Introduction to R

Bioinformatics, R & Bioconductor: Overview

Alexander Ploner

Medical Epidemiology & Biostatistics
Karolinska Institutet

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

2012-05-18

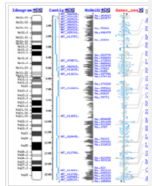
Bioinformatics I

`https://en.wikipedia.org/wiki/Bioinformatics`

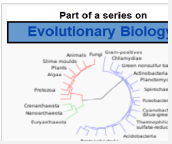
Bioinformatics /ˌbaɪ.oʊˌɪnfərˈmætɪks/ is the application of computer science and information technology to the field of biology and medicine. Bioinformatics deals with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing, information and computation 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



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



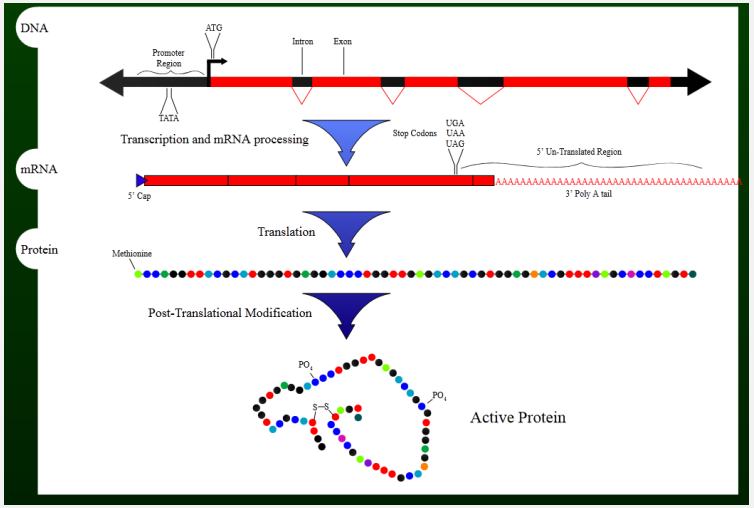
Bioinformatics II

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



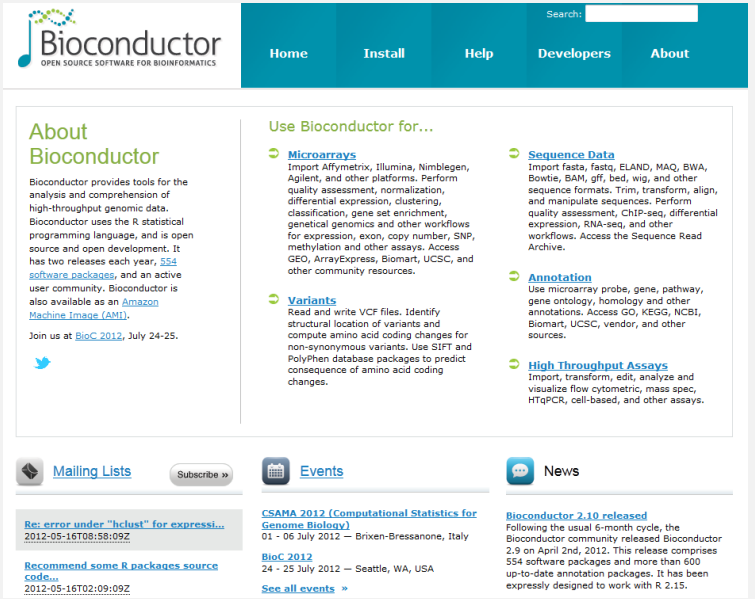
`https://en.wikipedia.org/wiki/Central_dogma`

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/`



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