**Solution notes**

## Data Structures and Algorithms 2005 (MR)
## Paper 3 Question 2, Paper 10 Question 3

Huffman and arithmetic coding are both covered in detail in the course.

(a) Bookwork.
The probability of the next symbol being x is independent of the sequence of symbols before (and after) it.

(b) A=00 B=01 C=1
So length in bits is 1000000*2 + 1000000*2 + 1000000*1 = 5000000

(c) A=1 B=00 C=01
So length in bits is 2000000*1 * 1000000*2 + 1000000*2 = 6000000

(d) The arithmetic encoding of the sting in (a) can be thought of as a number in the range 0..1 written in base 3 and a decimal point followed by 3000000 digits in base three (0,1 or2). There are 3**3000000 such numbers
Suppose x is the length of a binary expansion of such numbers, we require 2**x = 3**3000000.
ie log2(2**x) = log2(3**3000000)
ie x = 3000000 * log2(3)
ie x ~= 3000000 * 1.6 = 4800000

(e) Answer is the same as for (b) above, ie 6000000.
This is because the frequencies of A, B and C are 1/2, 1/4 and 1/4 which are all inverse powers of 2 so huffman does a perfect job with no wastage.
Arithmetic coding also has no wastage.
Alternatively, observe that arithmetic coding would generate a 1 for every A and 01 and 00 for the Bs and Cs, giving a total length of 6000000.