Bioinformatics
Introduction to genomics and proteomics II

ulf.schmitz@informatik.uni-rostock.de

Bioinformatics and Systems Biology Group
www.sbi.informatik.uni-rostock.de
Outline

1. Proteomics
   • Motivation
   • Post-Translational Modifications
   • Key technologies
   • Data explosion

2. Maps of hereditary information

3. Single nucleotide polymorphisms
**Proteomics**

**Proteomics:**
- is the large-scale study of *proteins*, particularly their structures and functions
- This term was coined to make an analogy with *genomics*, and is often viewed as the "next step",
- but *proteomics* is much more complicated than *genomics*.
- Most importantly, while the genome is a rather constant entity, the proteome is constantly changing through its biochemical *interactions* with the genome.
- One organism will have radically different protein expression in different parts of its body and in different stages of its life cycle.

**Proteome**:

The entirety of *proteins* in existence in an organism are referred to as the *proteome*. 
Proteomics

*If the genome is a list of the instruments in an orchestra, the proteome is the orchestra playing a symphony.*

R. Simpson
Proteomics

• Describing all 3D structures of proteins in the cell is called **Structural Genomics**

• Finding out what these proteins do is called **Functional Genomics**
Motivation:

• What kind of data would we like to measure?

• What mature experimental techniques exist to determine them?

• The basic goal is a spatio-temporal description of the deployment of proteins in the organism.
Proteomics

Things to consider:

• the rates of synthesis of different proteins vary among different tissues and different cell types and states of activity

• methods are available for efficient analysis of transcription patterns of multiple genes

• because proteins ‘turn over’ at different rates, it is also necessary to measure proteins directly

• the distribution of expressed protein levels is a kinetic balance between rates of protein synthesis and degradation
Why do Proteomics?

- are there differences between amino acid sequences determined directly from proteins and those determined by translation from DNA?
  - pattern recognition programs addressing this questions have following errors:
    - a genuine protein sequence may be missed entirely
    - an incomplete protein may be reported
    - a gene may be incorrectly spliced
    - genes for different proteins may overlap
    - genes may be assembled from exons in different ways in different tissues
  - often, molecules must be modified to make a mature protein that differs significantly from the one suggested by translation
    - in many cases the missing post-translational modifications are quite important and have functional significance
    - post-transitional modifications include addition of ligands, glycosylation, methylation, excision of peptides, etc.
  - in some cases mRNA is edited before translation, creating changes in the amino acid sequence that are not inferrable from the genes
- a protein inferred from a genome sequence is a hypothetical object until an experiment verifies its existence
Post-translational modification

- a protein is a polypeptide chain composed of 20 possible amino acids

- there are far fewer genes that code for proteins in the human genome than there are proteins in the human proteome (~33,000 genes vs ~200,000 proteins).

- each gene encodes as many as six to eight different proteins
  - due to post-translational modifications such as phosphorylation, glycosylation or cleavage (Spaltung)

- posttranslational modification extends the range of possible functions a protein can have
  - changes may alter the hydrophobicity of a protein and thus determine if the modified protein is cytosolic or membrane-bound
  - modifications like phosphorylation are part of common mechanisms for controlling the behavior of a protein, for instance, activating or inactivating an enzyme.
Post-translational modification

**Phosphorylation**

- phosphorylation is the addition of a phosphate (PO$_4$) group to a protein or a small molecule (usual to serine, tyrosine, threonine or histidine)
- In eukaryotes, protein phosphorylation is probably the most important regulatory event
- Many enzymes and receptors are switched "on" or "off" by phosphorylation and dephosphorylation
- Phosphorylation is catalyzed by various specific protein kinases, whereas phosphatases dephosphorylate.

**Acetylation**

- Is the addition of an acetyl group, usually at the N-terminus of the protein

**Farnesylation**

- farnesylation, the addition of a farnesyl group

**Glycosylation**

- the addition of a glycosyl group to either asparagine, hydroxylysine, serine, or threonine, resulting in a glycoprotein
Proteomics

**proteomics research**
- kind & quantity of proteins
- post-translational modifications
- protein-protein interactions
- cell

**protein synthesis**
- nucleus
- dna
- rna
- ribosome
- peptide chain
- protein
Key technologies for proteomics

1. **1-D electrophoresis** and **2-D electrophoresis**
   - are for the separation and visualization of proteins.

2. **mass spectrometry**, **x-ray crystallography**, and **NMR** (Nuclear magnetic resonance)
   - are used to identify and characterize proteins

3. **chromatography** techniques especially **affinity chromatography**
   - are used to characterize protein-protein interactions.

4. **Protein expression systems** like the **yeast two-hybrid** and **FRET** (fluorescence resonance energy transfer)
   - can also be used to characterize protein-protein interactions.
Key technologies for proteomics

High-resolution two-dimensional polyacrylamide gel electrophoresis (2D PAGE) shows the pattern of protein content in a sample.

Reference map of lymphoblastoid cell line PRI, soluble proteins.
- 110 µg of proteins loaded
- Strip 17cm pH gradient 4-7, SDS PAGE gels 20 x 25 cm, 8-18.5% T.
- Staining by silver nitrate method (Rabilloud et al.)
- Identification by mass spectrometry. The pink labels on the spots indicate the ID in Swiss-prot database

browse the SWISS-2DPAGE database for more 2d PAGE images
Proteomics

X-ray crystallography is a means to determine the detailed molecular structure of a protein, nucleic acid or small molecule.

With a crystal structure we can explain the mechanism of an enzyme, the binding of an inhibitor, the packing of protein domains, the tertiary structure of a nucleic acid molecule etc..

Typically, a sample is purified to homogeneity, crystallized, subjected to an X-ray beam and diffraction data are collected.
High-throughput Biological Data

• Enormous amounts of biological data are being generated by high-throughput capabilities; even more are coming
  – genomic sequences
  – gene expression data (*microarrays*)
  – mass spec. data
  – protein-protein interaction (*chromatography*)
  – protein structures (*x-ray crystallography*)
  – ......
Protein structural data explosion

Protein Data Bank (PDB): 33,367 Structures (1 November 2005)
28,522 x-ray crystallography, 4,845 NMR
Maps of hereditary information

Following maps are used to find out how hereditary information is stored, passed on, and implemented.

1. **Linkage maps of**
   - genes
   - mini- / microsatellites

2. **Banding patterns of chromosomes**
   - physical objects with visible landmarks called banding patterns

3. **DNA sequences**
   - Contig maps (*contiguous clone maps*)
   - Sequence tagged site (STS)
   - SNPs (Single nucleotide polymorphisms)
Eucalyptus globulus Linkage Map

Number and type of loci
169 cDNAs and gDNAs
40 microsatellites
31 expressed sequence tags
14 candidate genes
5 isozymes
Maps of hereditary information

**Variable number tandem repeats (VNTRs, also minisatellites)**

- regions, 8-80bp long, repeated a variable number of times
- the distribution and the size of repeats is the marker
- *inheritance* of VNTRs can be followed in a family and mapped to a pathological phenotype
- first genetic data used for personal identification
  - Genetic fingerprints; in paternity and in criminal cases

**Short tandem repeat polymorphism (STRPs, also microsatellites)**

- Regions of 2-7bp, repeated many times
  - Usually 10-30 consecutive copies
Microsatellite or Short Tandem Repeat (STR)

- Dispersed throughout genome
- Core repeat length: 2 – 7 bp
- Up to ~40 repeats per locus
- Detected by PCR amplification

Minisatellite or Variable Number of Tandem Repeats (VNTR)

- Commonly subtelomeric
- Core repeat length: 8 – 80 bp
- Variable total length of ~0.5 to 30 kb
- Detected by Southern hybridization

3bp

CGTCGTCGTCGTCGTCGTCGTCGTCGTCG
GCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGC

DNA gel electrophoresis gel showing fragments of DNA.
Maps of hereditary information

Banding patterns of chromosomes
Maps of hereditary information

Banding patterns of chromosomes

Schematic '1000-band' human karyotype & nomenclature, as detected with G-banding. Different shades represent varying intensities of G-bands, according to the key. The cross-hatched areas on chromosomes 1, 9, 16, and Y represent heterochromatic regions. (adapted from Griffiths et al. 1996)
Maps of hereditary information

Contig map (also contiguous clone map)

- Series of overlapping DNA clones of known order along a chromosome from an organism of interest, stored in yeast or bacterial cells as YACs (Yeast Artificial Chromosomes) or BACs (Bacterial Artificial Chromosomes)
- A contig map produces a fine mapping (high resolution) of a genome
- YAC can contain up to $10^6$ bp, a BAC about 250,000 bp

Sequence tagged site (STS)

- Short, sequenced region of DNA, 200-600 bp long, that appears in a unique location in the genome
- One type arises from an EST (expressed sequence tag), a piece of cDNA
Maps of hereditary information

Imagine we know that a disease results from a specific defective protein:

1. if we know the protein involved, we can pursue rational approaches to therapy
2. if we know the gene involved, we can devise tests to identify sufferers or carriers
3. whereas the knowledge of the chromosomal location of the gene is unnecessary in many cases for either therapy or detection;
   • it is required only for identifying the gene, providing a bridge between the patterns of inheritance and the DNA sequence
Single nucleotide polymorphisms (SNPs)

- **SNP** (pronounced ‘snip’) is a genetic variation between individuals
- single base pairs that can be substituted, deleted or inserted
- **SNPs** are distributed throughout the genome
  - average every 2000bp
- provide markers for mapping genes
- not all **SNPs** are linked to diseases

Figure 1 The most common sources of variation between humans are single nucleotide polymorphisms (SNPs) — single base differences between genome sequences. Fragments of two sequences, with eight SNPs, are shown.
Single nucleotide polymorphisms (SNPs)

• *nonsense mutations*:  
  – codes for a stop, which can truncate the protein

• *missense mutations*:  
  – codes for a different amino acid

• *silent mutations*:  
  – codes for the same amino acid, so has no effect
Outlook – coming lecture

• **Bioinformatics Information Resources And Networks**
  – EMBnet – European Molecular Biology Network
    • DBs and Tools
  – NCBI – National Center For Biotechnology Information
    • DBs and Tools
  – Nucleic Acid Sequence Databases
  – Protein Information Resources
  – Metabolic Databases
  – Mapping Databases
  – Databases concerning Mutations
  – Literature Databases
Thanks for your attention!