

EGT3  
ENGINEERING TRIPOS PART IIA

---

Wednesday 3 May 2017 14.00 to 15.30

---

**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

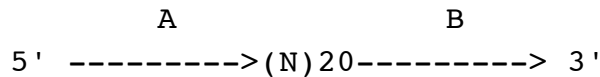
**SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM**

CUED approved calculator allowed

**10 minutes reading time is allowed for this paper.**

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

1 Suppose we instruct a DNA synthesis company to synthesise a population, P, of single-stranded DNA molecules of the form A(N)20B:



A and B are distinct sequences, TAATGTGCAATGTTCTTTATCCCCCC and CCCCCAGAATATACAGAAGGCAGAC respectively. (N)20 means that the DNA synthesis machine adds an equal mixture of all four bases twenty times in a row.

Sequences within this random (N)20 region are able to fold up through complementary base pairing and adopt specific shapes. It is known that the right shapes can bind other molecules tightly. We are interested in searching for such folded sequences that are able to bind to the molecule theophylline, which we have covalently attached to polymer beads. Therefore we carry out the following cycle of steps to enrich for sequences that bind to theophylline:

Step 1) the population P is made up in binding buffer.

Step 2) binding buffer containing P is added to the theophylline beads and incubated with end-over-end rotation for 1 hour at room temperature in a plastic test tube.

Step 3) the beads are washed several times: each time by letting the beads settle to the bottom of the tube, carefully removing the overlying liquid and then resuspending the beads in DNA-free binding buffer.

Step 4) the beads are allowed to settle and the buffer replaced again but this time the tube is heated to 95°C for 5 minutes so that any bound DNA is released from the theophylline beads. The buffer containing released DNA is then removed from the settled beads for use in the next step.

Step 5) The released DNA is amplified using PCR and (using a further step the details of which are not important) the single strands corresponding to the starting material, AN(20)B, are prepared.

This single-stranded material can again be bound to theophylline beads. Thus steps (2) through (5) are repeated another two times in succession so that three cycles in total of enrichment have been carried out. With each cycle we expect to progressively enrich for the specific sequences capable of binding the theophylline beads most tightly.

- (a) In Step (4) why does heating to 95°C cause the DNA to be released? [5%]

In much the same way that the high temperature step of PCR melts the strands apart, the high temperature will melt the base pairs forming the specific DNA shape that can recognise theophylline so releasing the bound DNA strands.

- (b) Paying careful attention to the sequences of A and B, write down the sequences of two primers that can be used in a PCR reaction to successfully amplify the sequences released in Step (4). Explain your design. [20%]

First candidates for the primers might be A and the reverse complement of B (B'):

A: TAATGTGCAATGTTCTTTATCCCCC

B': GTCTGCCTTCTGTATATTCTGGGGG

However the final 6 bases of these primers (underlined) match each other perfectly and are G/C (strong) base pairs and thus there is a high likelihood of self-priming, which would lead to a very short and useless PCR product. The solution is to shorten the primers by removing these self-priming bases:

A: TAATGTGCAATGTTCTTTAT

B': GTCTGCCTTCTGTATATTCT

The remaining sequences, at 20 bases, are sufficiently long to act as primers in PCR, with sequence A priming synthesis of the top strand and sequence B' priming synthesis of the bottom strand. Primers of length 15-20 are acceptable as long as they avoid the terminal runs of Cs and Gs.

- (c) It is likely that, in addition to sequences that can bind theophylline, the above procedure will select those sequences capable of binding the polymer beads themselves, and this is undesirable. Suggest and explain an extra step in each cycle of the above procedure that could reduce the abundance of such polymer bead-reacting sequences. [20%]

For each cycle of the process, an extra step could be inserted before step (2) (e.g. after steps (4) or (5): instead of binding the DNA to theophylline beads, the DNA would be bound to identical polymer beads but without covalently attached

theophylline. This step would deplete sequences that bind directly to the polymer beads from the population P. The remaining sequences present in the liquid phase can be used in step (2). Thus the extra step would be identical to step (2) except that the beads used would not have theophylline covalently attached to them and at the end of the incubation the beads would be allowed to settle and the liquid transferred to step (2) for use as (a depleted) population P.

- (d) Following the three cycles through the enrichment procedure described above it is likely that the diversity of sequences present is still high. Justify which sequencing method you would choose to most easily investigate the relative amounts of the most abundant (and thus presumably the tightest binding) sequences in the population. [10%]

Illumina sequencing would be a good choice as it is accurate, and can generate millions to billions of sequences per run, thus allowing a good sampling of the population. Also the material to be sequenced is of a suitably short length.

- (e) Describe how to prepare material from Step (4) so that it can be sequenced using the method you proposed in answer to (d). [5%]

The single-stranded DNA must be converted into double stranded form. This is done using a few cycles of PCR using primers A and B' (designed in part (b)). Once in double-stranded form, the Y-shaped Illumina adaptors are ligated and, after this, the material is ready to be sequenced on the Illumina platform.

- (f) After sequencing you count the number of times each distinct sequence is found. Perhaps the independent sequences that have been selected share some sequence similarity responsible for theophylline binding, and so you are curious to see whether the different sequences are related. Outline how you would approach this challenge. [25%]

A sequence alignment algorithm such as dynamic programming can be used to compare the sequences. You have no way of knowing where within the (N)20 region the theophylline-recognising sequence is found. Therefore you need to use a "find best sub-sequence" variant of dynamic programming. A Blosom matrix is not appropriate for scoring as we are comparing DNA sequences not amino acid sequences. Instead you can create a very simple scoring rule such as +2 for matching bases, and -1 for mismatching bases. To allow for the possibility for gaps, you should experiment with various gap penalties and observe which give plausible alignments. Note that just the central 20 bases of

sequence are of interest so the first and last 26 bases of each sequencing read, corresponding to sequences A and B should be removed. If this is not done then matches to these subsequences will dominate the results.

(g) Among your sequences you observe the following:

ACTATACCTATGGTATGACT (observed 200 times)

ACTATACCTGTGGTATGACT (observed once)

Provide three possible explanations for the origin of the sequence that is observed once. [15%]

The single copy sequence differs by a single base. This could be a sequencing error, a replication error introduced during PCR or it could represent an independent sequence, enriched from the starting population P.

2 Examine the reaction pathways shown in Fig. 1 below while reading the following: Organism 1 naturally produces the commodity chemical, C, using chemical A as a substrate. The pathway of interest is a linear pathway with two branches; one splitting off from intermediate B1 producing metabolite F through a series of reactions shown by a dashed arrow, and another branch splitting off from intermediate B4 similarly producing metabolite G. The yield of C on substrate A,  $Y_{C/A}$ , is 0.4g C per gramme of A. Enzyme  $x3$  has the highest flux control coefficient (FCC) in the linear pathway, with value 0.67. All reactions are non-equilibrium reactions with directionality indicated by the direction of the arrow. The enzymes catalysing each step are denoted in italic fonts as  $x1 - x5$ ,  $y1$ ,  $z1$ ,  $n1 - n2$ . For Organism 1, the numbers in parentheses indicate the number of carbon atoms contained within substrate A, intermediate metabolites B1 – B4, and the product of interest, C. The arrow from B3 to  $x1$  signifies a positive regulatory action, not a metabolic reaction. Similarly, the blunt-headed arrow from B2 to  $x1$  indicates an inhibitory regulatory action.

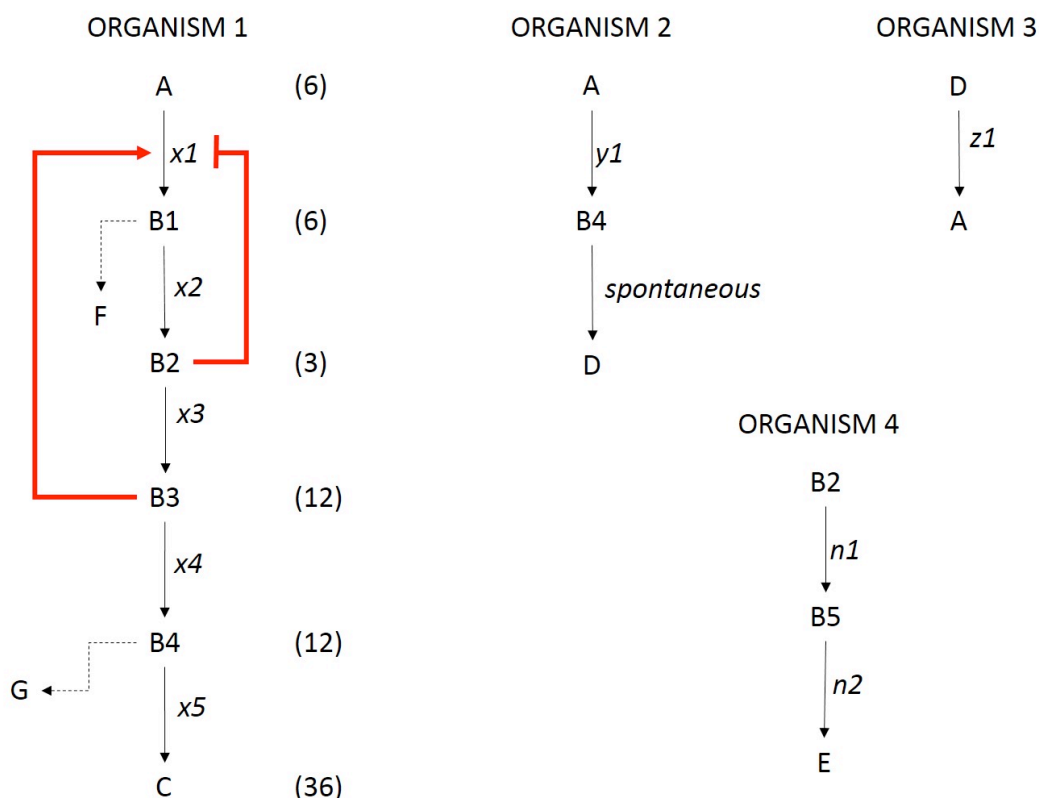


Fig. 1

- (a) Explain which enzymatic steps in the linear pathway in Organism 1 are more likely to be anabolic processes and which ones to be catabolic processes. [10%]

Reaction steps catalysed by  $x1$  and  $x2$  are more likely to be catabolic steps since decrease in number of carbon atoms may indicate the breaking down of the main substrate and the formation of simpler intermediates. The number of carbon atoms increasing in reaction steps from  $x3$  to  $x5$  may indicate the biosynthesis of more complex molecules, i.e. anabolic processes.

- (b) Identify the different types of metabolic regulation on this linear pathway. [5%]

Both mass action and allosteric regulation is observed.

- (c) You are consulted as a metabolic engineer to suggest alternative genetic engineering strategies to improve  $Y_{C/A}$ . Digging into the literature, you identify Organism 2 and Organism 3 containing potentially useful pathways, which are also shown in Figure 1.

Describe possible bottlenecks limiting production yield in the linear pathway in Organism 1, and explain whether or not an alternative strategy can be implemented with the help of genes from Organism 2 and Organism 3. [20%]

The two branching routes from the main pathway lead to the production of by-products F and G. Therefore, some of the substrate is converted into side products, decreasing the yield of product C on substrate A. The allosteric regulation for intermediate B2 inhibiting the first step of the reaction catalysed by enzyme  $x1$  is another bottleneck. Direct conversion of A to B4 appears to work advantageously bypassing both the inhibitory allosteric regulation and the by-product F formation. However, the next step being spontaneous is problematic. Introducing enzyme  $z1$  from Organism 3 brings the system back to its original metabolic state. Therefore, these pathways from Organism 2 and Organism 3 do not contribute to achieve the goal of improving the yield of product C on substrate A.

- (d) You investigate further and come across another pathway from Organism 4 (see Figure 1), which could also be useful. You are informed that enzyme  $n1$  has very high elasticity with respect to its substrate. Explain whether this pathway can be useful to increase the yield of compound C, and outline potential drawbacks or limitations. [15%]

The high elasticity of  $n1$  indicates that this reaction step can be useful in fine-tuning the main linear pathway in Organism 1 by maintaining the concentration of B2 constant, thus avoiding fluctuations in the allosteric regulation acting upon

the system. However, an additional branch synthesising another by-product E is introduced and this is a potential risk in reducing the yield of C on substrate A.

- (e) The pathway from Organism 4 is inserted into Organism 1 by genetic modification. The new  $Y_{C/A}$  is 0.56g chemical C/ g of chemical A. The new FCCs are calculated and  $x3$  no longer has the highest FCC. How do you think the metabolism re-wired itself? [20%]

Since the yield is increased, and  $x3$  does not exhibit the highest control over the linear pathway, it is likely that partial conversion of intermediate B2 to B5 through the catalytic action of  $n1$  relaxed the inhibitory action on enzyme  $x1$  and allowed more of B2 to be converted into B3. B3 in turn acted to activate enzyme  $x1$  and these events resulted in the improvement in  $Y_{C/A}$ .

- (f) Suggest three alternative methods to improve the yield of chemical C. [15%]

Any three of the following options will suffice: spatial engineering, titration of gene copy number, enzyme engineering, metabolic evolution, adaptive evolution, MAGE, optimisation and modulation of pathways, tuning promoter strength, augmenting activities of native enzymes, toggling the order of genes in the operon, creating synthetic protein scaffold, directed evolution, multiplex genome engineering and accelerated evolution.

- (g) Is it possible to calculate the new yields without proceeding with the genetic modifications and running experiments? What additional information is needed for this technique? What are the possible drawbacks? [15%]

Yes, flux balance analysis (FBA) / metabolic flux analysis may be conducted to calculate the yields. Exact stoichiometry for each involved reaction, and at least one mass balance constraint needs to be known about the system. FBA will provide one optimal solution from a convex space of possible solutions, so it may not be an exact and unique solution.



3 (a) Proteins that recognise specific DNA sequences are fundamental tools in Synthetic Biology. Give three examples of such proteins and brief descriptions of their molecular functions. [20%]

Restriction enzymes bind to more or less specific recognition sites and cut/hydrolyse phosphodiester bonds either in the DNA sugar-phosphate backbone.

Transcriptional repressors affect RNA polymerase transcriptional activity at promoters (e.g. recruitment, initiation) by binding to operator sequences.

RNA polymerases recognise promoter elements (e.g. -35 and -10 boxes in prokaryotic genomes) and initiate transcription of mRNA genes downstream of these promoters.

DNA recombinases catalyse exchange of DNA between target sequences that may be on the same or different DNA molecules. This can result in excision/insertion, inversion and cassette exchange reactions.

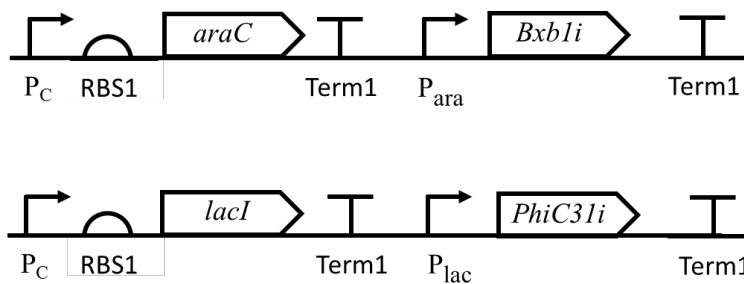
(b) Serine-type DNA recombinases catalyse unidirectional inversion or excision of DNA flanked by pairs of asymmetric recognition sites. Excision occurs where the orientations of the recognition sites are aligned. Inversion occurs where the recognition sites point towards each other. How might the essential properties of Serine-type DNA recombinases be used to control bacterial transcription? [20%]

Recombinase recognition sites can be placed either side of ORFs/CDSs or of transcriptional control elements such as promoters and transcriptional terminators. Control of the recombinase activity can thereby be linked to inversion or excision of these sequence elements, resulting in activation or repression of gene expression.

The unidirectional nature of the recombination reactions means that inversion reactions can tend toward completion rather than equilibrating at 50%. This is less important for excision reactions due to dilution effects.

(c) Using only components from Table 1 design a genetic AND gate based on DNA recombination. Draw a diagram to illustrate your design and explain how the gate functions. The inputs should be arabinose and IPTG. The output should be GFP fluorescence. [30%]

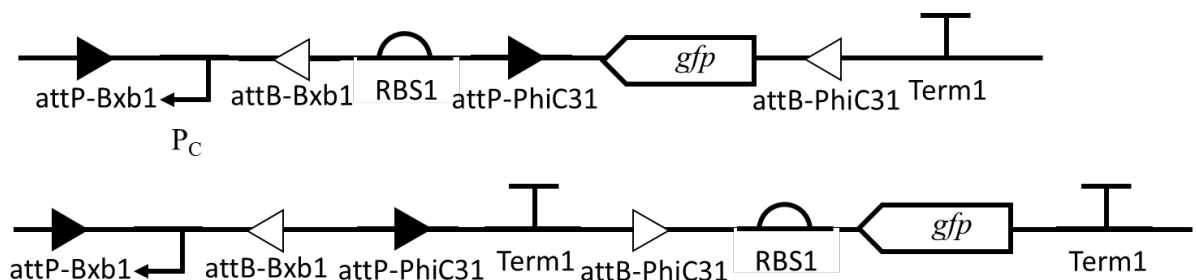
LacI and AraC independently controlling expression of different DNA recombinases:



A GFP expression cassette comprising the GFP CDS with a constitutive promoter and RBS upstream. ( $P_{ara}$  or  $P_{lac}$  would also be acceptable so long as the description of the gate's function is still accurate, i.e. GFP would be under control of an inducer even after recombination). A pair of Bxb1 attB and attP sites should flank one element of the GFP expression cassette (e.g. promoter/CDS/terminator) and a pair of PhiC31 attB and attP sites should flank a second element of the GFP expression cassette. Possible configurations include any two of the following:

- Excision of terminator between promoter and RBS.
- Inversion of an (inverted) promoter to drive GFP expression.
- Inversion of an (inverted) *gfp* CDS downstream of a promoter and RBS.

It would also be acceptable to describe the independent excision of two terminators between the promoter and RBS.



Accurate description of how the functions and behaviour of the gate: recombination events controlled independently by IPTG and arabinose result in high expression of GFP. If neither or only one recombination event is triggered then GFP expression should remain low. Descriptions of reaction should match their diagrams, describing inversion reactions (inward-or outward facing attachment sites) or excision reactions (both forward or both backward-facing attachment sites).

(d) How would you expect the behaviour of an AND gate based on serine DNA recombinases to differ from an AND gate based only on transcriptional repressor protein activity, such as *LacI* and *AraC* acting at the same promoter? [30%]

Memory/irreversibility: a unidirectional recombinase-based AND gate should show long term (multi-generational) memory/irreversibility, i.e. once triggered, the recombination events will not be reversed. Transcription factor binding is reversible and so the state of an AND gate based only on transcriptional repressor proteins should reset following the removal of inputs (e.g. IPTG or arabinose).

Response-time: The described recombinase-based system might be expected to have a longer response time to inducers, since an additional round/layer of gene expression is required between addition of inducers and GFP expression.

Discrete/continuous response: A single copy of the recombinase-based AND gate can only occupy one of four discrete expression states, corresponding to the various states of recombination, whereas the transcription factor AND gate would display a continuous (monotonic) response to inputs (e.g. IPTG and *araC*).






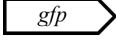
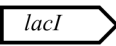
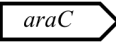
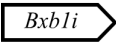
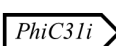




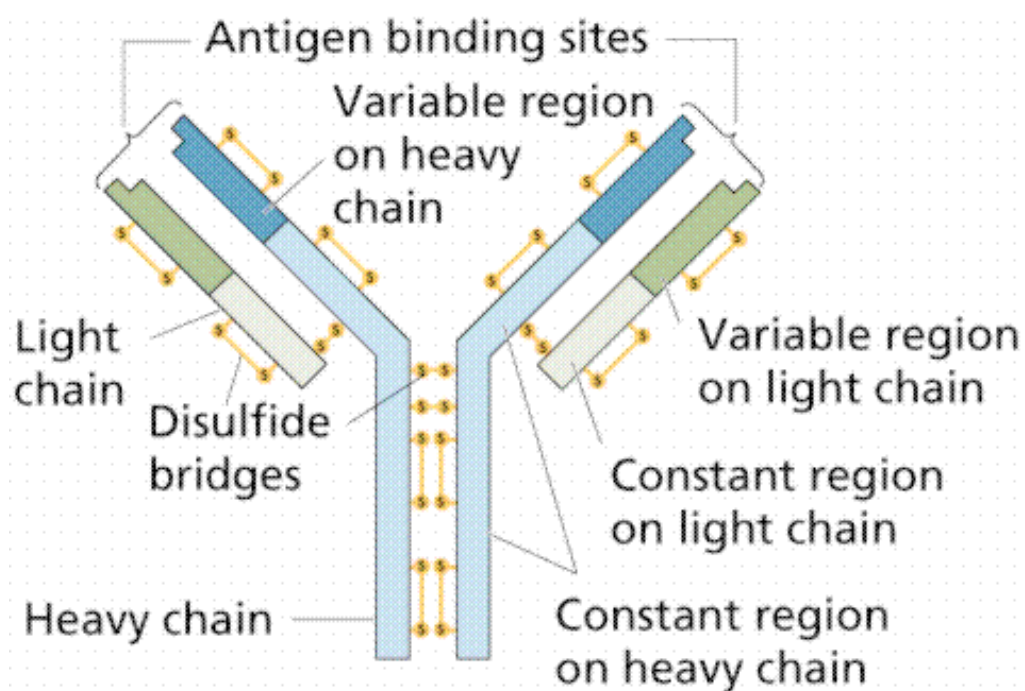
Abbreviation	Class	Description	Symbol
P <sub>C</sub>	Promoter	Constitutively active promoter.	 P <sub>C</sub>
P <sub>ara</sub>	Promoter	AraC binding represses transcription.	 P <sub>ara</sub>
P <sub>lac</sub>	Promoter	LacI binding inhibits transcription.	 P <sub>lac</sub>
RBS1	Ribosome binding site	Ribosome binding site. Responsible for recruitment of a ribosome during initiation of translation.	 RBS1
Term1	Transcriptional terminator	Bidirectional transcriptional terminator. Triggers release of mRNA from the RNA polymerase transcriptional complex. Function is independent of the direction from which RNA polymerase approaches.	 Term1
<i>gfp</i>	Protein coding sequence	Codes for Green Fluorescent Protein (GFP), which fluoresces green under 395 nm wavelength light.	
<i>lacI</i>	Protein coding sequence	Codes for LacI protein. LacI binds to the Plac promoter as a tetramer, forming a DNA loop that prevents RNA polymerase from binding. When IPTG is bound to the repressor, the loop is opened, allowing RNA polymerase to bind.	
<i>araC</i>	Protein coding sequence	Codes for AraC protein. AraC binds to the Para promoter as a dimer, forming a DNA loop that prevents RNA polymerase from binding. When arabinose is bound to the repressor, the loop is opened and the AraC dimer stabilises RNA polymerase binding to the promoter.	
<i>Bxb1-i</i>	Protein coding sequence	Codes for Bxb1 DNA recombinase. Bxb1 binds specifically to attachment sites attB-Bxb1 and attP-Bxb1. Depending on the orientation of the attachment sites, Bxb1 catalyses inversion or excision of DNA sequences flanked by these attachment sites.	
<i>PhiC31-i</i>	Protein coding sequence	Codes for PhiC31-i DNA recombinase. PhiC31-i binds specifically to attachment sites attB-PhiC31 and attP-PhiC31. Depending on the orientation of the attachment sites, PhiC31-i catalyses inversion or excision of DNA sequences flanked by these attachment sites.	
attP-Bxb1	Integrase attachment site	Attachment site for Bxb1.	 attP-Bxb1
attB-Bxb1	Integrase attachment site	Attachment site for Bxb1.	 attB-Bxb1
attP-PhiC31	Integrase attachment site	Attachment site for PhiC31.	 attP-PhiC31
attB-PhiC31	Integrase attachment site	Attachment site for PhiC31.	 attB-PhiC31

Table 1: Genetic parts for use in Question 3. Parts may be used more than once.

4 Antibodies form a fundamental part of the mammalian immune system. The average human will develop over 10 billion antibodies over their life time, each capable of detecting a distinct epitope.

(a) Describe the structure and function of the different regions of antibodies. [20%]



The antibody is composed of two light and two heavy chains. These are held together using disulphide bridges conferring high structural integrity. Both chains also contain constant and variable regions, as the name implies, one being highly conserved so as to preserve function and the other being highly variable to generate diversity of specificity.

In the variable regions, spanning the variable regions of the heavy and light chains, are the Complementarity Determining Regions (CDRs) that dictate the specificity and affinity of an antibody towards an antigen.

- (b) Describe the mechanisms that achieve such a high level of diversity. [20%]

The two mechanisms driving the generation of antibody diversity are recombination and splicing. In somatic recombination in B cells, stretches of DNA sequence are permanently removed from the genome of a cell: for light chain genes one of many V segments are joined with one of many J segments. For heavy chain genes there is an additional choice between D segments positioned between the V and J segments and this increases diversity to >10,000 possible VDJ combinations per B cell. The fact that this type of recombination is also error prone results in further diversity in the amino-acid sequence.

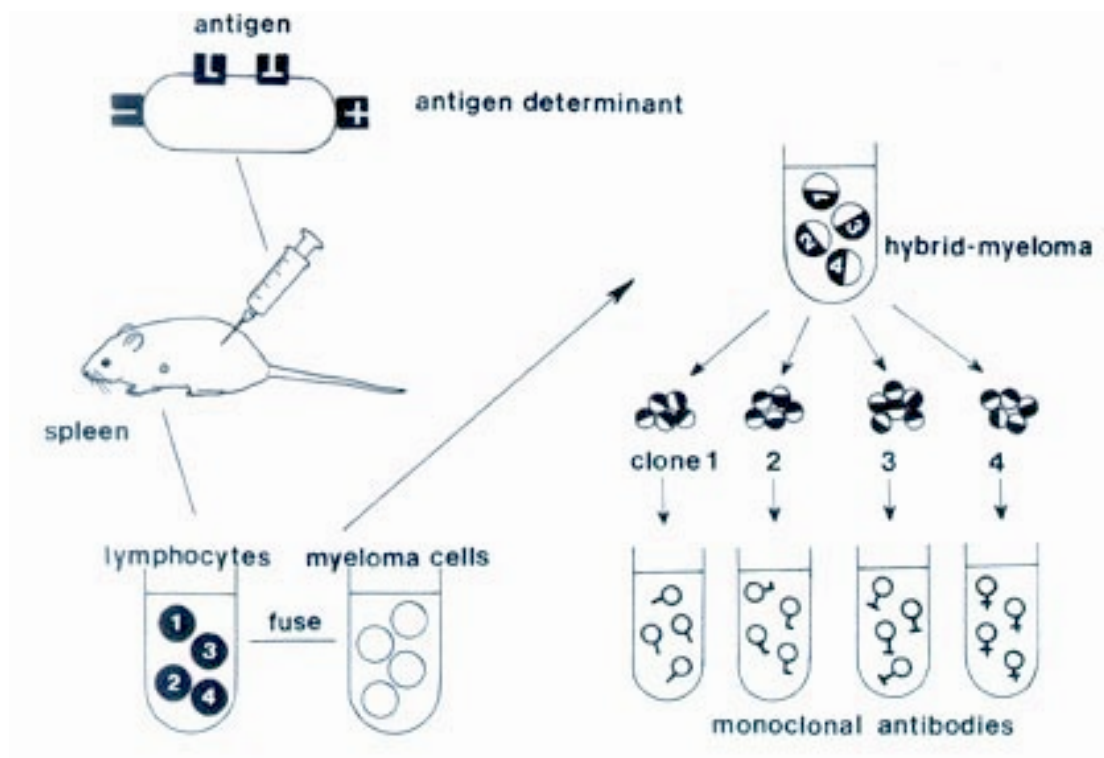
Hypervariability of the CDR regions, as evidenced by the much less conserved sequences within those areas, suggest that rates of mutation in these regions are substantially higher and therefore result in further increases in diversity.

Splicing occurs at the RNA level and therefore results in a one-time commitment to producing a multi-domain polypeptide, which can be produced in a different manner in subsequent transcription rounds within the same cell. This allows further diversity within a single cell. Heavy chain genes are alternatively spliced, “isotype switching”, allowing the C-terminus of the protein far from the antigen binding site to be changed to mediate different effector functions.

(c) Current therapeutic use of antibodies requires the use of a single highly specific antibody that must be produced monoclonally. Describe the process by which monoclonal antibodies are produced, starting from a mouse inoculated with a specific antigen.

[20%]

Innoculated mouse starting point



Lymphocytes are harvested from a mouse inoculated with an antigen. These are then fused to myeloma cells to create hybridomas, which are thereby immortalized and capable of perpetually producing antibodies as a cell line. Selection for successful hybridomas is very important: as lymphocytes are resistant to HGPRT but have a short lifespan and myelomas are sensitive to HGPRT and have a long lifespan, culturing a mixed population of lymphocytes, myeloma cells and hybridomas in HGPRT media will select for growth of hybridomas as long-lived HGPRT resistant clones.

As the mouse will most likely be producing several antibodies that bind to separate antigen epitopes on the antigen of interest, the population should be screened to find the clones producing the antibodies with highest affinity and specificity for the antigen, prior to clinical trials.

(d) Expressing antibodies in bacteria is a very attractive proposal but has so far proven incredibly difficult. What challenges can you identify with attempting this? [20%]

Heterologous expression, particularly between eukaryotes and prokaryotes poses many challenges.

One of the easiest challenges to overcome is the difference in codon usage between the organisms. This can be overcome by manipulating the sequence, for instance by resynthesis, so as to optimise codon use and so increase expression by making translation more efficient.

Post-translational modification is extremely rare within prokaryotes, but many eukaryotic proteins require such modifications or order to be functional. Therefore post-translational machinery from eukaryotes would have to be incorporated into the host expression chassis so as to overcome this challenge and this alone is a major undertaking, not so far undertaken successfully.

As prokaryotes do not have organelles, microenvironments that drive events such as folding, packaging and epitope removal do not exist and their absence may prevent successful expression of a functional protein. In addition, as chaperones that actively affect the protein folding process may differ between hosts, this can result in an increased chance of misfolding.



(e) A bacteriophage expression library has been made that expresses single-chain antibodies on its surface. It is screened against immobilised antigen and four clones each with high affinity for the antigen have been identified. In addition to high affinity, what other criteria have to be satisfied in order that these antibodies can be used as a therapy for humans?

[20%]

It is important to examine the specificity of each of the antibodies in order to demonstrate that they not only bind tightly to the antigen but that they do not bind to other antigens. In addition it would be necessary to determine that the antibody was not toxic when used as a therapy.

**END OF PAPER**

### **Q1 PCR & Sequencing**

A very popular question. Across all the candidates it was possible to find very good answers to all parts of the question. Notes:

- 1a, d, e, g were mostly answered well.
- 1b was often answered poorly.
- 1c identified students who really understood the overall question and could reason about it. Many good answers.
- 1f was mostly answered poorly with students failing to identify and discuss the important issues.

### **Q2 Metabolic Engineering**

This was a relatively popular question. The quality of the answers varied widely and a few students did achieve relatively good marks. On the other hand, one student performed poorly demonstrating only limited understanding of the concepts taught. Students did rather poorly on sections (a), (b), and (f), where they were asked to provide some knowledge on basic concepts. One student chose not answer sections (f) and (g). Sections (c), (d), and (e) were attempted by all students, but the marks for the answers were quite varied

### **Q3 Genetic circuit logic gates**

This was a less popular question with a large and relatively even spread of marks. The majority of candidates demonstrated good understanding of common genetic elements and how genetic networks might be used to approximate logic functions. However, many students were unable to extrapolate this understanding to using the less familiar DNA recombinases to control gene expression.

### **Q4 Antibody diversity and engineering**

This was a popular question, as all students answered it. There was a large and relatively even spread of marks. Overall the knowledge of detail regarding antibodies seems to be lower than in previous years despite the lecture material remaining the same.