

Introduction to R Bioinformatics, R & Bioconductor: Overview

Alexander Ploner

Medical Epidemiology & Biostatistics
Karolinska Institutet

www.meb.ki.se/~aleplo/IntroR

2012-05-18

Bioinformatics I

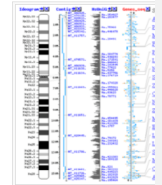
<https://en.wikipedia.org/wiki/Bioinformatics>

Bioinformatics (en.wikipedia.org/wiki/Bioinformatics) is the application of computer science and information technology to the field of biology and medicine. Bioinformatics deals with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing, information and computation theory, software engineering, data mining, image processing, modeling and simulation, signal processing, discrete mathematics, control and system theory, circuit theory, and statistics. Bioinformatics generates new knowledge as well as the computational tools to create that knowledge.

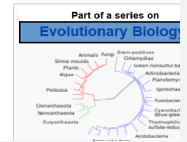
Commonly used software tools and technologies in this field include Java, XML, Perl, C, C++, Python, R, MySQL, SQL, CUDA, MATLAB, and Microsoft Excel.

Contents

- 1 Introduction
- 2 Major research areas
 - 2.1 Sequence analysis
 - 2.2 Genome annotation
 - 2.3 Computational evolutionary biology
 - 2.4 Literature analysis
 - 2.5 Analysis of gene expression
 - 2.6 Analysis of regulation
 - 2.7 Analysis of protein expression
 - 2.8 Analysis of mutations in cancer
 - 2.9 Comparative genomics
 - 2.10 Modeling biological systems
 - 2.11 High-throughput image analysis
 - 2.12 Structural Bioinformatic Approaches
 - 2.12.1 Prediction of protein structure
 - 2.12.2 Molecular Interaction
 - 2.12.2.1 Docking algorithms
- 3 Software and tools



Map of the human X chromosome (from the NCBI website). Assembly human genome is one of the great achievements of bioinformatics.



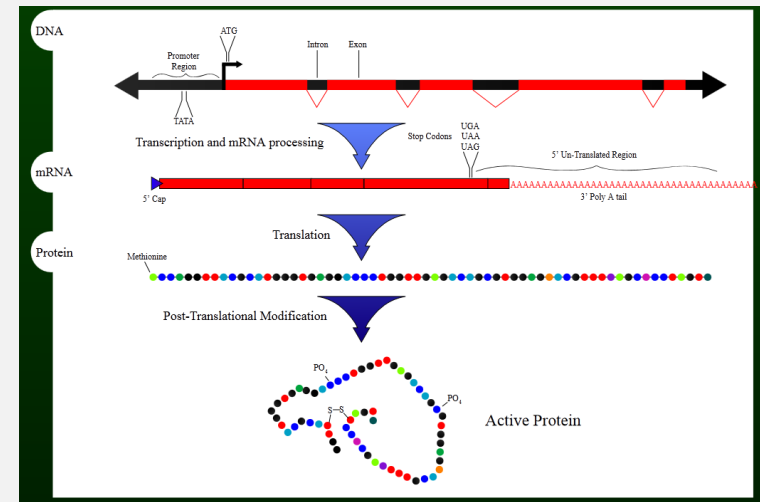
Bioinformatics II

More specifically

- ▶ Large data
- ▶ Molecular (sub-cellular)
- ▶ Technology-driven
- ▶ Web-enabled

The Central Dogma of Molecular Biology

I wish they had called it something else



https://en.wikipedia.org/wiki/Central_dogma

Bioinformatics and R

- ▶ Script language providing infrastructure (DB, connectivity, plotting)
- ▶ Data analysis: e.g. from corresponding task views at CRAN
 - ▶ Multivariate methods
 - ▶ Machine learning
 - ▶ Genetics
 - ▶ Medical image analysis

BioConductor

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

The screenshot shows the BioConductor website homepage. At the top, there is a search bar and a navigation menu with links for Home, Install, Help, Developers, and About. The main content area is divided into several sections. On the left, there is an 'About Bioconductor' section with a brief description of the project and a link to join the community. In the center, there is a 'Use Bioconductor for...' section with three sub-sections: 'Microarrays', 'Variants', and 'High Throughput Assays', each with a brief description of the tools available. On the right, there is a 'Sequence Data' section with a brief description of the tools available. At the bottom, there are links for 'Mailing Lists', 'Events', and 'News'. There are also several news items listed at the bottom, including 'CSAMA 2012 (Computational Statistics for Genome Biology)', 'BioC 2012', and 'Bioconductor 2.10 released'.

BioConductor: Structure

A set of related packages

Version 2.10 (April 2012)

- ▶ Software tools: 536 packages
 - ▶ Annotation data: 624 packages
 - ▶ Experimental data: 118 packages
-
- ▶ Core set of interacting packages
 - ▶ Contributed packages built on top/around/besides

Installation

From repository:

1. Download installer:

```
source("http://bioconductor.org/biocLite.R")
```
2. Download packages by name (automatic dependencies):

```
biocLite("limma")
```

Calling `biocLite()` installs three core infrastructure packages:

- ▶ Biobase
- ▶ IRanges
- ▶ AnnotationDbi

Documentation

<http://www.bioconductor.org/help/>

- ▶ (Books)
- ▶ Mailing lists (w. archives)
- ▶ Workflows
- ▶ biocViews (hierarchical tags)
- ▶ Vignettes (within packages)

A common workflow

Different focus from conventional statistical analysis

- ▶ Read
 - ▶ Pre-process
 - ▶ Analyze
 - ▶ Annotate
 - ▶ Report
- ... mostly in this order.

Outline for today

1. Gene expression microarrays
 - ▶ As example application
 - ▶ Demonstrate workflow
 - ▶ Highlight BioC generals (classes, annotation, repositories etc.)
 - ▶ Highlight BioC specifics in using R
2. RNA-seq
 - ▶ Of special interest for some
 - ▶ New and exciting
 - ▶ Pedestrian coverage