Bioinformatics and computational biology

Wikipedia:

Bioinformatics and computational biology involve the use of techniques including applied mathematics, informatics, statistics, computer science, artificial intelligence, chemistry, and biochemistry to solve biological problems usually on the molecular level.

Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, and the modeling of evolution.
Wikipedia:

The terms bioinformatics and computational biology are often used interchangeably. However bioinformatics more properly refers to the creation and advancement of algorithms, computational and statistical techniques, and theory to solve formal and practical problems inspired from the management and analysis of biological data. Computational biology, on the other hand, refers to hypothesis-driven investigation of a specific biological problem using computers, carried out with experimental or simulated data, with the primary goal of discovery and the advancement of biological knowledge.

NIH definition of Bioinformatics and Computational Biology:

Bioinformatics and computational biology are rooted in life sciences as well as computer and information sciences and technologies. Both of these interdisciplinary approaches draw from specific disciplines such as mathematics, physics, computer science and engineering, biology, and behavioral science.

Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful. Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology. Although bioinformatics and computational biology are distinct, there is also significant overlap and activity at their interface.
NIH definition of Bioinformatics and Computational Biology:

The NIH Biomedical Information Science and Technology Initiative Consortium agreed on the following definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.

**Bioinformatics:** 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, archive, analyze, or visualize such data.

**Computational Biology:** The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

2D gel electrophoresis
2D gel electrophoresis

Ingo Ruczinski  
Biostatistics and Computational Biology

![2D gel electrophoresis diagram](image)

In this diagram, the reduction of concentration as compared to background is plotted for the 1st trimester and 3rd trimester. The data is divided into two groups: folate and placebo. The x-axis represents the trimester, and the y-axis shows the percentage reduction of concentration.
2D gel electrophoresis
Karyotypes

Trisomy
General Cytogenetics Information  http://members.aol.com/chrominfo/

FISH

Courtesy of the Pevsner Laboratory
DNA changes

Types of mutation
- Deletion
- Duplication
- Inversion

Insertion

Translocation

Single nucleotide polymorphisms

SNP
Deletion
Amplification

Uniparental isodisomy
A versus B plots

SNP_A–4251622

SNP_A–8348190

plates ordered by median date
Samples from Aravinda Chakravarti and Betty Doan
Prediction regions for copy number

De novo deletion
The statistical environment R

- R is an environment for data analysis and visualization.
- R is both open source and open development.
- You can look at the source code and propose changes.
- R is not in the public domain.
- You are given a license to run the software (currently GPL).
The R software

- R is mainly written in C.
- R is available for many platforms:
  - Unix of many flavors, including Linux, Solaris, FreeBSD.
  - Windows 95 and later.
  - MacOS X.
- Binaries and source code are available from www.r-project.org.
- R “talks” to data bases, programming languages, and other statistical packages.
- R should be source code compatible with most of the Splus code written.
The R package system

- Packages are self-contained units of code with documentation.
- The packages are simple to obtain and to understand, and can easily be updated.
- You can write your own packages!
- All functions must have examples to run.
- There are automatic testing features built in.

CRAN packages
Pros and cons

Advantages

- Free
- Available for all major platforms
- Comprehensive
- Powerful graphics
- Well-designed programming language
- Unlimited extensibility
- Widely used by statisticians
- Increasingly used for genomic analyses (Bioconductor)

Disadvantages

- No dedicated support
- Complex syntax
- Not point-and-click
- Some simple tasks are rather hard

Bioconductor
JHSPH Biostatistics classes

3 140.615 Biostatistics for Laboratory Scientists I
MWF 10:30 – 11.20 (Ingo Ruczinski)

140.644 Practical Machine Learning
MW 1:30 – 2.50 (Rafael Irizarry)

4 140.616 Biostatistics for Laboratory Scientists II
MWF 10:30 – 11.20 (Ingo Ruczinski)

140.688 Statistics for Genomics
MW 10:30 – 11.50 (Jeff Leek)
What is statistics?
- Data exploration and analysis.
- Quantification of evidence and uncertainty.
- Inductive inference with probability.

What is probability?
- A branch of mathematics concerning the study of random processes.

### Diagnostics

<table>
<thead>
<tr>
<th>DISEASE</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>-</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

**TEST**

- TP: True Positive
- FP: False Positive
- FN: False Negative
- TN: True Negative
Assume that some disease has a 0.1% prevalence in the population. Assume we have a test kit for that disease that works with 99% sensitivity and 99% specificity. What is the probability of a person having the disease, given the test result is positive, if we randomly select a subject from the general population?
### Sensitivity
\[ \frac{99}{99+1} = 99\% \]

### Specificity
\[ \frac{98901}{999+98901} = 99\% \]

### Positive Predictive Value
\[ \frac{99}{99+999} \approx 9\% \]

### Negative Predictive Value
\[ \frac{98901}{1+98901} > 99.9\% \]

### Accuracy
\[ \frac{99+98901}{100000} = 99\% \]
Goal: Determine, by fluorescence, the concentration of quinine in a sample of tonic water.

1. Obtain a stock solution with known concentration of quinine.
2. Create several dilutions of the stock.
3. Measure fluorescence intensity of each such standard.
4. Measure fluorescence intensity of the unknown.
5. Fit a line to the results for the standards.
6. Use line to estimate quinine concentration in the unknown.

Question: How precise is the resulting estimate?
Summarizing data

Ingo Ruczinski
Biostatistics and Computational Biology

Summarizing data
Summarizing data

http://biostat.jhsph.edu/~iruczins/

ingo@jhu.edu