'Artificial Intelligence and Molecular Biology'
Chapter 1 'Molecular Biology for Computer Scientists' (Notes)

"Computers are to biology what mathematics is to physics" – Harold Morowitz.

Evolution has three components: inheritance is the main factor in determining an organism's structure and function as the parent organism's characteristics are passed down to the children; variation is the factor which makes child organisms different from their parents; i.e. a source of it must be found in the inheritance (ex. mutation, random changes in inherited material, sexual recombination, genetic rearrangements and even viruses); selection is the factor which favours certain organisms reproducing over others; i.e. it determines which variations will possibly persist. Natural selection is founded on an organism's reproductive fitness (deals with adaption to the environment).

Evolution can be described as a search through a precisely defined space of possible organism characteristics – a single messenger molecule called deoxyribonucleic acid or DNA. This is represented in a simple, linear, four-element code but with a translation which is complex. An organism's genetic encoding is called its genotype while its group of physical characteristics is called its phenotype.

Living things can be found in every type of environment; the same being said for the molecular level of life. Similar organisms can have different chemical compositions and genetic blueprints. All of an organism's genetic material is called its genome. A genome's size is not proportional to the corresponding organism's complexity. The size varies from 5000 elements in a simple organism to more than 10^{11} elements in certain higher plants. A human's genome has around $3x10^9$ elements.

In spite of this diversity, all organisms have a certain unity to them in that they nearly all have the same basic mechanisms – cells. Even the metabolic pathways (the reactions occurring in the cell) are similar across all organisms. The coded genetic material is written in approximately the same molecular language in every organism. Evolution is the reason for both the unity (due to inheritance from common ancestors) and the diversity (due to variation and selection) of living things.

Recognisable types of plants and animals make up only around 20% of living things. Eucarya (eucaryotes) have cells containing nuclei – a specialised area in the cell holding the genetic material. Other cellular areas called organelles exist, examples being mitochondria (where respiration takes place; i.e. oxygen is used for a more efficient exchange of food into energy) and chloroplasts (where energy is obtained from sunlight). All multi-cellular (animals and plants) and many single-celled (protists and fungi) organisms are Eucarya.

An important claim which underlies the majority of biological theorizing is that all organisms appear to have evolved from a common ancestor. All evolutionary theories state that the variety of life resulted from inherited variance through a constant descendant line. The stronger the relation between two species, the more recent their organisms diverged (the more recent their common ancestor). Knowledge of the DNA sequences of many genes in multiple organisms allows direct estimates of the genetic divergence time.

A typical vertebrate has more than 200 different specialised cell types. However, all the cells in a multi-cellular organism have the exact same genetic code. The differences arise from

differences in gene expression, a process whereby gene information is used in a gene product's synthesis.

Most cells have a lot of similar qualities such as genetic material and the ability to translate genetic messages into the protein (main type of biological molecule):

- Proteins are molecules which achieve most of the living cell functions;
- Genetic material is information which is generally stored in long strands of DNA. Note that Eucaryotic DNA is grouped into X-shaped structures; i.e. chromosomes;
- Nuclei contain the genetic material of Eucaryotic cells in the form of chromatin which has a variety of long stretches of DNA bounded by nuclear proteins.

Bio-molecules regulate networks of chemical reactions and include macromolecules (proteins, carbohydrates and lipids) along with multiple small molecules. The cell's genetic material specifies the process of protein creation and the quantity and time it will occur. The resulting proteins control the functions of the cell. The genetic material in question is known as a particular macromolecule, DNA.

The simplest cell has more than a thousand bio-molecules interacting. Humans likely have more than 100,000 types of proteins specified in their genome (present in different cells). The 'machine' language which encodes a set of instructions describing living systems' objects and processes contains four letters, with the text describing a person having around $3x10^9$ characters.

Materials found in the environment of an organism are broken down and reassembled into another organism following the instructions in the genome. The child organism will have instructions similar to the parent. The basic units of matter are proteins; the basic unit of energy is a phosphate bond in the molecule adenosine triphosphate (ATP); the units of information are four nucleotides which are combined into DNA and RNA.

Approximately 70% of any cell is water. Around 4% are small molecules like sugars (one being ATP). Proteins put together 15/20% and DNA & RNA make up 2-7%. The remaining 4-7% contains cell membranes and other similar molecules.

Proteins have many roles, one of them being of switches which control whether genes are turned on or off. They are made up of amino acids. The protein's primary structure is made up of a sequence of amino acid residues and is directly coded for in the genetic material – the individual elements of a DNA molecule form triples, unambiguously specifying an amino acid. A genetic sequence maps directly into a sequence of amino acids.

Nucleic acid control biochemical action. All genetic information is stored in deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), which are polymers of four simple nucleic acid units called nucleotides. There are four nucleotides found in DNA, each one consisting of a purine (adenine A or guanine G) or a pyrimidine (cytosine C or thymine T), a sugar (deoxyribose in DNA and ribose in RNA) and one or more phosphate groups. The length of a DNA sequence is measured in kb (thousands of bases). Nucleotides are also called bases and so, since DNA consists of two complementary strands bonded together, these units are called base-pairs.

Nucleotides are normally abbreviated by their first letter and appended into sequences; example: CCTATAG. They are linked to each other in the polymer by phosphodiester

bonds; the bond being directional with a strand of DNA having a head called the 5' end and a tail called the 3' end. DNA forms a double helix; two helical/spiral-shaped strands of the polypeptide running in opposite directions and held together by hydrogen bonds. Adenine binds solely with thymine (A-T) while guanine bonds solely with cytosine (G-C). Due to these bonding rules, although one strand's sequence is completely unrestricted, the complementary strand's sequence is completely determined. This property allows high consistent copying of DNA information and transcriptions into complementary strands of RNA which directs protein synthesis (with uracil U instead of T).

DNA can take other forms apart from the double helix B-DNA. It can become Z-DNA, for example, whereby it reverses its twist direction. Such alternative forms may turn certain genes on or off.

An organism's genetic information can be kept in one or more DNA molecules/chromosomes. Diploids are sexually reproducing organisms with two alike DNA molecules physically bound together (one from each parent) in their chromosomes while haploids are similar but with single DNA molecules in their chromosomes. Humans are diploid having 23 pairs of linear chromosomes. All the genetic information of an organism is referred to as its genome.

In DNA (for example the codons GCT and GCG), most codon synonyms vary in the last nucleotide only. This is referred to as the degeneracy of the code. The code is strongly conserved over evolution however there are still differences in codon to amino acid translation when comparing various organisms; in fact, there are a few systems which use a slight different code. Codons come in triples and thus the parsing of a segment of DNA (the synthesis of protein maps from codon sequences to amino acid sequences) can start at three possible places. This problem can be compared to the decoding of an asynchronous serial bit stream into bytes where each of the parsings is called a reading frame; an open reading frame (ORF) is a parsing with long string of codons having no interfering stop codons.

It is also possible to read a DNA sequence off either strand of the double helix. The second strand is the complements of the first so a sequence can be read inverted and in the opposite direction; i.e. reading from the antisense or complementary strand. This type of message can also be parsed three ways resulting in a total of six possible reading frames for every DNA sequence.

A DNA sequence which codes for a single protein usually has introns (inserted), noncoding sequences, inserted. They are removed before the sequence is mapped into amino acids. The DNA segments which do code for the protein are exons (expressed). DNA, in addition to protein coding, holds a lot more information as every cell in the body has the same DNA but each cell type generates a different set of proteins.

Every cell has the same DNA but they code for different proteins. This is due to the difference in the regulation of the genetic machinery – the regulatory mechanisms of the process of protein coding compose a parallel system with multifactorial feedback and control structure.

Genes are either expressed/not expressed (on or off). The process of production is controlled by a collection of proteins in the nucleus of eucaryotic cells which affect which genes are expressed. Histones are proteins which are tightly bound to the DNA in eucaryotic

chromosomes and are some of the most conserved proteins in life despite billions of years of divergence in their evolution. Topoisomerases are proteins which rearrange and untangle DNA and are the next prevalent proteins in chromosomes. Most regulatory proteins recognise and bind to specific DNA sequences called control regions (border the protein regions of genes). Promoters are sequences occurring towards the 5' end (upstream) of the region which encourage protein production, enhancers are similar sequences occurring either downstream or relatively far upstream of the region while repressors are sequences which tend to prevent protein production.

Molecules which are related are called homologous. The following are some aspects of molecular evolution: point mutation is the change of a single nucleotide in a genetic sequence; gene duplication is a chromosomal rearrangement where additional copies of a gene are inserted into the genome; pseudo-genes are similar to actual genes but they are not expressed; crossover is a process where the DNA from the parents of a sexually reproducing organism forms a type of combination which is passed to the child organism.

Most mutations have little effect. For example, mutations in the third position of most codons have little effect at the protein level due to genetic code redundancy. Neutral mutations are the support of genetic drift – the phenomena accounting for the DNA differences for functionally identical proteins in different organisms. On the other hand, some point mutations may lead to death or diseases; for example: cystic fibrosis. It is very rare that a mutation end up being advantageous.

Some areas of biomedical research involve humans directly or by just their cells which are grown in the laboratory, but it should be noted that not many human cell types can live outside the body. Human cancer cells can and so they are an important research tool.

Nearly every aspect of human biology can be correlated in some organism. The following six are the main models in molecular biology: *escherichia coli* (bacterium) – other organisms' genes are inserted into its genome to produce the mentioned genes in quantity; *saccharomyces cervesiae* (eucaryotes) – the yeast helps in making copies of moderate-sized pieces of DNA using yeast artificial chromosome YAS when sequencing large amounts of DNA; *arabidopsis thaliana* (common weed) – good model for higher plants with a genome with very little repetitive DNA; *caenorhabditis elegans* (worm) – one of the simplest creatures with a nervous system allowing tracing of genetic mutation effects; *drosophila melanogaster* (common fruit fly) – used in genetics research, mainly in genetic expression and control as well as genetic programs specification; *mus musculus* (basic laboratory mouse) – identical to people in terms of biochemistry with a relatively large genome.

A group of cells having identical genomes are clones. Identical child organisms are also clones but are sometimes called cell lines. Restriction enzymes are naturally produced by bacteria to attack foreign DNA; biologists use them to cut and paste specific DNA fragments into vectors. This is only effective a fraction of the time however; as cells and vectors are small and easy to grow, the process can be applied to many of them.

Hybridisation, a technique which measures the similarity between two related DNA sequences, tests how strongly the single-stranded versions of the molecules stick together; i.e. hybridise. The more easily they come apart, the larger the amount of differences there are between their sequences.

The Human Genome Project is the effort to produce a map and then the sequence of the human genome. A genetic map's purpose is the identification of the location and size of all of the genes of an organism on its chromosomes. Linkage analysis is a procedure which looks at genes' relationships (phenotypes) in large numbers of matings (crosses) to identify which are generally inherited together and thus, more likely to be near each other. It is possible to determine a medium-sized piece of DNA's sequence and thus, if a gene has been mapped its area sequence can be found and as such even the protein responsible for the genetic characteristics (this relates to inherited diseases.)

The ability to divide the genome is a requirement to determine its sequence. Thus, by taking this sequencing ability to sequence many different overlapping pieces and assembling them, the sequences of large pieces of DNA can be determined. The ordering of the pieces must be known and together, they must cover the entire genome. This process may involve polymerase chain reaction PCR. PCR makes possible the rapid production of large amounts of a specific region of DNA. It exponentially amplifies entire DNA molecules or regions of it for which bracketing primers (short pieces of DNA with a specific sequence) can be generated. A collection of short (easy to synthesise), unique (unambiguous DNA) sequences spread throughout the genome must be identified to be used as primers in order to use PCR for genome mapping and sequencing. The genome sites corresponding to these sequences are sequence tagged sites STSs. The more STSs known, the finer grained the genome map provided.

An early aim of the Human Genome Project is the production of a list of STSs spread at around 100kbp intervals over the whole human genome. Any DNA region can be identified by its two bracketing STSs. These STSs can then be stored in a database to maintain large clone collections. The project may also require sequencing of all the introns and other non-coding regions of DNA. One can target only coding regions for sequencing and thus be able to identify sequences used by a cell to produce proteins at a point in time. Thus, attention can be focused on the genome parts coded for expressed proteins.

There are several databases which maintain genetic sequences: Genbank, the European Molecular Biology Laboratory nucleotide sequences database EMBL and the DNA Database Japan DDBJ. The main protein sequence database is the Protein Identification Resource PIR. There are also several databases which maintain three dimensional structures of molecules: the Protein Data Bank PDB, BioMagRes BMR, CARBBANK, Chemical Abstracts Service (CAS) Online Registry File and Cambridge Structural Database. Genetic map databases (GDB) and a database of inherited human diseases and characteristics (OMIM) are maintained at the Welch Medical Library at Johns Hopkins University.