# 1 Basic Molecular Biology for statistical genetics

## 1.1 Mendelian genetics

Although it has been recognized for thousands of years that traits are passed from organisms to their offspring, the science of genetics began with the experiments of Gregor Mendel. By crossing various strains of peas, Mendel was able to deduce several principles of genetics. The most basic is *unit inheritance*, which states that inheritance is determined by discrete quantities known as *genes* and these genes don't mix. For example, when Mendel crossed round and wrinkled peas, all of the offspring were round, the offspring were not "kind of wrinkled". In this example, we think there is some gene which controls if the pea is wrinkled or not, and the gene takes 2 values; round and wrinkled. We call the different values taken by the gene *alleles*. Peas (and humans) have 2 copies of all genes, one from each parent. For every gene, the 2 alleles are on 2 different chromosomes. Hence humans have 2 alleles for each gene. The set of 2 alleles of a gene for an individual is known as the *genotype* of the individual. In contrast, a feature of the observable being of an organism is referred to as the *phenotype* of the individual. For example, if a pea plant has the allele for round (denoted R) and wrinkled (denoted r) at the pea shape gene, the genotype is Rr. If a pea plant has this genotype, it will appear round, hence we say it has the round phenotype. Clearly, phenotype is not as well defined a concept as genotype. The term genotype is also used to refer to the alleles an individual has at a large number of loci. If we also know which pairs of alleles are on the same chromosome, then we know the individual's *haplotype*. We use the term *genome* to refer to the set of all genes of some organism.

If the 2 alleles for a gene are distinct, we say that individual is *heterozygous*, while if the 2 alleles are the same, we say the individual is *homozygous*. For many traits, homozygous genotypes translate into definite phenotypes, for the pea example, if the genotype is rr, the pea will be wrinkled (i.e. the wrinkled phenotype) and if the genotype is RR the phenotype is round. The phenotypes of heterozygous individuals are more complicated. In classical Mendelian genetics, there are 2 ways in which a heterozygous genotype translates into a phenotype and this mapping of genotype to phenotype depends on the trait. Consider the pea example, the Rr genotype corresponds to a round phenotype, hence we say round is the *dominant* trait and wrinkled is the *recessive* trait. For some traits, heterozygous genotypes imply the individual displays both phenotypes. We call such traits *codominant*. From experimentation, Mendel also deduced that alleles *segregate randomly*, that is, if an individual has 2 distinct alleles for some gene then an offspring of this individual is equally likely to obtain each allele. Mendel's final deduction was that different genes combine independently of one another (this is the principle of *independent assortment*). We will see that Mendel's principles, while useful, are not entirely correct. Traits that seem to obey Mendel's first 2 principles are referred to as *Mendelian traits*, and several hundred medical disorders seem to be Mendelian traits (i.e. attributable to a single gene which is either dominant or recessive). Cystic fibrosis is a noteworthy example. Many other traits (e.g. obesity) are far too complex to fit neatly into the Mendelian framework, hence we refer to such traits as *complex traits*.

## 1.2 Genes and chromosomes

Complex organisms have tens of thousands of genes. These genes are located on *chromosomes*, and the spots on the chromosomes where genes are located are referred to as *loci* (locus is the singular). Chromosomes are segments of deoxyribonucleic acid (DNA), and are contained in the nucleus of the cells of eukaryotic organisms (a eukaryotic organism is one whose cells have nuclei, for example, humans). Humans have 46 chromosomes, arranged in 23 pairs of chromosomes. Of these 23 pair, 1 pair is known as the sex chromosomes and the other 22 are called *autosomal* chromosomes. Organisms receive 1 chromosome from each parent, hence, for autosomal chromosomes, the pairs all have the same gene loci, but possibly different alleles. One also receives one of the sex chromosomes from each parent, but these are not necessarily pairs. There are 2 sorts of sex chromosomes, denoted X and Y. An individual with 2 X sex chromosomes is a female, while an individual with an X and a Y chromosome is a male. Since females have 2 X chromosomes, they always transmit an X chromosome to their offspring, while males contribute either an X or a Y chromosome (which is actually transmitted depends on chance), hence it is the sperm cell which determines the sex of the offspring. There are several serious medical disorders attributable to a gene which resides on the X chromosome: such disorders are referred to as *X-linked* disorders (or X-linked traits more generally). The autosomal chromosomes are numbered according to their length, with chromosome 1 being the longest, and the sex chromosomes are chromosome pair 23. A surprising feature of chromosomes in higher organisms (like humans) is that most of the DNA is not part of a gene. In addition, most genes are not composed of contiguous segments of DNA, instead there are interruptions in the coding of a gene. Such interruptions are known as *introns*, while the segments of DNA which encode a gene are called *extrons*. It is currently not clear what is the purpose of this extragenic DNA, but at least some portions of this DNA regulates the working of the genes. Often the term locus is used to refer to a location on a chromosome even if there is no gene there, and similarly, allele is used to refer to the segment of DNA that an individual has at the locus. Since chromosomes are composed of DNA, further discussion of DNA is necessary.

## 1.3 DNA

All life is regulated by deoxyribonucleic acid (DNA). Cells use DNA to make proteins, and the resulting proteins carry out the functions necessary for life. Proteins are built up from a collection of amino acids (there are 20 amino acids), but often incorporate other molecules in their final form (for example, metals and lipids). DNA is composed of a sequence of sugar molecules (2-deoxyribose) joined by phosphate groups. These sugar molecules contain 5 carbon atoms, and these are referred to by a number between 1' and 5'. To each sugar molecule, at the carbon atom numbered 1', is attached one of four bases; adenine (A), guanine (G), cytosine (C) and thymine (T). All genetic information is encoded in these 4 bases. A *nucleotide* is a combination of a sugar, phosphate and a base. Phosphate groups attach to the carbon atoms numbered 3' and 5'. The linear chain of DNA is built by phosphate groups that link the 5' carbon atom of one sugar to the 3' carbon atom of the next sugar. The bases tend to form hydrogen bonds with one another in a specific manner; A binds with T and G binds with C. Due to this pairing, for every sequence of DNA there is a complementary sequence which is defined by substituting a T for an A, an A for a T, a G for a C and a C for a G. DNA has a tendency to bind to its complementary copy and form a double helix, but the hydrogen bonds which form between the nucleotides are easily broken (which is necessary for DNA to fulfill its function). While the cell is not dividing (a period of the cell cycle known as interphase), the DNA molecule actually exists as a complex with families of proteins called histones and a group of acidic nonhistone proteins-this complex is referred to as *chromatin*. Basically, the DNA winds around the histone complex, and is further condensed into a structure called a solenoid-these structures are far more compact than free DNA and help protect it from the sometimes harsh environs of the cell. In a DNA molecule, the nucleotides are arranged in a linear fashion along a chemical backbone composed of the alternating phosphates and sugars. If you start reading the bases as you travel along a DNA molecule, then groups of 3 nucleotides are referred to as *codons*. Of course which end you start reading from makes a difference. Recall the carbon atoms are numbered with connections from carbon 5' to carbon 3', so the ends of a DNA segment are referred to as 3' and 5', and you read codons by starting at the 5' end.

## 1.4 RNA

The process whereby DNA maps to proteins is composed of several steps. Proteins are produced by ribosomes, which are organelles found outside the nucleus of the (eukaryotic) cell, hence there needs to be a way of moving the information from the nucleus to the ribosome. Ribonucleic acid (RNA) is the molecule which allows for this communication. RNA is much like DNA, except the sugar group is ribose instead of 2-deoxyribose and uracil (U) takes the place of thymine (T). These chemical differences between DNA and RNA imply that RNA doesn't have a strong tendency to form a double helix, unlike DNA. The first step in generating a protein from DNA is when premessenger RNA makes a copy of a DNA segment. After this copying (known as *transcription*), some editing of the segment takes place. Certain sequences of nucleotides indicate that a gene is about to start or end (start and stop codons), so the editing deletes the portions of DNA which are not part of a gene. This editing yields messenger RNA (mRNA) which then associates with ribosomes outside of the nucleus. At this point transfer RNA (tRNA) helps coordinate the amino acids to form a chain based on the sequence of nuclueotides present in the mRNA (this process is known as *translation*). In this way a chain of amino acids is built up in the ribosome forming a *peptide* (a sequence of amino acids). The RNA molecule is released when translation is complete, at which point it may associate with proteins that degrade the RNA molecule, or it may interact with another ribosome and produce more protein. When there is high demand for a gene product, several ribosomes may join to produce protein from the same RNA molecule, thereby generating a *polyribosome.* While we can think of DNA as a linear polymer, RNA has a tendency to fold into more complicated structures, and these structures may impact the functioning of the RNA molecule. In addition, on the 3' end of the mRNA molecule is a frequently long stretch of adenine residues known as the poly(A) tail, while the 5' end of the molecule is said to be "capped" because a methylated nucleoside binds to the first nucleotide of the molecule.

## 1.5 Proteins

Since there are 4 distinct nucleotides, there are $4^3$ possible codons. In the process of creating a protein, each codon gets transcribed into one amino acid. Since there are 64 possible codons and only 20 (major) amino acids, there is some redundancy in the *genetic code* (i.e. the mapping of codons into amino acids). This redundancy is not completely without a pattern; it is least sensitive to mistakes in the third base and most sensitive to mistakes in the first. When a stop codon is reached, translation stops, and the protein is released into the cytoplasm of the cell. At that point, proteins undergo changes in shape (conformational changes) or incorporate other molecules so that the protein can serve its function. Some proteins change their shape depending on the environment, but all seem to undergo immediate changes once released from the ribosome. Unlike DNA (which we think of as a linear segment of bases), the actual shape of a protein has a tremendous impact on the function of the protein, and the final shape in a particular environment is not easily deduced from the sequence of amino acids, although certain rules do seem to hold.

An amino acid is composed of an amino group ($NH_2$), a carboxyl group (COOH), a hydrogen atom and a residue (or side chain) which bind at a common carbon atom (denoted $C_\alpha$). The particular residue determines the particular amino acid. There are 4 groups of amino acids: basic (the residue has a positive charge), acidic (the side chain has a negative charge), uncharged polar (electrically neutral side chains) and nonpolar neutral (these are hydrophobic, i.e. they avoid interacting with water). A peptide chain grows by forming bonds (known as *peptide bonds*) between the amino group of one amino acid and the carboxyl group of the adjacent amino acid (giving off water). These peptide bonds form a backbone along which the residues reside. All of the atoms in the peptide bond (NH-CO) lie in the same plane, so the 3 dimensional structure of a peptide depends on only 2 angles per $C_\alpha$. There are only certain combinations of these 2 angles which are allowed due to the residues interfering with one another in space (i.e. *steric constraints*), and these constraints give rise to the commonly encountered 3 dimensional structures in proteins. The amino acid glycine does not obey these restrictions because its residue is so small (it is a single hydrogen atom). Two major structures built from the backbone of peptide chains (referred to as *secondary structures*) are commonly encountered; $\alpha$-helixes (which are helixes of residues) and $\beta$-sheets (in which residues line up roughly co-linearly to produce a sheet of residues). There are also structures composed from these secondary structures which have been been documented (referred to as *tertiary structures*). The 3 dimensional structure of a protein is important for its functioning since proteins are the working parts of cells, hence the structure of a protein is generally conserved in evolution. As an example, since glycine is not as constrained in its angles, it is often in tight turns in the 3 dimensional structure, hence its location needs to be conserved in random mutations if the protein is to continue its function. Other residues are important for forming certain other 3-dimensional structures, hence they need to be preferentially conserved in mutation too.

## References

Lodish et al. (2001), *Molecular Cell Biology*, Freeman.

Thompson, M., McInnes, R., Willard, H. (1991), *Genetics in Medicine*, W.B. Saunders Co., Philadelphia.

Vogel, F., and Motulsky, A., (1997), *Human Genetics*, Springer, Berlin.