CMSC423: Bioinformatic Algorithms, Databases and Tools
Lecture 2

Molecular biology primer

Perl/Perl Modules

Administrative details

- Lecture notes and homework assignments can be found on Syllabus site.

RECAP

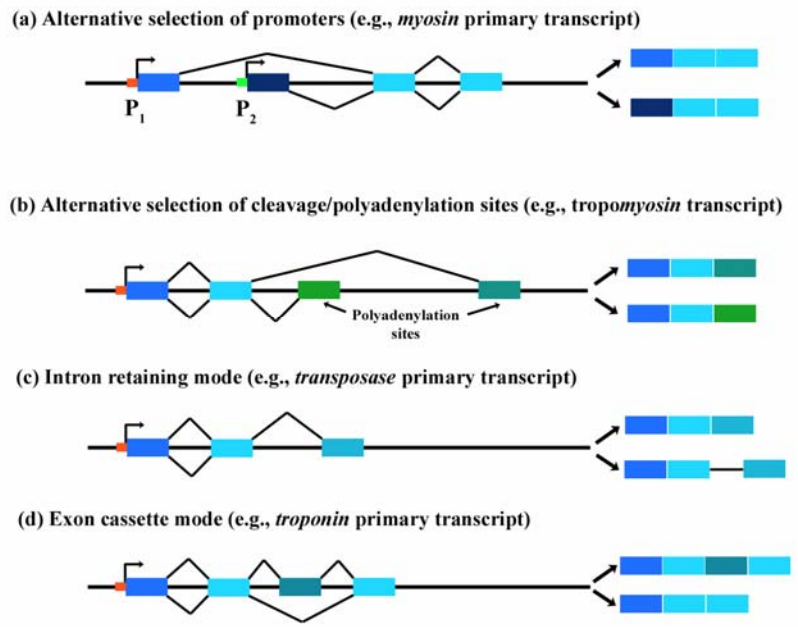
- DNA is a string formed with letters A, C, T, G (called nucleotides or bases)
- DNA is double-stranded – allows replication: transfer of genetic "code" from parents to offspring
- DNA is naturally oriented from 5' to 3' and the two strands are anti-parallel
- If you know the sequence of one strand, you can obtain the sequence of the other by reverse-complementation

```
5' AGACCTAGTGCACGGCTACTACC 3'
5' CCATCATCGGCACGTGATCCAGA 3'   Reverse
5' GGTAGTAGCCGTGCACTAGGTCT 3'   Complement
```

RECAP

- Central Dogma of molecular biology:
  - DNA – RNA (transcription)
  - RNA – Protein (translation)
- The transcribed segments of DNA are called "genes"
- Translation occurs in sets of 3 nucleotides – codons
- Each codon encodes one of 20 amino-acids and 3 stop-codons
- In many eukaryotes the genes are split into multiple exons, separated by introns: DNA segments that will not get translated
- The protein corresponding to a gene is translated from an RNA representing the concatenation of the exons of the gene

Alternative splicing examples



Playing with DNA

Biologists can:

- Cut the DNA – restriction enzymes (often palindromes) (Nobel prize – Arber, Nathans, Smith)

```
5'GAATTC          5'---G      AATTC---3'
3'CTTAAG          3'---CTTAA        G---5'
```

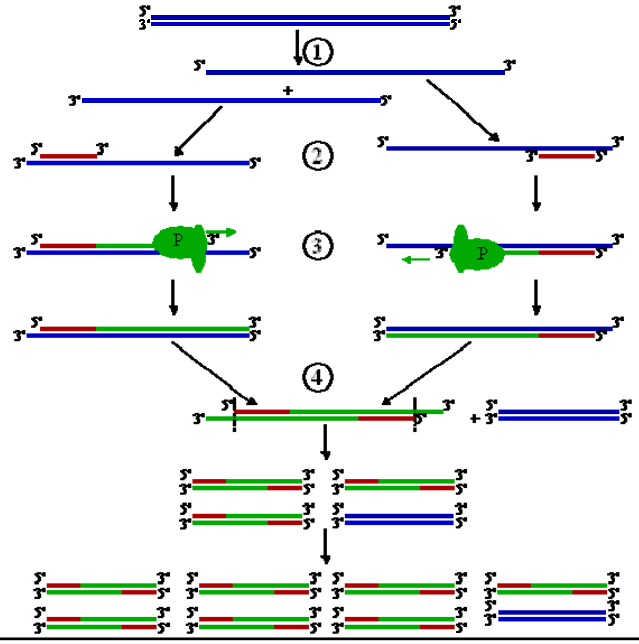
- Attach "things" to DNA (either single or double-strand)

```
TAGGCACGTTGCAACTACGGC

   TGCAACGT
```

- "Amplify" DNA – Polymerase Chain Reaction (Nobel prize – Mullis)

Polymerase chain reaction (PCR)



1. Denature

2. Anneal (attach primer)

3. Extend

4. Repeat

How does PCR work?

- 1. Start: 1 double-stranded molecule
- 1. Denature: 2 single-stranded molecules
- 1. Anneal: 2 single-stranded molecules with primers attached
- 1. Extend: 2 double-stranded molecules – one “long” (L) strand and one “short” (S) (terminated at a primer)
- 2. Start: 2 double-stranded molecules: L+S, L+S
- 2. Denature: 2 x L strands, 2 x S strands
- 2. Anneal: all strands with primers attached
- 2. Extend: 2 double-stranded molecules: L+S, L+S, 2 double-stranded molecules: S+SS, S+SS
  SS – strand terminated at both ends with a primer

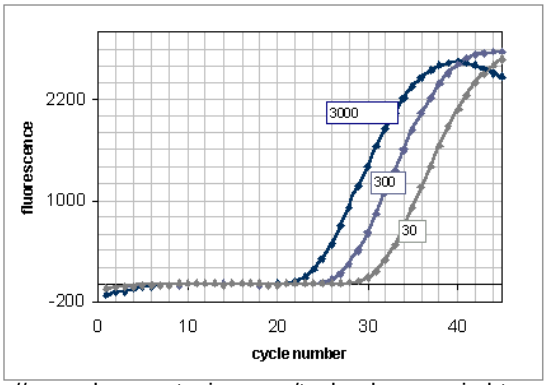
PCR Recurrences

- L_n, S_n, SS_n - # of strands of each type at cycle n
- $L_n = L_{n-1} = 2$
- $S_n = S_{n-1} + L_{n-1} = S_{n-1} + 2 = 2 * (n - 1) = O(n)$
- $SS_n = S_{n-1} + 2 * SS_{n-1} = O(2^n)$

- The sequence between the primers (SS) is amplified exponentially – will quickly overtake the solution

Quantitative PCR

- Measure # of PCR cycles needed to reach a certain concentration of DNA – depends on initial # of molecules
- Used in diagnostics: e.g. is this a random Anthrax spore from the environment or lots of spores from an attack



http://www.dxsgenotyping.com/technology_main.htm

DNA sequencing

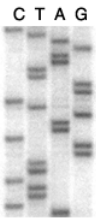
- Most techniques “trick” the polymerase into revealing the sequence
- The traditional method – Sanger sequencing – based on “terminator” bases – prevent the polymerase from extending the DNA
- Sanger sequencing is essentially PCR + terminator bases
- Other methods “spy” on the polymerase as it incorporates nucleotides

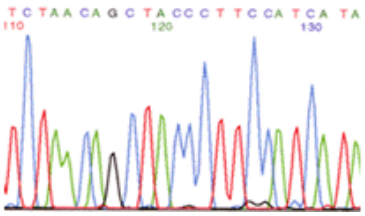
Sanger sequencing

Sanger, F, Coulson AR. *A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.* J.Mol.Biol. 94 (1975)

TCTAATAGA
AGATTATCTAACAGCTACCCTTCCATCA

TCTAATT
TCTAATTA
TCTAATTAG
TCTAATTAGA
TCTAATTAGAT



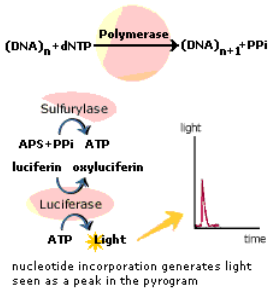


pictures from http://www.uvm.edu/~cgep/Education/Sequence.html

The future of sequencing

• Single molecule sequencing - current technology requires many copies of DNA being sequenced - requires DNA amplification
• Massively-parallel sequencing - 100k sequencing reactions occuring at the same time

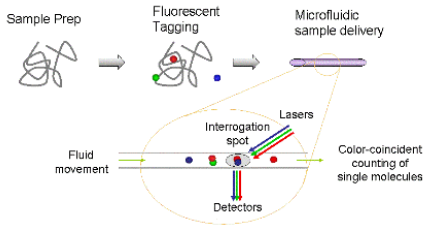
Sequencing by synthesis



TCTAATAGA
AGATTATCTAACAGCTACCCTTCCATCA

http://www.genetics.ucla.edu/sequencing/pyro.php

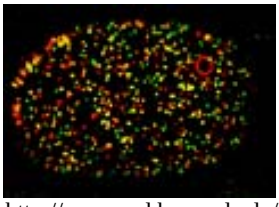
Micro-fluidics



http://www.usgenomics.com

The future of sequencing

Massively parallel sequencing



http://arep.med.harvard.edu/

• each spot is a molecule or amplified from one molecule
• image processing used to track molecules during sequencing by synthesis
• often micro-fluidics/lab-on-a-chip used

- 454 Life Sciences – approx. 60 Mbp in 200 bp reads / 4 hr run
- Solexa Ltd. – approx. 1 Gbp in 30-40 bp reads / 3 day run

Not yet available:

- Helicos – single molecule sequencing
- Agencourt
- Applied Biosystems
- etc.