PHYS 4xx DNA - Transcription and Replication

Copies of DNA needed in two roles:

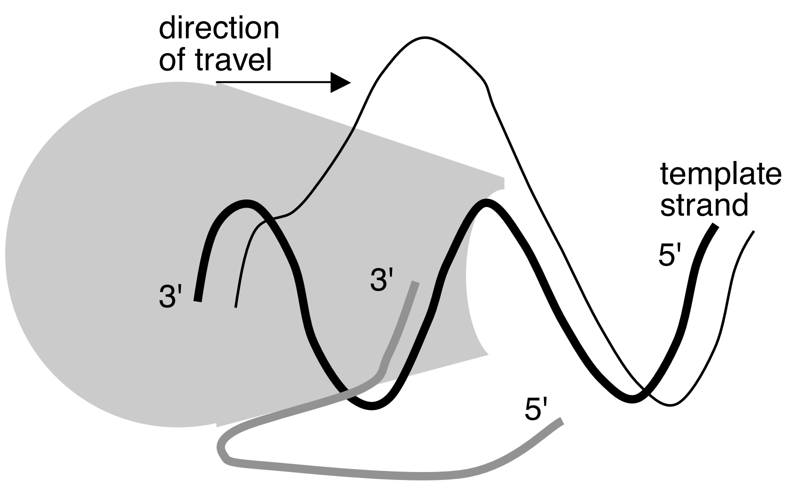
- segments *transcribed* to make mRNA (subsequently, mRNA is *translated* to make a protein)
- entire sequence *replicated* in preparation for cell division

DNA transcription

- as a construction blueprint, segments of DNA are copied to make messenger RNA, or mRNA, through **DNA transcription**
- about 1000 bp of DNA needed to code for an average protein, although the real sequence may be much longer because of the presence of non-coding sections called *introns*
- making a copy involves unwinding the helix in a local region, but much less than the few thousand bp needed for a typical mRNA

RNA polymerase (mass ~ 500 kDa) is the enzyme that directs production of RNA. In bacteria:

- following a random collision, RNA polymerase sticks lightly to DNA and slides along it
- binds tightly to DNA when it encounters a **promoter** sequence
- unwinds DNA helix as it moves along, then rewinds the helix when complete



- reads DNA from 3' to 5' end (defined in DNA Lec. 1), produces RNA from 5' to 3':

RNA 5' IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII3'
DNA 3' IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII 5'

- note: when a gene sequence is quoted for DNA, it is given from 5' to 3' on the DNA template strand, corresponding to 3' to 5' on the RNA.

- two strands in the DNA helix bound according to
  A ↔ T G ↔ C
  e.g.
  AATTTGCGCGTTAGAGACCTG
  TTAAACGCGCAATCTCTGGAC

- translates into RNA according to
  A ↔ U G ↔ C
  e.g.
  DNA template AATTTGCGCGTTAGA
  RNA copy UUAAACGCGCAAUCU

Where to start copying? In bacterial DNA, the consensus sequence in the promoter region (clearly in advance of the transcription region) is

```
                        -35 region                          -10 region          +1
          5'  NNNNNNNTTGACANNNNNNNNNNNNNNNNNTATATTNNNNNNA
(template)  3' NNNNNNNAACTGTNNNNNNNNNNNNNNNNNATATAANNNNNNT
```

At 0.34 nm per base pair, the -35 site is 10 nm from the +1 site, which is about the length of DNA polymerase. Where to stop? Again, in bacterial DNA

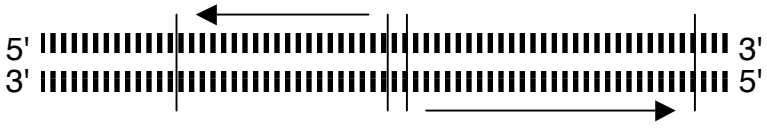
```
          5'  CCCACAGCCGCCAGTTCCGCTGGCGGCATTTTAACTTTC
(template)  3'  GGGTGTCGGCGGTCAAGGCGACCGCCGTAAAATTGAAAG

transcribed  5' CCCACAGCCGCCAGUUCCGCUGGCGGCAUUUUAACUUUC
  RNA
```

Physically, what does the tail end look like? RNA binds to itself, and forms a loop:

```
5' CCCACAGCCGCCAGUU
                 ||||||||||||||||||||    C
3' AUUUUACGGCGGUCGC
```

Not all of the DNA codes for RNA. Further, coding regions may be on either strand, although they do not overlap.



- transcription speed is 30 nucleotides per second at 37 °C (50 s^{-1} in *E. coli*)
- typical strand of primary RNA, 3000-4000 nucleotides including introns in a eucaryotic cell, takes 2 minutes; the excess RNA (only about 1200± nucleotides are needed to make the protein) is removed in the nucleus before release of the mRNA
- after releasing the primary RNA, the RNA polymerase may continue to transcribe useless RNA from non-coding DNA for hundreds of nucleotides in eucaryotes

Problems:
- how to ensure fidelity
- how to correct for gaps

The device by which successive tRNA molecules are linked and release their attached AAs (following the mRNA code) is a **ribosome**, a large complex of rRNA (r for ribosomal) and protein (mol. mass ~ 2-4 x 10^6 Da). The decoding process does not start right at the end of mRNA, rather, the end orients the ribosome as it attempts to find the start point.

But how does the ribosome know where to start? Not only may there be many "correct" AUG triplets coding for Met (the start codon), there may be a huge number of "accidental" AUG triplets arising from juxtaposed codons for other AAs.

In eucaryotes:
- a cap is added to the 5' end of mRNA in the nucleus; about 200 adenine residues are added to the 3' end
- ribosome recognizes 5' cap, then walks along mRNA until it finds an AUG
- only one protein coded per mRNA

binding site AUG coding region stop

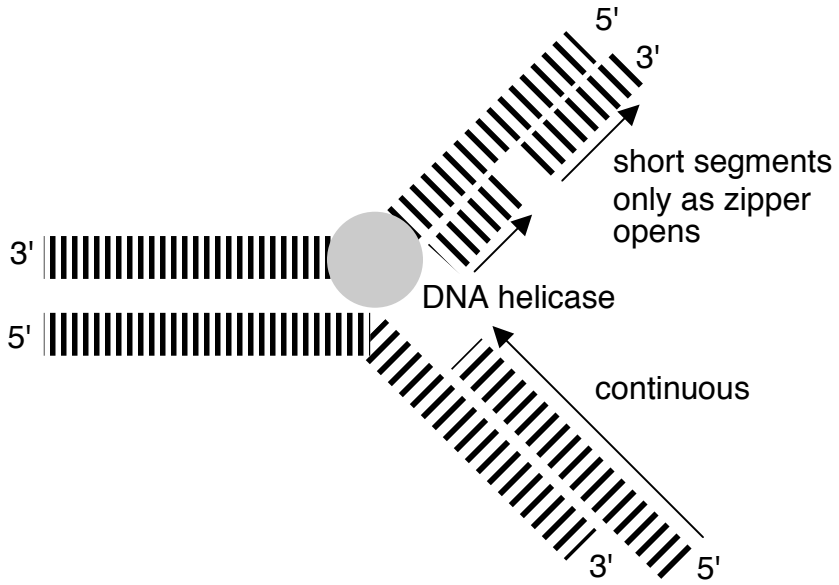
In procaryotes
- no 5' cap
- several proteins per mRNA -> several AUG starting points

binding site AUG coding region stop

DNA replication

replication rates are higher than transcription rates:
transcription - 30 nucleotides per second
replication - 50 nucleotides per second in mammals
500 nucleotides per second in bacteria
If replication proceeded at 50 nucleotides per second, then it would take
$3 \times 10^9 / 50 = 7 \times 10^7$ seconds = 2 years
to replicate a single strand of human DNA
HENCE, replication is not done linearly from one end to the other.

Mechanism
- "bubble" opens in the DNA zipper, created by DNA helicase
- the bubble ends at *replication forks*, both forks are active in bacteria (no counterexamples yet)
- replication performed by DNA polymerase (sliding ring keeps enzyme on track
- DNA polymerase works in one direction only: 3' to 5' on template strand
- copying at each fork is therefore asymetric:
  3' end is leading strand (fast, continuous)
  5' end is lagging strand (slower, performed in short segments)



- short segments of DNA on the lagging strand are referred to as Okazaki fragments; the small gaps between them must be filled in after the fact
- problem with unwinding is supercoiling (like a telephone cord); supercoiling may be relieved by DNA topoisomerase I, which clips one strand of the helix and allows it to rotate, releasing the torsional strain

DNA repair

DNA is damaged continuously:
- in transcription and replication
- by thermally activated reactions, even if otherwise inert

successful mutation rates are about **1 mutation per gene per 200,000 years**
- study the fossil record to determine when species diverged (evolutionary time is twice the elapsed time, since both species evolved
- compare DNA or protein sequences between existing species
- some proteins "evolve" faster than others, in that they can tolerate a higher variation without the host dying (*e.g.* tail end of fibrinogen is functionally discarded, therefore can tolerate greater variation)
- this rate is consistent with lab studies of 1 error per 10^9 base pairs per cell generation

unsuccessful mutation rates are much higher, as the affected individual is removed by natural selection

error rate is thought to limit an organism to 50,000 genes

most mechanisms for gene repair rely on there being 2 strands of DNA.