1 Some Molecular Biology

1.1 Basics

- Four bases: adenine (A), cytosine (C), quanine (G), and thymine (T)
- Two complementary base pairs: A — T and G — C
- DNA is double stranded, RNA is single stranded
  - Written from the 5′ to 3′ direction, e.g. 5′ ACCTGAC 3′

1.2 Central Dogma

- Central dogma of information flow, from Crick (1958)

  > The central dogma states that once 'information' has passed into protein it cannot get out again. The transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein, may be possivle, but transfer from protein to protein, or from protein to nucleic acid, is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein.

- DNA → RNA → Protein
  - DNA → DNA = replication
  - DNA → RNA = transcription
  - RNA → Protein = translation

1.3 The Genetic Code

- 20 amino acids are used to create protein sequences, as determined by Crick.
  - Assume sequences are read as blocks of nucleotides, and cannot be less than 3 letters long (since 4 and $4^2 < 20$ and $4^3 > 20$).
  - Therefore, if amino acids are encoded by triplets of nucleotides (called *codons*), and the code is comma–free, each reading frame is:

    $$\underbrace{x_1x_2x_3}_{\mathrm{R}_1}\ \underbrace{x_4x_5x_6}_{\mathrm{R}_2}\ \underbrace{x_7x_8x_9}_{\mathrm{R}_3}$$

    and not

    $$x_1\ \underbrace{x_2x_3x_4}_{\mathrm{R}_1}\ \underbrace{x_6x_7x_8}_{\mathrm{R}_2}\ \underbrace{x_8x_9x_{10}}_{\mathrm{R}_3}$$

    etc…
  - Each nucleotide needs a reading frame.
  - All possible amino acid sequences are possible. However, AAA, TTT, GGG, and CCC are not (since in AAAAAA, for example, there is no obvious reading frame, and four difference places to begin reading AAA). So we're left with $4^3 - 4 = 60$ combinations.
  - As to the others, let XYZ be a codon. to have a comma–free code, XYZXYZ must be read unambiguously. So if XYZ is a codon, YZX and ZXY cannot be. So we're left with $60 \times \frac{1}{3} = 20$.
- There are three stop codons, UAA, UAG, and UGA.
- There are many pairs of codons which code for the same amino acid that differ only in the third position, but relatively few which code for the same amino acid while differing in the first or second.
- Assume a sequence UUUUACUGCGGC…
  - There are three reading frames in the 5′ to 3′ direction and three in the opposite direction from the complementary DNA strand. So a possible of six reading frames for double–stranded DNA.
- Let $\mathbf{N} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$ be the set of nucleic acids, $\mathbf{C} = \{(x_1x_2x_3) : x_i \in N\}$, and $\mathbf{A}$ be the set of amino acids and termination codon. The genetic code is then just the map $g : \mathbf{C} \to \mathbf{A}$ (see Table **??**).

	2nd				
	U	C	A	G	
1st					3rd
U	Phe	Ser	Tyr	Cos	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	TC	TC	A
	Leu	Ser	TC	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Guy	A
	Val	Ala	Glu	Gly	G

Table 1: Genetic code. Termination codons indicated by TC.

1.4 Transfer RNA and Protein Sequences

- mRNA (messenger RNA) is read to make proteins.
- Amino acids are linked to tRNA (transfer RNA), which then brings them to the mRNA by interacting with codons on the mRNA.
- Since RNA is single stranded, it tends to fold back on itself to form helical regions (Figure **??**).

Figure 1: Example of RNA folding on itself to create a hairpin.