Genomics and Bioinformatics: Week 1

17 September 2013

1 Introduction and definitions

The discovery of the structure of DNA (1953) established a link between *biochemistry*, a large part of which was devoted the characterization of proteins and enzymes, and *genetics*, the study of heredity, by showing that the DNA molecule was at the same time the carrier of hereditary characters and the code for proteins. The term *molecular biology* describes the subsequent new area of research centered on the biology of the DNA molecule and its relationship with protein functions. Bioinformatics and Genomics both belong to this field.

Definition 1. ***Bioinformatics*** *is the science and technology of biological information. It is a theoretical science that develops mathematical, statistical, and algorithmic methods and makes use of large-scale experimental data (in digital form), in an effort to discover functions and modalities of information processing in the cell.*

Definition 2. *(Structural)* ***Genomics*** *is the study of genomes as static physical objects: the nucleotide sequence and the order and location of features on them (genes, bands, polymorphisms, regulatory elements, etc.).*

Definition 3. ***Functional Genomics*** *is the extension of genomics to quantitative and dynamical phenomena associated with the processing of the genetic information through transcription and translation, and their interactions.*

Remarks:

1. The management of biological data across large computer networks is an important, albeit technical, aspect of bioinformatics and is often presented as bioinformatics itself.

2. The "omics" suffix conveys the notion of systematic and comprehensive study (Genomics: simultaneous study of all genes, all discrete features in a genome, Proteomics: study of all the proteins in a cell, Lipidomics, Metabolomics, etc.)

3. The fundamental dogma of molecular biology is central to bioinformatics: the nature of DNA, RNA and protein sequences, the kinetics and mechanics of transcription and translation.

4. The genome is a repository of two types of information: the evolutionary history of the species, and the biochemical blueprint of the organism.

Functional genomics introduces a stochastic and dynamic component in genomics, because messenger RNAs and proteins are unstable on time-scales of minutes to hours, and such changes are relevant for the biochemical description of the cell. By contrast, the genome is stable on evolutionary time-scales and is generally considered as a fixed structure, (nearly) identical for all individuals of the same species. Changes in the genome are only considered when comparing species (evolutionary or comparative genomics) and more specifically when assessing the effect of genetic difference between individuals of the same species (association studies, linkage analysis).

Structural genomics works in a world of static, discrete, enumerable objects (genes, sequence motifs, protein domains, conserved blocks), while functional genomics deals with continuous and time-dependent values (expression levels, protein concentrations, rates of transcription), which implies noise and uncertainty.

2 Genome sequencing

The characterization of the DNA naively induces a reductionist program: we sequence the entire genome and then identify the complete set of protein-coding genes, allowing us to simply "read" the biology of an organism (in particular its development plan) from that information.

A number of large genome sequencing initiatives were conducted (Yeast: 1996, Worm: 1998, Fly: 2000, Mouse: 2002, Human: 2003) and concluded that:

1. the number of genes in the genome is rather small,
2. this number is similar between highly divergent species,
3. genes themselves are highly conserved between species.

Together with the sequencing programs came the development of high-throughput technologies (PCR, microarrays, sequencing, mass spectrometry) which contribute to shifting our understanding of the genome function by monitoring interactions and dynamical effects on a large scale and in many experimental conditions. Nowadays, DNA can easily be

1. Read (sequencing)
2. Written (recombinant synthesis)
3. Copied (PCR)
4. Stored (plasmids, inserts, etc.)

which makes an "information science" thinking experimentally and technologically relevant.

3 Emergence of systems biology and importance of modeling

One implication of the genome sequencing programs is clearly that the genome sequence alone cannot easily explain the biology of the cell. The reductionist program therefore needs to be reformulated: "Systems biology" can refer to this new program which aims to give a comprehensive description of cellular biology in terms of dynamic interaction networks. The state of the cell is described at a molecular scale by a set of dynamic interactions between transcription and translation events which modulate each others quantitatively.

Large scale quantitative data produced by current high-throughput molecular technologies often require to build models. These are defined as deemed appropriate for each dataset and serve to discriminate between alternative hypotheses given the available data. Models can be arbitrarily complicated but the parameters that underlie the switch between alternative observable behaviours must be simple.

There are generally two phases in genome-wide studies:

1. Discovery (data-driven):
   (a) Process data to make it accessible to a human interpretation (via computer visualizations and statistical analyses)
   (b) Summarize data and check for expected behaviour and known particular cases ("positive controls")
   (c) Display global properties of the data ("trends" valid for most features)
   (d) Identify systematic biases and design a procedure to eliminate them
   (e) Compare between different conditions and build a metric that shows (or fails to show) a significant difference
2. Modeling (hypothesis-driven):
   (a) Formulate precise hypotheses that would explain the observations
   (b) Build an algorithmic model which can generate data quantitatively comparable to the observations, and where some hypotheses can be modified
   (c) Perform computer "experiments": try different constructions until one shows a testable effect

4 Objectives of the course

1. Become familiar with the different types of data generated by functional genomics, in particular:
   (a) sequence data
   (b) models for genes and transcripts
   (c) evolutionary trees and models of evolution
   (d) quantitative expression and regulation data
2. Learn the standard algorithms for sequence analysis:
   (a) global and local sequence alignments
   (b) multiple sequence alignments
   (c) Hidden Markov Models
3. Learn to use and make data exploration tools
   (a) find trends and correlations in quantitative data
   (b) assess their significance
   (c) use basic classification algorithms
4. Analyze and infer gene regulation patterns from data
   (a) build simple models, simulate them and compare them to data
   (b) acquire a quantitative understanding of transcriptional mechanisms

5 Course plan and requirements

We start from the genome sequencing, which provides the nucleotide sequence in a form that can be stored and processed on a computer. We then progressively add layers of biological annotation on this sequence, first by identification of specific sequence features in the genome, then by comparing between species and inferring evolutionary constraints. Then we use additional quantitative data to evaluate the functions of these genomic features and to infer interaction networks between them.

To follow this course you must be comfortable with programming a computer. I recommend using R and/or Python languages, but you are free to use any your favorite programming languages.

You must also be prepared to use mathematical and statistical tools, and you must have prior knowledge of molecular biology and of genetics.