ENGINEERING TRIPOS PART IIA

---

*day * M** 2014 *.00 to *.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.*

Answers to questions in each section should be tied together and handed in separately.

There are no attachments

STATIONERY REQUIREMENTS	SPECIAL REQUIREMENTS
Single-sided script paper	CUED approved calculator allowed

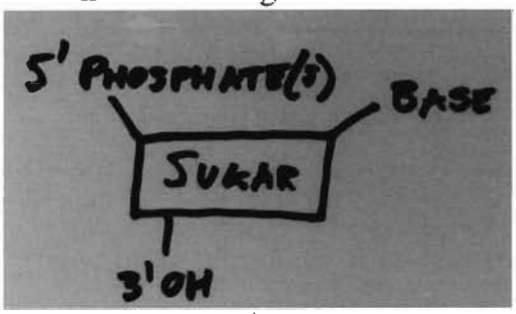
You may not start to read the questions printed on the subsequent pages of this question paper until instructed that you may do so by the Invigilator

1 (a) (i) Describe the structure of DNA [30%]

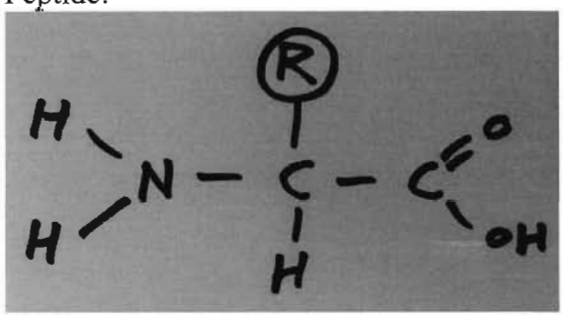
DNA is composed of two antiparallel strands, which twist together into a double helix. Each strand is a polymer of nucleotides joined with phosphodiester bonds. The nucleotides differ in their bases, the information storing components of DNA, which can be adenine (A), thymine (T), cytosine (C) or guanine (G). The two strands are held together by hydrogen bonds between the bases. A always pairs with T and C always pairs with G. As A and G are guanines (larger) and C and T are pyrimidines (smaller), the overall width of the molecule is constant.

(ii) Draw the generic monomers for RNA and proteins [10%]

RNA nucleotide: sugar = ribose

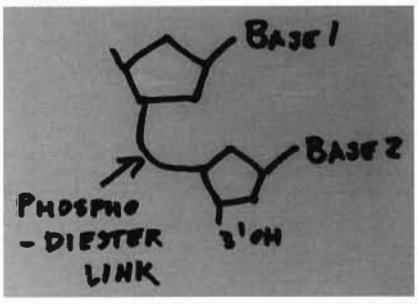


Peptide:

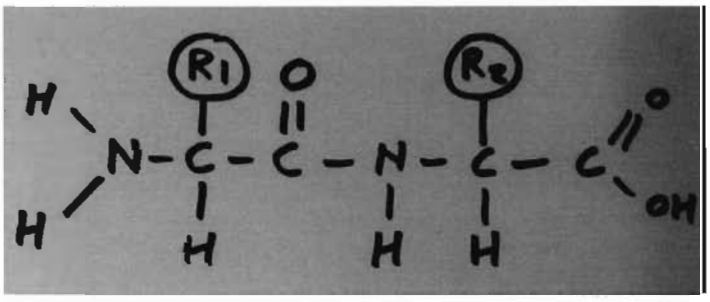


(iii) Show how two adjacent monomers are bonded together in RNA, and in proteins. [15%]

RNA:



Proteins:



(b) It has been found helpful to describe protein structures and protein complexes in terms of a hierarchy of different structural forms.

(i) Name and describe the different levels of protein structure hierarchy and say what the most important bonding types are for the first two levels. [30%]

Primary: *the sequence of different amino acids help together with peptide bonds.*
Secondary*: the folding of the primary structure into alpha-helices and beta-sheets held together with hydrogen bonds.*
Tertiary*: the overall, complex, packing of the secondary structural elements and any remaining primary structure into the mature protein.*
Quaternary*: the fitting together of tertiary structure into a protein complex*

(ii) The amino-acid side chains confer the structural and functional properties of proteins: describe three distinct types of property of amino acid side chains. [15%]

Examples include: Acidic and basic side chains
Hydrophobic and hydrophilic side chains
A side chain that can form disulphide bridges
Side chains that can be phosphorylated
Bulky and small side chains.

2 Limonene is a monoterpene with a strong orange smell found in citrus fruits and is used as a renewable solvent in cleaning products. Derivatives, such as perillyl alcohol, have medicinal uses. A recent paper describes the production of limonene and perillyl alcohol in *E. coli* (See Figure 1 below).

(a) In this study, the enzyme that produces limonene from GPP comes from the grand fir (*Abies grandis*):

(i) How might the authors have found this enzyme (assume that there are no published reports on this exact enzyme)?

The authors may have guessed as to the family of enzymes that perform reactions of this type, and then done a sequence comparison between species that do, or do not, produce limonene in order to find enzyme family members specific to those species that produce limonene.

(ii) How might this exogenous enzyme be expressed in *E. coli*?

To express this enzyme in E. coli, the authors would codon optimize the sequence for expression in E. coli then put it under the control of an appropriate promoter (either constitutive at an appropriate level, or inducible).

(b) IPP (along with other compounds in the pathway) is an important metabolite for many processes in *E. coli*.

(i) What potential problems would this raise for the design process?

If flux is diverted away from IPP by the heterologous pathway that produces limonene, cellular processes that require IPP may be inhibited, which may result in slow growth and therefore a reduction of overall yield. In addition, enzymes that convert IPP into other required products could divert flux away from the heterologous limonene pathway, reducing the rate of limonene synthesis.

(ii) How could these problems be solved?

To fix the slow growth, the entire pathway would have to be optimized to maximize both growth and synthesis of limonene by adjusting flux through the heterologous pathway. To fix diverted flux, endogenous E. coli enzymes could be put under the control of repressible promoters to reduce flux through undesirable side pathways.

(c) You think the authors may not have expressed all the enzymes in the pathway at the optimal levels. How would you go about determining which enzymes would be the best targets for changes in expression level to optimize the pathway.

You could perform metabolic control analysis: increase each enzyme in small increments and measure the resultant change in flux through the pathway. This would give you the control coefficients for the enzymes involved. You would want to manipulate the levels of those enzymes with the highest control coefficients. However, it is likely that control would be shared. This means the best strategy would probably be to increase the levels of all of the enzymes comprising the parts of the pathway through which the greatest flux is desired.

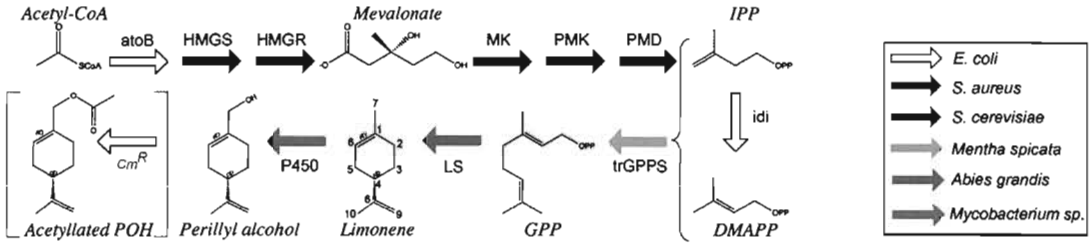


Figure 1: Production of limonene and perillyl alcohol in *E. coli*

3 (a) (i) Describe (with a diagram if appropriate) the Sanger dideoxy method of DNA sequencing.

Students should clearly explain how Sanger sequencing works – either the 'conventional' 4-pot reaction or the more modern dye-terminator method is acceptable. For instance, for the former: a sequencing primer is annealed to a population of identical template DNA molecules. The primer is a sequence that is complementary to the location in the template molecules from which the DNA sequencing reaction will start. The reaction also contains the four nucleotide monomers dATP, dTTP, dCTP and dGTP as well as DNA polymerase enzyme and ATP as an energy source. The above would be sufficient to allow elongation of a new DNA strand initiating at the primer but would not allow the reading of the DNA sequence. This is done by means of so-called chain-terminator nucleotides: these are 2,3-dideoxynucleotides that, when incorporated in the growing strand, do not provide the 3'hydroxyl group needed for further chain extension. If four separate reactions are set up which individually contain one each of the four possible dideoxynucleotides at an appropriate concentration, then, in a fraction of the molecules, chain termination will occur at every possible position. The resulting populations of molecules can be run on adjacent tracks of a polyacrylamide gel and the DNA sequence read from the ladder of bands.

(ii) What limits the length of sequence that can be read?

The length limitation is set by the ability of electrophoresis to resolve the reaction products according to size.

(b) Most current "next generation" sequencing methods require multiple copies of a template molecule for the actual sequencing step. Describe a method by which these copies can be made.

Could be Illumina-style surface-amplification or Ion Torrent style bead emulsion PCR. For instance, for Illumina, single template molecules are annealed to surface-immobilised primers. This is possible because the template molecules have had added, at either end, the appropriate generic sequences. A mixture that supports DNA replication is added and the immobilised primer is extended on the template. Heat cycling results in the original template being washed away. The free end of the copy of the molecule that is immobilised can itself anneal to another primer, and further DNA replication results in two complementary copies of the original template molecule. Subsequent rounds of heating and DNA replication double the number of molecules at each step (assuming unlikely 100% efficiency) and after ~10 cycles there are ~1000 molecules forming a spot that eventually will be made fluorescent and read as part of the sequencing reaction.

(c)~~ Some of the latest sequencing technologies, such as the Pacific Biosciences "SMRT" (single molecule real-time) system operate directly on single DNA molecules. What are the potential advantages of truly single-molecule methods, compared to methods that require multiple copies of the template molecule?

Advantages may include simpler template preparation as there is no need to add generic sequences at either end of the template molecules; reduced bias due to not requiring amplification steps; the possibility of longer reads because there is no need to keep multiple template molecules reacting in synchrony; ability to read through longer repeats due to relatively long reads compared to e.g. Illumina; and sensitivity to modified bases e.g. methylation.

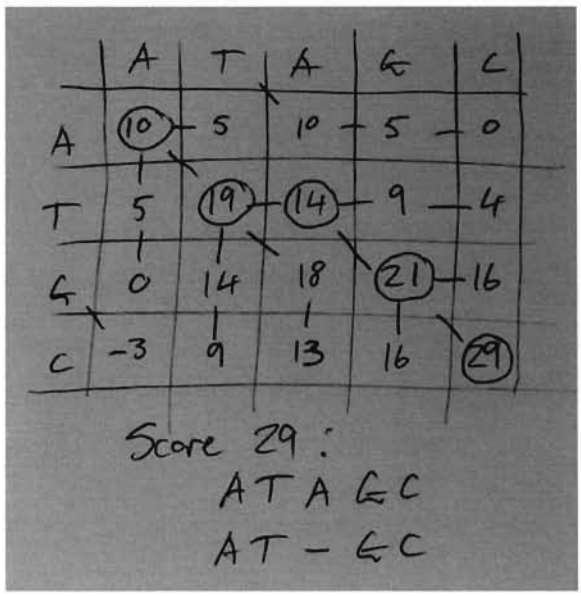
(d) In finding the best global alignment between two DNA sequences, the overall score of the finished alignment is calculated from a similarity matrix which assigns a positive score to each pair of aligned bases which are identical, a lower or negative score to each pair of aligned bases which are not identical, and a negative score (a *gap penalty*) for gaps which are introduced into either sequence.

The following is a typical similarity matrix giving scores for matched and mis-matched pairs of bases:

	Base in first sequence:	A	G	C	T
Base in second sequence:					
A		10	-1	-3	-4
G		-1	7	-5	-3
C		-3	-5	8	0
T		-4	-3	0	9

Using the above similarity matrix, and a linear gap penalty of -5 per nucleotide position, use dynamic programming to prove that the best global alignment between the two sequences **ATAGC** and **ATGC** is:

```
ATAGC
AT-GC
```



(e) In the similarity matrix above, the scores on the diagonal (for matched bases) range from 7 to 10. Likewise, the scores for non-matching (non-diagonal) pairs range from -5 to 0. Give possible reasons for this variation amongst the diagonal values and amongst the non-diagonal values.

Possible reasons include the base frequency in the genomes under consideration (for example, if the genome is rich in Gs but poor in Ts, an alignment of two Gs is less significant than an aligment of two Ts); and the tendency for some mutations to be more common than others (for example, C-to-T mutations may be more common, so that T/C mismatches carry less of penalty than other mismatches).

4 The formation of bacterial biofilms is important to disease pathogenesis. The signalling processes that underlie the formation of biofilms may be understood in terms of 2 dimensional pattern formation. The following three plasmids (Figure 2) were made to build up a genetic circuit that is based on acyl-homoserine lactone (AHL) signalling between individual bacteria.

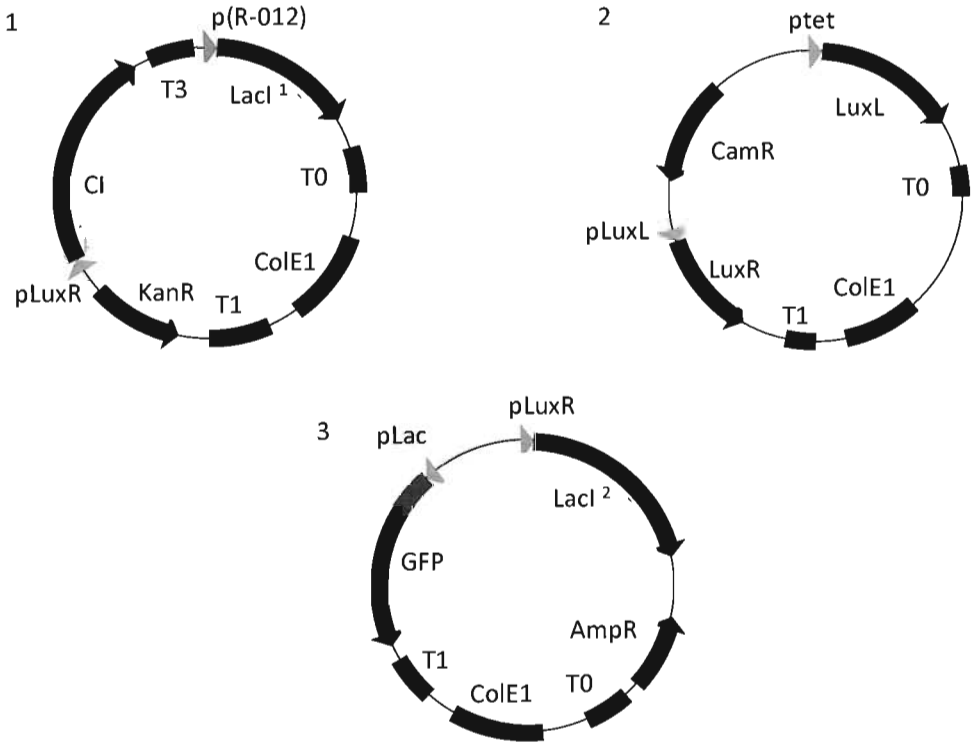
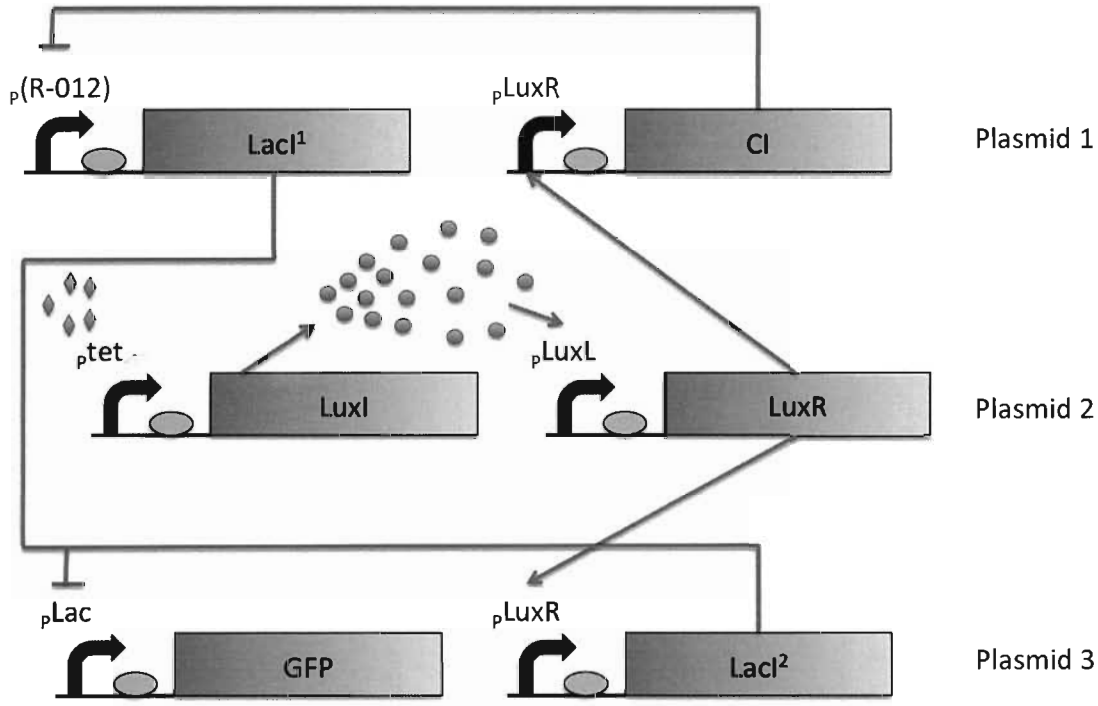


Figure 2: Three plasmids, which together can be used to investigate biofilm pattern formation in bacteria.

(a) Using the plasmid diagrams provided above (Figure 2), fill in the six promotor and gene names in Figure 3, as a first step in constructing a genetic circuit diagram.



Blue arrows represent activation; red bars represent inhibition.

Figure 3: Genetic circuit diagram. Based on the three plasmids shown in Figure 2, fill out the promotor and gene names in the blank boxes, as shown in the example at the bottom.

(b) Complete the genetic circuit diagram in Figure 3, by using the information shown in Table 1 (below), noting that the output of the circuit is the GFP gene product, which is fluorescent.

pLuxL	Promoter activated by LuxL gene product
pLuxR	Promoter activated by LuxR gene product
pLac	Promoter repressible by lacI gene product
ptet	Tetracycline activated promoter
p(R-012)	Promoter repressible by CI gene product
lacI^1	Wildtype Lac repressor gene. In the absence of Lactose or IPTG its gene product represses pLac.
lacI^2	Mutant Lac repressor gene. Its gene product binds the pLac promoter less efficiently than lacI^1. This means that low levels of lacI^2 will not repress pLac.
LuxL	Gene product activates the pLuxL promoter.
LuxR	Gene product activates the LuxR promoter.
CI	Gene product represses p(R-012) promoter activity.
AHL	Acyl-Homoserine Lactone, a molecule responsible for inter-cellular communication
GFP	Green Fluorescent Protein gene.
KanR	Kanamycin resistance gene.
AmpR	Ampicillin resistance gene.
CamR	Chloramphenicol resistance gene.
ColE1	*E.coli* origin of replication

Table 1: list of parts used in the plasmids, and their functions.

(c) A bacterial culture harbouring all the above plasmids is spread on an agar dish onto the centre of which is placed a small paper disc saturated with tetracycline. Taking into account the relationships between component activities shown in Table 2 (below), sketch the resulting pattern of bacteria expressing GFP.

A circle of GFP expressing bacteria will form around the tetracycline disc, and this GFP circle will be surrounded on both sides by zones in which GFP is not expressed.

AHL	CI	LacI[1]	LacI[2]	GFP
++	++	-	++	-
+	+	-	+	+
-	-	++	-	-

Table 2: relationship between different components of the genetic circuit when in different states.

(d) Explain how the genetic circuit gives rise to the pattern shown in your answer to (c).

At high concentrations of AHL the production of the LacI[1] will be high such that GFP expression is repressed. At low concentrations of AHL the production of LacI[2] will be high such that GFP expression is repressed. At intermediate concentrations of AHL the LacI genes are insufficiently expressed to overcome the pLac promotor that drives the GFP gene, which is therefore expressed. AHL activity varies with distance from the tetracycline disk due to the concentration gradient of tetracycline caused by diffusion acting on the ptet promotor that makes AHL.

END OF PAPER