INTRODUCTIONS

- Instructor: Mihai Pop (mpop at umiacs.umd.edu)
  http://www.cbcb.umd.edu/~mpop

  Class hours: TuThu 9:30-10:45, CSIC 3120
  Office hours: by appointment

- My offices:
  Biomol. Sci. Bldg. Room 3120F
  AVW 3223

Class webpage:
http://www.cbcb.umd.edu/confcour/CMSC701.html

Overview of course

- No knowledge of biology required
- Will cover basics of string matching algorithms and recent results in the field

Policies

- Attendance - follow University policy
  - written documentation of illness is required (from Dr.)
  - if possible inform me prior to the class you will skip

Students claiming an excused absence must apply in writing and furnish documentary support (such as from a health care professional who treated the student) for any assertion that the absence qualifies as an excused absence. The support should explicitly indicate the dates or times the student was incapacitated due to illness. Self-documentation of illness is not itself sufficient support to excuse the absence. An instructor is not under obligation to offer a substitute assignment or to give a student a make-up assessment unless the failure to perform was due to an excused absence. An excused absence for an individual typically does not translate into an extension for team deliverables on a project.

Policies...cont

- Disabilities
  - must inform me during the first 2 weeks of the semester if special accommodations necessary
  - request letter from Office of Disability Support Service

Any student eligible for and requesting reasonable academic accommodations due to a disability is requested to provide, to the instructor in office hours, a letter of accommodation from the Office of Disability Support Services (DSS) within the first two weeks of the semester.

- Religious observance
  - must inform me during first 2 weeks of class of any special accommodations
  - no additional requests will be accepted after this deadline

Grading & workload

- Homework (10%)
- Goal: small assignments
  - write lecture notes for ~ 2 lectures
  - other short assignments
- Programming project (30%)
  - your choice – should pick topic by spring break
  - can work in teams of up to 2 people
- Exams: midterm (25%) & final (35%)
- Late policy: 1 day late – 10 points off; 2 days late – 20 points off; 3 days late – 0 points

Policies...cont

- COMMUNICATION IS KEY!
  - talk to me about any issues whether covered or not by University policies
  - catch me after class, during office hours, or through email

- Note: add “CMSC701” to subject of your emails

Academic Honesty

http://www.shc.umd.edu/SHC/AICodeAndCaseProcess.aspx

- No cheating on homeworks/projects/exams
- No making up data/results
- No copying of other people's code
- You can work together on homeworks/projects but WRITE THE ANSWER BY YOURSELF

I pledge on my honor that I have not given or received any unauthorized assistance on this examination.

Flu/emergency preparedness

- flu may have a major impact on us
  - illness may make you miss classes, assignments, exams
  - if severe epidemic – classes may be canceled.
- University resources
  - http://www.helpdesk.umd.edu/emergencypreparedness
- For class
  - Don't come to class if you are sick!!! (wait for 24 hrs after fever/illness ends)
  - Inform me as soon as you can through http://grades.cs.umd.edu
  - Accommodations will be made so nobody falls behind – alternate homework/exam dates, lectures will be online, special office hours....etc.

Flu...cont

- Be prepared: Advil/Tylenol, Sudafed, Gatorade, chicken soup, Purel, tissues
- Get the vaccine (if you can, and it might not help...)
- Wash hands often (esp. before eating)
- If you have a fever (> 100 deg F, 38 deg C) or feel ill/have chills, etc. stay home (no classes, parties,etc.)
- See a doctor.
- Notify the campus: health@umd.edu

More about the class

- Each class will come with reading assignments: from textbooks, websites, scientific articles
- Read them ahead of time!

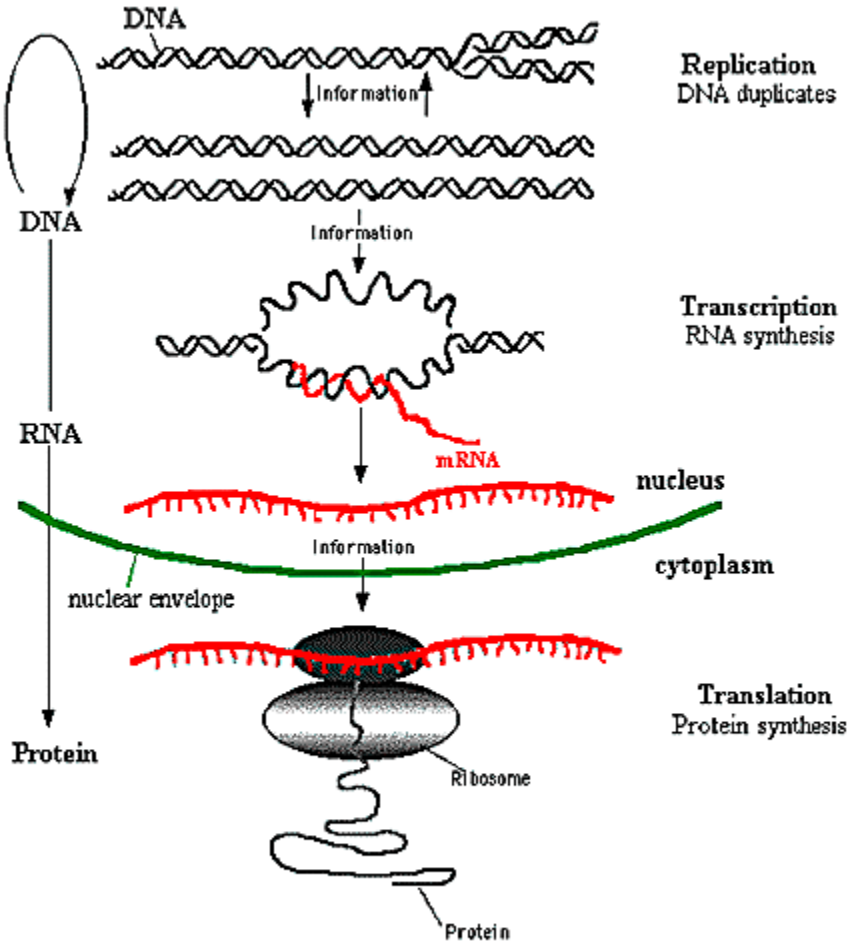
- PLEASE TAKE NOTES – most classes presented on board. No additional material provided.

String matching – what it's all about

- Which websites contain a (possibly misspelled) set of keywords?
- What is the rhyme structure of a poem?
- Do two documents written in different languages match/say the same thing?

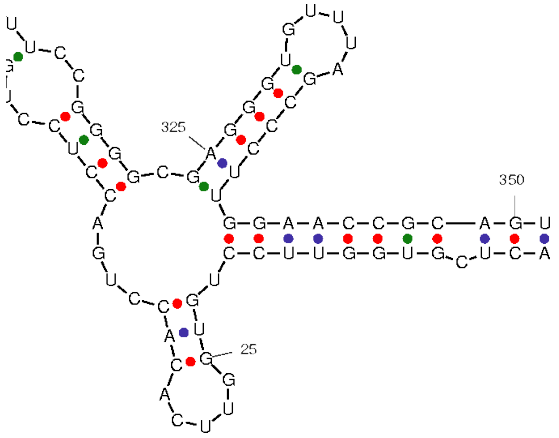
- What is the common evolutionary history of a string of DNA sequences?
- Is a newly discovered piece of DNA similar to anything in a database?
- What are the differences between two related bacteria, one that is pathogenic, the other that is not?

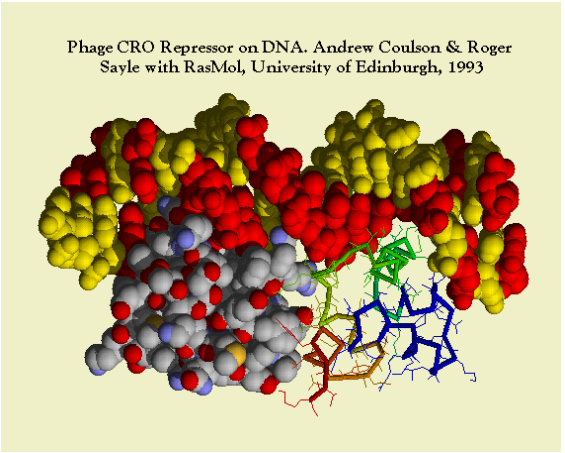
Central dogma of molecular biology



The Central Dogma of Molecular Biology

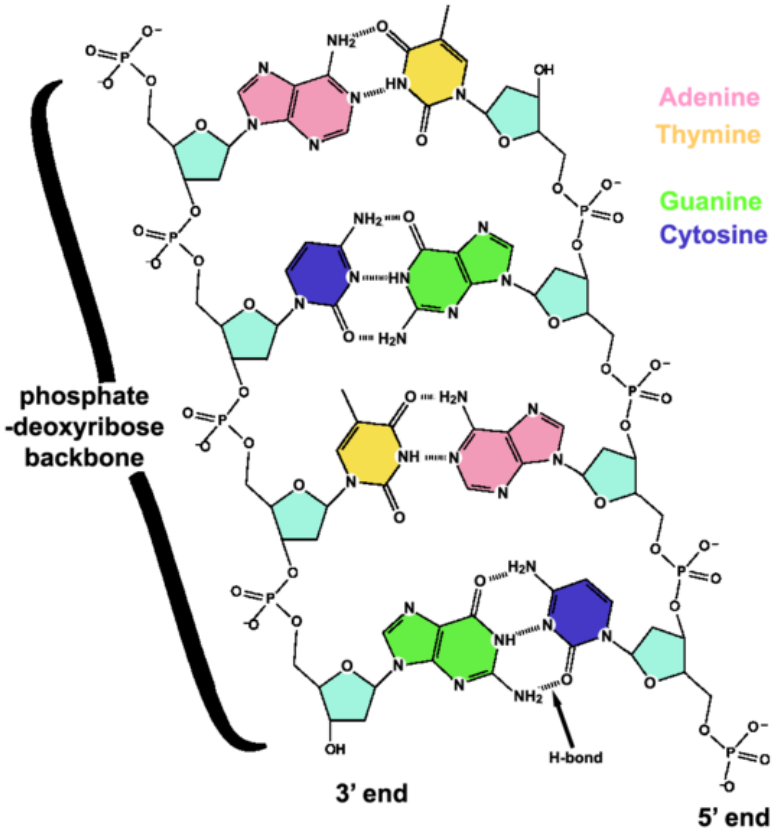
AGGTACGCGTACCTGACAGG





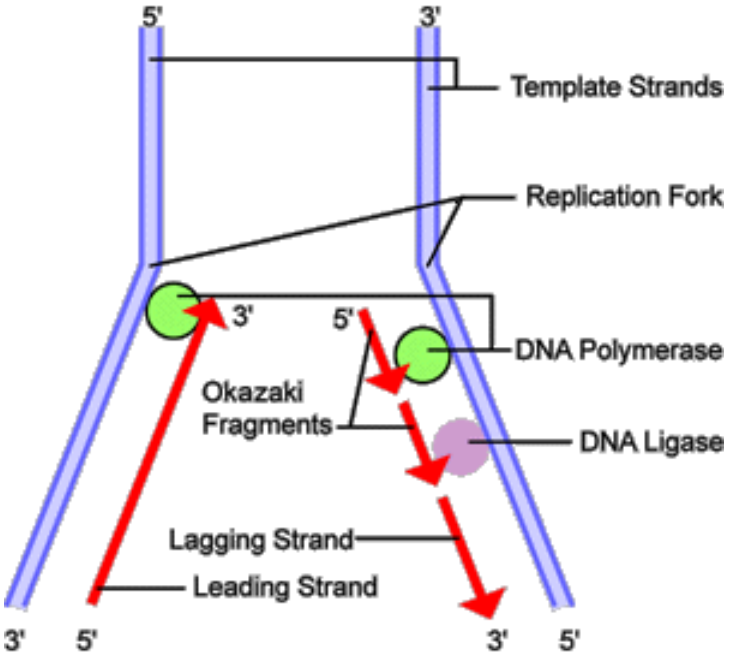
http://www.accessexcellence.org/RC/VL/GG/central.html

DNA – the code of life



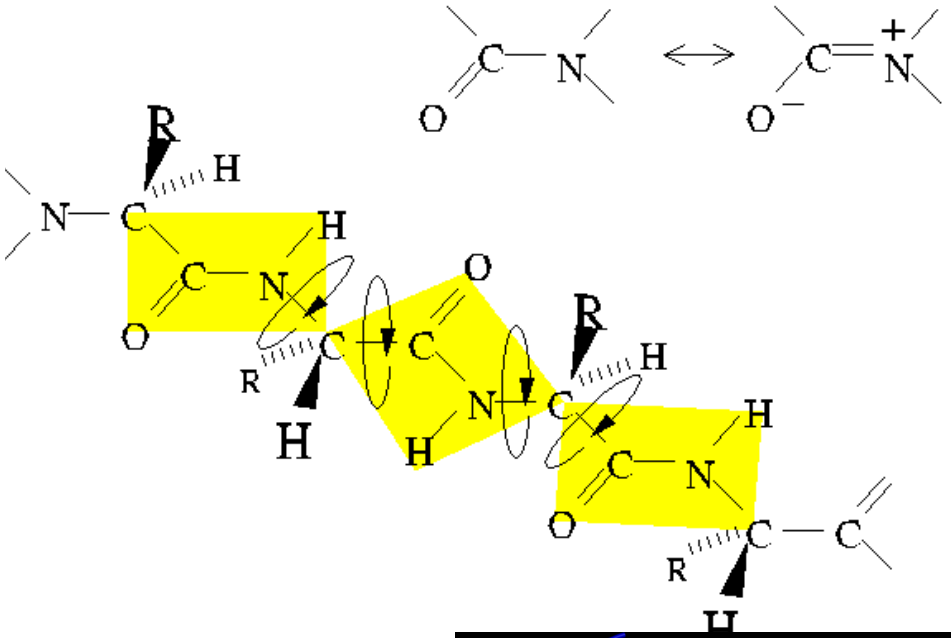
- Purines A, G, caffeine
- Pyrimidines C, T
- Sugar backbone (ticker tape)

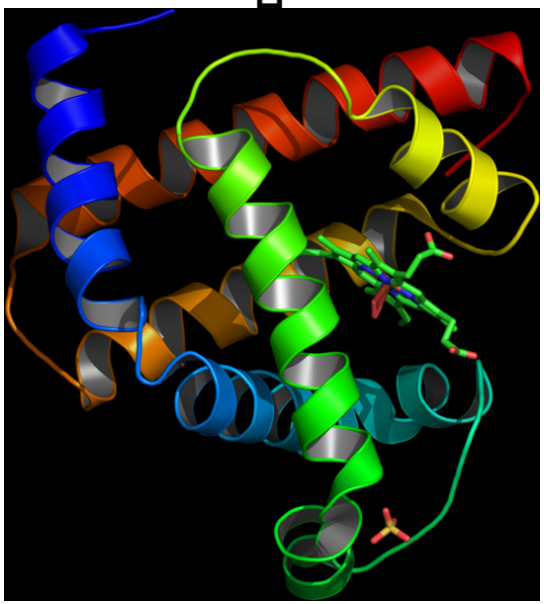
- Double-stranded – allows replication



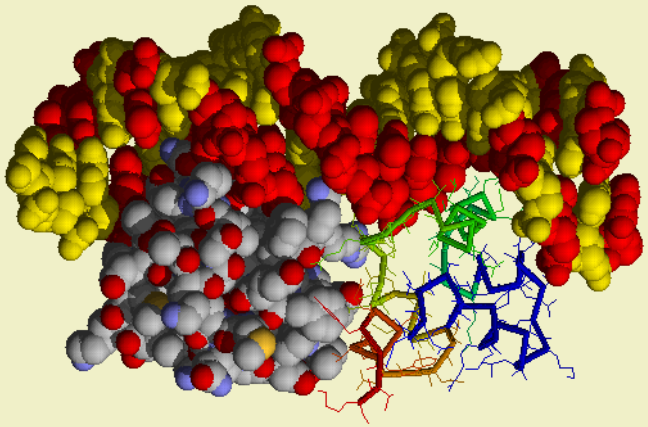
pictures from wikipedia

Protein strings





Phage CRO Repressor on DNA. Andrew Coulson & Roger Sayle with RasMol, University of Edinburgh, 1993



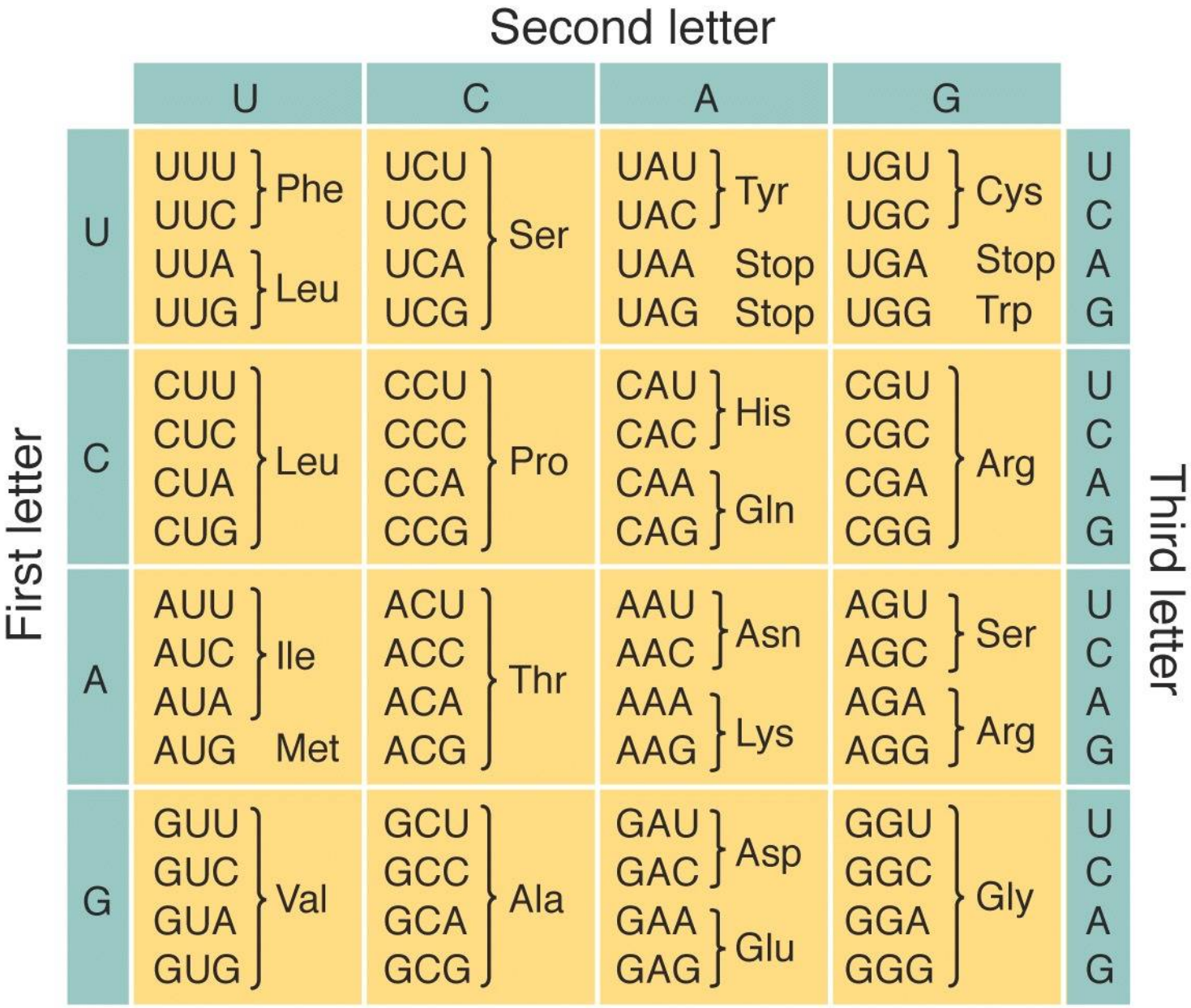
Genes, transcription, translation

- DNA – RNA - Thymine replaced by Uracil (T-U)
- The transcribed segments are called genes

ACCGUACC**AUG**UUA...AUAGGC**UGA**GCA

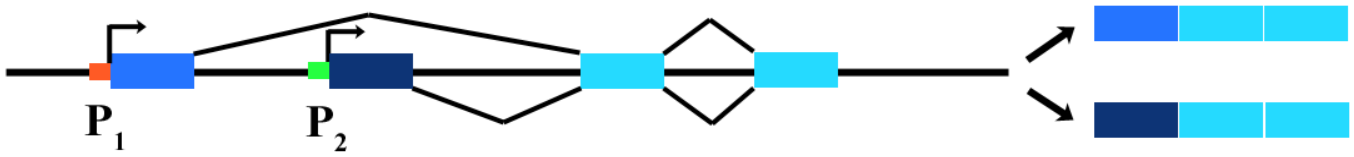
- AUG – start codon (also amino-acid Methionine)
- UAA, UAG, UGA – stop codons
- Genes are read in sets of 3 nucleotides during translation – $4^3 = 64$ possible combinations
- Each combination codes for one of 20 amino-acids – the building blocks for proteins

Amino-acid translation table

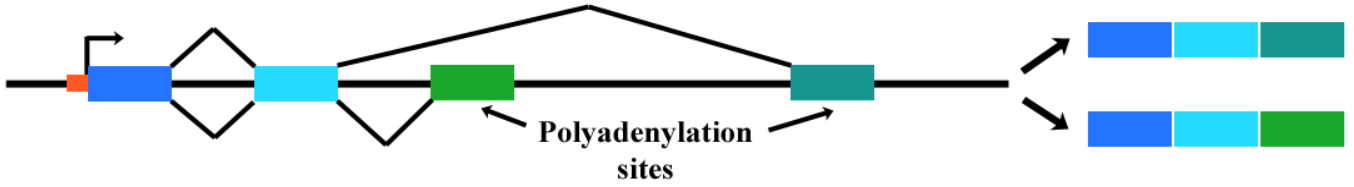


Alternative splicing examples

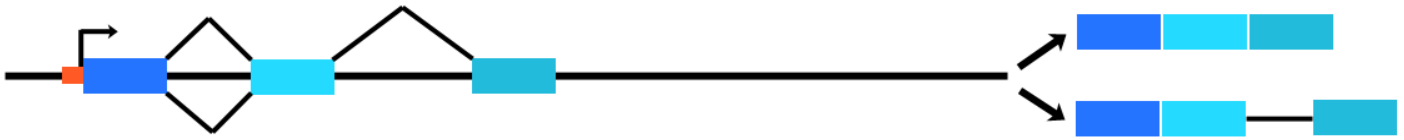
(a) Alternative selection of promoters (e.g., *myosin* primary transcript)



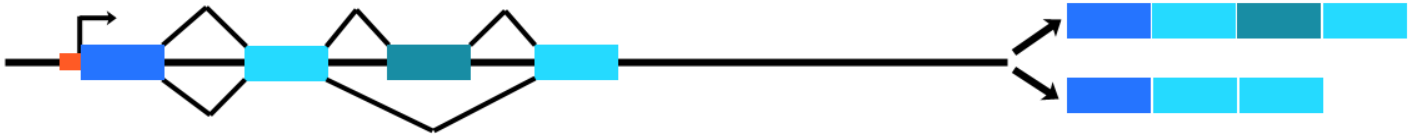
(b) Alternative selection of cleavage/polyadenylation sites (e.g., tropo*myosin* transcript)



(c) Intron retaining mode (e.g., *transposase* primary transcript)



(d) Exon cassette mode (e.g., *troponin* primary transcript)



Recap

- DNA – string of A,C,T,G.
- Base complementarity – A pairs with T, G with C
- The 2 strands of DNA are in reverse complement order – inverted w.r.t. each other and obey base complementarity
- Protein – string of amino acids
- Protein shape determines function
- Matching protein strings is tricky because aminoacids can be substituted without changing function
- Within genes – third codon can be changed without changing the amino-acid sequence

String matching questions in biology

- Compare human genomes
  - Single Nucleotide Polymorphisms (SNPs)
  - Copy number variants
  - Structural variants
- Find conserved regions (DNA or protein hasn't changed much during evolution)
- Find related sequence in a database
- Motif finding – find short patterns that occur more often than expected in "interesting" regions
- Multiple alignment of bacteria
- Phylogenetic analysis
- Sequence clustering
- Match RNA sequences to corresponding region in genome

Aren't these problems solved?

- Yes, but not efficiently enough

- Amount of data generated is rapidly increasing
- Amount of data available to scientists is rapidly increasing

- New questions keep popping up
- The characteristics of the underlying data keep changing

The evolution of DNA sequencing

Since	Technology	Read length	Throughput/run	Throughput/ hour	cost/run
1977-	Sanger sequencing	> 1000bp	4hr 400-500 kbp	100 kbp	$200
2005-	454 pyrosequencing	250-400bp	4hr 100-500 Mbp	25-100 Mbp	$13,000
2006-	Illumina/Solexa	50-100bp	~5 days ~12 Gbp	~100 Mbp	$15,000
2007-	ABI SOLiD	35-50bp	3 days 6-20 Gbp	75-250 Mbp	est. $3-5,000
2008-	Helicos **single molecule**	25-50 bp	8 days 10 Gbp	~50 Mbp est. 1Gbp/hour	~$18,000
2012	Pacific Biosciences **single molecule**	10-12 kbp	3 hours 3 Mbp	1-3 Mbp	$2,500
2014?	Oxford Nanopore	?	?	?	?