EXERCISE 1: PCR AND PRIMER DESIGN

Theory of PCR and primer design

Polymerase chain reaction (PCR) is a widely used technique for the selective amplification of particular DNA sequences, such as individual genes. You specify the sequence with two short "primers", which flank the region of DNA to be amplified. Primers are DNA oligonucleotides around 20-30 bases in length and the design of these is vital to the success of the PCR. PCR primers are important as they are complementary to the beginning and end of the DNA fragment of interest which one needs to amplify. Any sequence a primer is "complementary" to is repeatedly replicated by DNA polymerase in the reaction. In principle, a single molecule of DNA can be amplified to detectable levels. PCR is widely used in forensic medicine and clinical diagnostics. It can even be used to amplify DNA from the hides of stuffed museum specimens, to "bring back to life" genes from extinct species!

This exercise should aid in the understanding of this technique and be a guide for the design of PCR primers.

A. Understanding PCR

DNA polymerase uses single-stranded DNA as a template for synthesis of a complementary new strand. Both DNA strands of a double helix can serve as templates for synthesis, provided a primer is supplied for each strand. The PCR starts with heating the DNA to separate the two strands, the primers then anneal to their complementary binding sites, and new chains are synthesised. The cycle of heating, primer binding and extension is repeated many times to yield enough DNA that it can be easily visualised on an agarose gel (as you did on DAY 1). The following figure shows a diagrammatic representation of the PCR process, allowing you to visualise the exponential increase of DNA copies.

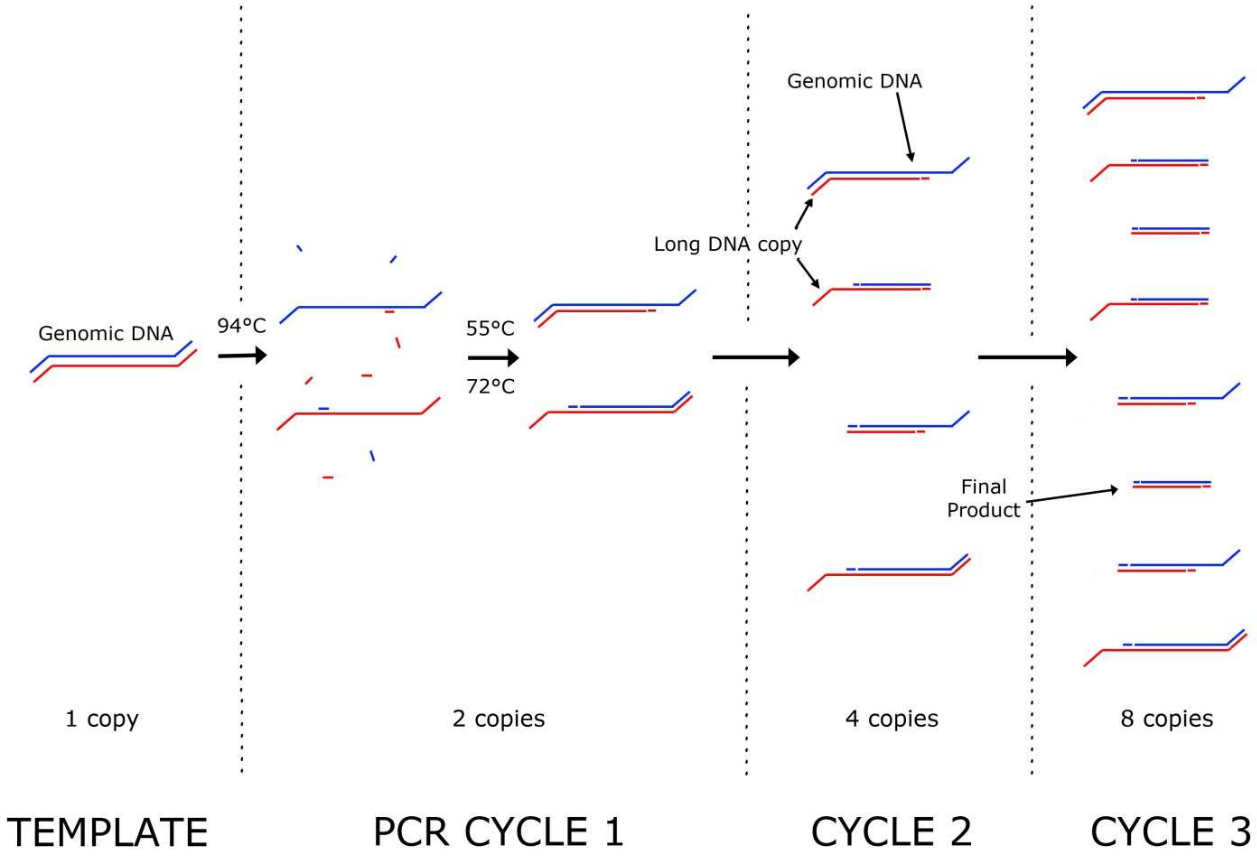


Figure 1 First Three Cycles of a Polymerase Chain Reaction

Cycle 1: After heating to 94°C, the double-stranded DNA dissociates. On cooling quickly to the annealing temperature (typically 40-55°C), the primers anneal to their complementary sequence flanking the target region. *Taq* DNA polymerase clamps on to the 3’ end of the primer annealed to the DNA template and extends it. So, on heating to the extension temp of 72°C, both our pieces of DNA will be extended, typically for about a kilobase. There are now 2 double-stranded copies containing the region between the primers.

Q1. **In cycle 1** does the *Taq* polymerase extend beyond the position of the primer on the opposite strand. **Why is this the case?**

Cycle 2: The reaction mixture is heated again; the original and newly synthesised DNA strands separate. Primers will bind again at the appropriate annealing temperature and then Taq polymerase synthesises new complementary strands. The extension of these chains is limited precisely to the target sequence. The two newly synthesised strands thus span exactly the region specified by the primers.

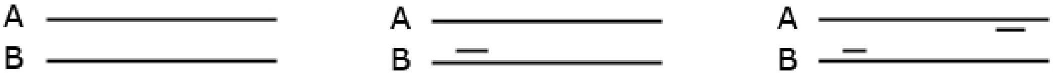
Cycle 3: The process is repeated, and primers anneal to the newly synthesised strands. Taq polymerase synthesises complementary strands, producing double-stranded DNA fragments that are identical to the target sequence. The process is repeated and the number of target fragments doubles for each subsequent cycle of the reaction.

Typically, a PCR programme continues for 30 cycles. Because the replicated strands subsequently act as templates in subsequent cycles, the number of copies increases exponentially - hence the name “chain reaction”.

What happens when there is a match for only one primer on the template DNA?

Imagine that you have 3 pieces of double-stranded DNA. One piece of DNA contains no matches to either primer, one contains a match for just the forward primer, and the third piece of DNA contains a match for the forward primer on one strand and the reverse primer on the other strand.

Q2. Calculate, for the cycle numbers shown below, how many copies (including the original) of strand A and strand B for each DNA exist at the end of each cycle



After PCR cycle…	**No match for either primer**		**Match for left primer only**		**Match for both primers**	
	A	**B**	**A**	**B**	**A**	**B**
1						
2						
3						
4						
5						
30						

Q3. If 30 cycles of the PCR are carried out on a sample containing 200 dsDNA templates, how many dsDNA copies are made after (a) 3 PCR cycles, and (b) 30 cycles?

B. PCR primer Design

To design primers we need some sequence information, and now that several genome projects have been completed, we can get the information we require from online databases. Once a suitable sequence has been chosen the normal practice is to order the required primer from a supplier – you send them the sequence you want made and they synthesise it using chemical technology. For about 25p per nucleotide (so roughly £6 per primer) you can get enough of a primer to carry out hundreds of reactions. But you have to make sure that the sequence you send is in the correct orientation, ie, 5′ to 3′, and this means that, for a reverse (or right) primer you need to deduce the complementary strand sequence.

In a PCR, there are two primers and you must order them in the correct orientation from 5′ to 3′ – see below for an example:

5′ AGGTCAGATA**CAGATGGATACGCAGTGCAGA**TCCGATACAGATCA 3′

REVERSE PRIMER **3′ CGTCT 5′**

FORWARD PRIMER **5′ CAGAT 3′**

3′ TCCAGTCTAT**GTCTACCTATGCGTCACGTCT**AGGCTATGTCTAGT 5′

Reverse (or Right) Primer: TCTGC (reverse of CGTCT)

Forward (or Left) Primer: CAGAT

To see if you understand here are a couple of primers to design. We use the Neuroglian genomic DNA sequence below which shows part of intron 2 followed by the start of exon 3 (**exon sequences are in bold face**).

Q4. Deduce the forward and reverse primer sequences, correctly orientated (5′ to 3′) to the underlined nucleotides below. Remember DNA is double stranded and we have only shown one of the strands to save space. It may help if you write out the underlined areas as double strands.

GCAAACTATT ATGTGTATTA GTATTCTATT GGTATTCATT ATCGACTGAC ACCCCCCACT ATTTCTACTT

TTAATCTACA G**ATATAGTTG GATCAAGAAC GGCAAGAAGT TCGATTGGCA GGCGTACGAT AACCGCATGC**

TGCGGCAGCC AGGACGTGGC ACCCTGGTGA TCACCATACC CAAGGACGAG GATCGCGGCC ACTATCAGTG

CTTTGCGTCC AATGAATTCG GAACGGCCAC CTCGAACTCA GTATATGTGC GTAAGGCCGA GCTGAATGCC

TTCAAGGATG AGGCGGCCAA GACACTGGAG GCCGTCGAGG GTGAGCCCTT TATGCTGAAA TGTGCCGCAC

CCGATGGTTT TCCCAGTCCG ACAGTCAACT GGATGATCCA GGAGTCCATC GATGGCAGCA TCAAGTCGAT

CAACAACTCT CGCATGACCC TCGATCCTGA GGGTAATCTC TGGTTCTCGA ATGTTACCCG TGAGGATGCC

AGCTCCGATT TCTACTATGC CTGCTCGGCC ACCTCGGTGT TTCGCAGTGA ATACAAGATT GGCAACAAGG

TGCTCCTCGA TGTCAAACAG ATGGGCGTTA GTGCCTCGCA GAACAAGCAT CCGCCCGTGC GTCAATATGT

TTCCCGTCGC CAGTCCTTGG CGTTGCGTGG CAAGCGAATG GAACTGTTTT GCATCTACGG TGGAACACCG

CTGCCGCAGA CCGTGTGGAG CAAGGATGGC CAGCGTATAC AGTGGAGCGA TCGAATAACG CAAGGACACT

ATGGCAAATC ACTGGTCATT CGGCAGACAA ATTTCGATGA TGCCGGCACA TACACCTGCG ACGTGTCCAA

CGGTGTGGGC AATGCCCAAT CCTTCTCCAT...

Figure 2 Part of the Neuroglian genomic sequence. This shows part of intron 2 followed by the start of exon 3 (in bold).

C. Rules of Primer Design

Good primer design is essential for successful PCR. The important design considerations are summarised below and are key to specific amplification with a high yield.

Sequence	• Avoid long runs of a single base eg. ACTGGGGGGGGCA • Have a G or C at the 3′ end • Avoid primer secondary structure eg self annealing and hairpin loop • Primer sequence must be unique to the DNA template. Actual DNA sequences are not random so some are more common that others and must be avoided.
Length	Optimal primer length is 18-30 nucleotides
GC content	40-60%
T_m	• Melting temperature (T_m) between 55°C and 65°C • T_m = 2°C x (A+T) + 4°C x (C+G)

Primer Length

Primers should ideally be 20 to 25 nucleotides long (range 18 to 30 nucleotides). This ensures that they will anneal to a unique sequence and therefore be specific. At the start of a PCR there will be very few template molecules, but sufficient primer to generate billions of copies of the template; in other words there will be billions of primer molecules. All these primers will be trying to locate a sequence to anneal to, and, since there will not be enough target DNA template, we have to ensure that there are unlikely to be other matches, or close matches, for our primer elsewhere in the template. This is ensured by using a sequence long enough that it has a very good chance of having a unique annealing site within the template. But we don't want the primer to be too long as efficiency of annealing decreases with increased primer length.

Q5. How many times are you likely to find (a) a particular 12 nucleotide sequence, and (b) a particular 18 nucleotide sequence within the human genome of 3.3 x 10^9 nucleotides (for the purposes of the calculation, assume a GC:AT ratio of 1:1).

Q6. What does the result from the previous question tell you about the importance of primer length?

GC content and T_m

The GC content should be 40 to 60% to ensure good annealing of the primer.

The GC content and melting temperature are clearly related: GC pairs have three H-bonds, whereas AT base pairs only have two H-bonds. So GC-rich sequence anneals more strongly, and we want our PCR primers to anneal efficiently.

The formula to determine the T_m of a primer uses nearest neighbour thermodynamic calculations, requiring enthalpy, entropy and molar gas constants to be considered. Fortunately, there are computer packages which will carry out the calculation for you! For sequences of the length of most PCR primers a reasonable approximation of the T_m can be generated by the equation: T_m = 2(A+T) + 4(G+C). Most importantly, it is crucial that the two primers have very similar annealing temperatures. If the T_m for one primer is, say 10ºC lower than the T_m for the second primer, then at the relatively low annealing temperature required for the first primer it is very likely that the second primer will be annealing to additional, close matches in the template DNA. For every PCR we need to choose an annealing temperature which is low enough to allow the primers to anneal, but high enough to prevent mismatches in primer annealing. This temperature is usually about 5ºC below that calculated by the above equation, and the best temperature to use for a primer pair is often determined experimentally.

Q7. Using the above equation, work out the T_m for the sequences chosen for the forward and reverse primers which you designed to the neuroglian sequence shown in Figure 2 (see previous question).

3’ end

The 3’ end base should be a G or a C, or even better, GC or CG.

Q8. Why is it best to have a G or a C at the 3’ end of the primer?

Sequence self complementarity

There should be no intra-primer or inter-primer homology, particularly at the 3’ends. The 3’ end bases of the two primers must not be complementary, and in other regions of the primer, there should be no homologies longer than 3 base pairs. As mentioned earlier, there will be billions of copies of primer and few copies of template during early reaction cycles. This means that the most likely molecule a primer will meet in the reaction tube is another primer molecule!

Q9. What would happen if, say, the forward primer had the sequence 5’GTTCGCATTCGAATGCGAAC3’? Draw below the structure a single primer molecule could form, and the structure formed by two of these primers together and suggest why a PCR using this sequence for one of the primers will be likely to fail.

Q10. What would happen if the forward primer had the sequence 5’GTTCGTCTCAACGAAGTC3’ and the reverse primer had the sequence 5’GGAAGACATCTGGTCGAC3’?

Polypurine and polypyrimidine sequences

These are to be avoided. Runs of Gs or Cs will generate primers which have a high propensity to anneal to GC-rich regions of the template DNA and therefore mis-prime. Runs of As or Ts will promote dissociation of that region of primer from the template due to the weaker H-bonding. Polypurine (A,G) and polypyrimidine (C,T) tracts should also be avoided. Aim for a pretty much random distribution of nucleotides.

D. Designing PCR primers which generate restriction sites at the ends of the PCR product

Cloning of PCR products is complicated by the fact that *Taq* polymerase tends to generate single base overhangs, usually (but not always) A, at the 3′ end of the product. Some kits for cloning PCR products actually take advantage of this, and produce cut vector, ready for use, that has a single T overhang at each 3′ end. However, this is expensive, so you are using the alternative strategy, which is to introduce restriction enzyme sites into the ends of your PCR product by adding the appropriate sequence to the 5′ end of your primer. By doing this, you produce a primer which is deliberately mismatched with the template at the 5′ end. But this is OK – the only position where mismatches are not tolerated by PCR is at the very 3′ end of the primer.

What happens when a primer mismatches the template at the 5′ end? When the primer anneals to the original template, the mismatched region will not pair with template as it will not be complementary. But each strand we make in PCR becomes a template for all the subsequent rounds of PCR (see the diagrams at the beginning of this exercise), and so the "mismatched" sequence of the primer is copied faithfully into every subsequent PCR product.

The only other point we need to bear in mind is that only some restriction enzymes will cut close to the end of a DNA molecule, and the ones that do generally require at least 3 or 4 bases **to the 5′ end** of the recognition site in order to cut the DNA efficiently.

Q11. Using the above information together with the selected forward primer sequence indicated on the sequence on Figure 2, write out a suitable primer (5′ to3′) which will incorporate a KpnI site into the PCR product (the KpnI recognition sequence is GGTACC) and allow cleavage of the end of the PCR product by KpnI.

Q12. Draw a diagram, with all 5′ and 3′ ends marked, showing how this left primer, with the additional sequence, will anneal to the *Nrg* template

Primer Design Helper

http://goo.gl/uroojK

The following part of this exercise is not done by all degree groups. Your lab leader will inform you if you are doing the next section.

Using RT-PCR to characterize products of alternative splicing

It has been estimated that approximately 60% of human genes exhibit alternative splicing of exons, with an average of about 3 alternatively spliced transcripts per gene, and that 70% of these alternative splicing events generate different versions of the protein product. Some genes generate large numbers of alternate transcripts, and thus obtaining the primary sequence of a gene is only the start of the task of characterizing how that gene functions. To see the alternate transcripts we must obviously analyse the mRNA, and one method is to employ reverse transcriptase PCR (RT-PCR).

Alternate splicing produces two versions of the neuroglian protein: there are two versions of the last exon, exon 7: 7a and 7b. mRNAs with exon 7a encode the shorter, non-neuronal neuroglian (167kDa) and mRNAs with exon 7b encode the longer, neuronal neuroglian (180kDa).

```
CGGATGGTTC ATTGGCATGA TGCTGGCCCT GGCCTTCATC ATCATCCTCT TCATCATCAT CTGCATTATC CGACGCAATC
GGGGCGGAAA GTACGATGTC CACGATCGGG AGCTGGCCAA CGGCCGGCGG GATTATCCCG AAGAGGGCGG ATTCCACGAG
TACTCGCAAC C/5..66nt..6/GTTGGATAA CAAGAGCGCT GGTCGCCAAT CCGTGAGTTC AGCGAACAAA CCGGGCGTGG
AAAGCGATAC TGATTCGATG GCCGAATACG GTGATGGCGA TACAG/6..1556nt..7a/GCATG AATGAAGATG
GATCCTTTA TTGGCCAATA TGGACGCAAA GGACTTTGAT TTAATTAGTA AGCAGCGCAC CGCAACAGCA ACTCAAAAAT
AATATCGAAA CCGAGCCCTT AACCCCAAAA ATCAAAAAAT CAACAAGACC AAACACCATC ACAGCAGAAA AATGAAAAAA
TTAATGAAAA TAATAGTAGC CTACATTTTA TTCGACTATA AGTGCAAACA CCACGACTAA TTTAAAGTAT ATATAAAAAT
AGAGGTTTTA TATATAACTA TTAAAATCTT AAAATGTGTA AAAAAAAAAA CAAACAA/7a...1356nt...7b/GACAA
TTTACCGAGG ATGGCTCCTT CATTGGCCAA TATGTTCCTG GAAAGCTCCA ACCGCCGGTT AGCCCACAGC CACTGAACAA
TTCCGCTGCG GCGCATCAGG CGGCGCCAAC TGCCGGAGGA TCGGGAGCAG CCGGATCGGC AGCAGCAGCC GGAGCATCGG
GTGGAGCATC GTCCGCCGGA GGAGCAGCTG CCAGCAATGG AGGAGCTGCA GCCGGAGCCG TGGCCACCTA CGTCTAAGAG
GCGTGGCTGG GATTCACTTG CCCCATTGTT CTCCTGATTT TCTACCAAAC GATTCAAACG CCTCTTAAAC AAAAAAGAAA
CTGTGTAATT CTATGTGTAA AACGAAAACT GCTTTAAGTG TTCTGCAAAA AA/7b
```

The coding sequences for *Drosophila* neuroglian exons 5 (3′ end only), 6, 7a and 7b. Splice junctions are shown (/), and intron length is indicated. Stop codons are underlined.

Q13. Explain how RT-PCR might be utilized to determine which *Drosophila* tissues express the non-neuronal neuroglian transcript. Mark on the sequence regions which would generate suitable primers for this, bearing in mind the rules for primer design.

Q14. How could you design your primers so that they will be specific for cDNA and not anneal to any genomic DNA which has contaminated your mRNA preparation?

From protein to primers

The similarity between genes may be rather limited, particularly if the shared function has emerged independently through convergent evolution. Can we use similarity at the protein level as the basis for the design of PCR primer? The answer is "yes" but as we'll see, the design of primers in this way demands careful thought.

Alignment of the protein sequence of *Drosophila* Nrg with other entries on the protein database reveals that the amino acid sequence FNEDGSFIGQ is well conserved amongst neural cell adhesion molecules from a range of species. Imagine we wanted to isolate part of the cDNA for an *Nrg* homologue from an animal about which little was known at the genetic level. Could we exploit the conservation of the amino acid sequence in designing a PCR primer? The deduction of the DNA consensus is not as straightforward as you might think. Although any given codon always codes for a given amino acid, a given amino acid may have more than one possible codon. So although any DNA sequence codes for a specific amino acid sequence, a given amino acid sequence can be coded by many different DNA sequences. This is known as **degeneracy**. You will see from the genetic code at the end of the manual that different amino acids can be coded by 1, 2, 3, 4 or even 6 codons. So choosing a **well-conserved** amino-acid stretch is not enough - one with **low degeneracy** (ie as few different ways of coding for the amino acids as possible) is necessary for successful primer design.

Q15. Why do we want to minimise degeneracy in designing a PCR primer?

Using the table below, evaluate if the nominated sequence would be useful for primer design. In Row 2 of the table, below each amino-acid in your consensus, write its degeneracy (ie the number of different codons for that amino acid), **then select the run of 8 amino-acids** with the lowest degeneracy. You will need to use the Table on codon usage (below) to help with this.

Q16. What would be the total degeneracy of your primer needed to code for these 8 aa (i.e. the number of different DNA sequences that could encode this same amino-acid sequence)?

Eg: **H**	1. Consensus (single letter aa code)	F	N	E	D	G	S	F	I	G	Q
2	2. Degeneracy										
CAC	3. Likeliest codon										
CAT	4. Second likeliest codon										
0.68	5. Prob (likeliest codon)										
0.32	6. Prob (second likeliest codon)										
1.0	7. Combined probability										
$CA\frac{C}{T}$	8. Degenerate DNA sequence										

Table of codon usage in 44 nuclear genes from our system of interest:

AA	Codon		Fraction
A	Ala	GCA	0.13
	Ala	GCC	0.57
	Ala	GCG	0.11
	Ala	GCT	0.19
R	Arg	AGA	0.06
	Arg	AGG	0.07
	Arg	CGA	0.06
	Arg	CGC	0.61
	Arg	CGG	0.07
	Arg	CGT	0.13
N	Asn	AAC	0.78
	Asn	AAT	0.22
D	Asp	GAC	0.68
	Asp	GAT	0.32
C	Cys	TGC	0.89
	Cys	TGT	0.11
*	STOP	TAA	0.40
	STOP	TAG	0.20
	STOP	TGA	0.40
Q	Gln	CAA	0.14
	Gln	CAG	0.86
E	Glu	GAA	0.13
	Glu	GAG	0.87
G	Gly	GGA	0.26
	Gly	GGC	0.49
	Gly	GGG	0.03
	Gly	GGT	0.22
H	His	CAC	0.68
	His	CAT	0.32
I	Ile	ATA	0.04
	Ile	ATC	0.71
	Ile	ATT	0.25
L	Leu	CTA	0.04
	Leu	CTC	0.16
	Leu	CTG	0.62
	Leu	CTT	0.03
	Leu	TTA	0.01
	Leu	TTG	0.14
K	Lys	AAA	0.14
	Lys	AAG	0.86
M	Met	ATG	1.00
F	Phe	TTC	0.77
	Phe	TTT	0.23
P	Pro	CCA	0.20
	Pro	CCC	0.49
	Pro	CCG	0.21
	Pro	CCT	0.10
S	Ser	AGC	0.21
	Ser	AGT	0.05
	Ser	TCA	0.05
	Ser	TCC	0.38
	Ser	TCG	0.24
	Ser	TCT	0.07
T	Thr	ACA	0.09
	Thr	ACC	0.65
	Thr	ACG	0.16
	Thr	ACT	0.10
W	Trp	TGG	1.00
Y	Tyr	TAC	0.77
	Tyr	TAT	0.23
V	Val	GTA	0.07
	Val	GTC	0.33
	Val	GTG	0.51
	Val	GTT	0.09

The next trick is to allow for ***codon preference*.** Not all the possible codons are equally likely. Some codons are used preferentially for a given amino-acid, even though others are theoretically possible. These codon preferences are species-specific. The codon preference table for our target species (above) will allow us to deduce the most likely codons to use. For each amino-acid it shows the frequency of usage for each of the possible codons.

For your chosen stretch of 8 amino acids, in Row 3 of the table, for each amino acid, write the preferred codon, and below that, in Row 5, the probability of its use.

Q17. Calculate the probability that this sequence (ie, using only the likeliest codon for each amino-acid) corresponds to the sequence that would be used in our target system to code for your chosen 8 amino-acid sequence, by multiplying the probabilities together (eg. 0.72 x 1.00 x 1.00 x 0.90 = 0.66).

You will find that the probability of even the most likely primer being a perfect match is extremely low. You can make this probability higher by making your primer **degenerate**. In other words, you could make primers with alternate possibilities at some positions. Fortunately, alternate codons for a given amino acid usually differ only at the third base, so you need only offer a choice of bases at one position in each triplet; we are going to limit the choices to two alternates.

For each amino-acid, write the second preferred codon in Row 4, together with its probability in Row 6, and the total probability that one or other of them will be used (by *adding* the individual probabilities) in Row 7.

Q18. Calculate the probability that a degenerate primer using both the likeliest and second most likely codons will contain the exact sequence used in our target system to encode this 8 amino-acid sequence. How big is the improvement compared to use of likeliest codon only?

Write your degenerate primer sequence in Row 8, placing degenerate bases one above the other, eg. ...AC$\genfrac{}{}{0pt}{}{T}{C}$ ACG...

Q19. What is the degeneracy of your primer now? (i.e. how many different DNA sequences will be present in your primer mix?)

You now have a primer but would it conform to the basic guidelines that we considered earlier? This is one reason why designing primers from protein sequences requires more skill than using nucleic acid sequences!

Assuming that the primer had to be designed using the amino acid sequence that we have considered, suggest