Lecture 2 (9/5/2006)

- Splicing/alternative splicing (the "truth" about transcription/translation)
  - In eukaryotes, a gene is usually represented by several disjoint genomic regions (exons) that are spliced into a single RNA molecule by removing the intervening regions (introns):

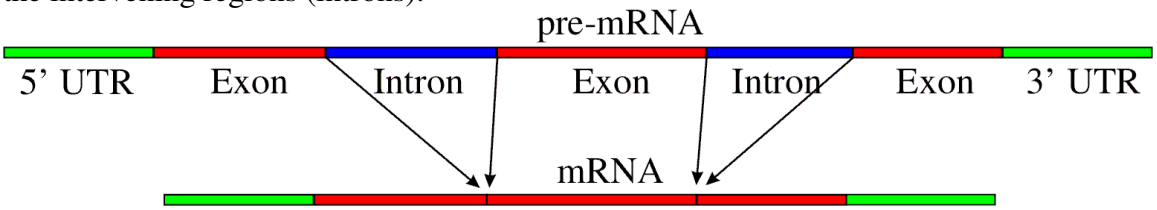


Figure 1 Gene splicing (from Wikipedia) (UTR- untranslated region: only the red parts will become a protein)

  - In many eukaryotes, multiple proteins are made by a single gene by selective splicing of exons:

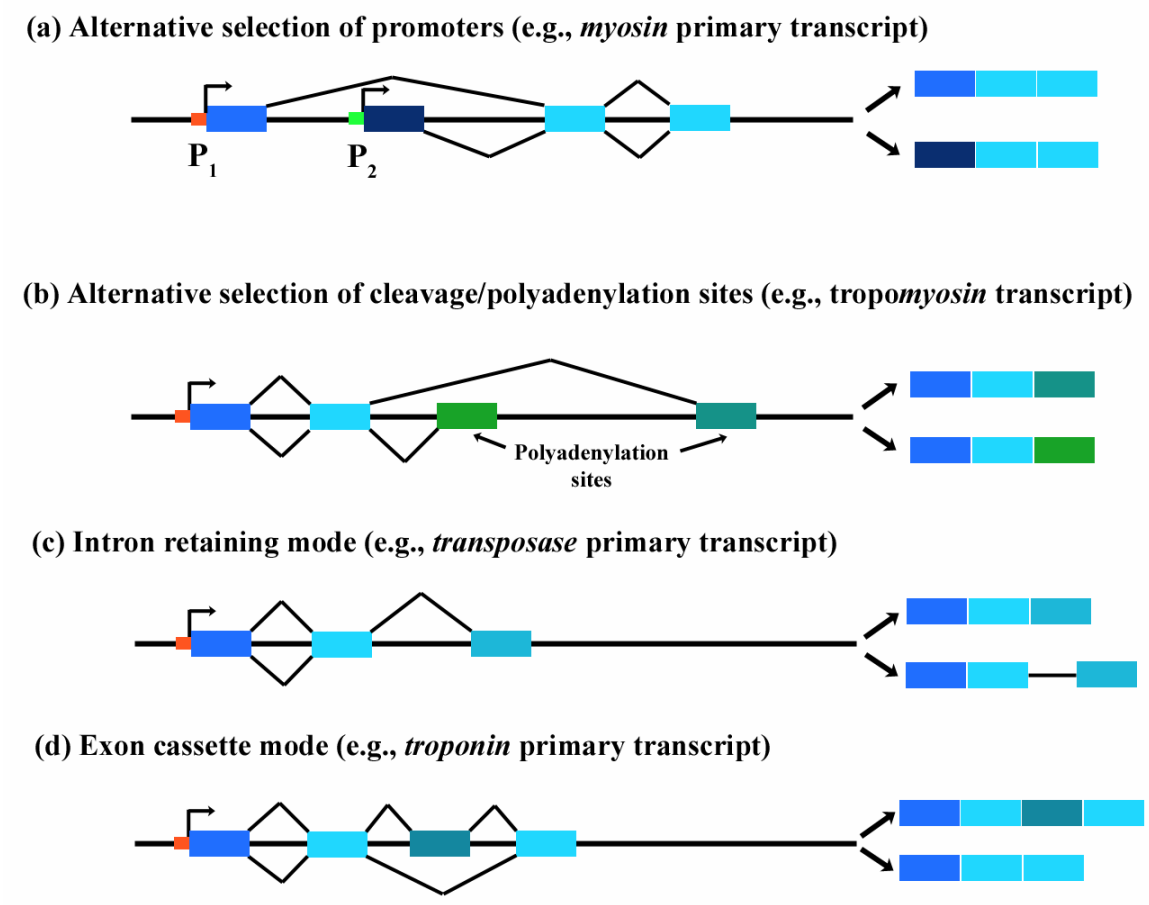


Figure 2 Alternative splicing examples (from Wikipedia).

  - Splicing/alternative splicing are one of the reasons why we need efficient tools for inexact matching. When matching RNA strings to a genome we

must tolerate gaps in the alignment, corresponding to the location of the introns.

- Manipulating DNA
    - Cutting DNA - restriction enzymes cut through double-stranded DNA. Usually they are short (6-8bp) palindromes (same sequence in reverse complement). E.g. EcoRI = GAATTC
    - How would you find palindromes in strings? How fast can you do it? Search at each location - 3 * n comparisons. Can we do better?
    - Amplifying DNA - Polymerase Chain Reaction (PCR)
        - Exponential amplification. Also targeted amplification of area of interest.
        - Double-stranded DNA - denature (make single stranded)
        - Primers attach and extend with polymerase
        - Repeat

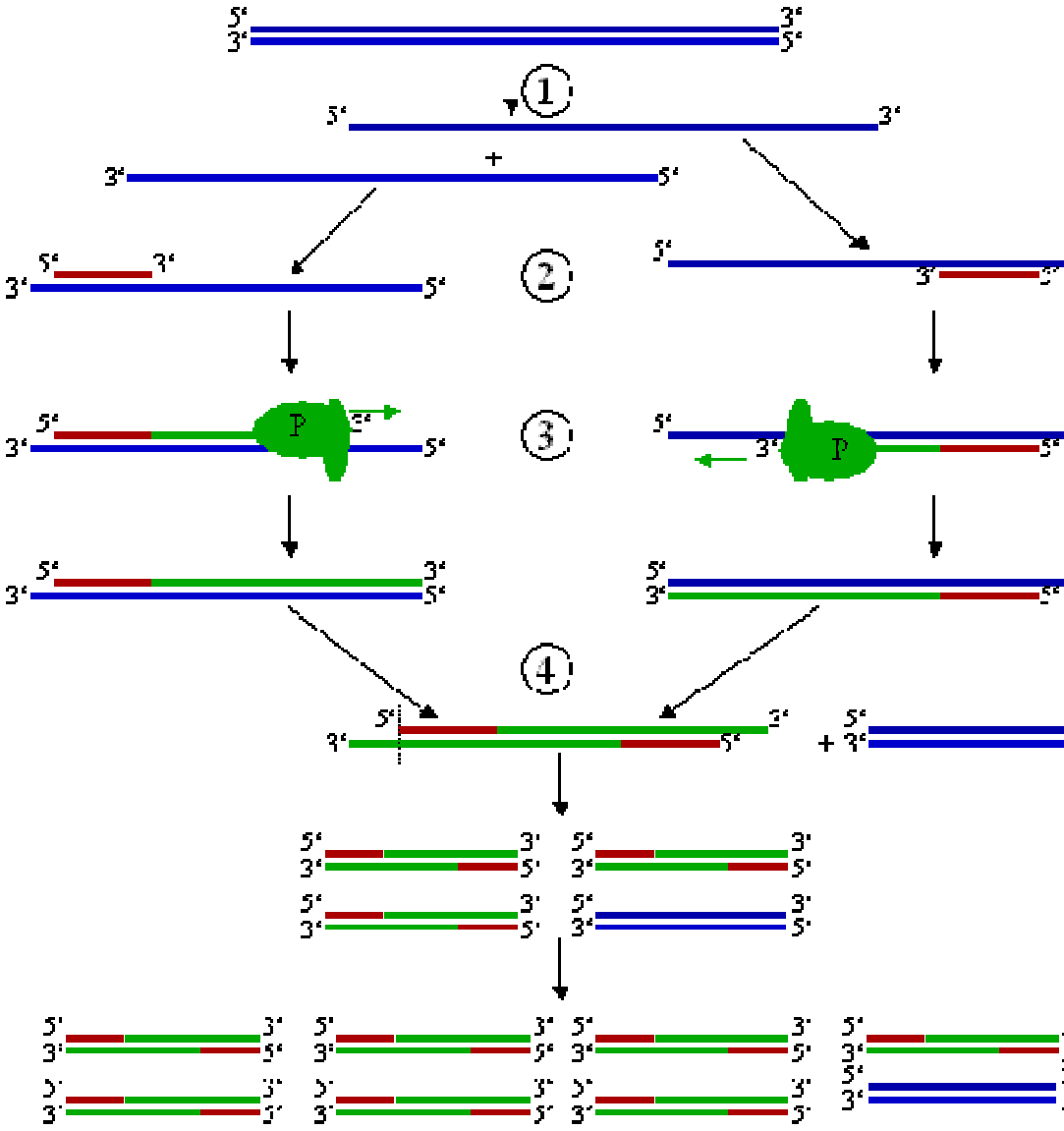


Figure 3 PCR (from Wikipedia) 1- denature, 2 - anneal, 3 - elongate, 4 - repeat

- quantitative PCR - how many rounds before you see a certain level of product (used in diagnostics)

- Sequencing DNA ("reading" the letters making up a piece of DNA)
  - Sanger sequencing (Maxam-Gilbert circa 1975, automated in the 90s)
    - Some bases terminate the extension (once you reach such a base you can no longer extend the DNA)
    - Similar to PCR, but linear (instead of exponential) amplification. Each resulting fragment ends at a different location
    - By sorting the fragments by size you can read the DNA
    - Main limitation - difficult to accurately separate large fragments by size (current limit 1000-2000 bp).

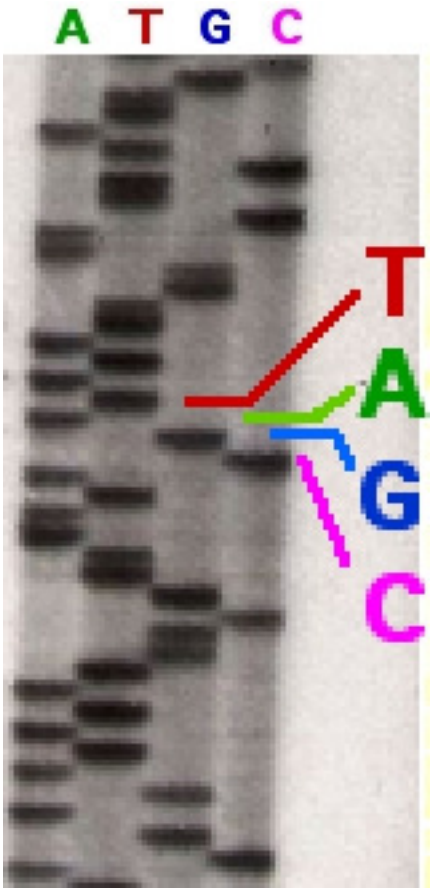


Figure 4 Electrophoresis gel sorting the four base-specific reactions.

- Newer methods - dyes attached to terminator bases. The sequence is read by laser

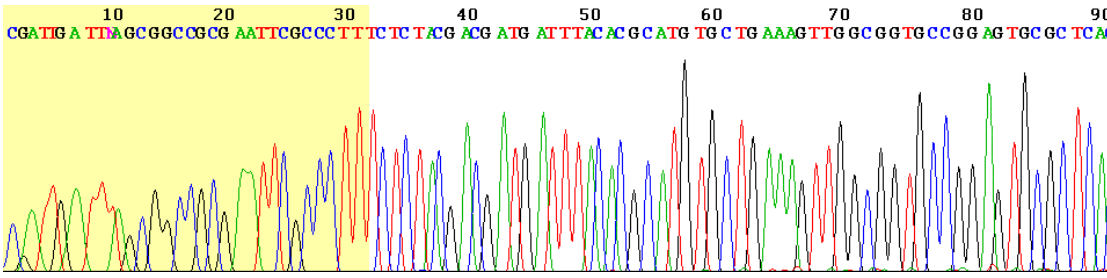


Figure 5 Electropherogram - ouptut of automated sequencer

- Sequencing by hybridization
  - Start with array containing all possible k-length (k-mers) DNA strings
  - Hybridize the DNA to be sequenced to array - identify all k-mers in string to be sequenced
  - Main problem: how do you reconstruct the original DNA?
  - Not really practical except for short strings.

- Pyrosequencing (late 90s, commercially available ~ 2004)
  - Chemical reaction emits light when a base is added (based on reaction used by fireflies). By controlling the addition of nucleotides can "read" DNA
    (see www.454.com)

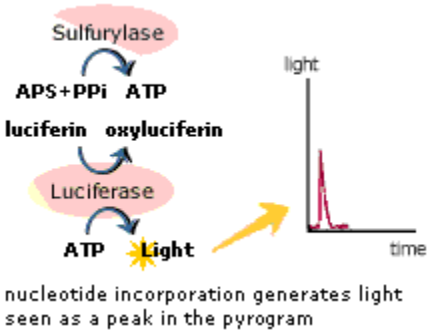
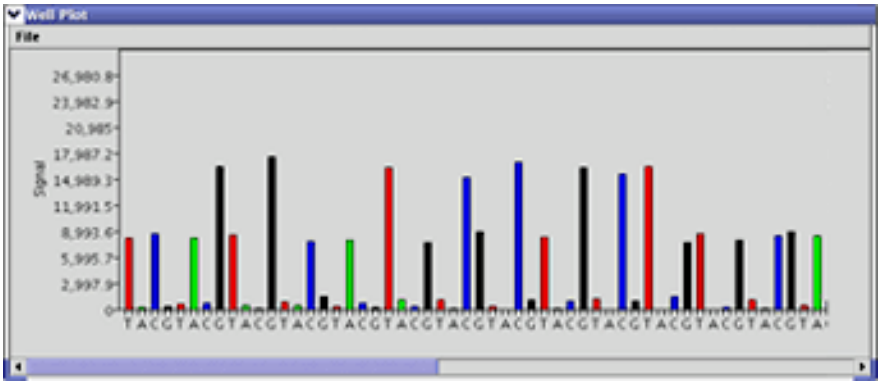


Figure 6 Pyrosequencing reaction

  - Output - intensity observed when adding a nucleotide:



  - Has problem with homo-polymers (2 or more repetitions of same base). Difficult to quantify difference in amount of light emitted.
  - Current length limitations - 100-200 bp.
- Sequencing by ligation
  - Probe DNA with tagged degenerate oligo-nucleotides interogating a same base:
    ANNNNNN
    NANNNNN
    NNANNNN
    ...
    (N stands for any of the other bases)
  - Each base has a different color. Test first base, denature, wash, and repeat with second base, etc.

        - Main limitation: can't make degenerate oligos that are too long (~6-8bp) due to # of combinations 4^6=4096, 4^8 = 65,536
    - Tagged nucleotides and reversible terminators (see www.solexa.com)
        - Similar to pyrosequencing (add one base at a time) except that bases are tagged with dyes like in Sanger sequencing.
        - Cycle: add base, interrogate with laser, remove termination and repeat.
        - No homopolymer issue (due to reversible termination)
        - Current length limitation 25-35 bp
    - Massively parallel amplification (needed by the sequencing techniques above)
        - Main idea: run PCR in a very small volume
        - 454 - attach DNA to beads, run PCR in water-in-oil emulsion (mayonnaise :))
        - Solexa - grow clusters on a microscope slide
        - George Church - immobilize DNA in a polymer and run PCR in the neighborhood

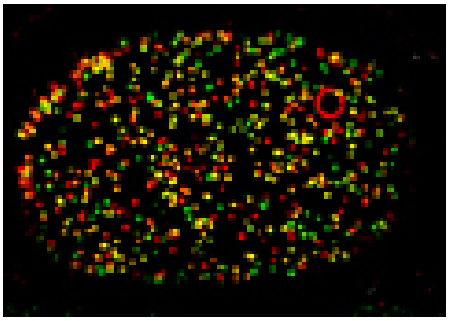


Figure 7 Many DNA clusters representing PCR amplifications of single DNA strands (color represents the base being interrogated).

    - Problem with massively parallel sequencing - phasing
        - One cluster represents n identical copies of DNA
        - If sequencing (base incorporation) efficiency is less than 100%, e.g. probability of mis-incorporation of a base is p = 0.001, at each step n * p DNA molecules fall behind (out of phase)
        - After many iterations, number of out of phase molecules becomes bigger than in-phase molecules - sequencing cannot proceed.
        - E.g.:
          First base addition: n(p -1) good, np bad molecules
          Second base addition: n(p - 1)^2 good, np + n(p - 1)p bad
          Third base addition: n(p - 1)^3 good, np + n(p - 1) p + np(p - 1)^2 bad
          ...

    - Nanopore sequencing

- pass DNA through a small hole and read the bases as they go through
- still at conceptual stage