CRIB

Version GM/final

EGT2
ENGINEERING TRIPOS PART IIA

---

2 May 2019 14.00 to 15.40

Module 3G1

INTRODUCTION TO MOLECULAR BIOENGINEERING

*Answer not more than **three** questions.*

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.*

*Write your candidate number **not** your name on the cover sheet.*

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed

10 minutes reading time is allowed for this paper at the start of the exam.

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

1 Phage display is a powerful method for isolating antibody genes of interest from high diversity libraries. The power of phage display arises from the fact that the gene encoding a particular displayed protein is carried in the phage and thus there is a physical association between the gene and the corresponding displayed protein. A related powerful technique is called mRNA display, a cell-free system in which DNA libraries are transcribed into a corresponding population of mRNA molecules and these are then translated into proteins. The association between "gene" and protein in mRNA display is made by means of a chemical linkage that fuses the mRNA and the protein.

Screening mRNA display libraries works in much the same way that phage display does. A target of interest, for instance a protein, is immobilised on beads. The target-carrying beads are mixed with the library and displayed proteins are allowed to bind. After washing to remove non-specifically bound displayed proteins, the specifically bound displayed proteins and their associated genetic material are eluted with e.g. high salt.

(a) Give two advantages that mRNA display might offer over phage display. [10%]

Crib: Various answers are possible with respect to the key difference that phage display has an *in vivo* component while mRNA display is completely *in vitro* e.g.:

(i) phage display requires the high complexity library to be transformed into *E. coli*, and this is a bottleneck. In contrast, mRNA display is a purely in vitro system and has no such bottleneck.

(ii) phage display requires proteins to be well expressed in *E. coli* and to be correctly assembled onto the surface of infectious viral particles.

(b) Camelids (llamas, dromedaries, camels) have immune systems like that of humans but produce a type of antibody that differs in structure to human antibodies: compared to human antibodies the light chain is missing and thus the mature antigen-binding protein is derived from a single variable immunoglobulin gene analogous to the human heavy chain gene.

(i) Give two reasons why an antibody based on a single gene might be better suited for mRNA display. [20%]

Crib: It makes the genetic engineering simpler:

A. only one kind of gene needs to be expressed correctly for the system to work

B. it is necessary to create a sufficiently complex representation of the natural immune system in the display library. This is hard and so doing it once is easier than twice.

(cont.

(ii) How would you obtain a set of single domain antibody DNA sequences with sufficient diversity, from an immunised animal? [10%]

Crib: Extract mature antibody producing cells (B cells) from camelids. Having gone through the process of maturation they will be producing antibodies with a range of complementarity determining region (CDR) sequences. Use reverse transcription followed by PCR to harvest the diversity in the population of cells.

(iii) An alternative approach to obtaining a diverse library would be through gene synthesis: to introduce diversity, random sequences of fixed length can be synthesised and inserted into the hypervariable regions of the antibody gene. What possible disadvantage might this approach have? [10%]

Crib: Random sequences will contain stop codons, thus producing non-functional proteins.

(c) Cells express proteins on their surfaces. Some of these proteins are found on all cells and others are specific to sub-categories of cells. Recently it has been discovered that some tumour cells express unique cell surface antigens and these are a route to destroying the cancer, for instance by directing the immune system to recognise these antigens.

You are given both normal and tumour cells from a patient. Describe modifications to the mRNA display process that will allow you to isolate mRNA from the display library that discriminate between normal and tumour by encoding proteins that bind only to the tumour cells. [30%]

Crib: Express the mRNA display protein library as usual. Select against those members of the library that bind non-tumour antigens by incubating the library with the non-tumour cells. This will deplete the library of proteins that bind antigens present on normal cells surfaces. Then take the remaining library and bind it to tumour cells, washing to remove non-specific binding proteins. Elute the bound mRNA-display libary to recover mRNA molecules that encode proteins that bind to tumour and not normal cells.

(d) Through your screen you find an mRNA that encodes a protein that strongly binds to tumour cells and not to normal cells. You want to find a way to display this protein on the surface of particular immune cells, which are capable of killing cells to which they tightly bind.

How might you engineer a human cell to display this protein, P, on its cell surface? [20%]

Crib: Various approaches are possible: for instance one could identify known human cell surface proteins from the literature. A variety of these could be fused to protein P and tested to discover which express P well on the cell surface. Alternatively a synthetic construct could be designed that includes both a secretion signal and a transmemberane

domain.

2 You are working with an oceanic microorganism (M), which lives in very low numbers in North Atlantic waters (average temperature < 20°C). It receives carbon-containing molecules (2-6 carbon atoms) from the water and, in the presence of energy provided by sunlight, converts them into useful building blocks. You would like to engineer this microorganism so that it consumes long hydrocarbon chains, hoping that the resulting microbes might be exploited to clear oil spills on the surface of the ocean. A paper has just been published, which describes a novel microbe (D) isolated from a desert environment. (D) can utilise long chain hydrocarbons efficiently as a carbon source, but is able to do this only in the presence of sunlight. You come up with a plan to make use of this microbe to engineer (M).

(a) Explain whether or not (M) becomes a cell factory once it is modified to acquire the desired ability to consume long hydrocarbon chains. [5%]

Crib: (M) will not be used to convert nutrients into commercially value-added products and therefore will not become a cell factory. This engineering approach aims to endow the cell itself with desired characteristics, hence to use the cell as a functional entity.

(b) What problems can be anticipated in utilising the hydrocarbon uptake and utilisation pathways from (D)? Suggest possible methods and approaches to overcome these problems. [20%]

Crib: The enzymes that need to be incorporated from (D) into (M) are likely to be most efficient at high temperatures, because (D) can break down hydrocarbons in the desert during daytime. The enzymes will likely require engineering to improve their thermal efficiency at low temperatures, and adaptive/metabolic evolution approaches could also be helpful to improve the overall efficiency of the recombinant (M).

(c) Suppose you would like to run preliminary modelling analyses to evaluate the feasibility of your approach. Assume that the metabolic network of (M) is available, while that of (D) is not. However, kinetic data measured in living cells are available for the hydrocarbon uptake of (D). Explaining your reasoning, suggest a modelling approach that would be useful for this situation, and one that would not be useful. Hint: pH and temperature are important parameters in determining reaction kinetics. [20%]

Crib: Constraint-based approaches such as flux balance analysis would be useful in determining bottlenecks in the material balances, but any reaction kinetics-based approach such as metabolic control analysis would not work since the kinetics of the enzymes available for (D) would be determined at their optimal working temperature suitable for daytime temperature in a desert. These enzymes will require modification in order to

successfully work at lower temperatures in (M), and therefore, the information available from (D) will not be a realistic representation for the existing problem.

(d) This project idea is pitched to an expert on metabolism in (M), who praises it by saying: "Very good choice of a host; this microbe can top flocculate when nutrient sources become scarce." Explain why this is relevant to your idea. Hint: Top flocculation occurs when the microbes clump together into a floc on the surface of the liquid environment. [20%]

Crib: If (M) flocculates on the surface of the water when the nutrients become depleted, these flocs can be collected from the surface of the ocean once the spill is cleared. Thus the overgrowth of the cells would have less impact on the existing marine microbiome.

(e) Why might flocculation not occur and what genetic modification(s) could help ensure that flocculation would take place at the right time? [15%]

Crib: If the organism normally can find sufficient nutrients in the ocean to survive, it will not flocculate. Therefore, alternative carbon uptake pathways will need to be blocked by approaches such as gene deletions, so that (M) would be more likely to suffer from nutrient limitation once the oil contamination is consumed.

(f) Assume that (M) has now been engineered to work in the North Atlantic. In the laboratory it has been tested on oceanic water samples that have been contaminated with oil. After successfully using (M) to clear up the oil contamination, you find that the rest of the marine microbiome has been severely and adversely affected. Explain a possible reason for this observation. [20%]

Crib: (M) is available at low abundance in the oceanic microbiome. Having an abundant nutrient source (i.e. the oil spill) available will rapidly increase its relative ratio in the microbiome. If (M) utilises other available nutrients that are essential for the organisms constituting the microbiome, or if it releases toxic by-products into the water, these will adversely affect the other organisms sharing the same environment.

3 (a) Bacteriophages, also known as phages, are viruses which infect bacteria. Like all viruses, phages commandeer the cellular machinery of their hosts to make copies of themselves. With the aid of a diagram, explain what functional elements encoded in the DNA are needed to encode a phage protein and cause it to be expressed. [5%]

Crib: Promoters allow initiation of transcription. Terminators allow termination of transcription. Ribosome binding sites allow translation to begin at the *start codon*. Translation ends when a *stop codon* is reached. The coding-sequence (open reading frame) of the gene encodes the protein. 4/6 gains the 5%.

(b) Some phage genomes are composed of RNA instead of DNA, which might seem to conflict with the central dogma of molecular biology that the *E. coli* host follows. Briefly explain two ways that phages with RNA genomes might avoid this problem. [10%]

Crib: The host organism will not carry out RNA-templated RNA replication. Therefore, the virus might be expected to encode additional enzymes that allow RNA to be replicated: 1) the virus may encode an RNA polymerase that uses RNA as its template. This would allow direct copying of the RNA genome. Alternatively, the virus may encode a DNA polymerase that uses RNA as its template (reverse transcriptase), so that the RNA genome can be converted to DNA. This DNA can then be replicated by the host DNA replication machinery as usual, and then transcribed to recreate the RNA genome.

(c) As a defence against phage infection, many bacteria produce restriction enzymes which cut foreign DNA at particular sequences. For instance, the NdeI restriction enzyme cuts at 5'-CATATG-3'.

(i) Show how often on average it would be expected that an NdeI restriction site will occur in a random DNA sequence with a 60% A-T base composition. [5%]

Crib:

$$\frac{1}{0.3^4 * 0.2^2}$$

so every ~3086 bases

(ii) A new enzyme is discovered that cuts 5'-CATCGG-3'. How often will this be expected to cut in the same random sequence as above? [5%]

Crib:

$$\frac{1}{0.3^2 * 0.2^4}$$

or every ~6944 bases, but this time the motif can occur on each strand independently, so will cut twice as often i.e. every ~3472 bases.

(cont.

(d) PhiX3G1 is a phage that infects the bacterium *Escherichia coli* and contains a single NdeI site in a gene essential for its ability to carry out infection. Wild type (WT) *E. coli* does not naturally express NdeI and is used as a host to grow a preparation of phage particles. A strain of *E. coli* (Eng) has been engineered to express NdeI along with a methylase that protects the *E. coli* genome from cleavage. Identical quantities of the phage preparation are used to infect a) the WT strain and b) the Eng strain. The efficiency of infection can be estimated by counting the number of phage "plaques", which are colony-like zones of killing formed on a lawn of bacteria.

(i) Explain why the number of plaques is lower when infecting the Eng strain. [15%]

Crib: The Eng strain expresses the NdeI enzyme and protective methylase. Infecting phage that contain recognition sites for NdeI may be cut and inactivated before they are recognised and methylated by the protective methylase. Thus the observed rate of infection will be lower in the Eng strain compared to WT.

(ii) Phage from many plaques arising from infection of the Eng strain above ("Round 1") are now used to repeat the experiment above ("Round 2"). Explain why there is now no difference in infection efficiency between the WT and Eng strains. [10%]

Crib: Phage that have successfully infected the Eng strain will now contain methylated NdeI sites, which will be protected from cleavage. Thus they will be able to infect the Eng strain with no loss of efficiency.

(iii) 10,000 plaques arising from the Round 2 infection of the WT strain above are now tested individually ("Round 3"), against both WT and Eng strains. 9,999 plaques tested behave as in Round 1 (low efficiency infection of Eng strain) but one behaves as in Round 2 (equally efficient infection of both strains). What molecular events across the three rounds explain the behaviour of the phage in this distinct plaque? [30%]

Crib: Round 1 progeny from Eng strain end up either methylated (common), or mutate their NdeI sites to be come resistant (rare). Passage through the WT strain in Round 2 restores their unmethylated status. Thus Round 3 can be used to identify the rare NdeI site mutations as those that are equally able to infect the WT and Eng strains. The mutation in the NdeI site must not affect the protein-coding capability of the gene essential for infection i.e. it must make no change to coding (through code redundancy) or else make an acceptable change to the protein that is essential for infection.

(e) PhiX3G1 possesses a DNA genome encoding 10 unique protein sequences. The total combined length of these 10 proteins is 1986 amino acids. The genome of PhiX3G1 is 5386 nucleotides in length. Why are these results discordant and what might the explanation be? [20%]

Crib: To code for 1986 amino acids one needs 5958 bases of DNA, which is longer than the genome. Therefore there must be some overlapping genes. The amino acid sequences coded by such overlaps still have some freedom to evolve due to redundancy in the genetic code, and the fact that different strands and reading frames code for different amino acids.

4 (a) Why is it useful to sequence the genomes of organisms? [15%]

Crib: Sequencing genomes provides baseline information on genes; provides information on genetic variation that can help with the idenfication of the underlying causes of disease; can aid comparision e.g. of pathogenic vs non-pathogenic bacterial strains; can be used to discover/survey the organisms in a given location.

(b) As part of the process of sequencing a genome, the genomic DNA is mechanically sheared in order to generated random fragments. Typically this proces results in there being single-stranded regions at the ends of each DNA strand and, unless repaired, these would interfere with subsequent processing steps. Thinking of cases in which the single-stranded region has a 5' end, outline the components needed in an enzymatic reaction that could repair this region back into double-stranded DNA. [15%]

Crib: A single stranded 5' end will have a recessed 3' end and this can be extended with a DNA polymerase. Thus the DNA would be added to a suitable buffer containing all four dNTPs along with DNA polymerase.

(c) Subsequent processing allows the sequencing of the above DNA. The resulting sequencing reads are compared to each other and those that overlap can be assembled to form longer sequences, perhaps ultimately the whole genome. What factors influence the extent to which a full and accurate genome assembly can be made? [20%]

Crib: Sufficient sequence reads; Repeated sequences in the genome create ambiguities that reduce the length of sequence contigs; Sequencing errors reduce the chance that the correct alignment will be found, but with sufficient sequence read coverage the consensus will over-rule random errors; The accuracy of the alignment algorithm will affect the assembly quality and length; Read length: shorter reads have a greater chance of matching incorrectly during assembly; Whether there is any bias for/against any region or regions of the genome.

(d) Is it typically easier to find protein-coding genes in prokaryotic or eukaryotic genomes? Explain why, by outlining the way in which such gene structure can be delineated in each case. [30%]

Crib: Prokaryotes are easier as the genes are typically intron-free and there is a high gene density. Thus looking for long open reading frames is sufficient to find many of the genes. Eukaryotes are harder as they typically have repeat-rich gene-sparse genomes and the genes themselves contain introns. For these reasons searching for open reading frames is insufficient. Instead one can use complex statistical models (e.g. GenScan).

Especially for eukaryotes, having cDNA sequences available from the same organism

improves annotation as they can be aligned to reveal intron/exon structure. Similarly comparision to known gene or protein sequences can also be useful. In all cases of sequence alignment, algorithm variants that are capable of allowing for introns are needed when analysing eukaryotic sequences.

(e) You have isolated and cultured a number of different prokaryotic and eukaryotic microbes that produce coloured pigments only when exposed to ultraviolet (UV) radiation. In each case you have reason to believe that the product of more than one gene is required in each microbe for pigment synthesis. Outline the approach you might take to discover the genes responsible for pigment synthesis. [15%]

Crib: Sequence cDNA prepared from cultures grown with and without exposure to UV radition: up-regulated genes are likely responsible for the response to UV exposure and these will likely include the pigment-producing enzymes.

(f) How might the genomic distribution of these genes differ in prokaryotes vs eukaryotes? [5%]

Crib: In prokaryotes, genes with related function are often grouped together into co-transcribed operons. In eukaryotes genes tend not to be grouped by function and are instead distributed around the genome.

END OF PAPER