Notes on Introduction to Computational Biology by Waterman

February 12, 2019

Contents

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 $\mathbf{N} = \{\text{A, C, G, U}\}$ 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 1).

	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 1).

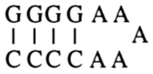

Figure 1: Example of RNA folding on itself to create a hairpin.

- tRNA often makes cloverleaf structures, with an *anticodon* at the bottom (Figure 2).
  - This anticodon is complimentary to the codon on mRNA that the tRNA brings amino acids to.
- mRNA is read at the ribosome, which then uses tRNA to create proteins.

1.5 Genes Are Not Simple

1.5.1 Start and Stop

- Three codons which code for ‘‘stop’’, one codon which codes for ‘‘begin’’ (AUG, coding for Met).
- RNA polymerase, which reads DNA and transcribes mRNA, binds to *promoter sequence* on DNA which sets it up to read down the sequence to the gene.
  - Binds to two sequences: TTGACA (-35) and TATAAT (-10), with numbers indicating nucleotide distance from start codon
  - Note: there is no position ‘‘0’’ when numbering sequences.

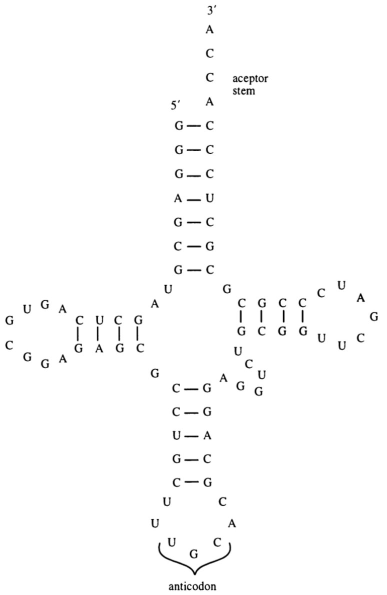

Figure 2: tRNA in cloverleaf structure, showing the anticodon on the bottom leaf.

1.5.2 Gene Expression Control

- Expression can be controlled at one of two points: DNA → RNA or RNA → Protein.
- Example of RNA → DNA control:
  - A molecule called a *repressor* binds to DNA stopping the expression of a protein which processes a particular molecule.
  - When the particular protein is present, it binds to the repressor and prevents it from binding to DNA.
  - When all of the molecule has been used in whatever process it's used in, the repressor is then free to bind to the DNA and stop transcription of its processor protein.

1.5.3 Split Genes

- DNA encoding proteins are interrupted by noncoding DNA which disappears in mRNA.
  - For example, a gene E is expected to be one contiguous region, but are split by noncoding regions I_1 and I_2 to form a noncontinuous gene E_1, E_2, and E_3
  - The split gene sections (E) are called *exons* and the noncoding regions (I) are called *introns*.
  - Introns are separated out during transcription, and the exons become a contiguous gene on mRNA.
- The vast majority of DNA is non–encoding – in humans, only 5% is used in protein coding

1.5.4 Jumping Genes

- There exist in both prokaryotic and eukaryotic genomes genes which move from place to place in the genome.
  - Called *transposable elements*.
  - Carry genes required for transposition, and are called ‘‘jumping genes’’.
- These can propagate themselves, creating identical segments of DNA
- Function is not known, some speculate they're ‘‘selfish’’ genes which exist only to propagate themselves, much like miniature organisms (fucking rad)

1.6 Biological Chemistry

- Skipping most of this, as it's just chemistry.

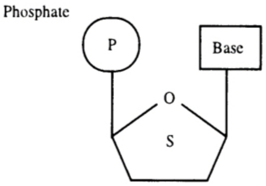

Figure 3: Nucleotide template.

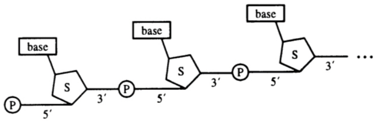

Figure 4: Single strand of DNA showing the nucleotides connected by phosphate bonds.

1.6.1 DNA

- DNA is composed of four different subunits, or *nucleotides*.
  - *Adenine*, *guanine*, *thymine*, and *cytosine*. Adenine and guanine have a purine as a base, and thymine and cytosine have pyrimidines as a base.
  - Each nucleotide is composed to a phosphate group and a base attached to a five–membered, oxygen containing ring (Fig. 1.6.1). In DNA, this sugar is 2–deoxyribose
- A single strand of DNA is formed by phosphate bonds connecting the sugars of the nucleotides (Fig 1.6.1).

1.6.2 RNA

- Ribonucleic acid (RNA) is quite similar to DNA, except its sugar is ribose, not 2–deoxyribose.
- The nucleotide thymine is replaced with uracil.