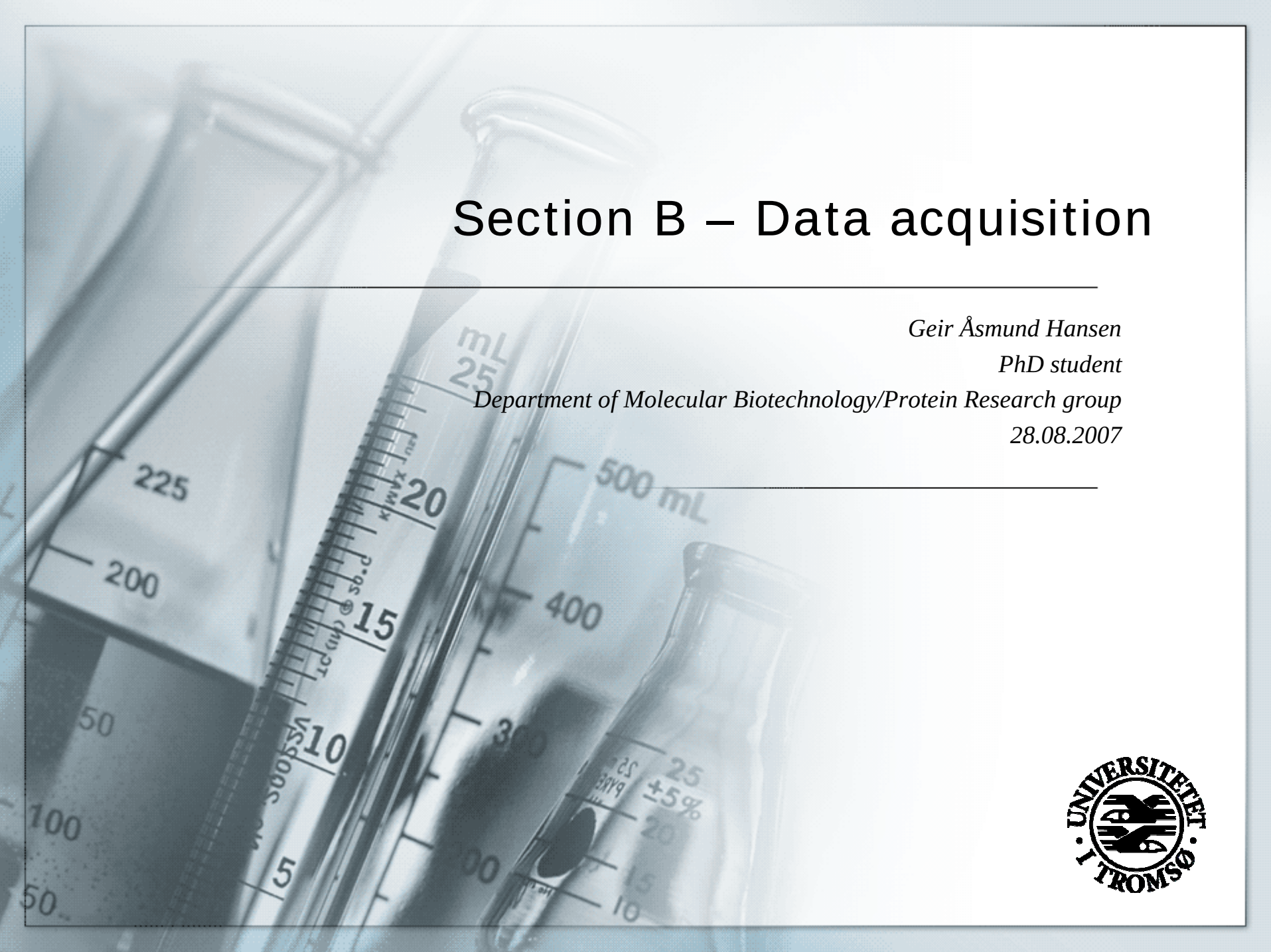
Section B – Data acquisition

Geir Åsmund Hansen
PhD student
Department of Molecular Biotechnology/Protein Research group
28.08.2007



Outline of lecture

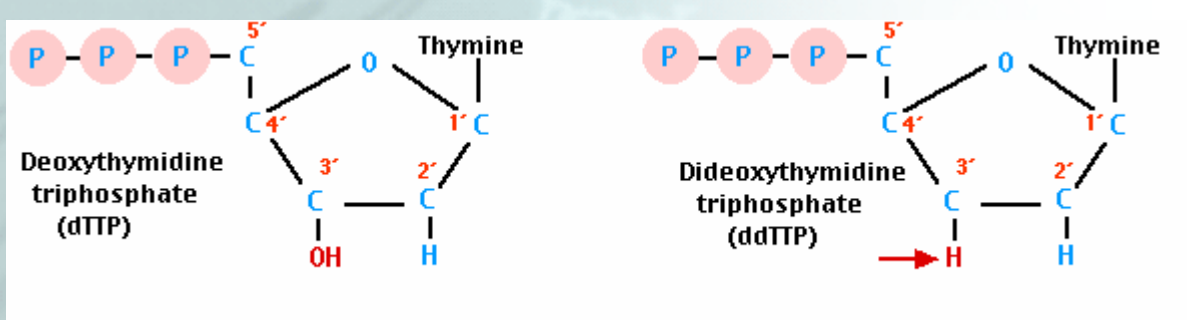
- B1: Sequencing DNA, RNA and proteins
- B2: Determination of protein structure
- B3: Gene and protein expression data
- B4: Protein interaction data

Section B1: Sequencing DNA, RNA and proteins

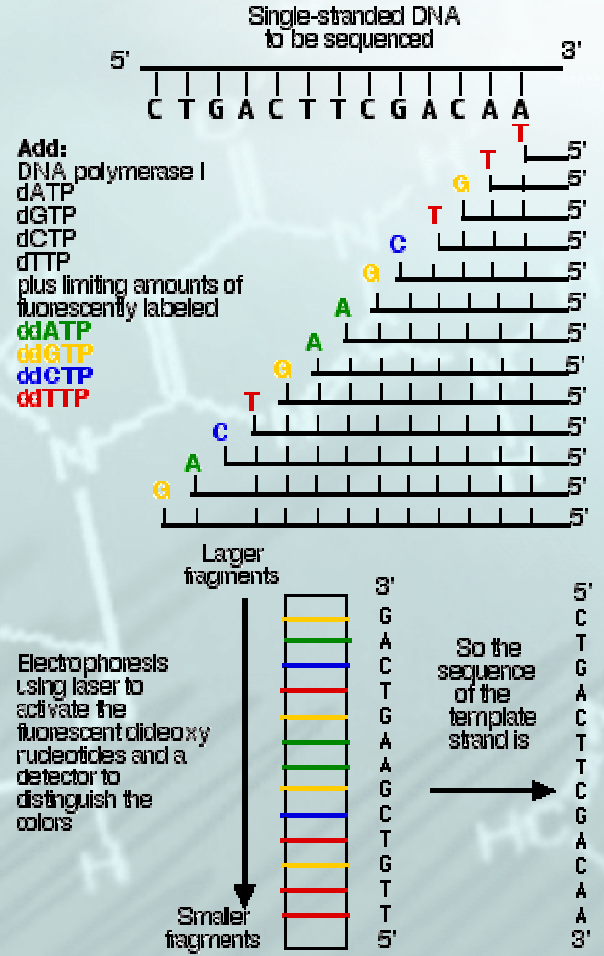
- Principles of DNA sequencing
- Types of DNA sequence
- Genome sequencing strategies
- Sequence quality control
- Single-pass sequencing
- RNA sequencing
- Protein sequencing

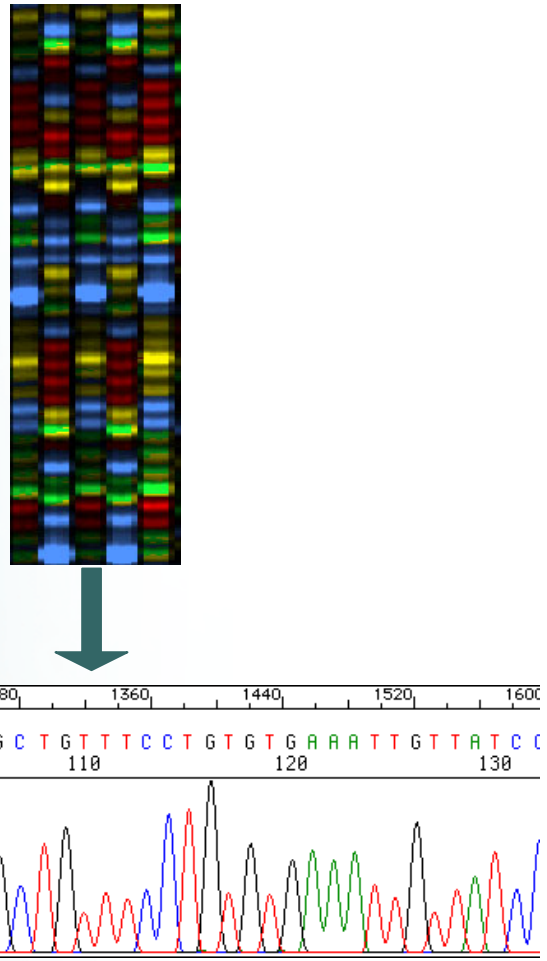
Section B1 – Principles of DNA sequencing

- Chain termination sequencing (dideoxy sequencing/Sanger method)
- Needed for the reaction: Single-stranded DNA template, primer, deoxyribonucleoside triphosphates (dNTP), DNA polymerase, limiting amounts of dideoxyribonucleoside triphosphates (ddNTP)
- Chain of nascent DNA terminated each time a ddNTP is incorporated
- Automated sequenced with fluorescent labelled ddNTPs
- DNA fragments separated by polyacrylamide gel electrophoresis
- Fluorescent tags read by a scanner near the bottom of the gel



Section B1 –
Principles of DNA sequencing





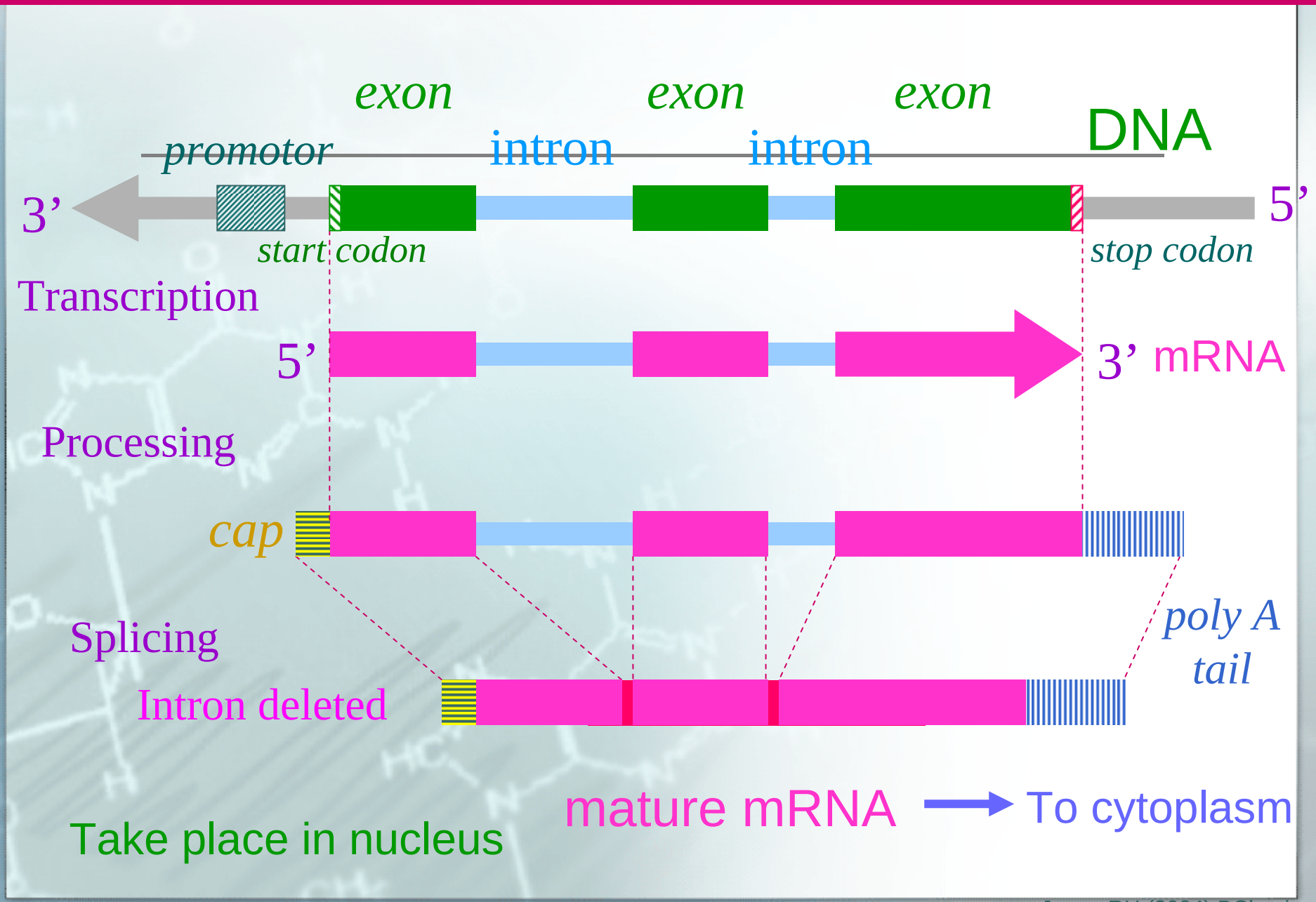
Section B1 –
Principles of DNA sequencing

- DNA sequencing animation

Section B1 – Types of DNA sequence

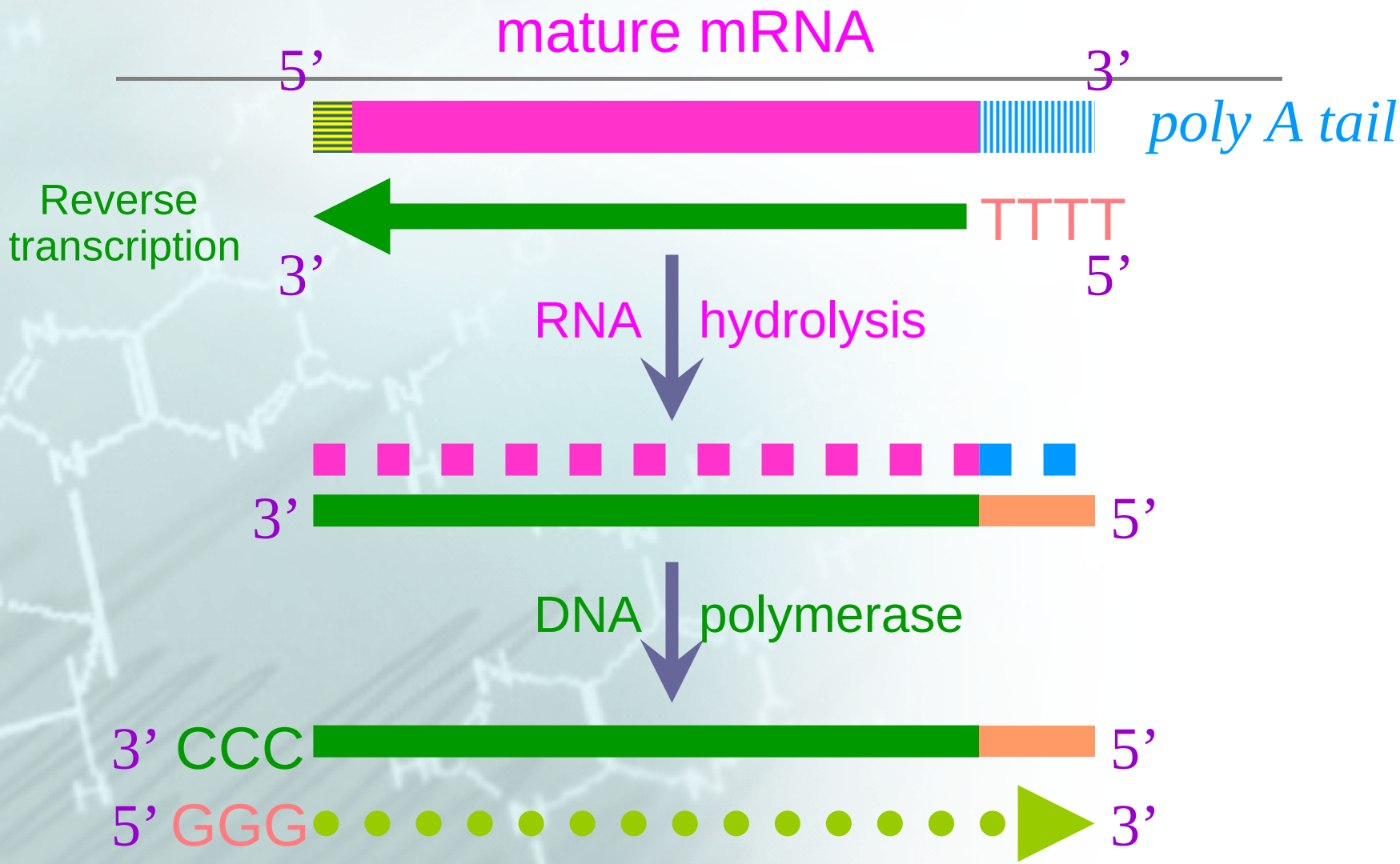
- Generally three types of DNA
  - Genomic DNA
    - Taken directly from the genome and contains genes in their normal state. In eukaryotes it include introns, regulatory elements and intergenic DNA.
  - Copy DNA (cDNA)
    - Reverse transcribed messenger RNA (mRNA). Gives direct access to genes.
  - Recombinant DNA
    - Sequences of vectors such as plasmids, modified viruses and other genetic elements used in the laboratory. Important to be removed from new genomic and cDNA sequences.

Intron and Exon in Eukaryotic Cells

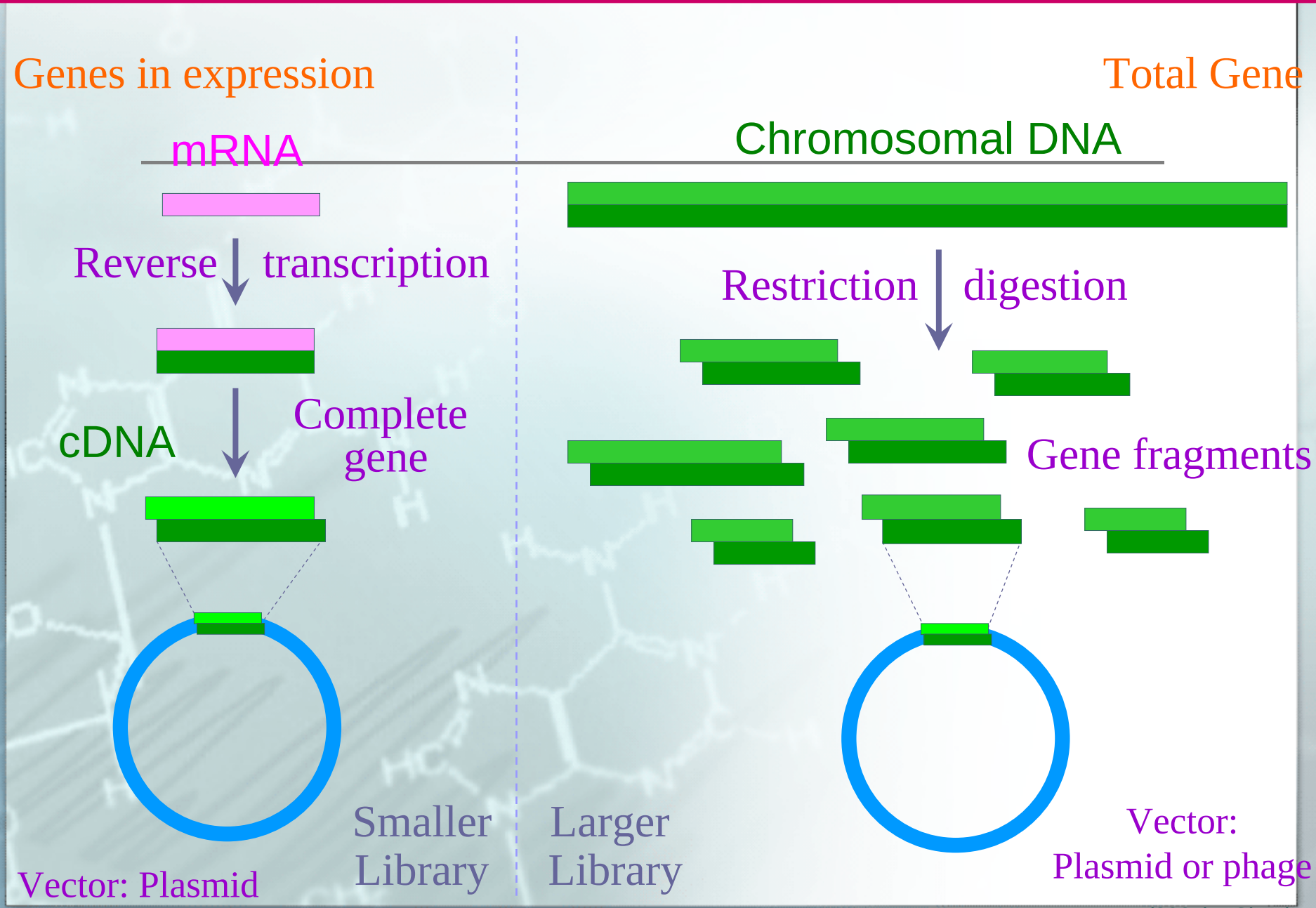


Juang RH (2004) BCbasics

cDNA Is Reverse Transcribed from mRNA



Two Libraries： cDNA Library vs Genomic Library



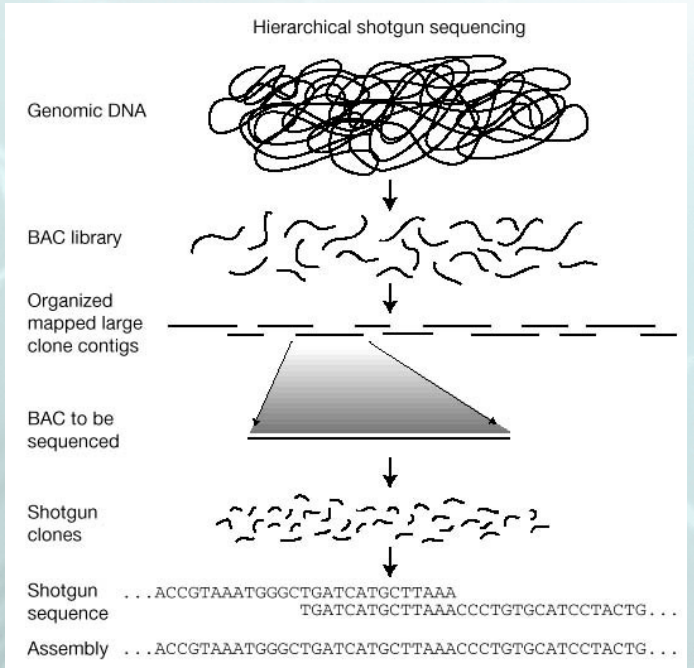
Juang RH (2004) BCbasics

Section B1 – Genome sequencing strategies

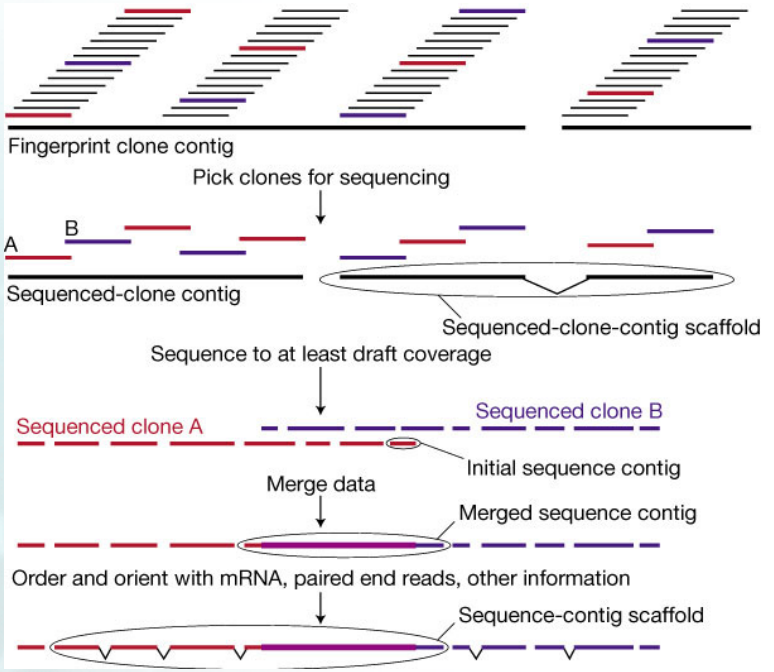
- In a single reaction only about 800 nucleotides of DNA can be sequenced
- To sequence larger DNA molecules they have to be broken up into smaller pieces and then assembled by:
  - Shotgun sequencing
    - DNA molecule randomly broken up and a large number of sequencing reactions are performed to cover the entire molecule
    - Sequences reassembled by computers searching for overlaps to get the sequences in the correct order
    - Generates data quickly, but time-consuming task to close the final gaps
  - Clone contig approach
    - DNA fragments are subcloned in a rational manner and systematically sequenced until the entire sequence is complete
    - Generates data slowly at the beginning, but finishes fast

Section B1 – Genome sequencing strategies

- Shotgun cloning



Clone contig approach



Section B1 – Sequence quality control

- When doing genome sequencing quality control is of great importance
- Sequence reaction is performed several times for each strand in each clone
- Pre-assembly process
  - Base calling and quality assessments (Phred)
  - Removal of vector sequences (vector_clip, CrossMatch)
  - Masking of repeats (RepeatMasker)
- Sequence reads assembled into contigs (contiguous sequence reads) and finally into one whole sequence (Phrap, Staden Gap4)
- Multiple sequencing reactions on both strands of a cloned DNA molecule are usually required to insure accuracy

Section B1 – Single-pass sequencing

- Single-pass sequencing is a quicker and cheaper sequencing method used where large amounts of data are generated
- Provides more data at the expense of accuracy
- Short (2-300 bp) and inaccurate sequences can be collected rapidly and inexpensively
- Used to identify genes in genomic DNA and to prepare large clone sets for DNA microarray
- Expressed sequence tags (ESTs) are made from cDNA clones, therefore containing only genes
  - Most ESTs from eukaryotic bacteria
    - Polyadenylate tail
- Genomic survey sequences (GSSs) are random, single-pass sequences of genomic data
  - Bacterial genomic DNA fragments very similar to cDNA

Section B1 – RNA sequencing

- Less straightforward than DNA sequencing because of the large number of minor nucleotides (chemically modified nucleotides)
- The RNA sequence can be deduced from either the DNA sequence of the corresponding gene or from a cDNA sequence (sequence of the reverse transcript)
- cDNA sequences are often more informative than gene sequences because of extensively processing during synthesis
  - Introns spliced out of the sequence
  - Extra nucleotides by RNA editing
- Direct RNA sequencing involves chemical characterization of modified nucleotides
  - Some need to be identified by mass spectrometry

Section B1 – Protein sequencing

- Protein sequences can be deduced from DNA sequences
  - Does not provide information on modified residues or other types of post-translational protein modifications (cleavage, disulfide bonds)
- Direct protein sequencing by Edman degradation
  - Labelling of terminal residue before it is removed and identified by chemical tests
- Mass spectrometry (MS)
  - Mass/charge ratio (m/e or m/z) of ions in vacuum accurately determined
  - Gives a spectrum that is specific to each protein
  - Large databases containing MS profiles from proteins

Section B2 – Determination of protein structure

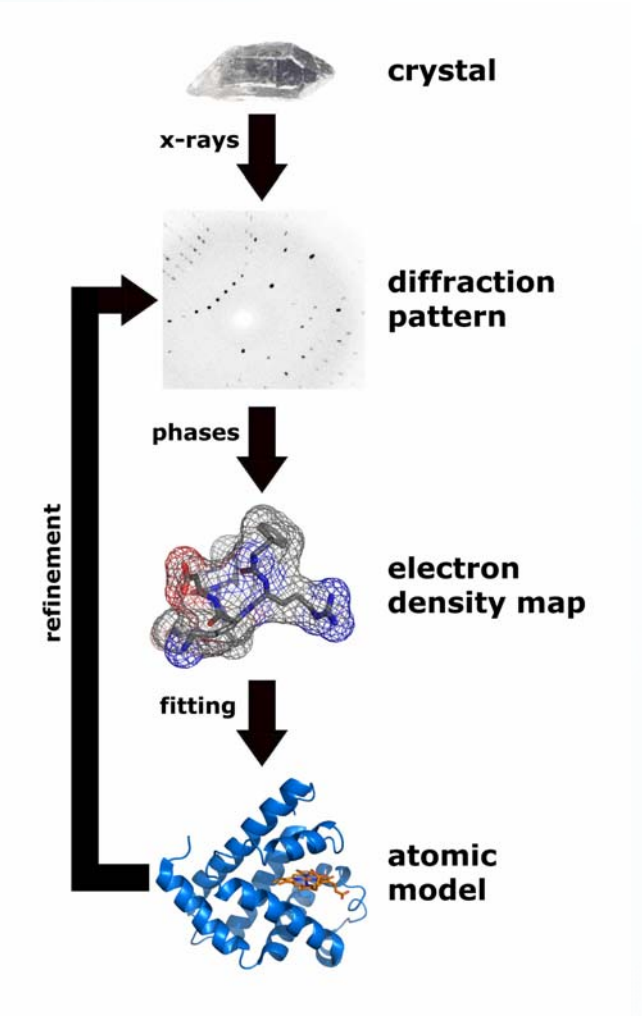
- X-ray crystallography
- NMR spectroscopy
- Other methods

Section B2 – X-ray crystallography

- Reconstruction of atomic positions based on diffraction patterns of X-rays through a protein crystal
- Positive and negative interference generating an ordered pattern of signals called reflections → Electron density maps
- Dependant of three variables
  - Wavelength of X-rays dependant of number of electrons – Known
  - Amplitude of scattering – Calculated from intensities of diffraction pattern
  - Phase of scattering – Cannot be calculated directly
    - Indirect strategies
      - Multiple isomorpheous replacement – Use of heavy atom derivatives
      - Multiwavelength anomalous diffraction – Synchrotron sources

Section B2 – X-ray crystallography

- X-ray crystallography

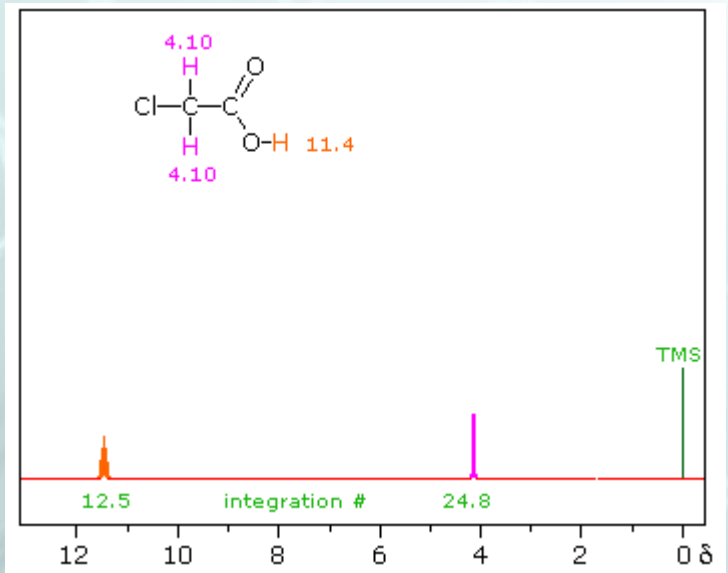


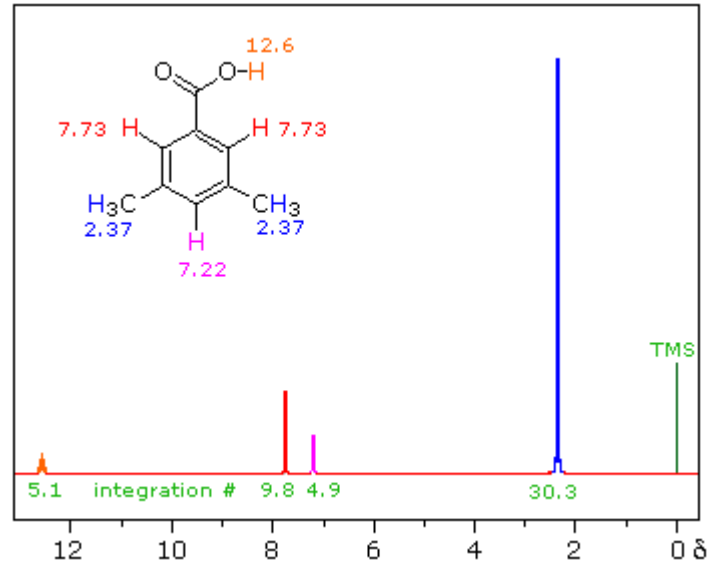
Section B2 – NMR spectroscopy

- Nuclear Magnetic Resonance is used to determine the structure of small proteins (Less than 25 kDa) in solution
- Often used on proteins that do not readily form crystals
- Natural isotopes of nitrogen, phosphorus and hydrogen behave as small magnets that can switch between magnetic spin states in applied magnetic fields
- Absorbance of low wavelength electromagnetic radiation generating NMR spectra
- NMR spectra for different atoms are unique, but is also influenced by surrounding electron density → Magnetic resonance changes by different chemical groups
- NMR spectra dependent of structure and spatial configuration of molecule
- Analysis of NMR spectra generates distance constraints which can be used to determine three-dimensional spatial arrangements

Section B2 – NMR spectroscopy

- NMR spectroscopy





Section B2 – Other methods

- Used when proteins cannot be crystallized or are to large for NMR
- Magnetic angle spinning NMR
  - Modification of regular NMR
- Electron microscopy
  - Low resolution
- Fiber diffraction
  - Two-dimensional crystal
- Circular dichroism spectroscopy
  - Optical activity of asymmetrical molecules gives differing absorption spectra in left and right circularly polarized light

Section B3 –
Gene and protein expression data

- Global expression analysis
- DNA microarrays
- 2D protein gels

Section B3 – Global expression analysis

- Global expression analysis is the study of all genes simultaneously, either at the RNA or protein level
- RNA level
  - DNA microarray
  - Direct sequence sampling
    - Sequencing of a number of randomly chosen cDNA clones. Abundant RNAs appear at a higher frequency → Determination of relative expression levels
  - Serial analysis of RNA expression (SAGE)
    - A SAGE tag (8-15 nt) uniquely identifies each gene. SAGE tags from a number of cDNA clones are ligated together and sequenced in a single reaction, and the relative expression levels can be detected by counting each of the unique SAGE tags
- Protein level
  - Two-dimensional polyacrylamide gel electrophoresis (2D-PAGE)
  - Protein microarray

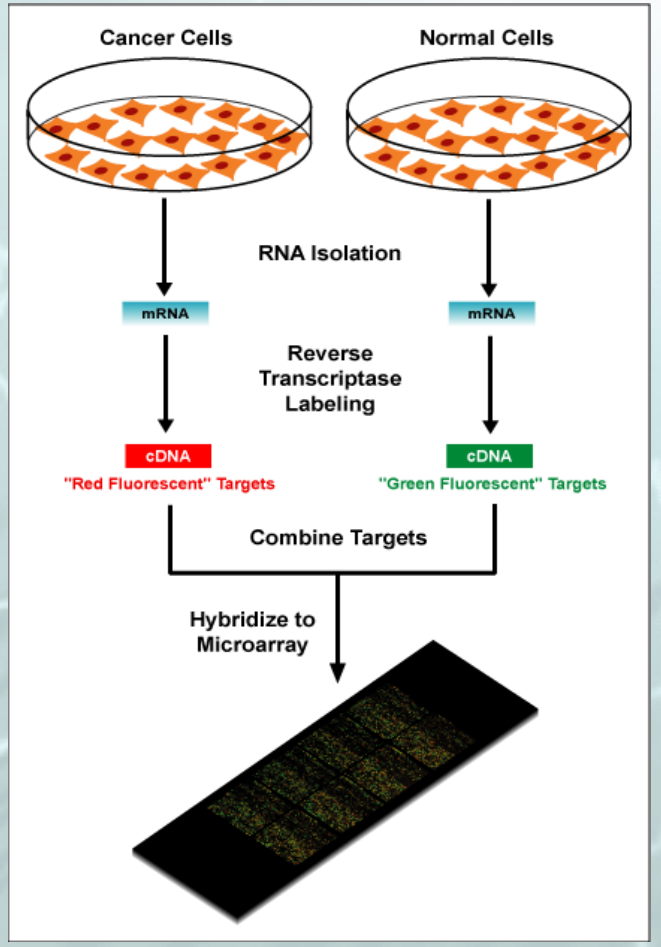
Section B3 – DNA microarray

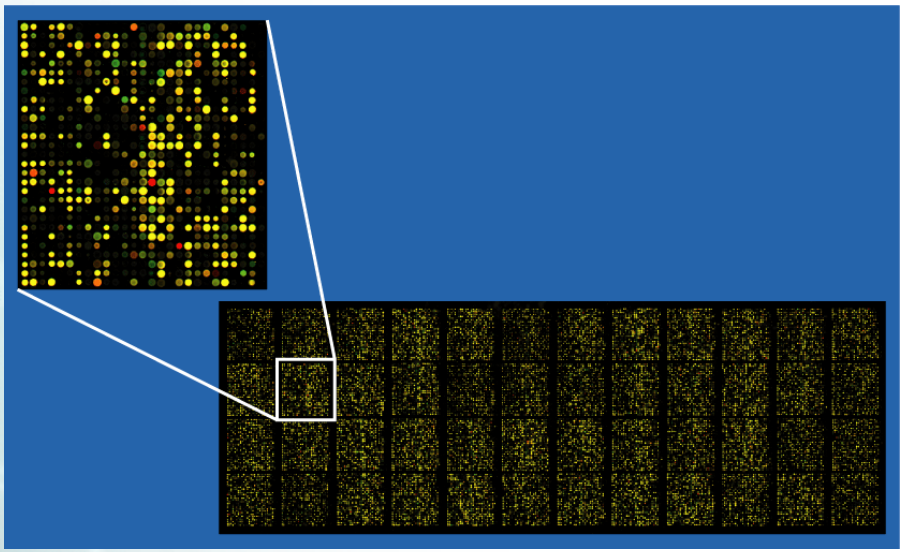
- Principle: Nucleic acids (RNA and DNA) can form double-stranded molecules by hybridization. By using a labelled (radioactive or fluorescent tag) probe one can isolate a complementary molecule from whole genomic DNA or RNA. The probes are specific so many hybridizations can run at the same time.
- A DNA microarray is a dense grid of DNA elements arranged on a miniature support. Each DNA element represents a specific gene
- The array is hybridized with a labelled complex RNA probe from a particular cell type, or from to different labelled with two different colours
- The intensity of the signal from each DNA element on the array gives the relative expression levels in that type of cell
- Can be used to distinguish healthy cells from not healthy ones, i.e. cancer cells from healthy ones
- Each signal on a DNA microarray is linked to a corresponding gene

Section B3 – DNA microarray

- Two major technologies
  - Spotted DNA arrays
    - Robotic spotting of DNA elements onto a coated glass slide
    - Up to 5000 elements per square cm
    - Double-stranded DNA molecules up to 400 bp in length that must be denaturated prior to hybridization
    - Two different samples labelled by different colour and hybridized simultaneously. Differences in colour gives differences in expression levels
  - On-chip photolithographic synthesis/GeneChips
    - Short oligonucleotides synthesized *in situ* during chip manufacture
    - Up to 1 000 000 elements per square cm
    - Single-stranded DNA molecules 25 nt in length
    - Because of short oligonucleotides each gene is represented by 20 elements
    - Separate probes on two identical chips

Section B3 –
DNA microarray





Section B3 –
DNA microarray

- DNA microarray animation

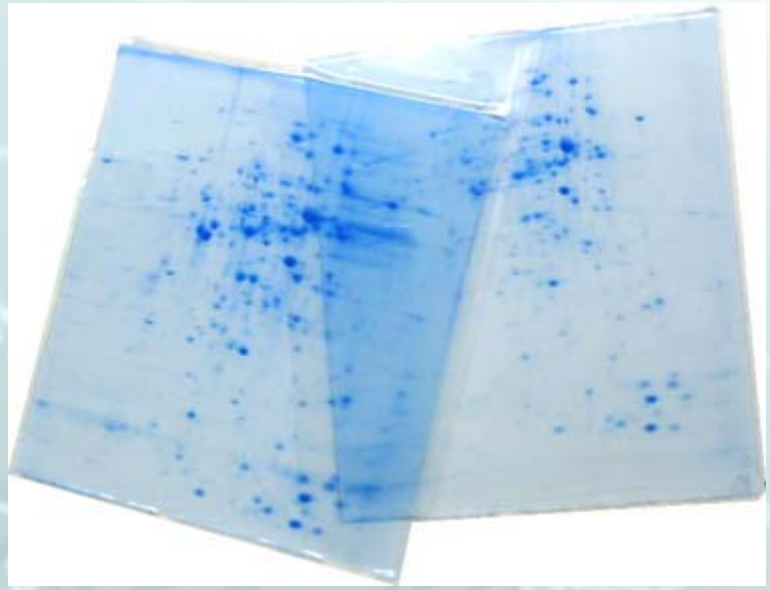
Section B3 –
2D protein gels

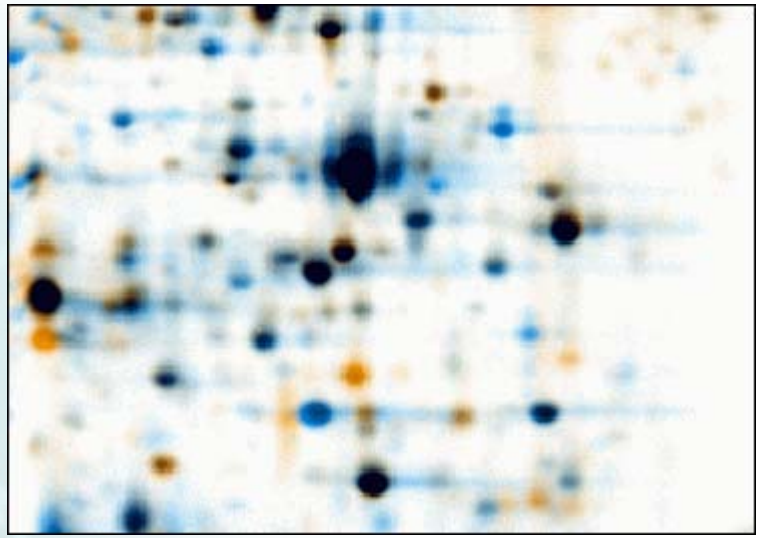
- Proteins separated on the basis of two separate properties
- Isoelectric point (pI)
  - First dimension separation in an immobilized pH gradient. The proteins migrate in a series of buffers in an electric field until the pH is the same as the pI of the proteins.
  - The pI of a protein is the pH at which it carries no net charge and therefore move anymore in the electric field
- Molecular mass
  - Before separation by molecular mass the proteins are equilibrated in sodium dodecylsulfate (SDS) which bind uniformly to proteins and gives them a negative charge
  - Separation as general gel electrophoresis in polyacrylamide gels

Section B3 –
2D protein gels

- After the second separation the gels are stained with a universal dye to reveal the position of all protein spots
- By reproduce the procedure for two different treated samples (or healthy/non-healthy cells) one can compare protein expression levels
- Characterization of each spot by x,y coordinates where x corresponds to pI and y to the molecular mass
- Position shifts or quantitative variations may be caused by post-translational protein modifications like phosphorylations and glycosylations
- Each spot on the gel has to be individually identified by mass spectrometry

Section B3 –
2D protein gels





Section B3 –
2D protein gels



"You've got one protein missing..."
"No, you've one extra protein!"

Section B4 –
Protein interaction data

- The importance of protein interactions
- Genetic methods
- Affinity methods
- Molecular and atomic methods
- Library-based methods

Section B4 –
The importance of protein interactions

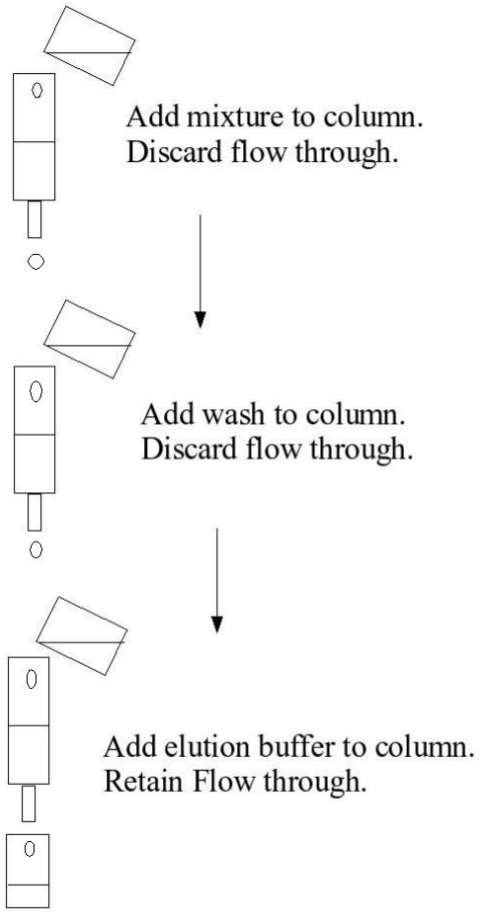
- Most cell functions depend on protein interactions; proteins interacting with each other and with other molecules
- Disruption of protein interactions can have ill effects while therapeutic drugs work by interacting beneficially with proteins
- Many proteins exist in permanent multi-subunit complexes or assembles into complexes when needed
- Proteins are also part of transient complexes like signalling proteins and proteins involved in regulation
- The ability to interact through diverse complexes and pathways is one way to give an organism its complexity
- Interactions between proteins and nucleic acids are of great importance to a number of cellular processes
- Proteins can also interact with a number of small molecules, i.e. receptor-ligand binding)

Section B4 – Genetic methods

- Used to infer protein interactions
- The interactions between two given proteins, X and Y, can be studied by looking at the behaviour of mutations in their corresponding genes, *X* and *Y*.
- By mutate specific locations in the gene one can look at the effect of the mutation in the expressed protein and proteins it interacts with
- Some mutations do not effect the expressed proteins, while a single mutation in a crucial part of the protein can make it
  - Not fold correctly
  - Lose all or some effect
  - Enhance the effect of the protein

Section B4 – Affinity methods

- Affinity is the ability of a protein to bind specifically
- Affinity chromatography
  - Attachment of a bait molecule to a matrix, i.e. in a column
  - Protein extracts are washed through with low-salt buffer
  - Proteins that interact with the bait molecules are retained while the others are washed out of the column
  - Bound proteins can be washed out using a higher salt concentration, and by using different salt concentrations it is possible to discriminate between proteins that bind weakly and those that bind strongly to the bait



Section B4 – Affinity methods

- Affinity chromatography animation

Section B4 – Affinity methods

- GST-pulldown
  - Modification of affinity chromatography
  - Bait molecule is expressed as a fusion with the enzyme glutathione-S-transferase (GST) which has a high affinity for its substrate glutathione
  - GST fusion protein binds glutathione-coated sepharose beads in the column
  - Proteins that interact with GST fusion protein will be retained on the column while noninteracting proteins are eluated
  - Any protein can be used as bait, provided it can be expressed as a GST fusion
  - Important to run control with GST to exclude proteins that interact with GST itself

Section B4 – Affinity methods

- Co-immunoprecipitation
  - Precipitation of protein complexes using an antibody specific for one of the components of the complex
  - The cell lysate is prepared under conditions that preserve protein-protein interactions
  - An antibody against a protein X is mixed with the lysate
  - Precipitation of protein X and protein complexes involving protein X
  - One problem is to distinguish between proteins interacting directly with te bait and proteins interacting indirectly
- Protein microarrays
  - Protein samples spotted onto a glass plate
  - Bait molecules labelled with colour and added to the glass plate

Section B4 – Molecular and atomic methods

- X-ray crystallography/NMR
  - Determination of structure of proteins
  - Either as uncomplexed or in complex with ligands, cofactors or substrates
- Fluorescent resonance energy transfer (FRET)
  - Interaction between two proteins shown by the transfer of energy from an excited donor fluorophore (conjugated to one protein) to an acceptor flourophore (conjugated to the other)
  - FRET only occurs when the two proteins are as close as in an interaction
  - Detected by a change in the emission wavelength
  - By incorporation of a naturally occurring bioluminescent protein (I.e. GFP) the technique is called bioluminescent resonance energy transfer (BRET)
  - FRET/BRET detects both stable and transient interactions
- Mass spectrometry
  - Composition of multi-subunit protein complexes

Section B4 – Molecular and atomic methods

- Surface plasmon resonance (SPR) spectroscopy
  - Optical resonance phenomenon occurring when plasmon waves become excited at the interface between a metal surface and a liquid
  - Protein X is immobilized on the metal surface and protein Y is free in the solution
  - Interaction between protein X and Y gives a change in the refractive index of the surface layer
- Surface-enhanced laser desorption/ionization (SELDI)
  - Interacting proteins are trapped by affinity to a capture agent (i.e. antibody) which is immobilized to a solid substrate
  - After washing and tryptic digestion of the interacting proteins the substrate is used for MALDI MS analysis

Section B4 – Library-based methods

- Standard methods can only analyse a few samples at the same time and the proteins cannot easily be assosiated with genes encoding them
- Library methods
  - Based on traditional DNA expression libraries
  - Proteins are expressed from their associated cDNA
  - Protein interactions are set up in vitro
  - Protein folding and recognition conditions do not reflect those in the cell