Seminar 5

Groep 2

13 maart 2013

What is Bioinformatics?

Bioinformatics is conceptualising biology in terms of molecules and applying **informatics techniques** to **understand** and **organise** the **information** associated with these molecules, on a **large scale**. There are **three** aims of bioinformatics: First, bioinformatics organises **data** in a way that allows researchers to access existing information and to submit new entries as they are produced. The second aim is to develop **tools** and **resources** that aid in the analysis of said data. The third aim is to use these tools to **analyse** the data and **interpret** the results in a biologically meaningful manner.

Much of the data found can be grouped together based on biologically meaningful similarities. For example, distinct proteins frequently have comparable **sequences**. There are common terms to describe the relationship between pairs of proteins or the genes from which they are derived: **analogous** proteins have related folds, but unrelated sequences, while **homologous** proteins are both sequentially and structurally similar. Homologues can be distinguished between **orthologues**, proteins in different species that have evolved from a common ancestral gene, and **paralogues**, proteins that are related by gene duplication within a genome.

From these observations, a finite '**parts list**' can be made for different organisms: an inventory of proteins contained within an organism, arranged according to different properties. With this list, categorising the proteins, for example by folds, results in a simplification of the contents of a genome, useful for future genomic analyses.

Protein sequence databases are categorised as **primary, composite** or **secondary**. Primary databases function as a repository for the raw data. Composite databases compile and filter sequence data from different primary database to produce combined non-redundant sets. Secondary databases contain information derived from protein sequences and help the user determine whether a new sequence belongs to a known protein family.

A source of **genomic-scale data** has been from **expression experiments**, which quantify the expression levels of individual genes. These experiments measure the amount of mRNA or protein products that are produced by the cell.

The most profitable research in bioinformatics often results from integrating multiple sources of data. However, it is not always easy to access and cross-reference these sources of data because of differences in naming and file formats.

In the end, two principal approaches form the basis of all studies in bioinformatics **(the three aims)**: **comparing** and **grouping** the data according to **biologically meaningful similarities** and **analysing** one type of data to infer and understand the observations for another type of data, so we can **understand** and **organise** the **information** associated with biological molecules on a **large scale**.

Bioinformatics Challenges for Personalized Medicine

Single Nucleotide Polymorphisms (SNPs) are now recognized as the main cause of human genetic variability. By combining these genetic associations with phenotypes and drug response, **personalized medicine** will tailor treatments to the patients' specific genotype. In the coming years, the bioinformatics world will be inundated with **individual genomic data**. This flood of data introduces significant challenges that the bioinformatics community needs to address and which fall in the following four main areas.

1: Processing large-scale robust genomic data

Sequencing technologies are becoming affordable and are replacing the microarray based genotyping methods. The error rate from these technologies is a source of significant challenges in applications, **including discovering novel variants**. A remaining challenge for short read assemblers is **reference sequence bias**: reads that more closely resemble the reference sequence are more likely to successfully map as compared with reads that contain valid mismatches. When the **diploid sequence** is known, reference sequence bias can be avoided by mapping the reads to both strands. Another challenge is **developing new methods** for novel SNP discovery. Finally, there is a pressing need to **improve quality control** metrics.

2: Interpreting the functional effect and the impact of genomic variation

After genomic data has been processed, the **functional effect and the impact** of the genetic variations

must be analyzed. In the last few years, several **computational methods** have been developed to predict deleterious missense SNPs. **Prediction methods** do not provide any information about the pathophysiology of the diseases and so experimental tests are required to validate genetic predictions. The methods for the **analysis** of SNPs are mainly limited to the prediction of the impact of missense SNPs.

3: Integrating systems data to relate complex genetic interactions with phenotypes

Given the **complex phenotypes** involved in personalized medicine, the simple "one-SNP, one-phenotype" approach taken by most studies is insufficient. Given the size of the genomic data sets, **dimensionality reduction methods** will be essential to make complexity algorithms tractable. **Systems biology** and **network approaches** address to the problem of complexity by **integrating molecular data** at multiple levels of biology. Combining disparate data sources can result in novel associations and provide insight into gene-gene and gene-environment interactions.

4: Translating these discoveries into medical practice

The ultimate challenge for this research is to apply the results for improved patient care. Most pharmaceutical development addresses medical problems with a "one drug fits all" approach. **Pharmacogenomics** connects genotype to patient specific treatment and has already been successful in improving drug prescription and dosing. Bioinformatics also translates discoveries to the clinic by **disseminating discoveries** through curated, searchable databases. Ultimately, bioinformatics needs to develop methods that **interrogate the genome** in the clinic and allow physicians to use personalized medicine in their daily practice.

Trends in computational biology

Next-generation sequence analysis

Computational biology has risen in prominence in recent years largely because of the increase in the data-generation capacity of high-throughput technology.

The advance: Two methods for de novo transcriptome assembly of short reads were published.

What it means: Advances in transcript assembly from RNA-Seq data should allow alternative splicing to be studied genome-wide across many biological conditions. In addition, RNA-Seq is now poised as a tool for discovering new RNAs that we may not have even known were transcribed from a genome.

Discovery from data repositories

Electronic medical records are becoming a reality, promising lower health-care costs, improved patient treatments and, perhaps, scientific advances.

The advance: A first study that demonstrates the feasibility of associating genetic modifications with data on phenotypic traits mined from electronic medical records.

What it means: The case of electronic medical records illustrates the potential value locked within unique biomedical databases and the challenges of realizing that value.

Learning to see

In biological research, one advantage of computational analysis is automation and fidelity.

The advance: Machine-learning algorithms that could accurately classify whether a pattern of fluorescent staining represents localization to one subcellular organelle or to a mixture of locations.

What it means: This represents the first step toward a new way of thinking about interpreting images that is generative rather than descriptive. Whereas descriptive approach may take an image of a cell and tell you that the protein is in the nucleus, a generative approach builds a model for other images.