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

R comprehensive network

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



4.3 Example 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.

Check out the Rpackage_download video

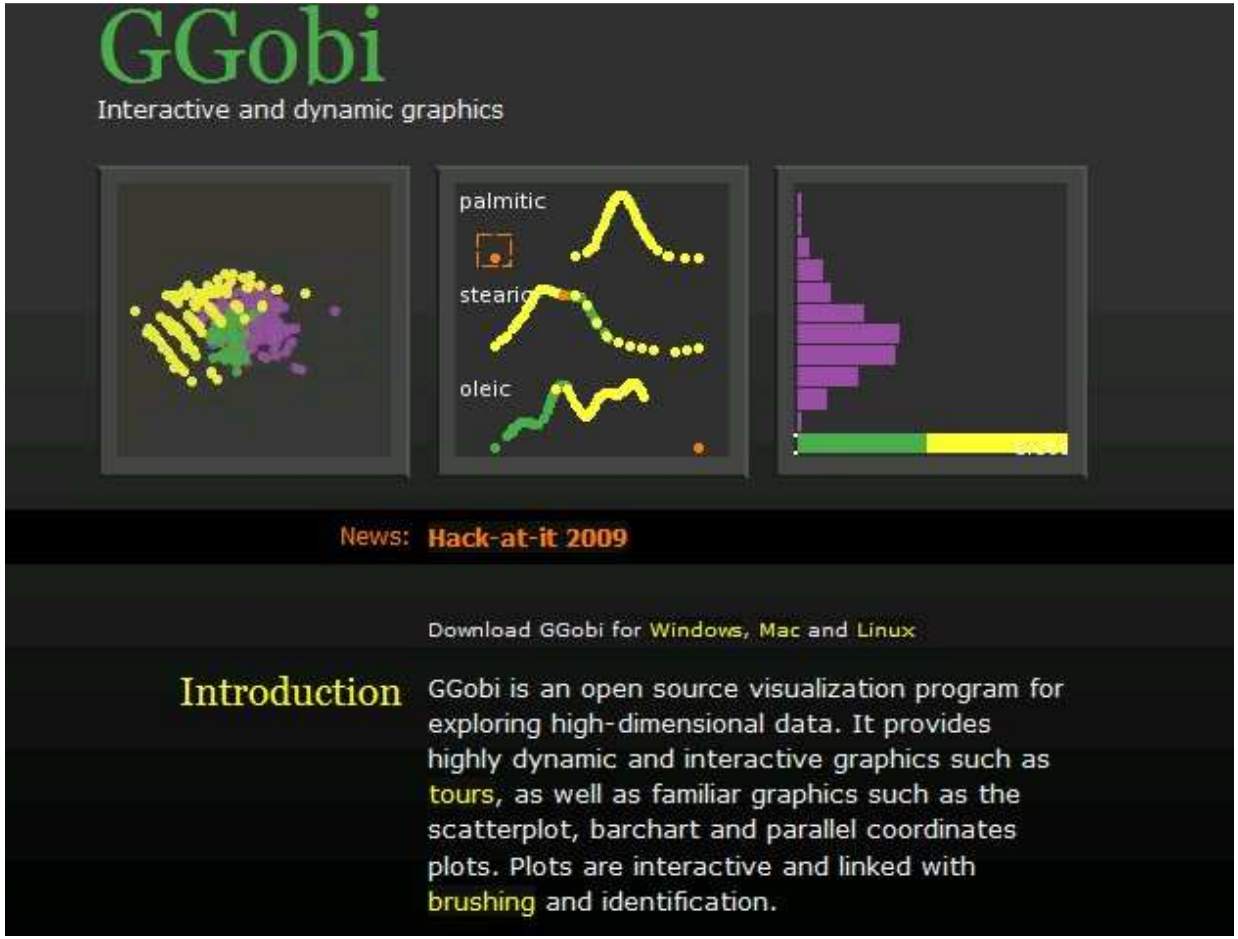
- A comprehensive R & BioConductor manual can be obtained via http://faculty.ucr.edu/~tgirke/Documents/R_BioCond/R_BioCondManual.html

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.

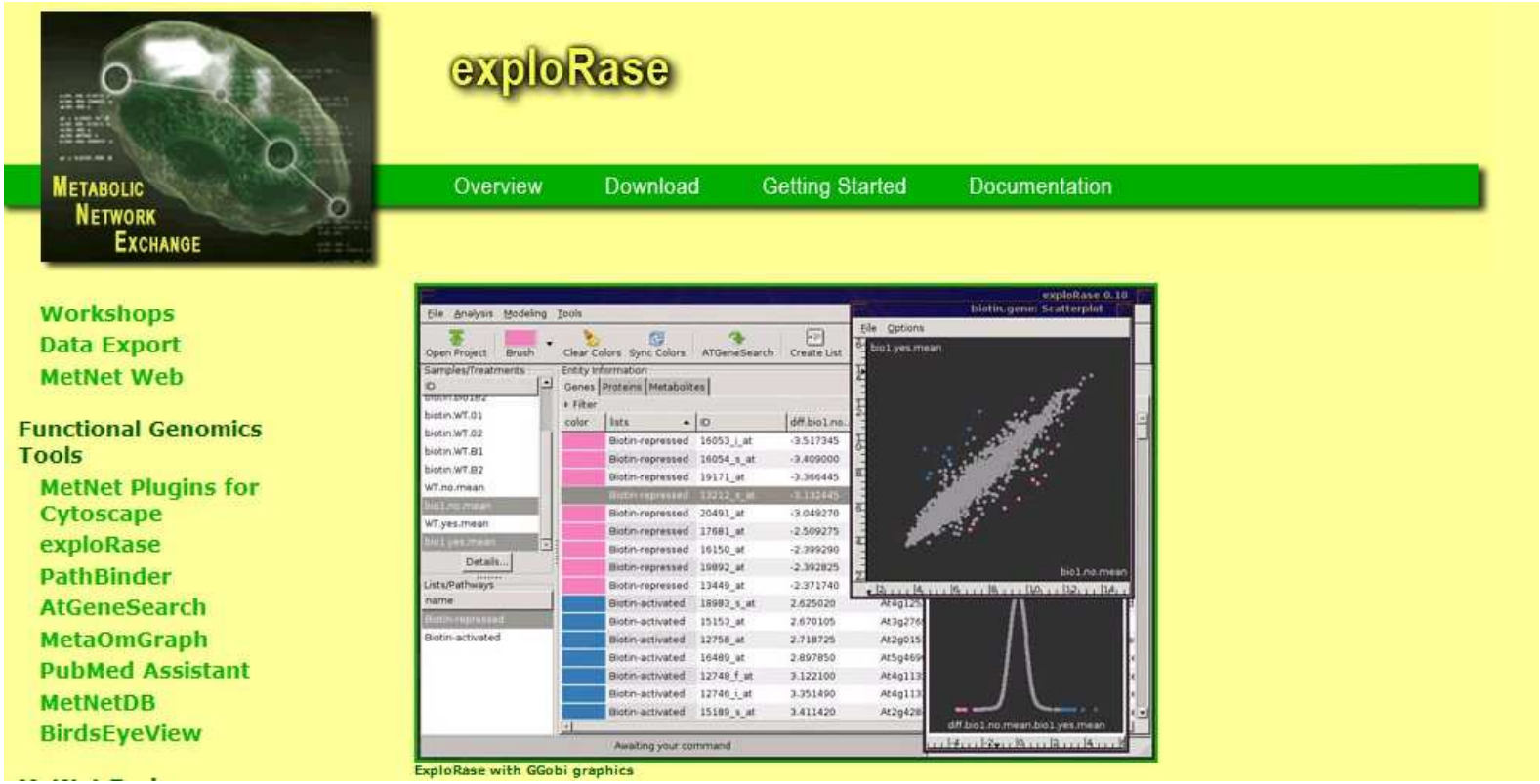
(https://secure.bioconductor.org/BioC2009/)

GGobi



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

exploRase



(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).

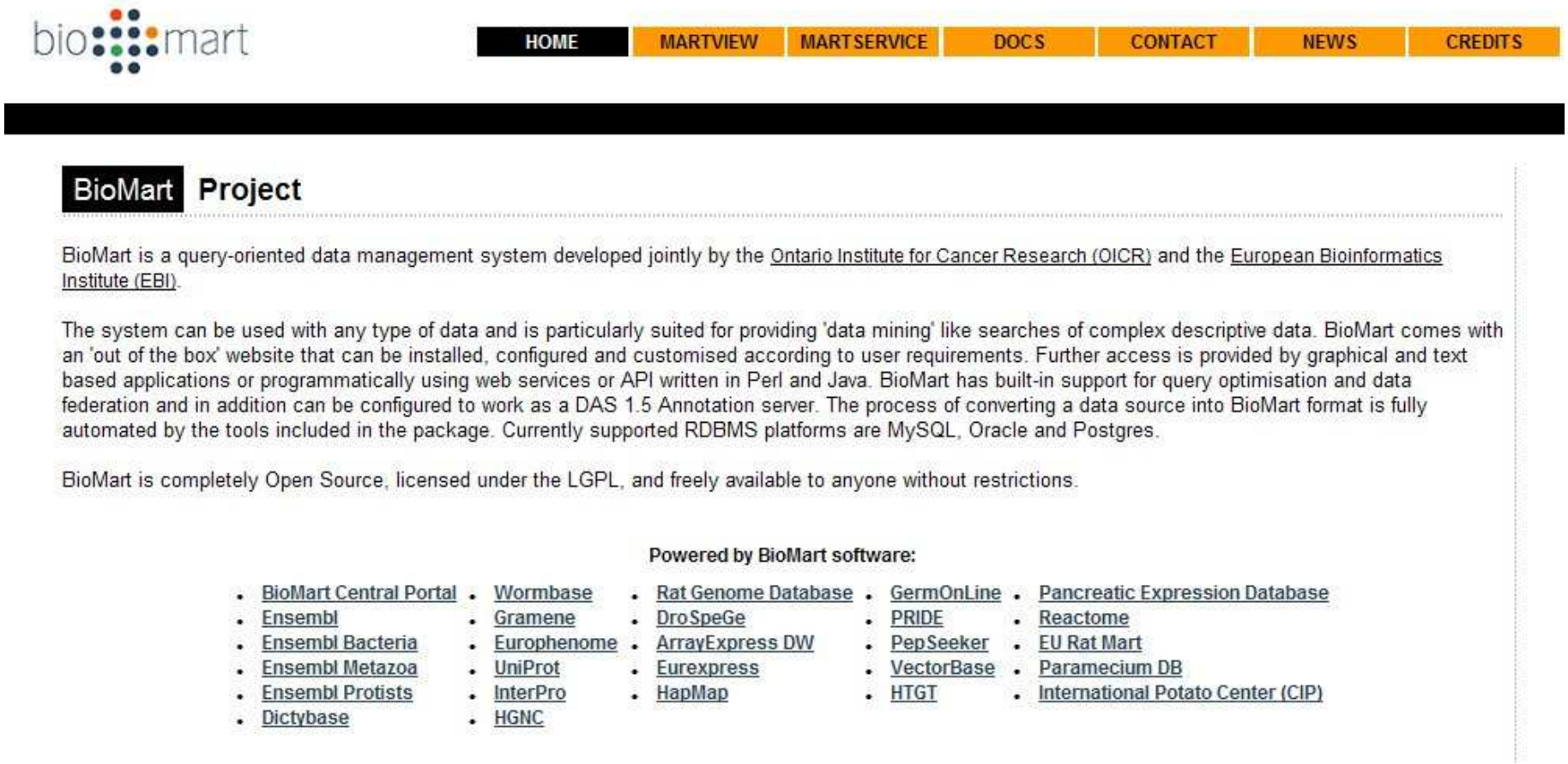
(https://secure.bioconductor.org/BioC2009/)

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.

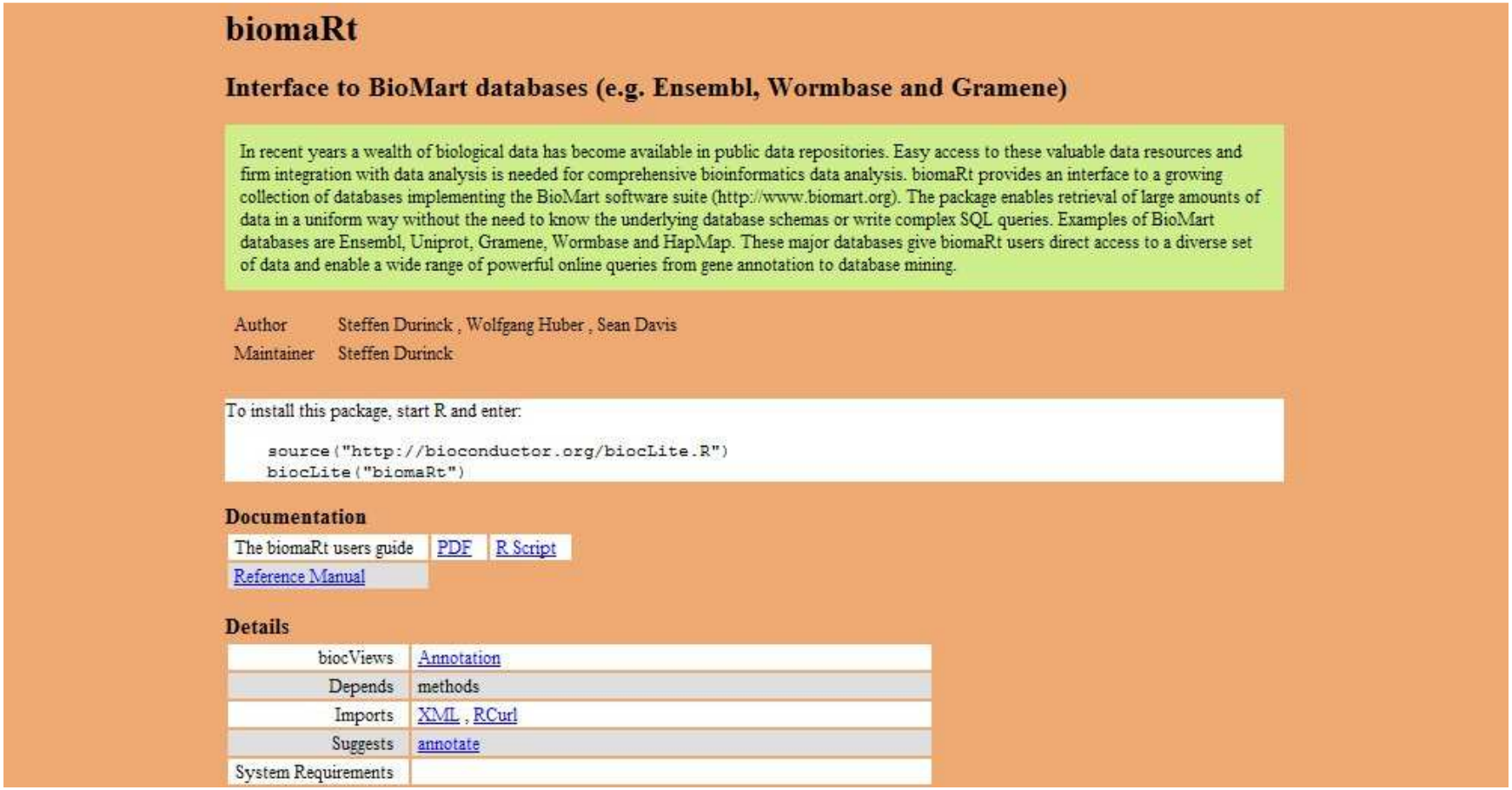
(https://secure.bioconductor.org/BioC2009/)

BioMart



(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

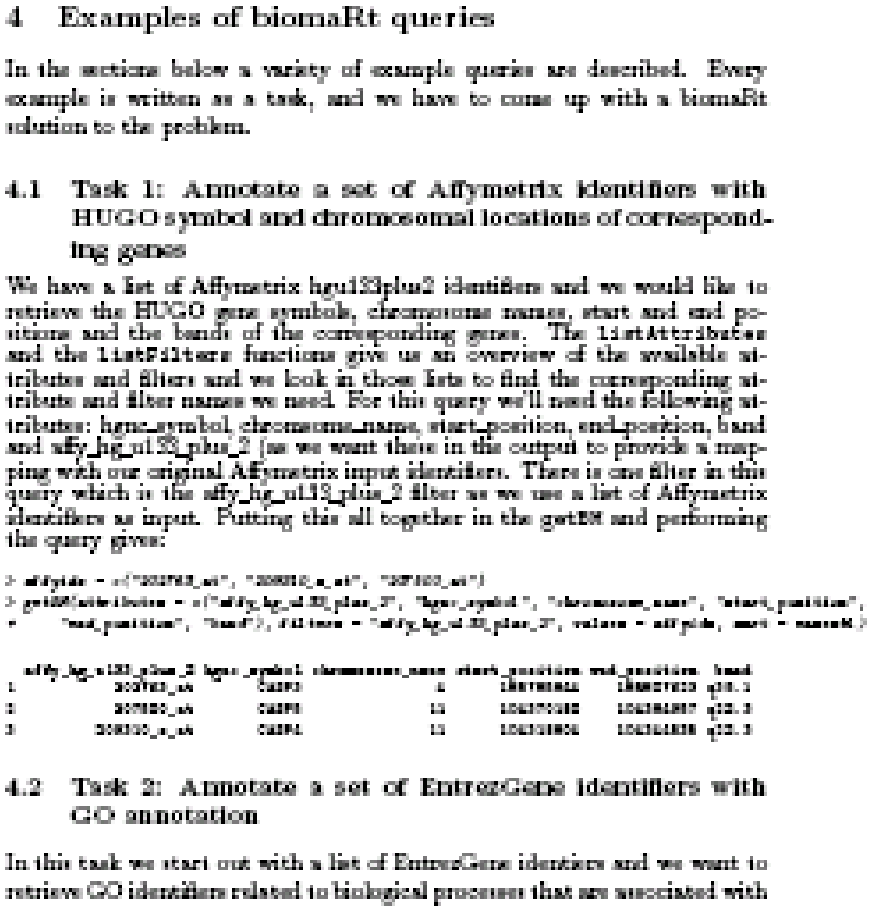
The biomaRt users guide	PDF	R Script
Reference Manual		

Details

biocViews	Annotation
Depends	methods
Imports	XML , RCurl
Suggests	annotate
System Requirements	

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

biomaRt



4 Examples of biomaRt queries

In the sections below a variety of example queries are described. Every example is written as a task, and we have to come up with a biomaRt solution to the problem.

4.1 Task 1: Annotate a set of Affymetrix identifiers with HUGO symbol and chromosomal locations of corresponding genes

We have a list of Affymetrix hgu133plus2 identifiers and we would like to retrieve the HUGO gene symbols, chromosome names, start and end positions and the bands of the corresponding genes. The `listAttributes` and the `listFilters` functions give us an overview of the available attributes and filters and we look in those lists to find the corresponding attributes and filter names we need. For this query we'll need the following attributes: hgnc_symbol, chromosome_name, start_position, end_position, band and affy_hg_u133_plus_2 (as we want these in the output to provide a mapping with our original Affymetrix input identifiers. There is one filter in this query which is the affy_hg_u133_plus_2 filter as we use a list of Affymetrix identifiers as input. Putting this all together in the `getBM` and performing the query gives:

```
> affyids = c("202763_at", "209310_s_at", "207500_at")
> getBM(attributes = c("affy_hg_u133_plus_2", "hgnc_symbol", "chromosome_name", "start_position",
+     "end_position", "band"), filters = "affy_hg_u133_plus_2", values = affyids, mart = mart36)
```

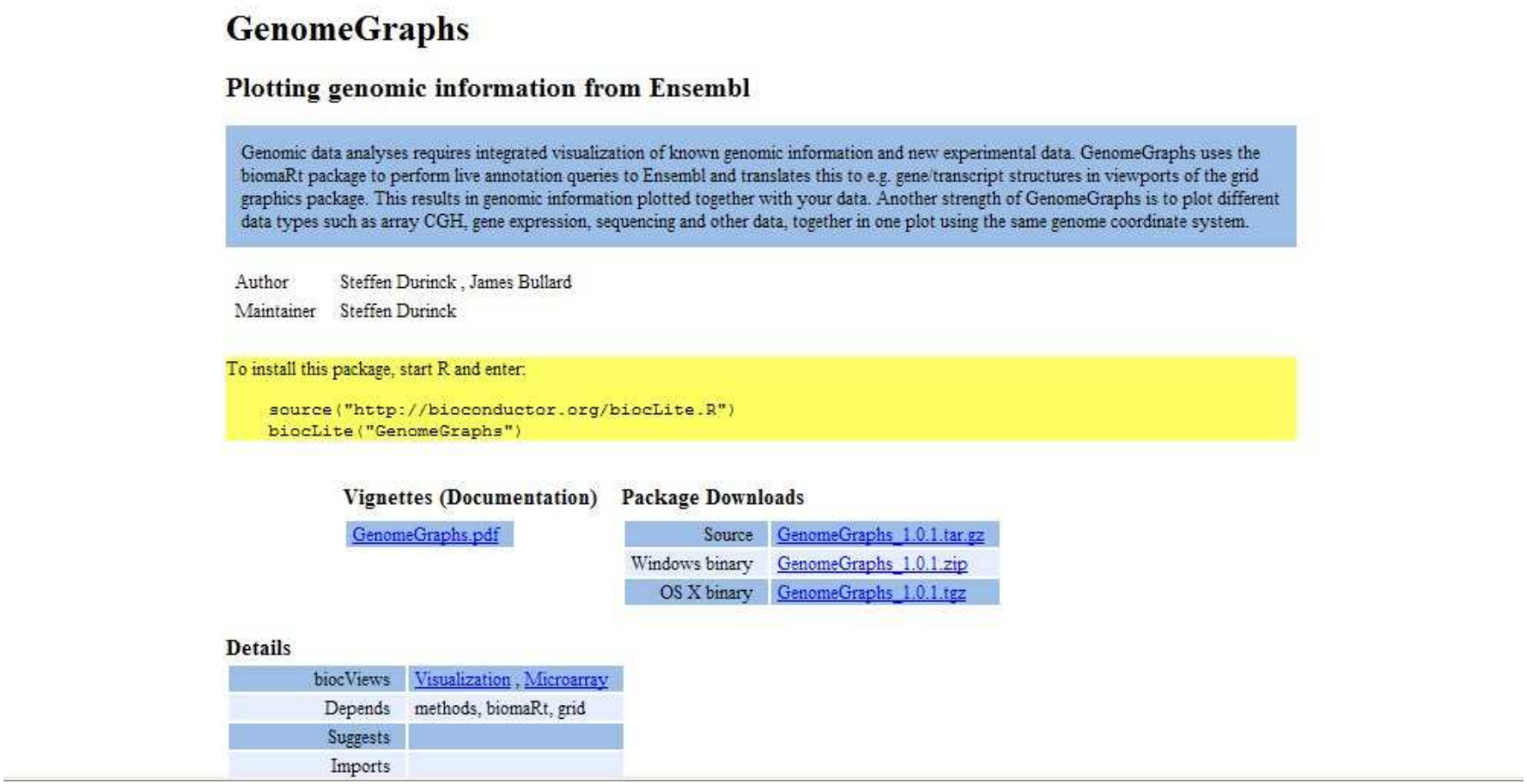
	affy_hg_u133_plus_2	hgnc_symbol	chromosome_name	start_position	end_position	band
1	202763_at	CASP3	4	184672832	184677933	q35.1
2	207500_at	CASP8	11	100370308	100384887	q22.3
3	209310_s_at	CASP4	11	100331804	100341838	q22.3

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

(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	GenomeGraphs_1.0.1.tar.gz
Windows binary	GenomeGraphs_1.0.1.zip
OS X binary	GenomeGraphs_1.0.1.tgz

Details

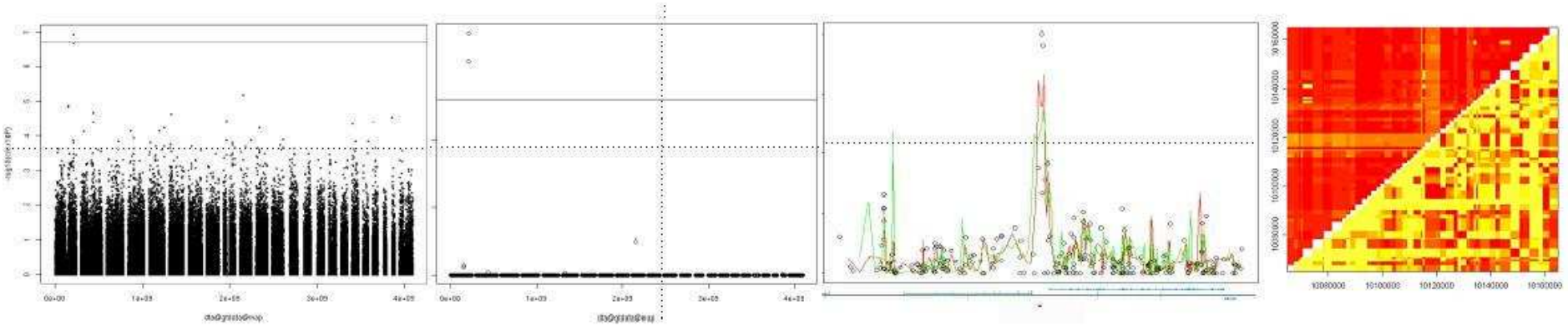
biocViews	Visualization , Microarray
Depends	methods, biomaRt, grid
Suggests	
Imports	

(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/)



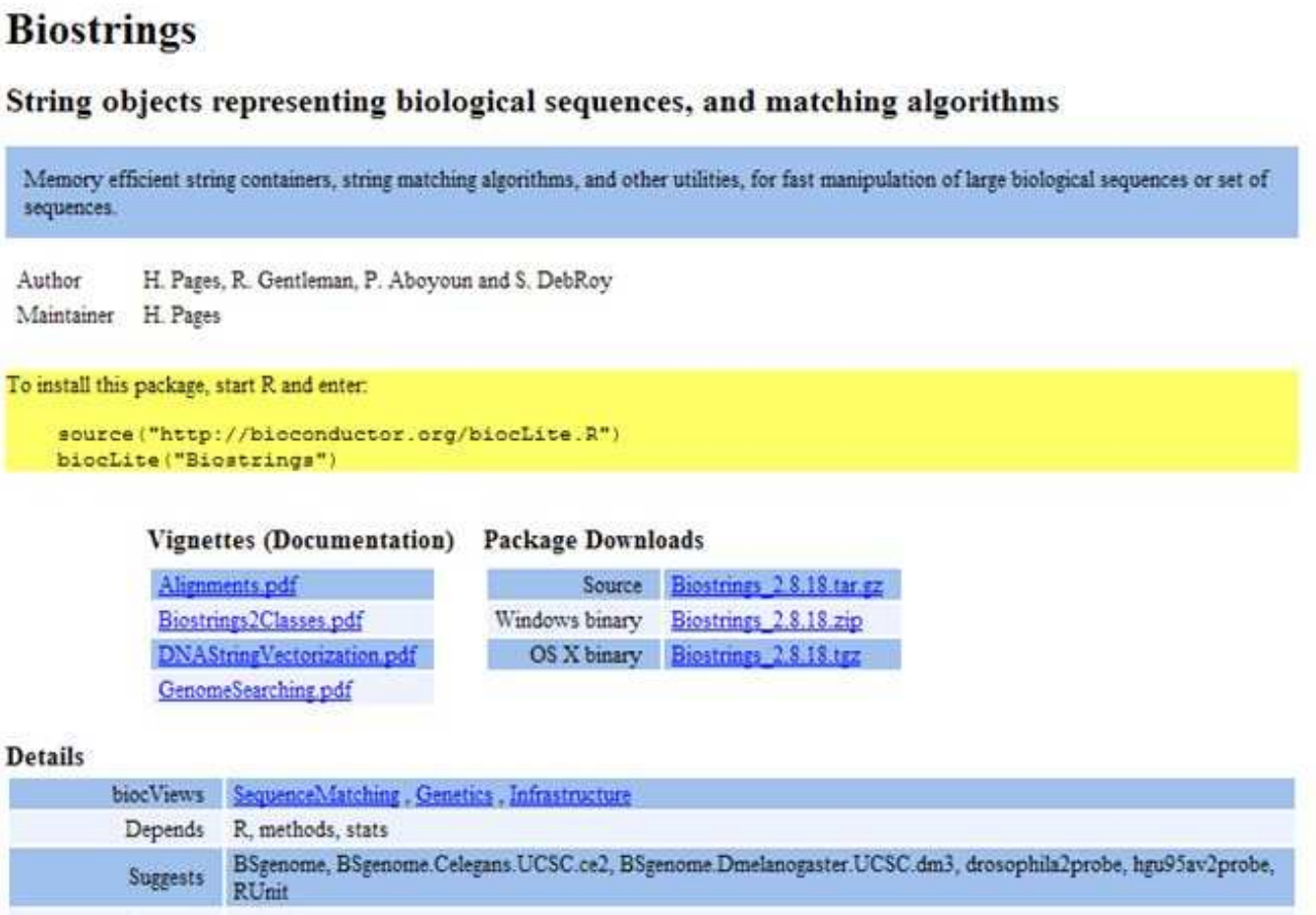
(http://mga.bionet.nsc.ru/~yurii/ABEL/GenABEL/)

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	SequenceMatching , Genetics , Infrastructure
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 `substituionMatrix` 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:

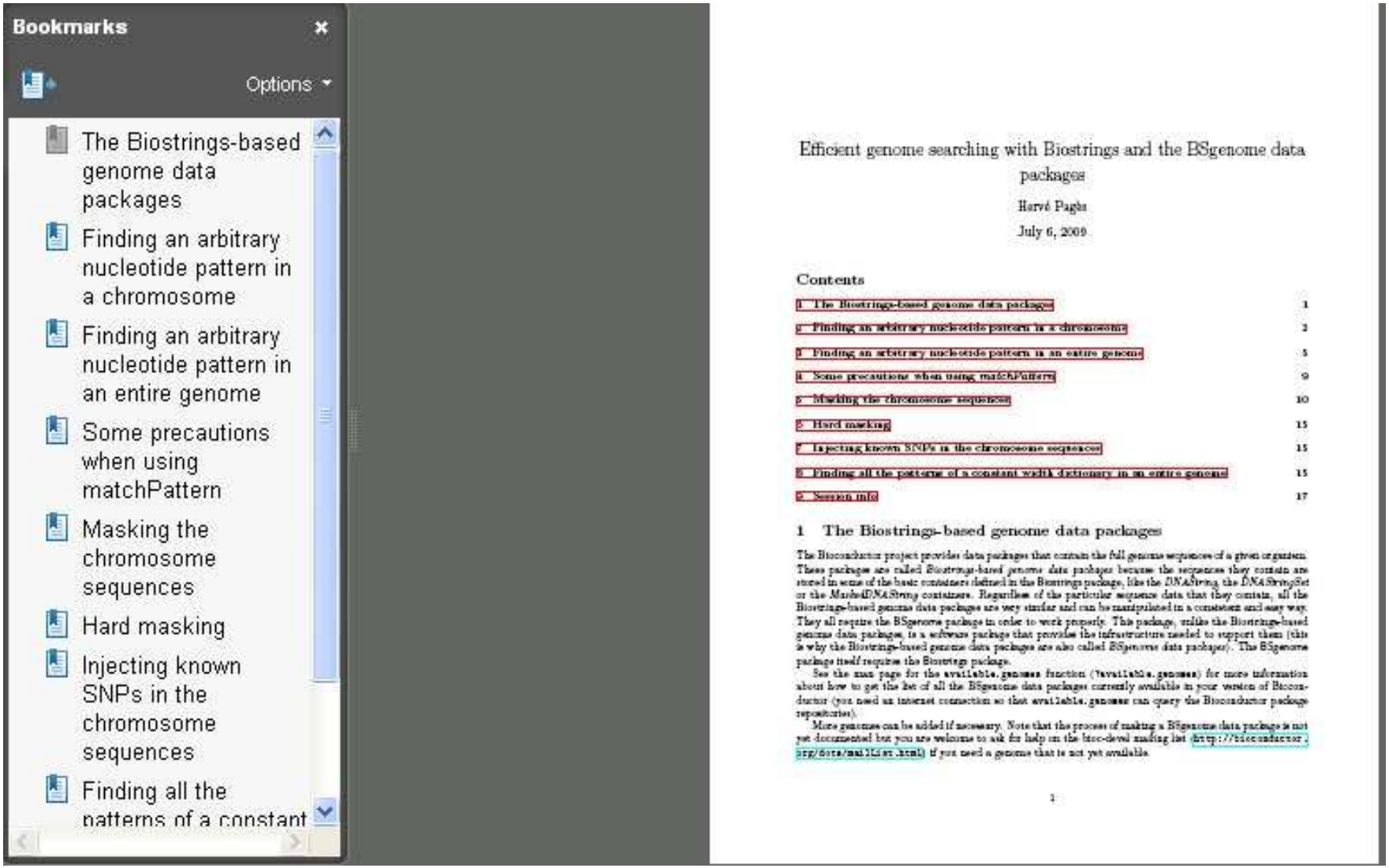
```
> data(BLOSUM50)
> BLOSUM50[1:4, 1:4]

   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/)



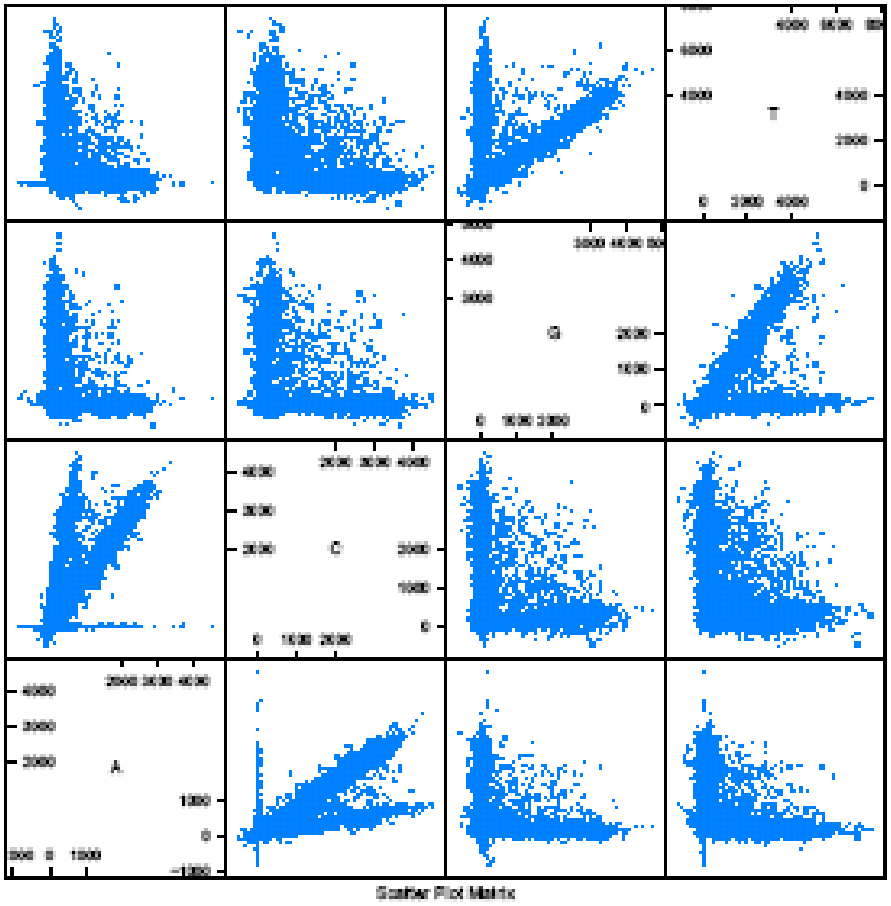
(http://www.bioconductor.org/packages/bioc/vignettes/BSgenome/inst/doc/GenomeSearching.pdf)

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



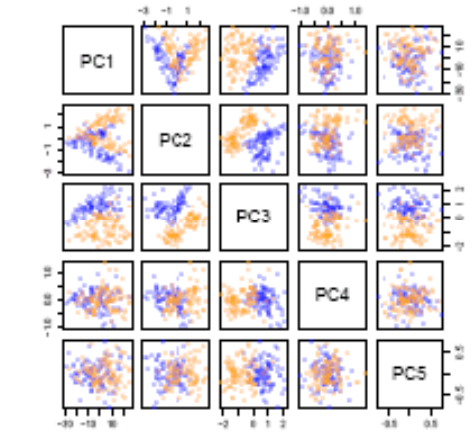
(http://www.bioconductor.org/workshops/2009/SSCMay09/ShortRead/IOQA.pdf)

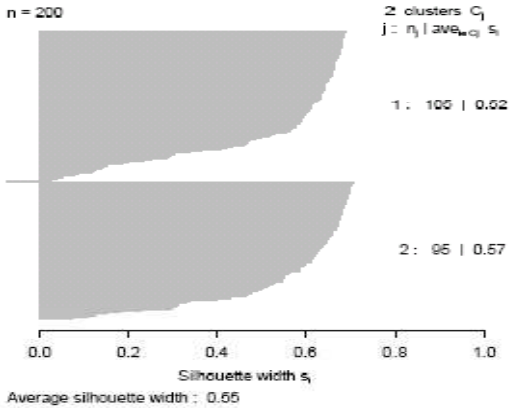
Machine learning with Bioconductor

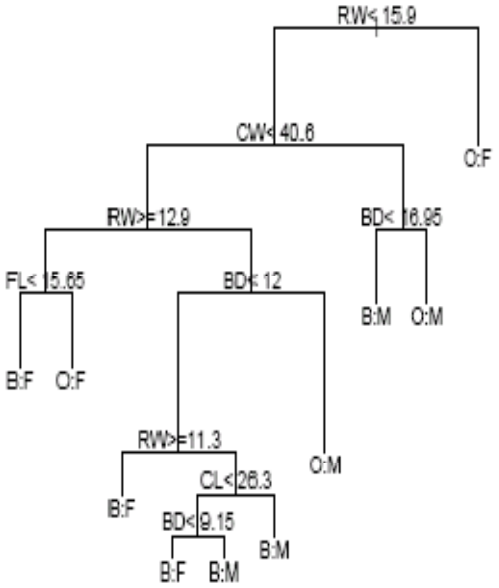
- The facilities of the MLInterfaces package are numerous.
- MLInterfaces facilitates answering questions like:
  - Given an ExpressionSet, how can we reason about clustering and opportunities for dimensionality reduction using unsupervised learning techniques?
  - For an ExpressionSet with labeled samples, 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 unform 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
adSplit	Claudio Lottaz	Annotation-Driven Clustering
clusterStab	James W. MacDonald	Compute cluster stability scores for microarray data
CORREP	Dongxiao Zhu	Multivariate Correlation Estimator and Statistical Inference Procedures.
ctc	Antoine Lucas	Cluster and Tree Conversion.
flowClust	Raphael Gottardo	Clustering for Flow Cytometry
geneRecommender	Greg Hather	A gene recommender algorithm to identify genes coexpressed with a query set of genes
hopach	Katherine S. Pollard	Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH)
maanova	Hyuna Yang	Tools for analyzing Micro Array experiments
made4	Aedin Culhane	Multivariate analysis of microarray data using ADE4
maigesPack	Gustavo H. Esteves	Functions to handle cDNA microarray data, including several methods of data analysis
MantelCorr	Brian Steinmeyer	Compute Mantel Cluster Correlations
Mfuzz	Matthias Futschik	Soft clustering of time series gene expression data
MLInterfaces	V. Carey	Uniform interfaces to R machine learning procedures for data in Bioconductor containers
puma	Richard Pearson	Propagating Uncertainty in Microarray Analysis
SAGx	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

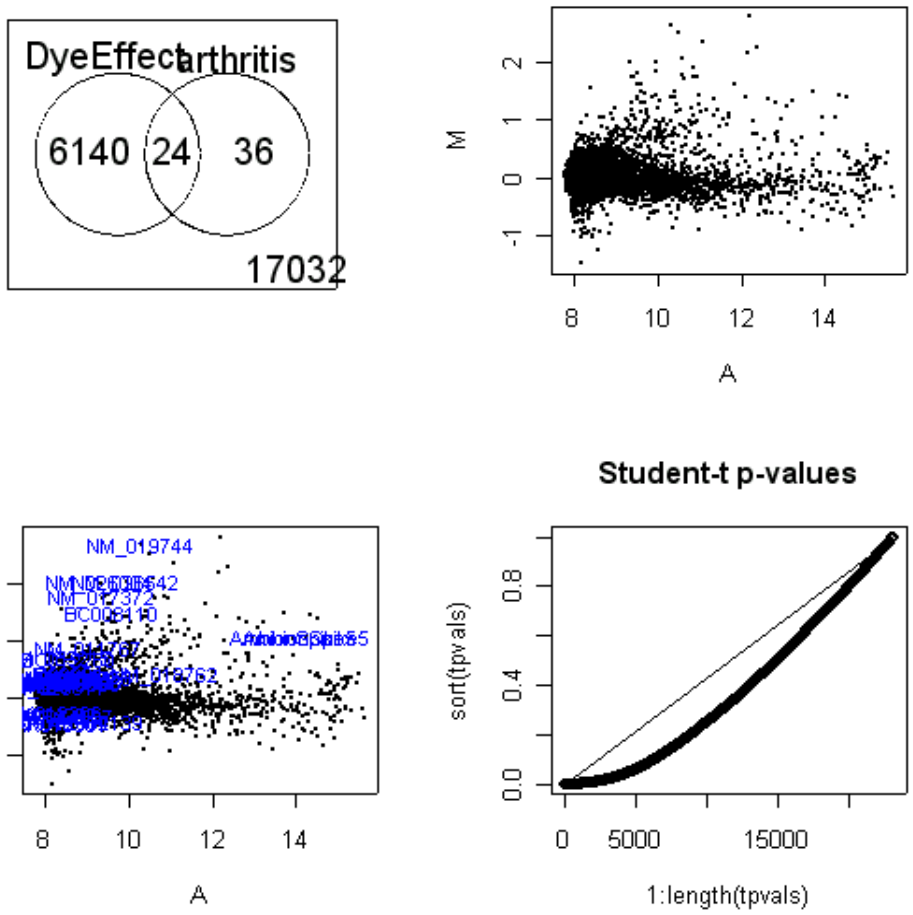
- One of the most common tasks when analyzing microarrays is to make comparisons between sample types, and the limma package in R is one of the more popular packages for this task.
- The limma package is quite powerful and allows users to make relatively complex comparisons.
- However, this power comes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

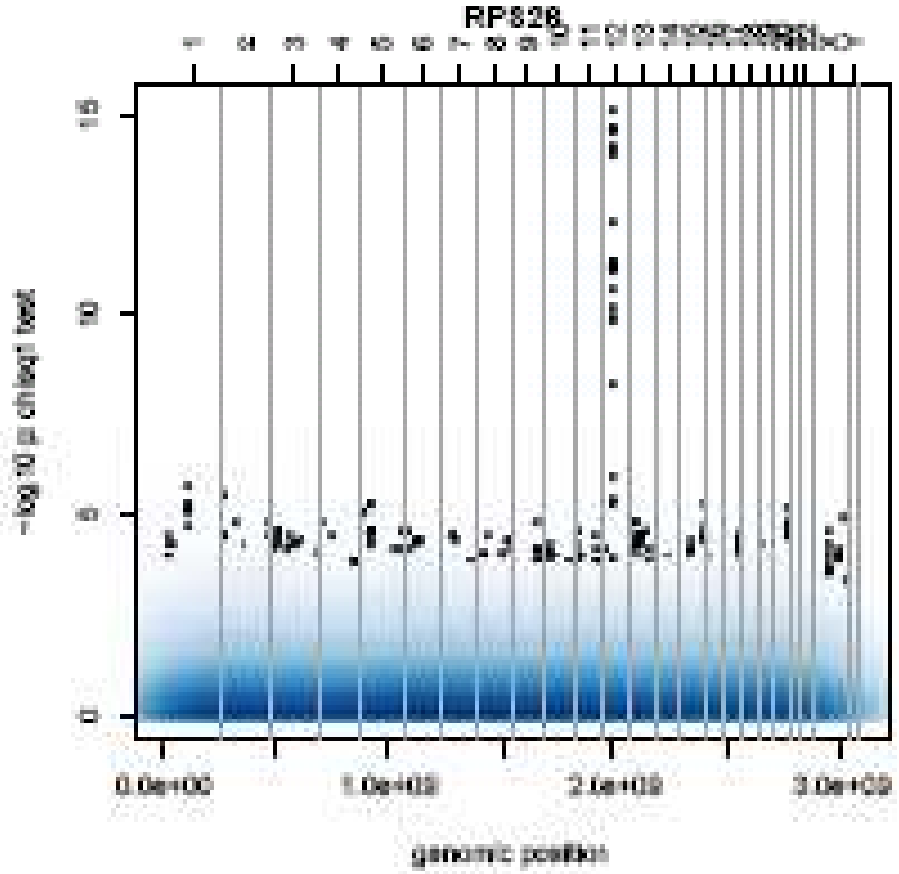
(https://secure.bioconductor.org/BioC2008/)

Limma



(Boer 2005)

GGtools



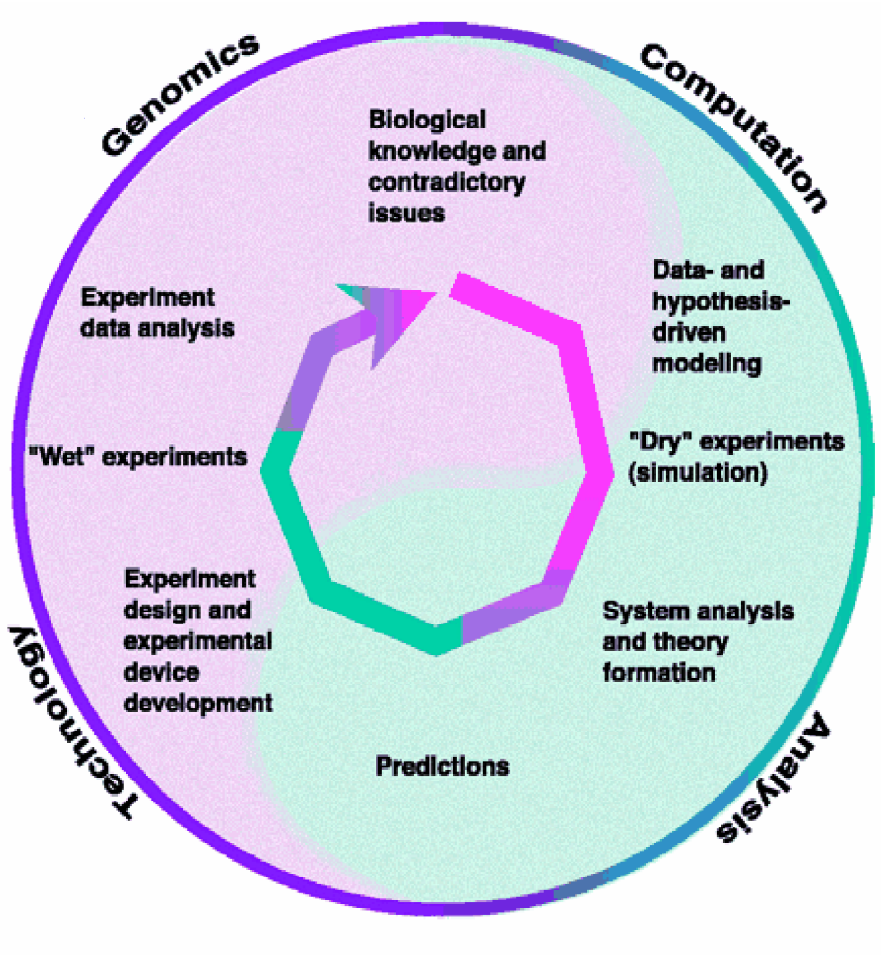
(http://www.bioconductor.org/packages/2.2/bioc/vignettes/GGtools/inst/doc/GGoverview2008.pdf)

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, mRAN, 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 DAN 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
- URL: http://www.ebi.ac.uk/2can/bioinformatics/

Background information / reading:

- 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)
- Gir Won Lee & Sangsoo Kim 2008. **Genome data mining for everyone** - http://bmbreports.org
- Elkin P (2003). **Primer on medical genomics. Part V: Bioinformatics**. Mayo Clin Proc, 78: 57-64

In-class discussion document

- "Dammit Jim, I'm a doctor, not a bioinformatician!"

  Academic Software, Productivity, and Reproducible Research

  by Christophe Lambert, CEO & President of Golden Helix [see course website]

Preparatory reading for next class:

- go over R Bioconductor installation guidelines [see Chapter 1, p73-76]