Lesson 1

A false start

After the famous discovery of the double helical structure of DNA by Watson, Crick and Franklin [Watson and Crick, 1953], attention turned immediately to determining the genetic code. At the time it was known that there are four nucleotides and twenty amino acids, but it was not clear how protein synthesis proceeded from DNA. The first ideas about how the code might work were based on suggestions of George Gamow to Francis Crick, who noted that while $20 > 4^2$, it was the case that $20 < 4^3$. He therefore postulated that the code might be based on DNA triplets. Gamow devised a beautiful "diamond code", which could explain protein synthesis. The diamond code was subsequently shown in to be "optimal" in a certain sense. We begin with this story for two reasons: the discovery of the genetic code is arguably the origin of computational genomics, and its discovery started with beautiful mathematics.

However beautiful mathematics do not always translate to a correct explanation of a biological phenomenon. Unfortunately, the diamond code of Gamow, while mathematically elegant, did not correspond to biological reality, and the optimality properties of the code turned out to be irrelevant. This is our first lesson in computational genomics: the genome is full of interesting and beautiful patterns, and it can be tempting to practice genome astrology. But the genetic code was not cracked by reading horoscopes. It was deciphered via a careful series of experiments which were collectively honored with a Nobel prize in physiology and medicine awarded to Har Gobind Khorana, Robert Holley and Marshall Nirenberg in 1968.

The genetic code continues to be a source of fascination for mathematicians. Its elegant simplicity and redundancy have been explored in depth. Recurring questions have been "has the code been optimized for something? If so, what?". In this lecture we begin to examine these questions with a false start.

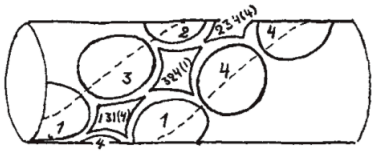

Fig. 1

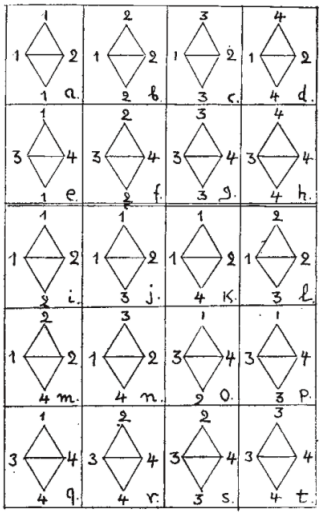

Fig. 2

Figure 1.1: The diamond code.

1.1 Codes without commas

Just a year after the publication of James Watson and Francis Crick's paper on the structure of DNA, the physicist George Gamow published a short paper in the journal Nature proposing a genetic code and translation mechanism [Gamow, 1954]. Gamow's idea, illustrated in Figure 1.1, is known as the "diamond code". It hypothesized that amino acids interlock inside "holes" formed by the DNA double helix.

Gamow's hypothesis was examined in more detail in a thought provoking article published in 1957 [Crick *et al.*, 1957], in which Francis Crick considered the question of how four objects (nucleotides) can be used to code for sequences made up of twenty objects (amino acids). Together with J.S. Griffith and L.E. Orgel, he focused on the overlapping property of Gamow's code, and formalized it with the notion of a *comma free code*, which is defined as follows:

Definition 1.1 Given finite alphabets Σ, a positive integer k and a subset $S \in \Sigma^k$, S is a comma free code if for any word w in a string $\sigma \in \Sigma^n$ $(n \geq k)$, $w \in S$ if the starting position of w is 1 (mod k) and $w \notin S$ otherwise.

Example 1.2 The following example is from [Crick *et al.*, 1957]: Let $\Sigma =$

$\{A, B, C, D\}$ and $k = 3$. The set

$$\begin{aligned} S \quad = \quad & \{ABA, ABB, ACA, ACB, ACC, BCA, BCB, BCC, ADA, \\ & ADB, ADC, ADD, BDA, BDB, BDC, BDD, CDA, CDB, \\ & CDC, CDD\} \end{aligned}$$

is a comma free code.

Crick, Griffith and Orgel thought that the genetic code might be a comma free code, because of the following theorem that they proved:

Theorem 1.3 ([Crick *et al.*, 1957]) *The largest comma free code for* $|\Sigma| = 4$ *and* $k = 3$ *has twenty elements.*

In other words, the largest comma free code that can be constructed from four nucleotides has size twenty, suggesting that this may be *why* there are twenty amino acids.

Exercise 1.4 Identify problems with Gamow's biophysical explanation for translation, and provide as many arguments as possible for why the genetic code is not, in fact, overlapping.

Exercise 1.5 Given $|\Sigma| = n$ and a positive integer k, let $M(n, k)$ be the largest size of a comma free code for Σ, k. Show that

$$M(n, k) \leq \frac{1}{k} \sum_{d|k} \mu(d) n^{\frac{k}{d}}, \tag{1.1}$$

where $\mu(d)$ is the Möbius function.

It is conjectured that $M(n, k) \sim \frac{k^n}{n}$.

Exercise 1.6 How many comma free codes are there for $|\Sigma| = 4, k = 3$?

The genetic code was finally solved experimentally, and in retrospect its clear that [Crick *et al.*, 1957] were led astray because of their focus on the number of amino acids as a clue for the mechanism of the genetic code. Ironically, the number *twenty* of amino acids that they were working with is also wrong!

Exercise 1.7 Who discovered selenocysteine? What is it and what is its biological role?

`GenBank` is a database containing freely available sequenced genomes [Benson and others, 2012]. The genomes can be directly downloaded from the web.

	T	C	A	G
T	TTT ↦ Phe	TCT ↦ Ser	TAT ↦ Tyr	TGT ↦ Cys
	TTC ↦ Phe	TCC ↦ Ser	TAC ↦ Tyr	TGC ↦ Cys
	TTA ↦ Leu	TCA ↦ Ser	TAA ↦ *stop*	TGA ↦ *stop*
	TTG ↦ Leu	TCG ↦ Ser	TAG ↦ *stop*	TGG ↦ Trp
C	CTT ↦ Leu	CCT ↦ Pro	CAT ↦ His	CGT ↦ Arg
	CTC ↦ Leu	CCC ↦ Pro	CAC ↦ His	CGC ↦ Arg
	CTA ↦ Leu	CCA ↦ Pro	CAA ↦ Gln	CGA ↦ Arg
	CTG ↦ Leu	CCG ↦ Pro	CAG ↦ Gln	CGG ↦ Arg
A	ATT ↦ Ile	ACT ↦ Thr	AAT ↦ Asn	AGT ↦ Ser
	ATC ↦ Ile	ACC ↦ Thr	AAC ↦ Asn	AGC ↦ Ser
	ATA ↦ Ile	ACA ↦ Thr	AAA ↦ Lys	AGA ↦ Arg
	ATG ↦ Met	ACG ↦ Thr	AAG ↦ Lys	AGG ↦ Arg
G	GTT ↦ Val	GCT ↦ Ala	GAT ↦ Asp	GGT ↦ Gly
	GTC ↦ Val	GCC ↦ Ala	GAC ↦ Asp	GGC ↦ Gly
	GTA ↦ Val	GCA ↦ Ala	GAA ↦ Glu	GGA ↦ Gly
	GTG ↦ Val	GCG ↦ Ala	GAG ↦ Glu	GGG ↦ Gly

Table 1.1: The genetic code. Each triplet is called a *codon*. There are $4^3 = 64$ codons of which 3 are called *stop codons* as they terminate translation. Two codons are called *synonymous* if they code for the same amino acid.

Exercise 1.8 Visit GenBank and determine how many bacterial genomes are currently available for download. Download one of the genomes and determine its codon usage.

Exercise 1.9 Explain how you might learn the genetic code directly from data in `GenBank`.

There are many theories about how selection may have led to the mechanism and specification of the genetic code. One theory posits that the genetic code is "optimal" in that it is most flexible in allowing for arbitrary sequences to appear within coding regions [Itzkovitz and Alon, 2007]. The implication is that the coding sequences can serve a dual purpose: both coding for protein sequences but also containing regulatory sequences allowing proteins to bind to the DNA for regulatory purposes.