egt2

ENGINEERING TRIPOS PART IIA

---

2 May 2018 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	SPECIAL REQUIREMENTS
Single-sided script paper	CUED approved calculator allowed

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: GM/3

1 The open reading frame from a human gene has been expressed in *E. coli* and the purified protein used to immunise a mouse. The resulting polyclonal antibody preparation, PAb-1, recognises different-sized proteins, derived from differentially spliced transcripts, in human tissue samples: in heart a 270-amino-acid form is present, while in liver there is both a 270-amino-acid form and a larger 300-amino-acid form. It is known that the N-terminal 100 amino acids and the C-terminal 120 amino acids are common to both the 270 and 300-amino-acid forms and each of these regions are derived from single exons.

(a) Explain the difference between polyclonal and monoclonal antibodies. [10%]

Crib: Polyclonal antibody preparations consist of many different antibodies, with varying affinities and specificities, each originating from a distinct B cell clone. In contrast monoclonal antibodies are produced by hybridoma cells that are the clonal offspring of the fusion of a myeloma cell with a single B cell and thus produce only a single type of antibody, recognising a single epitope i.e. antigenic site

(b) PAb-1 is used to test brain tissue samples and in all cases a 220 amino-acid form is observed. What is the simplest explanation for this observation? [10%]

Crib: mRNA splicing from the end of the N-terminal exon (codes for 100 amino acids) to the start of the C-terminal exon (codes for 120 amino acids) would generate a product of 220 amino acids.

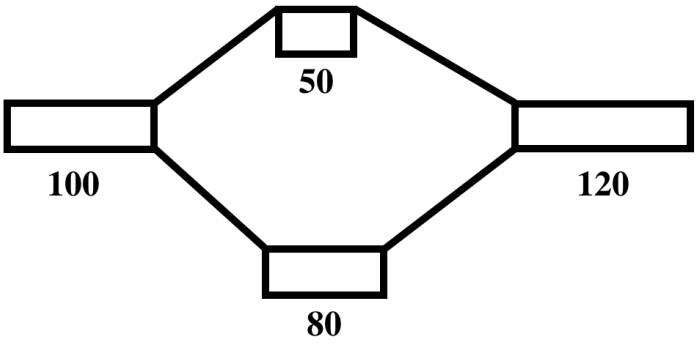
(c) Monoclonal antibodies, MAb-1, MAb-2 and MAb-3, are derived from the immunised mouse and tested against tissues as shown in Table 1 below. Explain, with the use of a diagram, possible mRNA splicing variations that could give rise to the observations in Table 1. [40%]

Table 1: Length in amino acids of proteins recognised by various antibody preparations. NR means nothing is recognised.

Antibody	Heart	Liver
PAb-1	270	270, 300
MAb-1	270	270
MAb-2	NR	300
MAb-3	270	270, 300

Crib: There are various possibilities, a relatively simple one of which is that there are two, alternatively used, internal exons. One codes for a length of 50 amino acids

and together with the terminal exons yields the 270-amino-acid form (100+50+120). The other exon codes for a length of 80 amino acids and yields the 300-amino-acid form. MAb-1 recognises an epitope encoded by the 50-amino-acid exon; MAb-2 recognises an epitope encoded by the 80-amino-acid exon; MAb-3 recognises an epitope coded by either the N or C-terminal exon. The mRNA splicing that gives rise to the 300-amino-acid form only takes place in the liver.



(d) Patients with a genetic liver disorder are tested with each of the antibodies and with the same results as given in Table 1, except that MAb-2 no longer recognises the 300-amino-acid form that it recognises in normal tissue samples.

Give a possible explanation for this observation. [20%]

Crib: We can tell the 300-amino-acid form is still present because PAb-1 and MAb-3 still identify it. Thus something is preventing MAb-2 recognising its target epitope. It is possible that a mutation (could be missense, insertion or deletion) is disrupting the epitope that MAb-2 recognises so preventing binding. The fact that we are dealing with a genetic liver disorder and the 300-amino-acid form is liver-specific is consistent with this.

(e) Outline what would be necessary in order to use the mouse monoclonal antibodies for human therapy and why this is important. [20%]

Crib: Mouse antibodies are recognised as foreign and thus cause an immune response in humans, preventing their use in therapy. Therefore it would be necessary to a) clone the genes coding for the heavy and light chains, b) genetically engineer them to change the mouse sequence scaffold into the corresponding human one, while preserving the diversity regions that form the antigen binding site and c) express the engineered construct in an appropriate human cell line in order to ensure human-specific

glycoslyation.

2 As part of an investigation, bacterial cells are broken open, and their contents extracted and purified to prepare a "cell-free" extract. These cellular contents are separated to remove the bacterial DNA, mRNA, cell membranes and the cell wall. Therefore the cell-free extract is composed of the remaining enzymes and tRNAs, which can be concentrated and remain fully functional. This mixture is not capable of sustaining transcription and translation until certain missing components, lost during the extraction process, are added as supplements.

(a) (i) What enzyme in the cell-free extract catalyses transcription? [5%]

Crib: RNA polymerase

(ii) In order to carry out transcription, what missing components must be added to the above cell-free extract and what roles do they play? [20%]

Crib: It is necessary to add the building blocks of RNA that are polymerised by RNA polymerase: ATP, UTP, CTP, GTP. Transcription uses DNA as a template and so a suitable substrate DNA must also present. This must contain a promotor sequence at which transcription initiates.

(b) Assuming you added the critical molecules for transcription, a messenger RNA (mRNA) in principle can be made.

(i) Along with the ribosomes, tRNAs and enzymes already present in the cell-free extract, what molecules must be added for translation also to occur, and what is their role? [15%]

Crib: the twenty amino acids that are the building blocks of proteins must be added. They become covalently attached to their corresponding tRNA molecules, which deliver them for polymerisation into a protein (polypeptide) by the ribosome.

(ii) What must be encoded on the mRNA for translation to take place? [15%]

Crib: a ribosome binding site, an AUG initiation codon, followed by an open reading frame of codons (triplets) that code for amino acids.

(c) Bacterial cells can detect arsenic at concentrations as low as 10 parts per billion (ppb) using the arsenic (Ars) operon. The Ars operon is regulated by a simple repressor molecule (ArsR) that binds to the promotor region (pArsR) in order to prevent

transcription. If the concentration of arsenic reaches 10 ppb then ArsR no longer binds to the promotor and the operon is transcribed.

The cell-free extract was used to make a biosensor in order to detect arsenic. Unfortunately ArsR is lost during purification of the cell-free extract.

Design a simple genetic circuit that will detect arsenic at levels similar to the Ars operon and explain how it works. The gene encoding Green Fluorescent Protein (GFP) should be used to report the presence of arsenic. [15%]

Crib: pArsR could be used to drive GFP expression whether by inserting GFP into the Ars operon, or by using pArsR as the promotor of a different construct, though in the latter case it will be important that DNA containing the Ars operon is also provided in order that ArsR is expressed. In both cases, in the absence of arsenic, the ArsR bound pArsR would not allow transcription. If the arsenic concentration reached 10ppb the repressor would stop binding, and transcription of the GFP reporter would start. Following translation of the GFP mRNA, the GFP protein will produce a fluorescent signal under appropriate illumination.

(d) When you test your ciruit, the signal-to-noise ratio isn't high enough to easily detect arsenic at 10 ppb. The noise is due to the random unbinding of the repressor, which generates a low but detectable background signal. Discuss approaches you could take to avoid this problem. [30%]

Crib: Increasing the expression of the repressor (for instance by putting it under the control of a stronger promotor) should increase pArsR occupancy and thus decrease noise, but it might also decrease sensitivity. Similarly using genetic engineering to make a multimeric form of ArsR (for instance by making a tandem duplication of the open reading frame, or fusing ArsR to a dimerisation domain) should increase affinity for pArsR much as antibody affinity is higher due to having two antigen binding sites. This might also reduce sensitivity. Finally, an amplification circuit could be used (for instance pArsR could drive expression of a strong RNA polymerase, and this would drive expression of GFP via its cognate promotor) with the risk that this would just amplify the noise too.

3 The schematic representation of a hypothetical metabolic pathway with two branches is given below. Substrate A is converted into products E and G. The uppercase letters denote metabolites and lowercase letters denote enzymes catalysing the adjacent reaction step. Route I denotes the pathway converting A to E, and Route II denotes the pathway converting A to G. ADP: Adenosine diphosphate, P: phosphate, ATP: adenosine triphosphate.

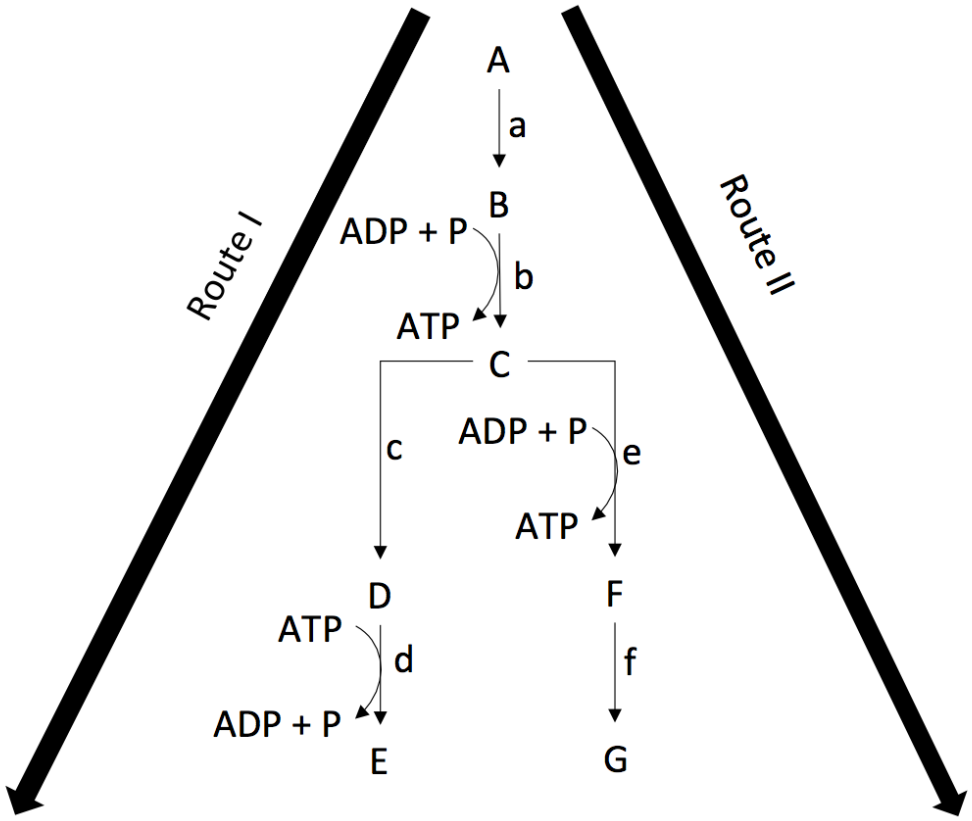


Fig. 1

(a) Are the reactions catalysed by the enzymes b, d, and e likely to represent catabolic or anabolic processes? Explain your answer. [10%]

Crib: Reactions catalysed by b and e are releasing energy, and are more likely to be associated with catabolic processes, whereas the reaction catalysed by d requires energy, and is likely to represent an anabolic process.

(b) Explaining your answers, identify which metabolite in each pair below is

expected to have a greater molecular weight:

B or C ;

D or E ;

C or F . [10%]

Crib: Assuming that the reactions catalysed by b and e represent catabolic processes, the molecular weight of B is expected to be greater than that of C, and of C to be greater than that of F. The reaction catalysed by d is likely to be an anabolic process; therefore the molecular weight of E is likely to be greater than that of D.

(c) What does it mean for a reaction to have a low saturation constant (K_m)? [10%]

Crib: A reaction catalysed by an inefficient enzyme with a low saturation constant would indicate that it would have a high affinity for its main substrate

(d) Suppose that metabolite G is the main precursor for a product, which is of commercial interest. You are given the task of improving the metabolic capability of this pathway for producing G (and consequently the product). The overall saturation coefficient (K_m) for Route I and Route II are provided in the table below (Table 2) under aerobic and anaerobic conditions. Assuming that these enzymes are inefficient, would you choose to operate this system under aerobic or anaerobic conditions? Explain your answer. [15%]

Table 2: K_m values

CONDITION	K_M FOR ROUTE I	K_M FOR ROUTE II
AEROBIC	0.06mM	0.3mM
ANAEROBIC	0.9mM	0.1mM

Crib: The preferred route should have a higher affinity for the production of precursor metabolite G from the main substrate A than for the production of E. Route II has a lower K_m value (and therefore higher affinity) than Route I under anaerobic conditions. Therefore the system should be operated under anaerobic conditions to assist the preferential selection for the production of metabolite G rather than metabolite E.

(e) The Response Coefficient (R) is defined as the measure of how the pathway flux (J) responds to an effector (P). Mathematically it is defined as:

$$R_p^{J_{ydh}} = \frac{\delta J_{ydh}}{\delta P}.\frac{P}{J_{ydh}} = \frac{\delta lnJ_{ydh}}{\delta lnP}$$

where *ydh* represents any pathway.

Identifying oxygen as an effector of this pathway, how do the pathway fluxes in Route I and II respond to varying oxygen levels? [20%]

Crib: K_m value (consequently the affinity) is an indicator of enzyme activity (affecting reaction flux) for inefficient enzymes. The overall K_m for Routes I and II would consequently represent a measure of the activity of these routes (and since no other parameter is modified a measure of the fluxes in Routes I and II). The amount of oxygen (effector) available for both Routes are identical under aerobic and anaerobic conditions, but Route I is observed to respond more drastically as seen by the change in its affinity with oxygen availability. Therefore the response coefficient of Route I for oxygen is higher than that of Route II.

(f) Suggest an experimental metabolic engineering approach, or a combination of approaches, to improve the selective production of the precursor metabolite G. [15%]

Crib: With the information provided above, the goal would be to increase the difference between the overall K_m values for Route I and Route II under anaerobic conditions. This can be facilitated by adopting either one of more of the following approaches:

Adaptive evolution, metabolic evolution or directed evolution to improve pathway flux in Route II (for example by promoting feedback activation via product G), enzyme engineering to improve the efficiencies of individual enzymes (particularly of those below the branching point in Route II), genome engineering, employing synthetic protein scaffolds to facilitate the co-localisation of the enzymes in Route II.

Note: The examples provided above are not exhaustive. All suggestions employing one or more of the approaches and methodologies discussed in the lectures can be acceptable provided that a justifiable strategy is proposed.

(g) You are now given the following information. Enzyme h from another organism is functionally equivalent to enzyme f in Fig. 1. Replacing the gene encoding enzyme f by a gene encoding enzyme h changes the overall K_m for Route II. The saturation coefficient for Route II becomes 0.2 mM under aerobic conditions and 0.5 mM under anaerobic conditions. Assuming that the K_m values for Route I remain unchanged,

is this a feasible genetic modification for improving the selective production of the precursor metabolite G? Explain your answer, referring to Table 2 as required. [20%]

Crib: The incorporation of the gene encoding enzyme h into this organism would not render neither aerobic nor anaerobic options feasible as modes of operation. Although it improves the affinity for Route II slightly under aerobic conditions, Route I would still remain as the preferred route for the direction of fluxes. Under anaerobic conditions, the affinity for Route II is lower than that employing the original enzyme, so h would be a poor choice under anaerobic conditions.

4 Plants produce 200 billion tons of non-food lignocellulosic (woody) biomass each year, with huge potential as a carbon-neutral fuel source. The microbes that live in the gut of the termite, the termite microbiome, are responsible for digesting the cellulose-rich wood that termites consume. Motivated by this, you extract DNA from the termite microbiome and generate a metagenomics dataset of billions of sequencing reads using the Illumina platform, with the hope of discovering the enzymes responsible for cellulose digestion.

(a) (i) Sequence analysis shows that the extracted DNA is dominated by sequences with high similarity to plants. What has happened and what simple modification to the experiment might avoid this problem? [10%]

Crib: The plant contamination likely originates from partially digested woody material collected with the microbiome. A solution could be to starve the termites before collecting the microbiome.

(ii) Furthermore, you also find sequences with high similarity to humans within the dataset. What is the likely explanation for this observation and how might the experiment be repeated to avoid this problem? [10%]

Crib: At some point in the process human material, for instance a minute flake of skin, has contaminated the sample or reagents. The solution would be to take measure to prevent such contamination (e.g. gloves, mask, hair covering) and to carefully prepare fresh reagents.

(b) Having repeated the experiment and produced a DNA sequence dataset free of plant and human sequences, you carry out sequence assembly. This is done by comparing the sequences and combining those with extensive overlaps of near-identical sequence to form longer sequence *contigs*.

(i) Representatives of which kingdoms of life may reasonably be present in the termite gut microbiome? [15%]

Crib: prokaryotes, archae and eukaryotes might well be present.

(ii) Thus, when examining the assembled sequence contigs to find genes, what features would you expect the protein-coding genes to have? [20%]

Crib: Prokaryotes and archea typically have intron-free protein-coding genes and so genes will have ATG start codons followed by long open-reading frames and may be present in operons. In contrast the genes of eukaryotic microbes will likely contain introns and will not be arranged as operons.

(iii) Examining the length distribution of the sequence contigs you observe that longer contigs are rarer, shorter ones more abundant, and unassembled individual sequencing reads most abundant of all. Discuss reasons for these observations and approaches to increasing the fraction of long contigs. [30%]

Crib: When sequencing a single genome, insufficient sequence coverage will give rise to fewer long contigs and many unassembled reads. In this cases the dataset is derived from an unknown and possibly large number of genomes so it will be harder to obtain sufficient coverage of all genomes. This is compounded by the fact that different microbes will have different abundances in the population so the reads from rare microbes are unlikely to assemble and will contribute to the population of single reads. Furthermore, the eukaryotic microbial genomes present are likely to contain repeated sequences and these will prevent unambiguous sequence assembly leading to shorter contigs.

Aproaches to increasing the fraction of long contigs include: increasing the sequence coverage by carrying out further large-scale sequencing reactions, and using single-molecule long-read technologies (PacBio, Oxford Nanopore), which have more chance of reading through repetitive region.

(iv) An alternative approach to extracting and sequencing DNA from the termite gut microbiome would be to extract RNA, reverse transcribe it to DNA and sequence that. What possible advantages and disadvantages might this have compared to the previous approach?

[15%]

Crib: Disadvantage: genes are expressed at different levels which may compound the fact that the microbes are present at different abundances i.e. bias towards observing highly expressed genes from abundant microbes. Advantages: cellulose digesting genes are likely to be expressed at high levels; easier to find genes as don't have to consider introns in the eukaryotic portion of the dataset.

END OF PAPER