EGT2
ENGINEERING TRIPOS PART IIA

Friday 6 May 2022 2 to 3.40

Module 3G1

MOLECULAR BIOENGINEERING I

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.

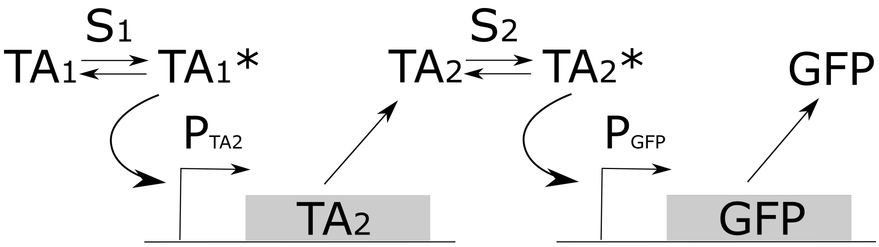
You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

You may not remove any stationery from the Examination Room.

1 Consider a cascade circuit, where a transcription activator TA1 activates the expression of a second transcription activator TA2, which in turn activates the expression of a GFP reporter gene. TA1 is constitutively expressed and is activated by signal S1. Actived TA1 causes TA2 to be expressed at the rate of E2, and TA2 is activated by signal S2. GFP expression starts when the level of TA2 exceeds the threshold K2.

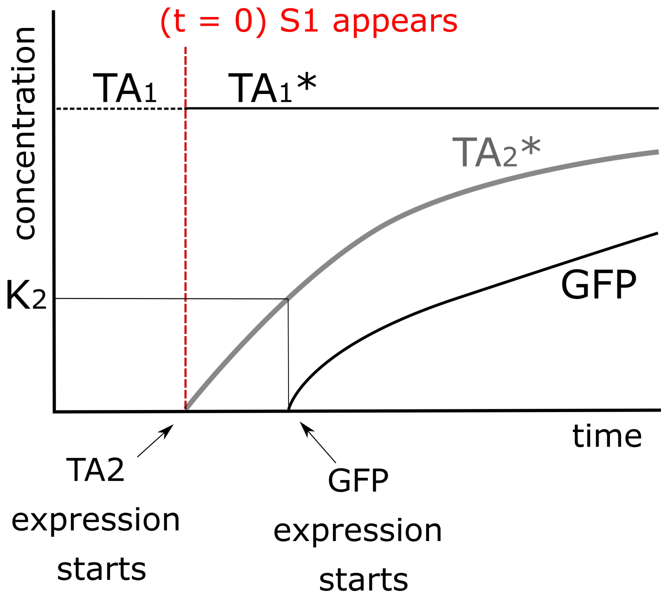
(a) Sketch a schematic of the circuit showing the different genetic elements (promoters and coding sequences), how the different signals regulate the transcription factors, and how the activated transcription factors regulate the different genes. [25%]

Crib:



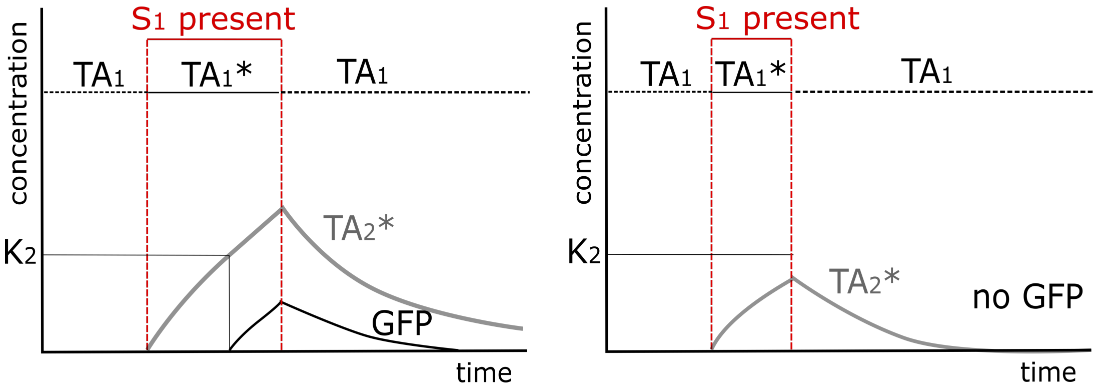
(b) Sketch a graph of the levels of TA1, TA2 and GFP in the situation where signal S1 appears at t = 0, and signal S2 is present throughout the experiment. [25%]

Crib:



(c) If the signal S1 occurs as a pulse of duration P and then vanishes, sketch how the levels of TA1, TA2 and GFP change. [25%]

Crib:



Left: medium pulse, Right: short pulse
Full marks for either or both

(d) What is the minimum value of P required for the activation of GFP expression? [10%]

Crib:

The concentration of active TA2 needs to exceed K2 for GFP to be activated. Since it is being produced at the rate of E2, the required duration is K2/E2. So, the minimum pulse duration is K2/E2.

(e) Cells carrying the circuit are exposed to the S1 signal for the minimum required pulse duration as above. However, no GFP signal is detected. What is the most likely explanation for this discrepancy? [15%]

Crib:

Treating cells with S1 for the minimum required pulse duration will lead only to a brief pulse of GFP expression. However, GFP takes a long time to become fluorescent (maturation time ~10 mins) after expression. Therefore the expressed GFP will be diluted before it can become fluorescent, attenuating the signal below the level at which it can be detected.

2 You have designed a synthetic genetic circuit to produce a substance of value. You know that such a circuit will impose a resource burden on the carrying host. Therefore you intend to quantify the effect of this circuit on host fitness. You design an experiment in which you co-culture the circuit-carrying strain (strain 1) with a circuit-free strain (strain 2), monitoring the culture to see how one competes with the other. To distinguish the two strains by microscopy, you express yellow fluorescent protein (YFP) from promoter P1 in the circuit-carrying strain, and cyan (blue) fluorescent protein (CFP) from promoter P2 in the circuit-free strain. Both of these fluorescent reporter genes are integrated into the bacterial chromosome.

(a) Strain 1 (carrying the genetic circuit on a plasmid backbone and expressing YFP from P1) has a doubling time of 24 minutes, and strain 2 (carrying no circuit and expressing CFP from P2) has a doubling time of 20 minutes, both in rich medium. What is the expected proportion of these two strains after 6 hours of growth in a mixed culture that had equal proportions of these strains at t = 0? [30%]

Crib:

When the growth of these two strains is not limited due to nutrients, we can approximate their growth as exponential. Therefore, after time = t, the ratio of the number of cells from strain 2 and strain 1 is:

$$\frac{S_2}{S_1} = \frac{2^{\frac{t}{\tau_1}}}{2^{\frac{t}{\tau_2}}}$$

τ_1 and τ_2 are the respective doubling time of strain 1 and strain 2

$$\implies \frac{S_2}{S_1} = 2^{t.(\frac{1}{\tau_2}-\frac{1}{\tau_1})}$$

For t = 360 minutes, τ_1 = 24 minutes and τ_2 = 20 minutes, we get

$$\frac{S_2}{S_1} = 2^{360.(\frac{4}{20.24})}$$

$$\frac{S_2}{S_1} = 2^3 = 8$$

Therefore, we expect to see 8 times more blue cells (CFP expressing) from strain 2 than the yellow cells (YFP expressing) from strain 1.

(b) If you were to perform the experiment from part (a) above five times, would you expect to see blue and yellow cells consistently in this proportion? Explain your answer. [10%]

Crib:

No.

Growth and division of individual cells are noisy processes and therefore we expect variable outcomes at the end of each experiment.

(c) When you actually carry out the experiment in part (a) above you are surprised that the proportions of colours observed are the opposite of those expected.

(i) Given reasons to explain this observation. [20%]

Crib:

The opposite of our expected observation is seeing a higher frequency of yellow (YFP expressing) cells than blue (CFP expressing cells) in the microscopy images. The possible explanation is that during the experiment, the cells of strain 1 lost the plasmid containing the synthetic genetic circuit and grew faster than those of strain 2. Since the circuit was the main source of slow growth and division, without the circuit strain 1 could grow faster. This observation also suggests that the burden from P1-YFP is actually lower than P2-CFP, possibly due to the higher strength of P2 as a promoter, and therefore without the circuit the strain 1 could grow faster than strain 2.

(ii) How can your explanations be verified? [15%]

Crib: If we place a fluorescent protein expression cassette (e.g. RFP) on the plasmid that carries the synthetic circuit, we could use its signal to detect when the plasmid is lost. So, we could check if the majority of the YFP cells at the end of the experiment are plasmid-free or not.

(d) The unexpected observation in (c) above suggests that there are problems in the design of this experiment. How could you change the strain construction to avoid these problems? [25%]

Crib:

The main problem arises due to loss of the circuit-carrying plasmid. We could integrate the synthetic genetic circuit into the chromosome so that it is not lost due to partitioning errors, or add antibiotic to the culture medium to maintain selection for the plasmid carrying the circuit. In this latter case the circuit-free strain would carry the empty vector.

The other problem is the differential burden arising from the different promoters used in the two strains. To estimate the actual burden of carrying this circuit, we should use the same promoter for expressing YFP and CFP in both strains, and ensure that they have comparable growth rates without the circuit being present.

3 The bacterial species *E. coli* is commonly used to produce proteins for medical or industrial applications. Plasmids can be used to carry the genetic circuits that allow expression of these proteins. Human insulin was one of the first recombinant proteins expressed in *E. coli*. In humans, insulin is expressed as a single polypeptide that is proteolytically processed such that two polypeptide chains (designated polypeptides A and B), which are bound together by two disulfide bonds, form the mature single protein unit. *E. coli* does not have the cellular machinery to process a single polypeptide in this way. Therefore the A and B polypeptides have to be expressed independently, purified, and then joined via an oxidation reaction.

We wish to make human insulin by synthesising DNA encoding a genetic circuit, rather than cloning the parts from elsewhere.

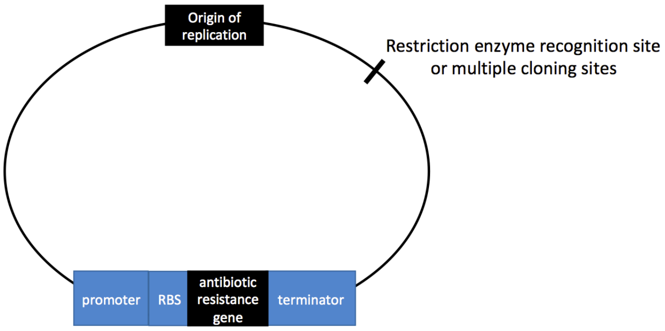
(a) First, we need to build a plasmid into which this synthetic DNA will later be cloned. Using a diagram, show the key DNA elements that should be present on the plasmid, and describe the function of each of them. [15%]

Crib:

(i) origin of DNA replication in order to allow propagation in the host cell

(ii) restriction sites/ multiple cloning sites in order to allow insertion of the synthetic DNA

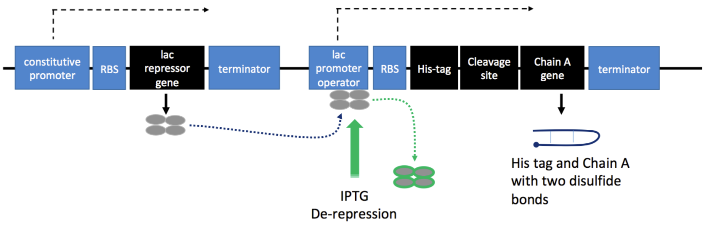
(iii) selectable marker i.e. an antibiotic resistance gene in order to allow the selection of host cells that contain plasmids.


Origin of replication
Restriction enzyme recognition site or multiple cloning sites
promoter
RBS
antibiotic resistance gene
terminator

(b) The original scheme for making human insulin A and B chains was based on making gene fusion polypeptides where, in separate constructs, part of the lacZ protein was fused with either the A or B polypeptides. These lacZ::A and lacZ::B fusion coding sequences were expressed under the control of the lac promoter, and the resulting proteins were purified using antibodies that reacted against the lacZ part of the fusion proteins. After purification, this lacZ protein fragment was cleaved off using proteases.

An adaptation of this approach would be to use a peptide tag such as six consecutive histidine residues (a 'His tag') instead of the lacZ component of the fusion, as such residues allow tight binding to a nickel column facilitating purification. Adjacent to the His tag would be a cleavage site for a specific protease. Draw the full genetic circuit for this approach applied to polypeptide A, showing all the parts and how they interact. [30%]

Crib:



The overall construct allows for controllable induction (de-repression) of the His tag::polypeptide A fusion protein gene expression. This requires a cell that expresses the lac repressor, which is the first part of the circuit. The *E. coli* cells can be grown to high density and then induced with IPTG to produce the His tag::polypeptide, so avoiding potential problems with the fusion protein expression being toxic or inhibiting cell growth. RNA polymerases (producing mRNA) and ribosomes (translating the mRNA) should also be shown. Note that the tag can in principle be at either the N-terminus or C-terminus of the protein, separated from the body of the protein by the protease cleavage site.

(c) To put the above scheme into practice, you would have to design a DNA sequence carrying the insulin polypeptide A under the control of the lac promoter. How would your design be influenced by the fact that you are intending to express a human open reading frame in *E. coli*? [5%]

Crib: You would codon optimise the construct by replacing the codons used in the human gene with ones favourable for expression in *E. coli*.

(d) Your designed DNA arrives from the manufacturer as linear double-stranded DNA. Using the plasmid vector from (a), describe the steps required for inserting the synthesised DNA into the vector. [15%]

Crib: Assembling the synthesised DNA into the plasmid vector: this could be done by ligation or e.g. Gibson assembly but either way the vector will have to be linearised either by cutting it with (an) appropriate restriction enzyme(s), or by PCR amplification.

(e) The circular DNA resulting from (d) above is transformed into *E. coli*. Describe the steps following this that are required to isolate clones, and to check that the DNA constructs carried by the clones are correct. [15%]

Crib:

The transformation mixture is spread on selective media (i.e. a nutrient agar plate containing an appropriate antibiotic for the selectable marker on the plasmid), and allowed to grow overnight. The resulting colonies are clones. Clones can be picked into liquid selective medium, grown by shaking at 37°C before DNA is isolated.

The structure of the clones might be checked either by PCR reactions across the vector/insert boundary to confirm the correct insertion has taken place, or by appropriate restriction enzyme digestions. In either case the products are run on agarose gel electrophoresis with molecular weight standards to check the size of the products, and the positive clones sequenced.

(f) *E. coli* cells containing the correct genetic construct identified above were grown by shaking at 37°C in rich medium but do not express polypetide A. You prepare plasmid DNA from such cultures but it is only present in vanishingly low quantities instead of being abundant. Assuming the plasmid preparation was carried out correctly, what is the simplest explanation with regards to the growth conditions? [10%]

Crib: An appropriate antibiotic needs to be added to the rich medium in order to select for the presence of the plasmid via its antibiotic resistance gene.

(g) You solve the problem in the previous step and repeat the experiment. The expression level of polypeptide A is now measurable but low. Analysis of mRNA shows that the gene is also being transcribed at low levels. What is the simplest explanation with regards to the growth conditions? [10%]

Crib: It is necessary to add IPTG to the medium in order to de-repress the lac promotor. Credit also if high lactose/low glucose.

4 In the central dogma of molecular biology, translation is the process by which information encoded in the sequence of four nucleic acid bases (A, C, G and T) determines the sequence of 20 different possible amino acids in a protein.

(a) Explain why a triplet coding system is needed to encode the 20 different amino acids. [10%]

Crib: 4^2 bases are only capable of encoding 16 amino acids. 4^3 enables 20 via a redundant code of 64 codons.

(b) During protein translation it is possible to insert a non-standard fluorescent amino acid at defined locations within an engineered protein by means of an 'orthogonal' aminoacyl-tRNA synthetase (aaRS)/tRNA pair. This aaRS adds ('charges') the fluorescent amino acid to its own target tRNA rather than to the native *E. coli* tRNA molecules. These fluorescently charged tRNAs are able to take part in protein translation. In order to make space in the genetic code the anticodon loop of the fluorescently charged tRNA recognises the least-used stop codon, UAG. Thus, by engineering this codon into the desired location in a protein's gene it is possible to introduce the fluorescent amino acid into a precisely defined location in the protein.

Using an established aaRS/tRNA system, a researcher wants to modify a protein, P, using this approach. They have engineered the gene for P by introducing a 5'-TAG-3' codon at the desired position. This construct has then been introduced into the *E. coli* strain that expresses the appropriate aaRS/tRNA pair. The engineered strain is grown correctly in culture medium containing the non-standard amino acid, X, which is taken up into the cell. However, the researcher is disappointed to find that around half of protein P variants are truncated at the site where amino acid X should be present.

(i) What is the most likely explanation for the observed protein truncation? [10%]

Crib: The conventional translation termination mechanism is still operating and in only half the cases does translation continue using the fluorescently charged tRNA.

(ii) In addition, the strain grows more slowly than the parent strain. Explain why this is the case. [10%]

Crib: Read-through of other TAG stop codons in the genome will lead to proteins with abnormal C-terminal extensions, and this will impact on their function and thus the growth rate of the cells. Partial credit was given to other credible explanations.

(iii) It is possible, but slow and expensive, to resynthesise bacterial genomes completely by progressively replacing segments of the bacterial chromosome in living cells. If time and money were not limiting, how would you redesign the genome to avoid the truncation problem above? [15%]

Crib: resynthesis the *E. coli* genome to eliminate one of the redundant amino-acid codons from the genome so it can be used instead of TAG. The corresponding tRNA would also have to be eliminated so one which recognises more than one codon would not be appropriate.

(c) Short peptide affinity tags are often added to proteins of interest in order to allow purification. For instance, the FLAG tag is an eight amino acid tag that is tightly bound by a particular antibody, which in turn can be attached to a solid substrate, so allowing purification. Explain where in the gene coding for the protein described in part (b) above you would place the tag and why you would choose this location. [15%]

Crib: At the C-terminus, as otherwise peptides generated by termination at the UAG will also be purified.

(d) Agarose gel electrophoresis is used at several stages when constructing recombinant DNA. Outline how it works. [15%]

Crib: a slab of agarose is cast with a comb making slots into which samples will be loaded. A voltage applied across the gel causes the negatively charged DNA to migrate away from the cathode. Smaller fragments are able to migrate through the gel more quickly and so mixtures of fragments are separated by size. The DNA is visualised by means of an intercalating dye such as ethidium bromide under UV illumination.

(e) After constructing a clone, it is normal practice to determine its sequence. Outline how Sanger capillary sequencing works when determining the sequence of a 300 base pair fragment inserted into a cloning vector of known sequence. [25%]

Crib:

A synthetic DNA primer complementary to sequences adjacent to the cloning site and directed 5’ to 3’ towards the cloning site is used to prime DNA synthesis. The DNA synthesis takes place in a buffer including all four dNTPs as well as DNA polymerase. In addition, the four distinct dye terminators are present at low concentrations. When incorporated into the growing strand these terminators prevent further extension while at the same time labelling the molecule with a dye, the colour of which indicates the base incorporated at that position. As all the synthesised molecules start with the primer and termination is at random locations, a population of molecules of different sizes is generated, corresponding to the locations at which each distinct dye terminator is incorporated. Capillary gel electrophoresis allows the sequence to be read from smaller molecules to larger as a series of peaks of different colours.

END OF PAPER

3G1 exam 2022

One of twenty marks for each question was given for each 5% of marks noted in the margin of each question part.

Q1 Cascade circuit: 32 candidates with mean 15.0/20 (SD 3.2)

The most-popular question. Some of the students seem to have confusion about signals acting directly as activators and got answers to (a) wrong. But most others got it right, with some difference in the details in the diagram. Answers to (b) and (c) were usually good and demonstrated that most students have clear understanding of the dynamics, but some students failed to consider the dilution aspect of the concentration control. (d) most students got this right and some students went the extra mile to calculate the limit without the assumption of ignoring dilution at short times. Answer to (e) was variable and only few of them identified maturation delay being the main reason for the discrepancy.

Q2 Host fitness: 31 candidates with mean 12.7/20 (SD 3.1)

Almost all the students answered (a) correctly, except a few that confused growth with dilution. Some students got answer to (b) right, while a significant fraction confused the noise in division times with the noise in gene-expression. While these two are related, in most of the answers the explanation was wrong. (c) Very few students got this part right, but some of them got very close, as was the case for (d), which requires a test design for verification. The answer to (e) depends on their understanding of the problem in (c), so many students got it wrong. But some of them provided many alternative explanations and got close to part of the solution.

Q3 Insulin production: 19 candidates with mean 11.1/20 (SD 2.8)

The least popular question. For (b) students tended not to consider the lac repressor as part of the circuit and lost marks accordingly and many failed to consider that they were sketching an adaptation of the system described in the question. For (e) there was a tendency not to explain how clones are isolated. (f) and (g) needed to have been read carefully and understood properly and were often answered poorly. Across the cohort every part of this question was answered correctly, yet no individual student did strikingly well.

Q4 Non-standard amino acid incorporation: 31 candidates with mean 12.2/20 (SD 3.1)

One of the most popular questions. Almost all students answered (a) correctly. (b) and (c) were good at identifying students who really understood the lecture material and could reason with their knowledge. (d) was answered well, with performance on (e) being more patchy and a tendency to describe the original four-pot Sanger sequencing.