Classical Cryptography

JOACHIM VON ZUR GATHEN, JÉRÉMIE DETREY

3. Tutorial: Information entropy (Hand in solutions on Monday, May 26th, at the beginning of the tutorial)

Exercise 3.1 (Entropy and randomness).

We are given a biased coin whose probability for flipping heads is $p_{\rm H} = 30\%$.

- (i) Compute the information entropy of such a coin toss.
- (ii) What is the maximal entropy which can be expected of a coin toss?
- (iii) How can you transform this biased coin into a fair(er) coin?
- (iv) What would be the value of the entropy, then?

Exercise 3.2 (Entropy and Huffman trees).

We are given an alphabet $\mathbb{A} = \{A, B, C, D, E, F\}$ with the following frequency distribution:

Letter	A	В	С	D	Е	F
Frequency	5	18	10	15	45	7

- (i) Compute the corresponding entropy.
- (ii) Using the same number of bits to encode each letter, how many do we need?
- (iii) What is the expected length of an *n*-letter message with this encoding?
- (iv) How can you do better?

Exercise 3.3 (Entropy of the English language).

(13 points)

The following table gives the frequency distribution of the letters in English.

Letter	А	В	С	D	Е	F	G	Н	Ι	
Frequency	8.04	1.54	3.06	3.99	12.51	2.30	1.96	5.49	7.26	
Letter	J	Κ	L	Μ	N	0	Р	Q	R	
Frequency	0.16	0.67	4.14	2.53	7.09	7.60	2.00	0.11	6.12	
Letter	S	T	Ū	V	W	X	Y	Ż		
Frequency	6.54	9.25	2.71	0.99	1.92	0.19	1.73	0.09		

- (i) What is the entropy of English?
- (ii) What is the maximal entropy for a 26-letter alphabet?
- (iii) Compute the *redundancy* of English, *i.e.* the entropy distance between English and a uniformly-distributed 26-letter language.
- (iv) Give a Huffman encoding of English according to this frequency distribution.
- (v) Have fun with the Java applet available at this URL: http://math.ucsd.edu/~crypto/java/ENTROPY/ What entropy do you obtain?
- (vi) Why is it lower than the previously computed entropy?
- (vii) Does it mean that we can compress an English text in approximately 1.2 bits per letter? Why aren't such extreme compression techniques not used in practice?

Exercise 3.4 (Entropy of amino acids).

(10 points)

In order to build proteins, our living cells use ribosomes to translate strands of *messenger ribonucleic acid* (mRNA) into sequences of amino acids. The mRNA encodes amino acids using *tri-nucleotide codons, i.e.* sequences of three nucleotides among Adenine (A), Cytosine (C), Guanine (G) and Uracil (U).

So here, instead of encoding an alphabet of 26 letters (A, ..., Z) using bits (0 or 1), proteins are seen as words over an alphabet of 20 amino acids (Alanine, Cysteine, ...) encoded using nucleotides (A, C, G, or U).

Amino acidAlaCysAspGluPheCodon(s)GCU, GCCUGU, UGCGAU, GACGAA, GAGUUU, UUCGCA, GCGIIIIIFrequency7.81.95.36.33.9Amino acidGlyHisIleLysLeuCodon(s)GGU, GGCCAU, CACAUU, AUCAAA, AAGUUA, UUGGGA, GGGCAU, CACAUAAAA, AAGCUU, CUCGGA, GGGIIIIIIIIIFrequency7.22.35.35.99.1Amino acidMetAsnProGlnArgCodon(s)AUGAAU, AACCCU, CCCCAA, CAGCGU, CGCFrequency2.34.35.24.25.1Amino acidSerThrValTrpTyrCodon(s)UCU, UCCACU, ACCGUU, GUCUAU, UACFrequency6.85.96.61.43.2										
GCA, GCGI.95.36.33.9Frequency7.81.95.36.33.9Amino acidGlyHisIleLysLeuCodon(s)GGU, GGCCAU, CACAUU, AUC AUAAAA, AAGUUA, UUG CUU, CUC CUA, CUGFrequency7.22.35.35.99.1Amino acidMetAsnProGlnArgCodon(s)AUGAAU, AACCCU, CCC CCA, CCGCAA, CAGCGU, CGC CGA, CGGFrequency2.34.35.24.25.1Amino acidSerThrValTrpTyrCodon(s)UCU, UCCACU, ACC UCA, UCGGUU, GUC GUA, GUGUGGUAU, UAC UAU, UACAmino acidSerThrValTrpTyrAmino acidSerThrValTrpUYAmino acidSerThrValUGGUAU, UAC UAU, UACAGU, AGCACU, ACCGUU, GUC GUA, GUGUAGUAU, UAC	Amino acid	Ala	Cys	Asp	Glu	Phe				
Frequency7.81.95.36.33.9Amino acidGlyHisIleLysLeuCodon(s)GGU, GGCCAU, CACAUU, AUCAAA, AAGUUA, UUGGGA, GGGCAU, CACAUUAAUACUU, CUCCUU, CUCFrequency7.22.35.35.99.1Amino acidMetAsnProGlnArgCodon(s)AUGAAU, AACCCU, CCCCAA, CAGCGU, CGCFrequency2.34.35.24.25.1Amino acidSerThrValTrpTyrCodon(s)UCU, UCCACU, ACCGUU, GUCUAU, UACAmino acidSerThrValTrpUAU, UACAmino acidGerAC, ACGGUU, GUCUAU, UACUAU, UACAmino acidSerThrValTrpTyrCodon(s)UCU, UCCACA, ACGGUU, GUCUAU, UACUAU, UACAGU, AGCACA, ACGGUA, GUGUAU, UACIAU, UAC	Codon(s)	GCU, GCC	UGU, UGC	GAU, GAC	GAA, GAG	UUU, UUC				
Amino acidGlyHisIleLysLeuCodon(s)GGU, GGCCAU, CACAUU, AUCAAA, AAGUUA, UUGGGA, GGGCAU, CACAUACUU, CUCCUU, CUCFrequency7.22.35.35.99.1Amino acidMetAsnProGlnArgCodon(s)AUGAAU, AACCCU, CCCCAA, CAGCGU, CGCFrequency7.22.35.35.99.1Amino acidMetAsnProGlnArgCodon(s)AUGAAU, AACCCU, CCCCAA, CAGCGU, CGCFrequency2.34.35.24.25.1Amino acidSerThrValTrpTyrCodon(s)UCU, UCCACU, ACCGUU, GUCUAU, UACUCA, UCGACA, ACGGUA, GUGUAU, UACAGU, AGCICA, ACGGUA, GUGIAU, UAC		GCA, GCG								
Codon(s)GGU, GGC GGA, GGGCAU, CAC AUU, CUC AUAAUU, AUC AUAAAA, AAG CUU, CUC CUA, CUGFrequency7.22.35.35.99.1Amino acidMetAsnProGInArgCodon(s)AUGAAU, AACCCU, CCC CCA, CGGCAU, CAGCGU, CGC CGA, CGGFrequency2.34.35.24.25.1Amino acidSerThrValTrpTyrCodon(s)UCU, UCC UCA, UCGACU, ACC ACA, ACGGUU, GUC GUU, GUCUAU, UACAmino acidSerThrValTrpTyrCodon(s)UCU, UCC UCA, UCGACU, ACC ACA, ACGGUU, GUC GUA, GUGUAU, UACAmino acidSerThrValTrpTyrAmino acidSerThrValIrpIuAmino acidSerThrValIuIuAmino acidSerThrValIuIuAmino acidSerThrValIuIuAmino acidSerIuIuIuIuAmino acidSerIuIuIuIuAmino acidSerIuIuIuIuAmino acidSerIuIuIuIuAmino acidSerIuIuIuIuAmino acidSerIuIuIuIuAmino acidSerIuIuIuIuAmino acid<	Frequency	7.8	1.9	5.3	6.3	3.9				
Codon(s)GGU, GGC GGA, GGGCAU, CAC AUU, CUC AUAAUU, AUC AUAAAA, AAG CUU, CUC CUA, CUGFrequency7.22.35.35.99.1Amino acidMetAsnProGInArgCodon(s)AUGAAU, AACCCU, CCC CCA, CGGCAU, CAGCGU, CGC CGA, CGGFrequency2.34.35.24.25.1Amino acidSerThrValTrpTyrCodon(s)UCU, UCC UCA, UCGACU, ACC ACA, ACGGUU, GUC GUU, GUCUAU, UACAmino acidSerThrValTrpTyrCodon(s)UCU, UCC UCA, UCGACU, ACC ACA, ACGGUU, GUC GUA, GUGUAU, UACAmino acidSerThrValTrpTyrAmino acidSerThrValIrpIuAmino acidSerThrValIuIuAmino acidSerThrValIuIuAmino acidSerThrValIuIuAmino acidSerIuIuIuIuAmino acidSerIuIuIuIuAmino acidSerIuIuIuIuAmino acidSerIuIuIuIuAmino acidSerIuIuIuIuAmino acidSerIuIuIuIuAmino acidSerIuIuIuIuAmino acid<										
GGA, GGGAUACUU, CUC CUA, CUGFrequency7.22.35.35.99.1Amino acidMetAsnProGlnArgCodon(s)AUGAAU, AACCCU, CCC CCA, CCGCAA, CAGCGU, CGC CGA, CGGFrequency2.34.35.24.25.1Amino acidSerThrValTrpTyrAmino acidSerThrValUCGUAU, UACAmino acidSerThrValInpTyrAmino acidSerThrValInpInpAmino acidSerThrValInpInpAmino acidSerThrValInpInpAmino acidSerInpInpInpAmino acidSerInpInpInpAmino acidSerInpInpInpAmino acidSerInpInpAmino acidSerInp	Amino acid	Gly	His	Ile	Lys	Leu				
Frequency7.22.35.35.99.1Amino acidMetAsnProGlnArgCodon(s)AUGAAU, AACCCU, CCCCAA, CAGCGU, CGCFrequency2.34.35.24.25.1Amino acidSerThrValTrpTyrCodon(s)UCU, UCCACU, ACCGUU, GUCUAU, UACAmino acidSerThrValIrpIyrAmino acidSerACU, ACCGUU, GUCUAU, UACUAU, UACAddu, AGCACU, ACCGUA, GUGUAGUAU, UACAGU, AGCACA, ACGGUA, GUGUAU, UACIuru	Codon(s)	GGU, GGC	CAU, CAC	AUU, AUC	AAA, AAG	UUA, UUG				
Frequency7.22.35.35.99.1Amino acidMetAsnProGlnArgCodon(s)AUGAAU, AACCCU, CCCCAA, CAGCGU, CGCCodon(s)AUGAAU, AACCCU, CCCCAA, CAGCGU, CGCFrequency2.34.35.24.25.1Amino acidSerThrValTrpTyrCodon(s)UCU, UCCACU, ACCGUU, GUCUGGUAU, UACAdu, AGCCGU, ACCGUA, GUGUAU, UACIndicated of the text of text		GGA, GGG		AUA		CUU, CUC				
Amino acidMetAsnProGlnArgCodon(s)AUGAAU, AACCCU, CCCCAA, CAGCGU, CGCCodon(s)AUGAAU, AACCCU, CCCCAA, CAGCGU, CGCFrequency2.34.35.24.25.1Amino acidSerThrValTrpTyrCodon(s)UCU, UCCACU, ACCGUU, GUCUGGUAU, UACAdu, AGCACU, ACCGUA, GUGUAU, UACInterval						CUA, CUG				
Codon(s)AUGAAU, AACCCU, CCCCAA, CAGCGU, CGCCAA, CAGCGU, CGCCAA, CAGCGU, CGCCGA, CGGFrequency2.34.35.24.25.1Amino acidSerThrValTrpTyrCodon(s)UCU, UCCACU, ACCGUU, GUCUGGUAU, UACUCA, UCGACA, ACGGUA, GUGUGGUAU, UACAGU, AGCIIIII	Frequency	7.2	2.3	5.3	5.9	9.1				
Codon(s)AUGAAU, AACCCU, CCCCAA, CAGCGU, CGCCAA, CAGCGU, CGCCAA, CAGCGU, CGCCGA, CGGFrequency2.34.35.24.25.1Amino acidSerThrValTrpTyrCodon(s)UCU, UCCACU, ACCGUU, GUCUGGUAU, UACUCA, UCGACA, ACGGUA, GUGUGGUAU, UACAGU, AGCIIIII										
Image: series of the series	Amino acid	Met	Asn	Pro	Gln	Arg				
Frequency2.34.35.24.2AGA, AGGAmino acidSerThrValTrpTyrCodon(s)UCU, UCCACU, ACCGUU, GUCUGGUAU, UACUCA, UCGACA, ACGGUA, GUGIndicationIndicationAGU, AGCIndicationIndicationIndicationIndication	Codon(s)	AUG	AAU, AAC	CCU, CCC	CAA, CAG	CGU, CGC				
Frequency2.34.35.24.25.1Amino acidSerThrValTrpTyrCodon(s)UCU, UCCACU, ACCGUU, GUCUGGUAU, UACUCA, UCGACA, ACGGUA, GUGUAUUAUAGU, AGCImage: Complex of the second se			(7)	CCA, CCG		CGA, CGG				
Amino acidSerThrValTrpTyrCodon(s)UCU, UCCACU, ACCGUU, GUCUGGUAU, UACUCA, UCGACA, ACGGUA, GUGUAUUAUAGU, AGCImage: Comparison of the second se						AGA, AGG				
Codon(s)UCU, UCCACU, ACCGUU, GUCUGGUAU, UACUCA, UCGACA, ACGGUA, GUGUAU, UACUAU, UACAGU, AGCAGUAGUAGUAGUAGU	Frequency	2.3	4.3	5.2	4.2	5.1				
Codon(s)UCU, UCCACU, ACCGUU, GUCUGGUAU, UACUCA, UCGACA, ACGGUA, GUGUAU, UACUAU, UACAGU, AGCAGUAGUAGUAGUAGU	·	•								
UCA, UCG ACA, ACG GUA, GUG AGU, AGC	Amino acid	Ser	Thr	Val	Trp	Tyr				
AGU, AGC	Codon(s)	UCU, UCC	ACU, ACC	GUU, GUC	UGG	UAU, UAC				
		UCA, UCG	ACA, ACG	GUA, GUG						
Frequency 6.8 5.9 6.6 1.4 3.2		AGU, AGC								
	Frequency	6.8	5.9	6.6	1.4	3.2				

You will find in the following table the actual codon encodings of each amino acid, along with their occurrence in proteins¹:

- (i) Compute the entropy of amino acids in proteins in terms of nucleotides. 4 *Warning!* The definition of Shannon's entropy as seen in the lecture measures the entropy in terms of bits (*i.e.* encoding the values with only 0's and 1's). But here, we want to measure the entropy in terms of nucleotides (*i.e.* encoding the values with A's, C's, G's and U's). Think carefully and modify the entropy formula accordingly!
- (ii) Give a Huffman encoding for amino acids, still using nucleotides. *Warning!* Once again, you can encode 4 different values with a single nucleotide. This should give you a hint as to the arity of the Huffman tree.
- (iii) Look more closely at the redundancy of the codons. Discuss.

¹Source: http://en.wikipedia.org/wiki/List_of_standard_amino_acids

4

2