ENGINEERING TRIPOS PART IIA

Day Month 2010 start to finish

Module 3G1

INTRODUCTION TO BIOSCIENCE

Answer all of Section A and ***two*** *questions from Section B.*

All questions carry the same number of marks.

The ***approximate*** *percentage of marks allocated to each part of a question is indicated in the right margin.*

There are no attachments.

STATIONERY REQUIREMENTS	SPECIAL REQUIREMENTS
Single-sided script paper	None

You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

Version: DRAFT 18jan10

SECTION A

Answer all of this section.

1 (a)

(i) What is a chromosome?

An organised structure of a single piece of DNA and accompanying proteins. A chromosome is essential to the survival of the cell, unlike a plasmid.

(ii) How is chromosomal DNA packaged in prokaryotes and eukaryotes?

In prokaryotes the chromosome can be circular or linear and is a dynamic supercoiled mass, unbound within the cell. In eukaryotes the linear DNA is wrapped tightly around histones to form chromatin, with a centromere in the middle and telomeres at the ends. The eukaryotic chromosomes are bound within the nucleus of the cell.

(iii) What is meant by the *genome* of an organism? [30%]

The genome is all of the hereditary information of an organism, including variation between individuals. In prokaryotes this is the chromosome and any mobile genetic elements. In eukaryotes this is all of the chromosomes in a diploid cell, excluding the mitochondrial DNA, which is considered a separate genome.

(b)

(i) How do the two strands of the DNA double-helix fit together?

The two strands run anti-parallel to each other and twist around each other to form the double-helix. Each strand is made of subunits called nucleotides, which contain one of four bases. Each base pairs uniquely with one other base: Adenine with Thymine, Cytosine with Guanine. By complementing the base sequence of one strand, the other strand forms hydrogen bonds between bases to hold the structure together.

(ii) How does this facilitate DNA replication?

Because these hydrogen bonds are relatively weak, it is easy to 'unzip' the two strands and allow DNA Polymerase to access a single

strand for replication. A and T form two hydrogen bonds; C and G form three hydrogen bonds, so are slightly more difficult to break apart. Also, the unique base complementation allows for high fidelity replication of the original DNA sequence.

(iii) Explain why the width of the DNA double-helix is constant. [30%]

The purines (A and G) are the same width; the pyrimidines (C and T) are the same width. Since a purine always pairs with a pyrimidine, the combined width will be constant throughout the double helix.

(c)

(i) What is a clone library?

A library is a set of cloned DNA fragments. The DNA fragments are typically derived from a fragmented genome (genomic library) or mRNA molecules that have been copied back to DNA (cDNA library).

(ii) When aligning biologically related sequences of different lengths, it is necessary to include gap characters in the alignment. These are often found to occur in runs. Why is this the case?

Alignment of nucleotide sequences: cDNA/genomic DNA alignments have long gaps to match the introns. In genome/ genome comparisons or cDNA/cDNA comparisons, coding regions will tend to have insertions/deletions in runs of three corresponding to insertion/deletion of amino-acids. The insertion of a transposon in one version of a genome sequences would lead to a run of gaps. Alignment of protein sequences: the core of the protein molecule is tightly packed and variations in length (insertions or deletions) compromise function. In contrast, the surface of the protein is more amenable to insertions or deletions of residues. Where such a length change does not compromise function then it usually does not matter how great the change is. This is reflected in pairwise protein sequence alignments as a run of gaps in one or other sequence.

(iii) A *C. elegans* nematode worm gene is cloned into a bacterial plasmid. The plasmid contains promoters that, when transformed into bacteria, result in the production of a double-stranded RNA version of the gene. If such bacteria are fed to the worms, what happens? [40%]

Remarkably, the double stranded RNA survives disruption of the

bacteria in the gut of the C. elegans. *It is absorbed and distributed to all cells of the worm, where it will knock-down the function of the gene.*

SECTION B

*Answer any **two** of the following questions with a short essay, using diagrams where appropriate.*

2 How do we deduce the evolutionary origin of different species? How did the eukaryotic domain emerge, and how are its members distinct from the prokarya and archaea? [100%]

Using traditional phenotypic taxonomy we can construct a skeletal map of how different species are related. By comparing the DNA sequences of different organisms we can more accurately determine how similar they are - this is phylogenetics. Often, highly conserved regions of DNA are compared, such as rRNA or mitochondrial RNA, as too many differences make similarity hard to assess. If we want to determine the relatedness of organisms within a local group, other essential or 'housekeeping' genes may be used. This technique can also be used to determine the evolution of a single gene within a group of organisms. By considering geological data, what we know about the organisms in question and the likelihood of bases changing from one to another, we can attempt to consider the timescale over which two species emerged from a common ancestor.

The eukaryotic domain emerged from the bacteria and archaea in one of three ways: the complete fusion of a bacteria and an archaea, an archaea that developed bacterial traits from the proto-mitochondrion or at the same time as all of the archaea diverged from the bacteria. The timeline for the development of key structures is also unclear. Either the archaea developed their membrane structure and then consumed the bacteria (phagotrophic hypothesis) or the archaea relied on the bacteria as a food source and grew to surround them, gaining membranes later (syntrophic hypothesis).

Eukaryotes are distinct from the other two domains of life principally because of their mitochondria, which provide energy for the cell. They have highly compartmentalised many of the cell functions with membranes, in particular the cell's DNA is kept inside the nucleus to protect it from violent chemistry. They also have developed a cytoskeleton allowing for larger cells without an exterior cell wall. Eukaryotes have specialised even further by developing multicellularity and specialised cells within a single organism.

3 A researcher has cloned a gene that expresses GFP into *E. coli.* However, despite confirming that the transformation of the DNA was successful, the bacteria do not fluoresce. Describe how the polymerase chain reaction is used to clone a gene. What are the different types of mutation that might have occurred during the PCR of this gene? Which of these are likely to be responsible for the inactivity of the clone? How would the scientist show that it was indeed a mutation that had caused the inactivity, rather than a problem with the host strain itself? [100%]

Polymerase chain reaction, or PCR, is an in vitro *method to amplify a specific sequence of DNA. First, two oligonucleotide primers are needed which will complement the two ends of the sequence to be amplified. They need to be around 20-30 base pairs long and specific to the target sequence. Then the DNA template, primers, dideoxynucleotide triphosphates (dNTPs), a thermotolerant DNA polymerase and an appropriate buffer are mixed together. By varying the temperature at which the mixture is held, the activity of the polymerase and primers can be controlled. First the DNA template strands must be separated with heat, then the primers must anneal to the template DNA to trigger the replication of the strand by the DNA polymerase. It will extend this replication all the way through to the target site for the other primer, such that each replication has produced a new primer site in addition to the old one. In subsequent cycles, the sequence between the primer sites (inclusive) is amplified at an exponential rate, creating many copies of the target sequence.*

During PCR, the most likely mutation is a point mutation caused by the inaccuracy of the DNA polymerase. This could be any of the following:

- *Silent - a base changes, but does not change the amino acid coded for*
- *Neutral - the amino acid coded for changes to one with similar properties*
- *Missense - the amino acid coded for changes to one with different properties*
- *Nonsense - the amino acid coded for changes to a stop codon*
- *Frameshift - an insertion or deletion changes many of the amino acids coded for*

The inactivity of the gene could be due to a missense, nonsense or frameshift

mutation (though the latter is less likely during PCR). To show for sure that the inactivity is due to a mutation the scientist should sequence the gene and confirm whether or not it is identical to the gene he started with. He could also test the construct in another host strain or under different conditions to see if there is a more general problem with expressing the gene.

4 What is the central dogma of molecular biology? Describe each of the processes involved. Has any violation of the central dogma been discovered? [100%]

The central dogma of molecular biology describes the transfer of information between DNA, RNA and protein. Specifically, that information generally flows from DNA to RNA to protein, and once in protein form, it cannot be transferred back to either protein or nucleic acid.

The process by which information is transferred from DNA to RNA is called transcription. In transcription, RNA polyermase, prompted by a gene's control structure (principally including its promoter), separates the strands of the DNA and then moves along one of them (the template strand), producing a complementary RNA strand version of the gene. The key difference in the structure of this messenger RNA is that Thymine is replaced with Uracil in the sequence.

The process by which information is transferred from RNA to protein is called translation. In translation, the ribosome reads along the RNA strand, constructing a protein from amino acids that are chosen on the basis of the RNA sequence. Every three bases equates to a codon that determines which amino acid is next in the sequence. The protein is built one amino acid at a time, linked together with peptide bonds. Once complete, the protein will fold itself and possibly link with other proteins before becoming a functional unit.

The other general case in the central dogma is replication, wherein DNA polymerase unzips the DNA, moves along one strand and produces a complementary DNA strand whilst the second, lagging, strand is replicated in short fragments. These fragments are ligated together and the two new strands base pair with their templates (or each other) to create two copies of the original double-stranded DNA.

There are three special cases allowed by the central dogma, though they require special circumstances or a laboratory. RNA sequence can become DNA sequence through reverse transcription. Many viruses replicate themselves by direct RNA

replication. Direct DNA to protein translation can also occur under laboratory conditions. The three impossible information transfers (protein to DNA, RNA or protein) have not been found, although prions are proteins that can cause other proteins to become more like them - this is not strictly a transfer of genetic information however.

5 Compare and contrast the evolution of the original Sanger DNA sequencing method with the latest developments in sequencing technology. [100%]

The Sanger method operates on a large population of identical molecules derived from a cloned template molecule. Synthetic oligonucleotide primers are hybridised adjacent to the cloning site and used to initiate a DNA synthesis reaction that proceeds through the cloned insert. The synthesis reaction is poisoned with chain-terminating nucleotides. The earliest version was radioactive and required 4 reactions (one per base) and 4 lanes on gel. Later, dye-labelled terminators allowed a single reaction for 4 bases and required only one lane. Later still, capillary electrophoresis was used to avoid lane tracking errors, which is important in whole genome shotgun sequencing.

Latest developments are converging on single-molecule sequencing. At the moment you start with single molecule which is amplified in a PCR-like reaction to generate a cluster of identical molecules on a slide (Illumina) or on a bead (Roche 454). "Sequencing by synthesis" is then carried out in which each base (or set of identical bases for Roche) is added and detected before addition of the next. Illumina uses reversible dyed chain termination. Roche flow each base one at a time and generate a flash of light when they are incorporated. Current machines can generated in excess of 10GBases of sequence data per run. In the near future, machines will actually sequence single molecules directly.

6 Describe the key steps in genome annotation. Consider differences between the approach taken for prokaryotes and eukaryotes. What kinds of conservation can help with genome annotation? [100%]

The objective of genome annotation is to find and record the positions of biologically meaningful sequences in the genome. For the most part this means genes, though it can also be of interest to find sequence motifs near genes that may

confer regulation. For prokaryotes, there are no introns, therefore large open reading frames (sets of adjacent amino-acid triplet codons with no translation stop codons) are immediate candidates for being genes.Genes tend to be tightly packed and can be found in co-transcribed sets with related function. For eukaryotes, the problem is harder as gene density tends to be lower, so that just looking for open reading frames is too naive. You have to allow for introns and the fact that the genome is repeat rich.

The steps are as follows:

- *Find and mask repetitive sequences (common in eukaryotes)*

- *Find genes (prokaryotes) - search for open reading frames, can also use statistical gene-finding programs. In both cases compare the resulting putative protein sequences to all known proteins to collect functional evidence.*

- *Find genes (eukaryotes) - compare genome sequence to library of repeat sequences and mask the matches. Compare masked sequence to a library of known and related mRNA transcript sequences and to all known proteins to try to find relatives of known genes. Finally, use a statistical gene finder such as genscan to find genes that are missed by the above.*

Protein-coding gene finding is hard in eukaryotes. Ribosomal RNA genes are highly conserved and thus easy to find. The tRNA gene finding program tRNAscan-SE is extremely effective at finding tRNA genes. Other classes of small regulatory RNA molecules are extremely hard to find.

Conservation of genes between organisms at the DNA or protein level means that one can identify genes during genome annotation by comparison with libraries of known mRNA and protein sequences. Conservation of gene order between organisms can allow you to find the location of genes that may have been missed during genome annotation. Regulatory elements can be conserved between organisms, therefore comparison of regions upstream of genes can lead to their identification.

END OF PAPER

Version: DRAFT 18jan10