Chapter 13

All You Need to Know about Molecular Biology

Well, not really, of course, see Lewin, 1999 [220] for an introduction.

DNA is a string in the four-letter alphabet of *nucleotides* A, T, G, and C. The entire DNA of a living organism is called its *genome*. Living organisms (such as humans) have trillions of cells, and each cell contains the same genome. DNA varies in length from a few million letters (bacteria) to a few billion letters (mammals). DNA forms a helix, but that is not really important for this book. What is more important is that DNA is usually double-stranded, with one strand being the *Watson-Crick complement* (T pairs with A and C pairs with G) of the other, like this:

$$\begin{array}{ccccccccc} A & T & G & C & T & C & A & G & G \\ | & | & | & | & | & | & | & | & | \\ T & A & C & G & A & G & T & C & C \end{array}$$

DNA makes the workhorses of the cell called *proteins*. Proteins are short strings in the *amino acid* 20-letter alphabet. The human genome makes roughly 100,000 proteins, with each protein a few hundred amino acids long. Bacteria make 500—1500 proteins, this is close to the lower bound for a living organism to survive. Proteins are made by fragments of DNA called *genes* that are roughly three times longer than the corresponding proteins. Why three? Because every three nucleotides in the DNA alphabet code one letter in the protein alphabet of amino acids. There are $4^3 = 64$ triplets (*codons*), and the question arises why nature needs so many combinations to code 20 amino acids. Well, genetic code (Figure 13.1) is redundant, not to mention that there exist *Stop* codons signaling the end of protein.

Biologists divide the world of organisms into *eukaryotes* (whose DNA is enclosed into a nucleus) and *prokaryotes*. A eukaryotic genome is usually not a single string (as in prokaryotes), but rather a set of strings called *chromosomes*. For our purposes, the major difference to remember between prokaryotes and eukaryotes is that in prokaryotes genes are continuous strings, while they are broken into pieces (called *exons*) in eukaryotes. Human genes may be broken into as many as 50 exons, separated by seemingly meaningless pieces called *introns*.

A gene broken into many pieces still has to produce the corresponding protein. To accomplish this, cells have to cut off the introns and concatenate all the exons together. This is done in *mRNA*, an intermediary molecule similar to short, single-stranded DNA, in a process called *transcription*. There are signals in DNA to start transcription that are called *promoters*. The protein-synthesizing machinery then *translates* codons in mRNA into a string of amino acids (protein). In the laboratory, mRNA can also be used as a template to make a complementary copy called *cDNA* that is identical to the original gene with cut-out introns.

Second position

First position	T		C		A		G	
T	TTT	**PHE**	TCT	**SER**	TAT	**TYR**	TGT	**CYS**
	TTC		TCC		TAC		TGC	
	TTA	**LEU**	TCA		TAA	**Stop**	TGA	**Stop**
	TTG		TCG		TAG		TGG	**TRP**
C	CTT	**LEU**	CCT	**PRO**	CAT	**HIS**	CGT	**ARG**
	CTC		CCC		CAC		CGC	
	CTA		CCA		CAA	**GLN**	CGA	
	CTG		CCG		CAG		CGG	
A	ATT	**ILE**	ACT	**THR**	AAT	**ASN**	AGT	**SER**
	ATC		ACC		AAC		AGC	
	ATA		ACA		AAA	**LYS**	AGA	**ARG**
	ATG	**MET**	ACG		AAG		AGG	
G	GTT	**VAL**	GCT	**ALA**	GAT	**ASP**	GGT	**GLY**
	GTC		GCC		GAC		GGC	
	GTA		GCA		GAA	**GLU**	GGA	
	GTG		GCG		GAG		GGG	

Figure 13.1: Genetic code.

Over the years biologists have learned how to make many things with DNA. They have also learned how to copy DNA in large quantities for further study. One way to do this, *PCR* (polymerase chain reaction), is the Gutenberg printing press of DNA. PCR amplifies a short (100 to 500-nucleotide) DNA fragment and produces a large number of identical strings. To use PCR, one has to know a pair of short (20 to 30-letter) strings flanking the area of interest and design two

PCR primers, synthetic DNA fragments identical to these strings. Why do we need a large number of short identical DNA fragments? From a computer science perspective, having the same string in 10^{18} copies does not mean much; it does not increase the amount of information. It means a lot to biologists however, since most biological experiments require using a lot of strings. For example, PCR can be used to detect the existence of a certain DNA fragment in a DNA sample.

Another way to copy DNA is to *clone* it. In contrast to PCR, cloning does not require any prior information about flanking primers. However, in cloning, biologists do not have control over what fragment of DNA they amplify. The process usually starts with breaking DNA into small pieces. To study an individual piece, biologists obtain many identical copies of each piece by *cloning* the pieces. Cloning incorporates a fragment of DNA into a *cloning vector.* A cloning vector is a DNA molecule (usually originated from a virus or DNA of a higher organism) into which another DNA fragment can be inserted. In this operation, the cloning vector producing an does not lose its ability for self-replication. Vectors introduce foreign DNA into host cells (such as bacteria) where they can be reproduced in large quantities. The self-replication process creates a large number of copies of the fragment, thus enabling its structure to be investigated. A fragment reproduced in this way is called a *clone.* Biologists can make *clone libraries* consisting of thousands of clones (each representing a short, randomly chosen DNA fragment) from the same DNA molecule.

Restriction enzymes are molecular scissors that cut DNA at every occurrence of certain words. For example, the *Bam*HI restriction enzyme cuts DNA into *restriction fragments* at every occurrence of GGATCC. Proteins also can be cut into short fragments (called *peptides*) by another type of scissors, called *proteases.*

The process of joining two complementary DNA strands into a double-stranded molecule is called *hybridization.* Hybridization of a short *probe* complementary to a known DNA fragment can be used to detect the presence of this DNA fragment. A probe is a short, single-stranded, fluorescently labeled DNA fragment that is used to detect whether a complementary sequence is present in a DNA sample. Why do we need to fluorescently label the probe? If a probe hybridizes to a DNA fragment, then we can detect this using a spectroscopic detector.

Gel-electrophoresis is a technique that allows biologists to measure the size of DNA fragments without sequencing them. DNA is a negatively charged molecule that migrates toward a positive pole in the electric field. The speed of migration is a function of fragment size, and therefore, measurement of the migration distances allows biologists to estimate the sizes of DNA fragments.