A Class Note on DNA, Proteins, Genes and Genomes

Kun-Mao Chao[1,2,3]

[1]Graduate Institute of Biomedical Electronics and Bioinformatics
[2]Department of Computer Science and Information Engineering
[3]Graduate Institute of Networking and Multimedia
National Taiwan University, Taipei, Taiwan 106
Email: kmchao@csie.ntu.edu.tw

October 30, 2007

1 Introduction

Deoxyribonucleic acid (DNA) is the genetic material of cells. It carries information in a coded form from cell to cell and from parent to offspring. A gene is a linear array of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product (a protein or RNA molecule). When a gene is active, its information is copied first into another nucleic acid, ribonucleic acid (RNA), which in turn directs the synthesis of the gene products, the specific proteins. This lecture introduces some basic concepts of DNA, proteins, genes and genomes.

2 The Nucleic Acids: DNA and RNA

Each nucleic acid contains four types of base. DNA is made up of four similar chemicals: adenine, guanine, cytosine and thymine that are repeated millions or billions of times throughout a genome. The human genome, for example,

has about three billion pairs of bases. RNA is made of four chemicals: adenine, guanine, cytosine and uracil. The bases are usually referred to by their initial letters: A, G, C, T for DNA and A, G, C, U for RNA.

The particular order of As, Gs, Cs, and Ts is extremely important. The order underlies all of life's diversity, even dictating whether an organism is human or another species such as yeast, rice, or fruit fly, all of which have their own genomes and are themselves the focus of genome projects.

In the late 1940s, Erwin Chargaff noted an important similarity: the amount of adenine in DNA molecules is always equal to the amount of thymine, and the mount of guanine is always equal to the amount of cytosine (A = T and G = C). In 1953, based on the x-ray diffraction data of Rosalind Franklin and Maurice Wilkins, James Watson and Francis Crick proposed a model for DNA structure. The Watson-Crick model states that the DNA molecule is a double helix (two strands twisted together). The only two pairs that are possible are AT and CG. This yields a molecule in which A = T and G = C. The model also suggests that the basis for copying the genetic information is the complementarity of its bases. For example, if the sequence on one strand is AGATC, then the sequence of the other strand would have to be TCTAG - its complementary bases. For their ground-breaking theory, Watson and Crick shared the Nobel Prize in 1962.

There are several different kinds of RNA made by the cell. In particular, mRNA, messenger RNA, is a copy of a gene. It acts as a photocpoy of a gene by having a sequence complementary to one strand of the DNA and identical to the other strand. Other RNAs include tRNA (transfer RNA), rRNA (ribosomal RNA) and snRNA (small nuclear RNA). RNA is too bulky to form a stable double helix. In fact, RNA exists as a single-stranded molecule. However, regions of double helix can form where there is some base pair complementation (U and A , G and C), resulting in hairpin loops. The RNA molecule with its hairpin loops is said to have a secondary structure.

3 Proteins

The building blocks of proteins are the amino acids. Only 20 different amino acids make up the diverse array of proteins found in living things. Each protein differs according to the amount, type and arrangement of amino acids that make up its structure. The chains of amino acids are linked by peptide bounds. A long chain of amino acids linked by peptide bonds is a

polypeptide. Proteins are long, complex polypeptides.

The sequence of amino acids that makes up a particular polypeptide chain is called the primary structure of a protein. The primary structure folds into the secondary structure, which describes the path that the polypeptide backbone of the protein follows in space. The tertiary structure describes the organization in three dimensions of all the atoms in the polypeptide chain. The quaternary structure consists of aggregates of more than one polypeptide chain. The sequence of amino acids, as well as shape, is crucial to the functioning of a protein.

4 Genes

Genes are the fundamental physical and functional units of heredity. Genes carry information for making all the proteins required by all organisms. These proteins determine, among other things, how the organism looks, how well its body metabolizes food or fights infection, and sometimes even how it behaves.

A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product (a protein or RNA molecule). Expressed genes include those that are transcribed into mRNA and then translated into protein and those that are transcribed into RNA but not translated into protein (e.g., tRNA and rRNA).

How does the sequence of a strand of DNA correspond to the amino acid sequence of a protein ? This concept is explained by the central dogma of molecular biology. Information flow (with the exception of reverse transcription) is from DNA to RNA via the process of transcription, and then to protein via translation. Transcription is the making of an RNA molecule off a DNA template. Translation is the construction of an amino acid sequence (polypeptide) from an RNA molecule.

How does an mRNA specify amino acid sequence? The answer lies in the genetic code. It would be impossible for each amino acid to be specified by one nucleotide, because there are only 4 nucleotides and 20 amino acids. Similarly, two nucleotide combinations could only specify 16 amino acids. The final conclusion is that each amino acid is specified by a particular combination of three nucleotides, called a codon.

To code for the 20 essential amino acids a genetic code must consist of at least a 3-base set (triplet) of the 4 bases. If one considers the possibilities

of arranging four things 3 at a time (4x4x4), we get 64 possible code words, or codons (a 3-base sequence on the mRNA that codes for either a specific amino acid or a control word). The genetic code was broken by Marshall Nirenberg and Heinrich Matthaei, a decade after Watson and Crick's work.

5 The Genomes

A genome is all the DNA in an organism, including its genes. In 1990, the Human Genome Project was launched by the U.S. Department of Energy and the National Institutes of Health. The project originally was planned to last 15 years, but rapid technological advances have accelerated the draft completion date to 2001. The goals of the project are to identify all the genes in human DNA, determine the sequences of the 3 billion base pairs that make up human DNA, store this information in databases, develop tools for data analysis, and address the ethical, legal, and social issues that may arise from the project.

.....

Wouldn't you agree that the genomes are the largest programs written in the oldest language, and are quite adaptable, flexible, and fault-tolerant?

6 A Brief History of Genetics

See http://www.csie.ntu.edu.tw/∼kmchao/seq07fall/Introduction.ppt.

7 Milestones of Bioinformatics

See http://www.csie.ntu.edu.tw/∼kmchao/seq07fall/Introduction.ppt.

Acknowledgements

Kun-Mao Chao was supported in part by NSC grants 94-2213-E-002-018 and 95-2221-E-002-126-MY3 from the National Science Council, Taiwan.