Introduction to computational and systems biology

Lecture 3: Elements of DNA biotechnology for computer scientists

Ion Petre
Computer Science, Abo Akademi
http://users.abo.fi/ipetre/?n=Teaching.Intro-CompSysBio
Fall 2018
• Introduce some of the methods used in biotech labs (as well as processes in the cell itself) to manipulate DNA
  - The availability of such tools made the rapid development of the field possible
  - It is not a lab manual, only present a number of available tools and the ideas supporting them
  - Essential to know about elementary lab tools also for mathematicians and computer scientists

• Topics:
  - measuring DNA,
  - testing for/isolating DNA molecules,
  - separating and fusing DNA strands,
  - lengthening,
  - shortening,
  - cutting,
  - linking DNA,
  - modifying nucleotides,
  - multiplying,
  - sequencing,
  - microarrays
Measuring the length of DNA molecules
Measuring the length of DNA molecules

- **Problem**: measure the length of a molecule without sequencing it

- **Length – definition:**
  - Single stranded DNA: the number of nucleotides (ex: 12 mer)
  - Double stranded DNA: the number of base pairs (ex: 12 bp)
Measuring the length of DNA

- **Central facts:**
  - DNA molecules are negatively charged: in electric fields, they migrate towards + electrodes
  - The force needed to move DNA molecules – proportional to its length (due to friction)

- **Idea:** gel electrophoresis technique
The gel electrophoresis technique

- Pour some gel in a rectangular container with electrodes on the sides
- In the cooling process: a comb is inserted on the negative side
- Gel cooled down, comb removed: row of small wells

The gel electrophoresis technique

- The solution with the DNA to be measured brought in the wells
- Activate the electric field: DNA molecules travel towards +
- Gel acts as a molecular sieve – big friction
  - small molecules move faster than big molecules
The gel electrophoresis technique

- Deactivate the field when the first molecules have reached +
- Small molecules – longer distance than bigger molecules
- Molecules of same length – same distance
Gel electrophoresis

Figure 1.9: Gel electrophoresis

The gel electrophoresis technique

- Reading the resulting gel
- DNA molecules are colorless: mark them
- Example of two techniques for marking:
  - Staining with ethridium bromide \(\rightarrow\) fluorescent mark under ultraviolet light
  - Attaching radioactive markers to the ends of DNA molecules \(\rightarrow\) expose a film
Computing the length

- Based on the distance traveled

- Using a calibrating solution in one of the wells → compare the location of bands on the other paths with the calibrating path
Testing for known DNA molecules in a solution
Testing for known molecules

- **Problem**: test if there are in a given DNA solution single stranded molecules containing a known sequence \( s \)
  - Variant: in that case, isolate some of them

- **Solution**: Attach the complementary strands (*probes*) to a filter and pour the solution through the filter
  - \( s \) molecules bind to their complements to form double strands (*annealing*) and stay on the filter
Testing for/isolating known molecules

- **Other methods:**
  - Attach probes to tiny glass beads, place them in a glass column and pour the solution over them.
  - Attach probes to tiny magnetic beads, throw them in the solution and stir; attract them to one side using a magnet.
  - Microarrays.
Operations on DNA molecules

- Separating and fusing DNA strands
- Lengthening DNA
- Shortening DNA
- Cutting DNA

- Linking DNA
- Modifying nucleotides
- Multiplying DNA (cloning, PCR)
- Sequencing
Separating and fusing DNA strands
Separating and fusing DNA strands

- Hydrogen bond between complementary bases is weaker than the covalent bond between consecutive nucleotides within one strand.
- One can separate the two strands of a DNA molecule without breaking the single strands.
- Separating the two strands: *denaturation*
- Fusing two complementary strands: *annealing/renaturation/hybridization*
Denaturation and annealing

- **Denaturation**: heat the DNA solution until DNA “melts”
  - Melting temperature: 85 – 95°C
  - Use of chemicals (e.g., formamide) lowers the melting temperature
  - Melting temperature of DNA = the temperature at which half of the molecules separate

- **Annealing**: solution cooled down (slowly), the separate strands fuse again by hydrogen bonds (also called renaturation or hybridization)
Lengthening & shortening DNA
Lengthening DNA

- Polymerases: enzymes whose central function is to synthesize polymers of nucleic acids
  - DNA polymerases - add nucleotides to an existing DNA molecule

- Requirements for a DNA polymerase to function:
  - Existing single stranded template describing the chain of nucleotides to be added
  - Existing sequence (primer) bonded to a part of the template

Polymerase with error-proof capability
Lengthening DNA - polymerases

- Polymerases only extend the DNA molecules in the 5’-3’ direction
  - Note: they move in the 3’-5’ direction on the template strand

- They extend repeatedly the 3’ end, provided that the required nucleotides are available in the vicinity

- Given one strand, to produce the corresponding double stranded molecule, one needs the primer and the polymerases
Lengthening DNA

- Exceptions: *terminal transferase* does not need a template: extends the 3’-end of double strands by some single stranded tails

  - *terminal transferase* does not need a template: extends double strands at both ends by some single stranded tails
Lengthening DNA

Figure 1.12: Transferase activity

Application: engineering DNA strands

- Engineering double strands if we have one of the strands (template)
  - Prime the given strand and then use polymerase to extend the primer according to the template

- Engineering single strands: a procedure leading to automation exists → “synthesizing robots”
  - Short such single strands are called oligonucleotides (oligos in short); they are essential in PCR
Shortening DNA

- The enzymes: *DNA nucleases* – degrade DNA
- Two types of enzymes:
  - *DNA exonucleases*: cleave (remove) one nucleotide at a time from the ends of the DNA
  - *DNA endonucleases*: destroy internal bonds in the DNA molecule (cutting DNA)
Shortening DNA

- **Exonucleases:**
  - some remove nucleotides from 5’, others from 3’
  - some degrade single strands, others on double strands, other operate on both

- **Example:**
  - Polymerases can act as exonucleases: error correction in DNA replication
Exonucleases

Figure 1.13: Exonuclease III in action

Exonucleases

Figure 1.14: Exonuclease \( Bal31 \) in action

Cutting DNA

- *Endonucleases*: destroy internal covalent bonds in DNA molecules
- Can be quite specialized: what/where/how they cut
- Examples:
  - S1 – not site specific
  - DnaseI – not site specific
Endonucleases

Figure 1.15: S1 endonuclease in action (i)

Figure 1.16: S1 endonuclease in action (ii)

Restriction enzymes

• Restriction endonucleases: much more specific
  - Cut only double stranded molecules, only at very specific set of sites
  - Bind to the recognition site and then cut (sometimes even outside the binding site)
  - The cut: blunt or staggered

• Examples:
  - EcoRI: 5’-GAATTC
  - XmaI: 5’-CCCGGG
  - SmaI: 5’-CCCGGG
Restriction enzymes


Figure 1.17: EcoRI in action
Restriction enzymes

Figure 1.18: Multiple cut by $E_{coRI}$

Restriction enzymes

Figure 1.19: XmaI in action

Figure 1.20: SmaI in action

Restriction enzymes

Figure 1.21: \textit{PstI} in action

Restriction enzymes

Figure 1.22: \( Hgal \) in action

Linking DNA

- “Fuse” together two pieces of DNA
- Add the necessary chemical links

**Examples**
- Hydrogen bonding
- Ligation (enzyme: \textit{ligase})
- Hybridization: link two complementary single strands
- Blunt end ligation
Hydrogen bonding

- Complementary base pairing

Figure 1.23: Complementary base pairing

Ligation

- Fix the phosphodiester bonds in single strands (*nick*)

![Diagram of DNA ligation](image)

*Figure 1.24: Ligation*

Hybridization

- Double strands produced out of two single strands (sticky ends)

![Diagram of hybridization process](image)

**Figure 1.25: Hybridization**

• **Note:** blunt end ligation is much less efficient than sticky end ligation
  - Typically about 100 x slower
  - Happens through random collisions of the two molecules
Blunt ligation

![Diagram of blunt ligation process]

**Figure 1.27:** Joining blunt ended molecules using homopolymer tailing

Modifying nucleotides
Modifying nucleotides of DNA

- The enzymes: *modifying enzymes*

- Examples:
  - *Methylases*: used *in vivo* as partners of restriction enzymes – the same recognition site
    - It adds a methyl group to one of the nucleotides
  
  - *Alkaline phosphatase*: modifies the 5’ end so that the DNA molecule cannot self-ligate
    - removes the phosphate group adding a hidroxil group

  - *Polynucleotide kinase*: reverses the effect of alkaline phosphate
Example

Figure 1.28: Alkaline phosphatase in action

Figure 1.29: Polynucleotide kinase in action

Multiplying DNA
Multiplying DNA

- Major problem in biochemical research: obtain sufficient quantities of some substance
- Solution: cloning techniques, PCR
Cloning

- **Clone**: a collection of identical organisms, all replicas of a single ancestor

- **Cloning**: insert the DNA fragment in the DNA of some fast reproducing cell and the fragment will be replicated with the cell
  - Use viruses or hybridization
Cloning

- Screen the cell with the DNA of interest from the others using an antibiotic
- This technique provides high quantities of DNA fragments
- It also gives a mean to preserve the fragments for long time (keeping alive the “host”)
Polymerase chain reaction - PCR

- Developed in 1983 by Kary Mullis
  - Awarded the 1993 Nobel prize in chemistry for it, along with Michael Smith

- *In vitro* multiplication method

- Made possible by the technique of designing oligos

- Very efficient: produce in a short time billions of copies of even a single strand
The PCR technique

- We want to amplify a DNA molecule alpha with known borders beta and gamma

The PCR technique

- We want to amplify a DNA molecule alpha with known borders beta and gamma
- **Process**: repeat a cycle of *denaturation*, *priming*, *extension*
- **Start**: prepare a solution containing alpha (the target), oligonucleotides beta’ and gamma’ (complements of beta and gamma), polymerase enzymes, and many nucleotides (A, C, G, T)

The PCR technique

• **Denaturation**: heat the solution to 85-95 C: alpha denatures into two single strands $\alpha_1$ and $\alpha_2$

![Diagram](image)

Figure 1.31: Denaturation

The PCR technique

- **Priming**: Cool down the solution to 55°C: the β- and the γ-primers anneal to their complementary borders

![Diagram showing priming process](image)

• **Extension**: Heat the solution to 72°C: polymerase extend the primers to produce two double stranded DNA molecules alpha

![Diagram of PCR reaction stages]

**Figure 1.33: Extension**

The PCR technique

- Repeat the cycle $n$ times: $2^n$ copies (in principle): highly efficient bio-copymachine!

- A single cycle takes about 5 minutes: obtain billions of copies in several hours (days for cloning)

- Note: The polymerase must be heat resistant – nature’s solution: thermophilic bacteria

- Important observation:
  - one needs to know the borders of the DNA segment to be copied
Sequencing
Sequencing

• Learning the exact sequence of nucleotides of a DNA molecule

• The most popular method of sequencing: the Sanger method
  - 1940s: Sanger, insulin sequencing (Nobel prize)
  - 1977: Sanger, Gilbert, DNA sequencing (Nobel prize)
  - 1985: Mullis, PCR (Nobel prize)

• Main tool: nucleotide analogues – nucleotides chemically altered so that no other nucleotide can attach to their 3’ end: ddA, ddC, ddG, ddT (dideoxynucleotides)
Sequencing – Sanger method

- **Problem:** sequence a single stranded molecule $\alpha$
- Assume we know it short 3′-end $\gamma$ and we add the $\gamma$-primer (labeled): let $\beta$ be the new molecule

![Diagram of DNA sequencing](image)

**Figure 1.34: $\beta'$ molecule**

Sequencing – Sanger method

- **Problem**: sequence a single stranded molecule $\alpha$
- Assume we know it short 3’-end $\gamma$ and we add the $\gamma$-primer (labeled): let $\beta$ be the new molecule
- Prepare 4 tubes (A,C,G,T) containing:
  - $\beta$ molecules
  - $\gamma$-primers
  - Nucleotides
• If only nucleotides are used: produce full duplex

![Full duplex diagram]

Sequencing

- If only nucleotides are used: produce full duplex
- **Solution:** tube X contains a *limited amount* of nucleotide analogues $ddX$, for all $X$ in the set $\{A, C, T, G\}$

Example: alpha=3’-AGTACGTGACGC

- Tube A:
  - 5’- γ’ TCATGCACTGCG
  - 5’- γ’ TCA
  - 5’- γ’ TCATGCA

- Tube T:
  - 5’- γ’ TCATGCACTGCG
  - 5’- γ’ T
  - 5’- γ’ TCAT
  - 5’- γ’ TCATGCA

- Tube C:
  - 5’- γ’ TCATGCACTGCG
  - 5’- γ’ TC
  - 5’- γ’ TCATGC
  - 5’- γ’ TCATGCAC
  - 5’- γ’ TCATGCACTG

- Tube G:
  - 5’- γ’ TCATGCACTGCG
  - 5’- γ’ TCATG
  - 5’- γ’ TCATGCACTG
Sequencing – Sanger method

- Gel electrophoresis using 4 wells - one for each tube
- Read the bands (the primers were marked)
- Obtain the sequence
Result: 5’-TCATGCACTGCG

Observations

• This was only a very simplified presentation; many details skipped:
  - Use polymerase not having the associated exonuclease activity: do not cut the ddX
  - Limited amount of nucleotide analogues: too many make the polymerase before then end of the strand
  - ...

• Impossible to sequence long molecules with this method

• Sequencing a longer molecule: divide it in overlapping short segments, sequence them and then reassemble the original sequence: computational problem!
Microarrays
Microarrays

- Microarrays: microscope slides containing an ordered series of samples (DNA, RNA, protein, tissue)
  - DNA microarray (most commonly used)
  - RNA microarray
  - ...
  - Since the samples are ordered on the microarray, data can be traced back to the sample it came from
  - Tens of thousands of samples can be placed on one microarray

- Usage of DNA microarrays
  - determine gene expression levels of genes in a sample (expression profiling)
  - determine the gene sequence of a sample (minisequencing)
  - mutation (or SNP) analysis for single nucleotide reads
Microarrays – principle

• Core principle: hybridization between two DNA strands
• High number of complementary base pairs: tighter hydrogen bonding between the sequences
• Wash-off: only strongly paired strands remain
• Fluorescent labelling of the target sequences
• Total strength of the signal from a spot depends on the amount of target sample binding to the probes on that spot

• Widely available both from commercial companies, as well as from university core facilities
  - Finnish Microarray and Sequencing Centre in Biocity Turku
  - Biomedicum Biochip Centre in Helsinki
  - ...

• Typical workflow
  - Isolation of RNA
  - Label the RNA by a reverse transcription procedure with fluorescent markers
    - produce cDNA, then label it with markers (most common are Cy3, Cy5)
  - Purification of the labeled products
    - remove all the non-labeled products
  - Hybridize to microarray
  - Scan
    - the amount of signal emitted by the fluorescent markers is directly proportional to the amount of dye at that spot on the microarray
    - most often two-color data (two dyes used)

Figure 1.2: Work flow of a typical expression microarray experiment.

Microarrays

- **Typical applications**
  - gene expression profiling: expression level of genes in comparison between two conditions. Examples:
    - up- or down-regulation during various conditions;
    - several genes regulated in the same way during heat shock → heat shock responsive genes
    - identify genes involved in the ripening of tomatoes: isolate RNA from raw and ripened fruits
    - map the localization of genes
    - pharmacological studies: discern the mechanism of action of therapeutic agents
    - comparative genomics: identify the function of unknown genes by comparing their activity with that of known genes
    - SNP analysis: detect the presence of single nucleotide differences between genomic samples (SNPs occur approximately 1 in every 1000 bases in humans)