# Problem Set 1 Solutions

COMP 411 Spring 2012

## 1 Problem 1

### 1.1 A

If all nucleic acids are equally likely, a 3-nucleic acid sequence can be represented with 6 bits, as such:

$$\log_2\left(\frac{4*4*4}{1}\right) = \log_2\left(\frac{4*4*4}{1}\right) = 6 \tag{1}$$

Given that there are 20 amino acids and 1 stop code, the minimal number of bits to encode a single item in a protein chain is:

$$\log_2\left(\frac{21}{1}\right) \tag{2}$$

The numbers do not agree. Reasons why should include something along the lines of less information being stored in amino acid representation, the smaller set size of the amino acids, or such.

### 1.2 B

Since there are 64 codons, and 61 are amino acids. Since the only information given is that the codon represents an amino acid, the bits conveyed are:

$$\log_2\left(\frac{64}{61}\right) = 0.07 \tag{3}$$

Similarly for the 3 stop codes:

$$\log_2\left(\frac{64}{3}\right) = 4.42 \tag{4}$$

Since there are 6 possible codons for Serine:

$$\log_2\left(\frac{64}{6}\right) = 3.42 \tag{5}$$

## 1.3 C

There are 37 codons that contain the T nucleotide and 64 possible codons. Thus, the bits conveyed are:

$$\log_2\left(\frac{64}{37}\right) = 0.79 \tag{6}$$

To figure out how many bits of information are added by knowing that the codon is a stop code, subtract the original amount of information conveyed from the new amount of information conveyed.

$$\log_2\left(\frac{64}{3}\right) - \log_2\left(\frac{64}{37}\right) = \log_2\left(\frac{37}{3}\right) = 3.62 \tag{7}$$

## 1.4 D

There are 20 amino acids and 3 bases:

$$\log_2\left(\frac{20}{3}\right) = 2.74 \tag{8}$$

There are 64 possible codons and 10 of them encode to bases:

$$\log_2\left(\frac{64}{10}\right) = 2.68 \tag{9}$$

There are 2 ways to encode Lysine, so:

$$\log_2\left(\frac{64}{2}\right) - \log_2\left(\frac{64}{10}\right) = \log_2\left(\frac{10}{2}\right) = 2.32 \tag{10}$$

## 1.5 E

Across the set of 64 codons, there are 32 available transitions. Thus, let us consider each position in a codon and its influence on the final value. For the right-most position, there is only one transition that changes the resulting amino acid: Isoleucine  Methionine. For the middle position, all 32 of the possible 32 transitions change the protein. For the right-most position, 30 of the 32 transitions change the protein (TTA  CTA and TTG  CTG do not). Thus, the number of bits conveyed is:

$$\log_2\left(\frac{32+32+32}{1+32+30}\right) = \log_2\left(\frac{96}{63}\right) = 0.61a \tag{11}$$

(Most attempts to identify the transition set and valid transition set should have gotten full credit.)

## 1.6 F

In the case of Glycine, since the first 2 nuclic acids are all that is needed to identify the resulting amino acid, no bits of information are conveyed in the last nucleic acid.

## 1.7 G

Using the entropy formula given in the lecture $\sum_i p_i * \log \frac{1}{p_i}$ the entropy of the codes is:

$$(0.24\log_2(0.24)+0.14\log_2(0.14)+0.12\log_2(0.12)+0.5\log_2(0.5)) = 1.77 \tag{12}$$

The bits wasted using a fixed length scheme are: 2 - 1.77 = 0.23

## 1.8 H

The string 0011011101010 can be decoded as follows (only the last acid in the codon is shown): GGC, GGC, GGA, GGT, GGC, GGG, GGG.

### 1.9 I

The expected length is:

$$1000(0.51 + 0.242 + 0.143 + 0.123) = 1760 \tag{13}$$

The worst case is 3*1000 = 3000, as GGT and GGA have an encoding length of 3.

## 2 Problem 2

### 2.1 A

Using our earlier definition, Information per digit = log2(10/1) = 3.3219

### 2.2 B

F(d) = 1 => d = 0
F(d) = 2 => d = 1/3/7
F(d) = 3 => d = 2/5/8
F(d) = 5 => d = 4/9
F(d) = 7 => d = 6

Information in 1 (1 possibility only) = log2(10/1) = 3.3219 (approx.)
Information in 3 (3 possibilities) = log2(10/3) = 1.737 (approx.)
Information in 5 (2 possibilities) = log2(10/2) = 2.3219 (approx.)

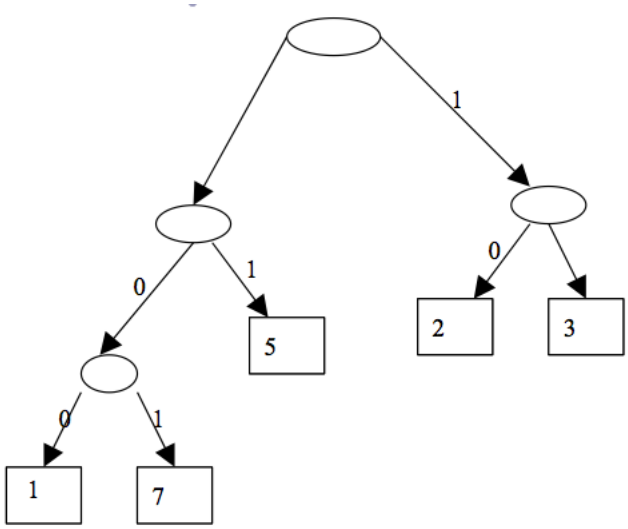
### 2.3 C

Average amount = weighted sum of all f(d) output values = 2(0.1)(3.3219) + 2(0.3)(1.737) + (0.2)(2.3219) = 0.6643 + 0.4643 + 1.0422 = 2.1708

### 2.4 D

It will take n-1 iterations  each step removes 2 members and adds one member.

### 2.5 E




| Symbol | Probability | Encoding |
|---|---|---|
| 1 | 0.1 | 000 |
| 2 | 0.3 | 10 |
| 3 | 0.3 | 11 |
| 5 | 0.2 | 01 |
| 7 | 0.1 | 001 |

### 2.6 F

Average length = 3(0.1) + 2(0.3) + 2(0.3) + 2(0.2) + 3(0.1) = 0.3 + 0.6 + 0.6 + 0.4 + 0.3 = 2.2

## 3 Problem 3

### 3.1 A

Add $21, $0, $5
0x0005A820

### 3.2 B

Ori $t0, $t0, 0xff
0x350800FF

### 3.3 C

Lui $s0, 0xdd
0x3C100DAD

### 3.4 D

Lw $v0, 53($gp)
0x8F820035

### 3.5 E

Sw $sp, -4($sp)
0xAFBDFFFC

### 3.6 F

Sll $t9, $a0, 3
0x0004C8C0

### 3.7 G

Loop: bne $s0, $s0, loop
0x16100000

### 3.8 H

Not affected. $s0 always equals $s0, so the branch never occurs.

# 4 Problem 4

## 4.1 A

0x100 to 0x12C are as follows: Lui $gp, 0x1000
Lw $t0, 0x0100($gp)
Addi $t1, $zero, 0x0001
And $t2, $zero, $zero
Addi $t2, $t2, 0x0001
Sub $t0, $t0, $t1
Addi $t1, $t1, 2
Slt $t3, $t0, $t1
Beq $t3, $zero, -4
Sll $zero, $zero, 0
Sw $t2, 0x0100($gp)
Sw $t0, 0x0104($gp)

## 4.2 B

Computes Square root.

## 4.3 C

0x124

## 4.4 D

No, since branches are relative to PC