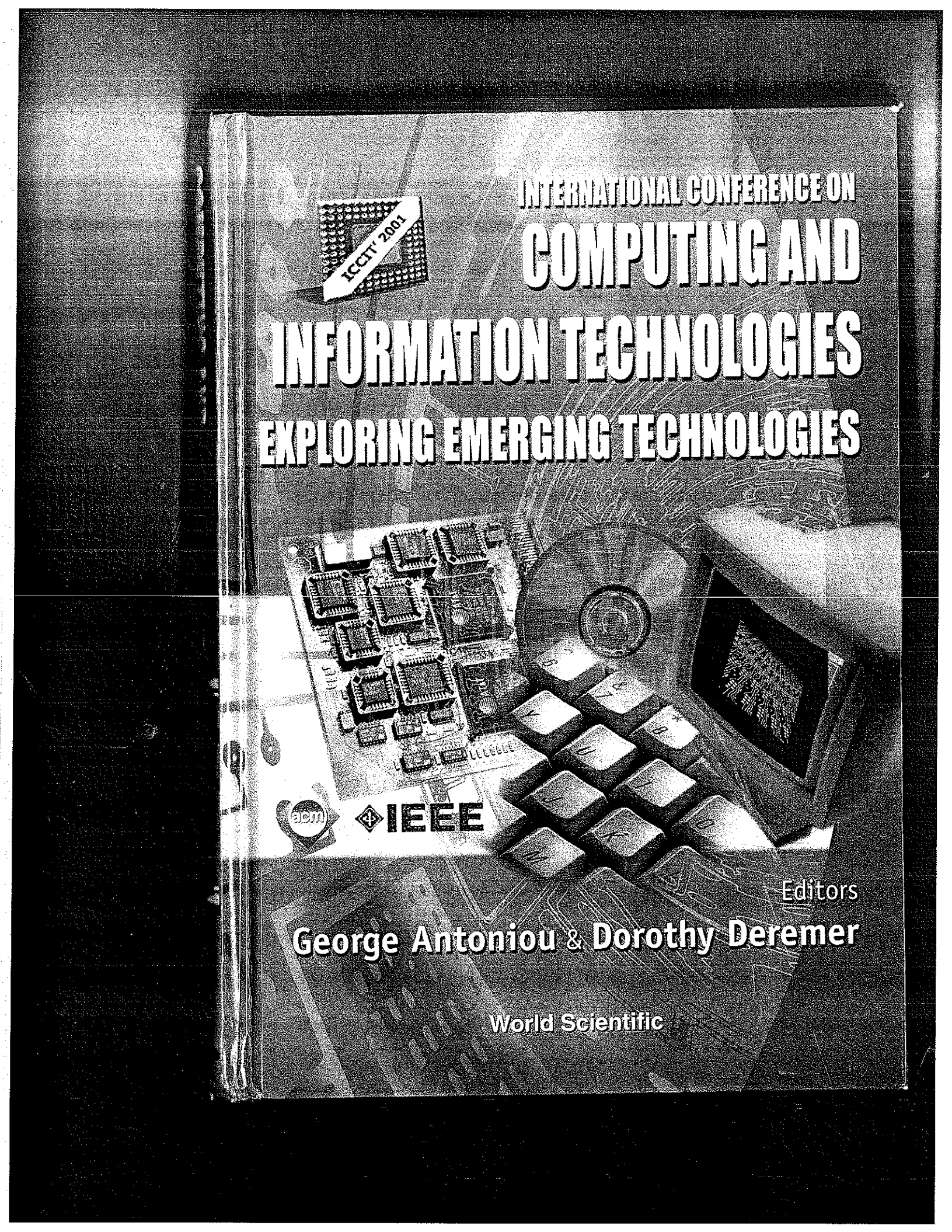
ICCIT 2001
INTERNATIONAL CONFERENCE ON
COMPUTING AND
INFORMATION TECHNOLOGIES
EXPLORING EMERGING TECHNOLOGIES
acm
IEEE
Editors
George Antoniou & Dorothy Deremer
World Scientific

INTERNATIONAL CONFERENCE ON

COMPUTING AND INFORMATION TECHNOLOGIES

EXPLORING EMERGING TECHNOLOGIES

Montclair State University, NJ, USA

12 Oct 2001

Editors

George Antoniou
Dorothy Deremer

Montclair State University

World Scientific
New Jersey • London • Singapore • Hong Kong

A NATURAL CODON SPACE AND METRIC

H.M. HUBEY
Department of Computer Science, Montclair State University, Upper Montclair, NJ 07043
E-mail: hubeyh@mail.montclair.edu

The building blocks of life are the nucleutides adenine (A), cystosine (C), guanine (G) and thymine (T). These can be coded in binary and displayed in a table similar to the chemical table of elements. A particular form of coding, the Gray code, shows the relationship of these nucleotides in better ways than a simple binary code.

1 Introduction: the Nucleotides

Cells are the building blocks of life. Simple organisms like bacteria do not have cells with nuclei (prokaryotic cells). Most other cells have a nucleus distinct from the rest of the surrounding cytoplasm (eukarytic cells). DNA (the basis of life) is located in the nucleus (in eukoryates) and carries genetic material (genes) that determines the structures of proteins which are the building blocks of life. It is a an amazing fact that the DNA molecules (which are linear polymers) consist of only four types (bases of nucleotides): adenine (A), cystosine (C), guanine (G) and thymine (T). These are linked together by sugar-phosphates which form the 'backbone' [1].

The spatial structure of the DNA as is well-known is a double-helix of two complementary linear strands of nucleotides are held together by weak hydrogen bonds in which a G in one strand complements with a C in another and an A complements a T. DNA replication takes place via a process in which the two strands separate and then act as templates for new complementary strands. At this level DNA may be considered to be a language with four symbols in the alphabet. Enzymes, which are also proteins, are catalysts for cellular chemical reactions.

2 The Amino Acids

They are built up of only 20 different types of amino acids, linked to each other by peptide bonds. Such polypeptide molecules are not linear but have unique three-dimensional structures specific to their functions. The simplest part of the genetic code, then, is to map the four-letter alphabet code (the bases) of the linear DNA molecules into the 20 letter alphabet code (amino acids) of the three-dimensional protein molecule.

In the barest essentials, a triplet of bases (called a *codon)* codes for a distinct amino acid. In other words, one of the DNA strands read in a particular direction

(specified by its ends) three bases at a time, translates into a unique sequence of amino acids. A single strand of DNA can have many genes, each coding for a different protein (one gene for one protein). Because of the unique start and termination of the strands, there are codons which also code for start and termination of the amino acid sequence pertaining to a protein. This code was deciphered in 1966.

	G 00	A 01	C 11	U 10
GG 0000	GLY	GLY	GLY	GLY
GA 0001	GLU	GLU	ASP	ASP
GU 0010	VAL	VAL	VAL	VAL
GC 0011	ALA	ALA	ALA	ALA
AG 0100	ARG2	ARG2	SER2	SER2
AA 0101	LYS	LYS	ASN	ASN
AU 0110	MET	ILE	ILE1	ILE1
AC 0111	THR	THR	THR	THR
UG 1000	TRP	STOP2	CYS	CYS
UA 1001	STOP1	STOP1	TYR	TYR
UU 1010	LEU2	LEU2	PHE	PHE
UC 1011	SER1	SER1	SER1	SER1
CG 1100	ARG1	ARG1	ARG1	ARG1
CA 1101	GLN	GLN	HIS	HIS
CU 1110	LEU1	LEU1	LEU1	LEU1
CC 1111	PRO	PRO	PRO	PRO

Table 1

Plain Binary Coded Codon Table

	G 00	A 01	C 11	U 10
GG 0000	GLY	GLY	GLY	GLY
GA 0001	GLU	GLU	ASP	ASP
GC 0011	ALA	ALA	ALA	ALA
GU 0010	VAL	VAL	VAL	VAL
AU 0110	MET	ILE	ILE1	ILE1
AC 0111	THR	THR	THR	THR
AA 0101	LYS	LYS	ASN	ASN
AG 0100	ARG2	ARG2	SER2	SER2
UG 1000	TRP	STOP2	CYS	CYS
UA 1001	STOP1	STOP1	TYR	TYR
UC 1011	SER1	SER1	SER1	SER1
UU 1010	LEU2	LEU2	PHE	PHE
CU 1110	LEU1	LEU1	LEU1	LEU1
CC 1111	PRO	PRO	PRO	PRO
CA 1101	GLN	GLN	HIS	HIS
CG 1100	ARG1	ARG1	ARG1	ARG1

Table 2

KH-map Coded Codon Table

There are 64 possible codons; $(4 \times 4 \times 4)$. Of these three codons code for STOP or termination of the translation into amino acids. The mapping from the 61 codons to 20 amino acids determines the three-dimensional folding of the protein molecule. This process is called *gene expression.*

A part of one strand of DNA is first transcribed into a molecule called messenger RNA or mRNA (ribonucleic acid). The *mRNA* is single-stranded and pos-

que sequence of
coding for a dif-
start and termi-
and termination
as deciphered in

C	U
11	10
GLY	GLY
ASP	ASP
ALA	ALA
VAL	VAL
ILE1	ILE1
THR	THR
ASN	ASN
SER2	SER2
CYS	CYS
TYR	TYR
SER1	SER1
PHE	PHE
LEU1	LEU1
PRO	PRO
HIS	HIS
ARG1	ARG1

ıle 2

Codon Table

s code for STOP
ing from the 61
ng of the protein

cule called mes-
tranded and pos-

sesses bases complementary to the ones on the DNA with the exception that the adenine (A) in the DNA is paired with uracil (U) on the *mRNA* (instead of thymine, T). *mRNA* is an intermediary which carries the genetic information from the nucleus to the cytoplasm where the actual translation into protein occurs. The *mRNA* alphabet consists of A, C, G, and U, and it too can fold into distinctive three-dimensional structures along some of its parts via base pairing.

Only a fraction of the DNA strand called exons actually code for proteins. The rest of it (called introns) are non-coding regions which are also called junk DNA, and which have to be removed. The edited version of the *mRNA* is the template that is translated into protein. This translation is carried out by two other types of RNA, ribosomes and *tRNAs* (transfer RNAs). The *tRNA* has a dual role; it is able to recognize both the amino acid and its corresponding codon. Because sometimes only the first two bases are read (the third base being irrelevant for *tRNA* selection), it is not necessary for there to be 61 types of *tRNAs*.

3 Binary Coding of Codons

These processes, DNA replication, DNA transcription into *mRNA*, and translation of *mRNA* into proteins is the basis of molecular biology and genetics. This article looks at the patterns not seen until now in the coding of the 64 codons into 20 amino acids. The genetic code can be said to be constructed from codons. It has been shown that the codon code is essentially a combination of a fixed-length and variable-length code [2], and the same authors have shown that there is a binary coding scheme for this. This code is a combination of fixed length code and a variable length code, and is given in Table 1. The coding is A=01, G=00, C=11 and U=10. The above authors' contribution is to determine the efficiency of the genetic code from an algorithmic complexity viewpoint. In this article, we look into remarkable patterns this code possesses.

In Table 1 (Appendix) , the binary code uses x as a 'don't care' condition of digital design. Therefore the first 8 amino acids use only a 4-bit code; the last two bits which represent one of {A,G,C,U} can be any of them. The second and third sets of amino acids in the table, are 5-bit codes. In the second set the third codon can be only one of {C,U}. In the third set, the third codon can be only A or G. Therererefore in the binary code, the last bit of the second set is always 1, and the last bit of the third set is always 0. The last set is a 6-bit code. Thus, using z as one of the four bases, and b as a bit, the codes are of the type zzbb, zz1b, zz0b, and zzz. We can put these in a table using this binary code as done in Table 1.

Table 1 shows the various amino acids, at most one per row. In some cases there are two per row and in two cases, single occurrences are adjacent to doubles. The table shows the code graphically similar to a chemical table. However the relationships of the various amino acids are not clearly shown. An amino acid

that differs from another by a single nucleutide (such as the codons numbered 9-22) should somehow be shown next to each other.

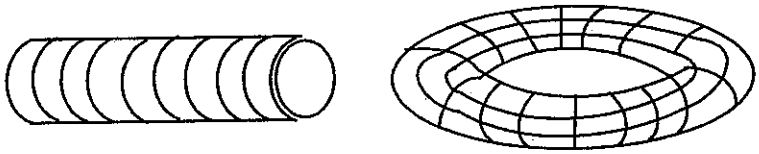

Figure 1: *The Codon Torus*: the KH-table can be wrapped on a torus as shown

However the codon code exhibits other striking regularities. If we map out this binary code into a KH-map (which is a K-map of size greater than 4×4) [4] in a particular shape as shown in Table 2, we see a pattern in which every row is taken up by either a single codon, two codons or three codons. The coding scheme for this KH-map is created by ordering the binary codes in a specific way, called the Gray code, or the reflected code [2,3]. The binary codes are ordered in such a way that each cell differs from any of its neighbors by one bit.

A neighbor is defined as a cell either vertically or horizontally adjacent. This map then is a very specific ordering of the codons. Because of its construction, the cells along the top are also neighbors of the cells along the bottom, and the cells along the left edge are neighbors of the cells along the right edge. Therefore, this planar map can be mapped onto a torus (as in Figure 1) in such a way that the cells are physically adjacent to the cells with which they are neighbors [2,3].

The number of nucleotides at each row can be expressed as 1211212322131121 (see Fig.2.) Other combinations could have also occured but there might be a reason why this particular pattern has evolved. We can shift the acids around on the table by reassigning the binary codes to G,A,C, and U, however it is best to think of the table as being wrapped on the surface of the torus shown in Figure 1. Then the row or column shifts will merely result in the rotation of the torus in different ways, one of them being rotating it so that the surface facing the inside circle is on the outside. The distance of the nucleotides to each other can then be read off the torus.

From the table we can see that the 4 bit codes are the most robust in the sense that the last two bits can be anything and still represent the same nucleotide. The 6 bit codes are the most volatile in the sense that any change in any of the bits will result in turning the nucleotide into another one.

Therefore the 'distance metric' is row dependent, along the length of the torus for all of them, and it is row and column dependent for the others.

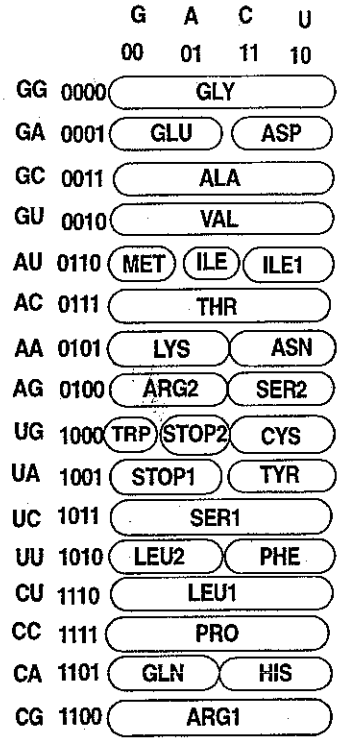

Figure 2: Another View of the Codons

4 Conclusion

In addition to the codon code being a combination of a fixed length code and a variable length code [4,5], the code also displays regularities in the form of Gray-coding and thus can be displayed as a KH-map/table [2], and wrapped on a torus [3]. Thus the torus is the natural space for this code, and a metric can be constructed on this space.

References

1. Lewin, B, Genes V, (Oxford University Press, Oxford, 1995)
2. Hubey, H.M. A Complete Unified Method for Taming the Curse of Dimensionality in Datamining and Allowing Logical-ANDs in ANNs, submitted to *Journal of Knowledge Discovery and Datamining,* June 2001.
3. Hubey, H.M. Mathematical and Computational Linguistics, (Moscow, Mir Domu Tvoemu, 1994)
4. Naranan, S. and V.K. Balasubrahmanyan, Information Theory and Algorithmic Complexity: Application to Linguistics Discourses and DNA sequences as Complex Systems, Part I: Efficiency of the Genetic Code of DNA, *Journal of Quantitative Linguistics*, **Vol 7** (2000), No 2, August, pp. 129-152.
5 Naranan, S. and V.K. Balasubrahmanyan, Information Theory and Algorithmic Complexity: Application to Linguistics Discourses and DNA sequences as Complex Systems, Part II: Complexity of DNA Sequences, Analogy with Linguistic Discourses, *Journal of Quantitative Linguistics*, **Vol 7** (2000), No 2, August, pp. 153-183.

APPENDIX I

Table 1: The Genetic Code

#	Codon Set	Name	Abbrev	Binary Code
1	AC*	Threonine	THR	0111xx
2	CC*	Proline	PRO	1111xx
3	GC*	Alanine	ALA	0011xx
4	GG*	Glycine	GLY	0000xx
5	GU*	Valine	VAL	0011xx
6	CG*	Arginine	ARG1	1100xx
7	CU*	Leucine	LEU1	1011xx
8	UC*	Serine	SER1	1011xx
9	AA#	Asparagine	ASN	01011x
10	CA#	Histidine	HIS	11011x
11	GA#	Aspartic Acid	ASP	00011x
12	UG#	Cysteine	CYS	10001x
13	UU#	Phenylalanine	PHE	10101x
14	AG#	Serine	SER2	01001x
15	AU#	Isoleucine	ILE1	01101x
16	UA#	Tyrosine	TYR	10011x
17	AA?	Lysine	LYS	01010x
18	CA?	Glutamine	GLN	11010x
19	GA?	Glutamic Acid	GLU	00010x
20	AG?	Arginine	ARG2	01000x
21	UU?	Leucine	LEU2	10100x
22	UA?	Stop	STOP1	10010x
23	AUG	Methionine	MET	011000
24	UGG	Tryptophan	TRP	100000
25	AUA	Isoleucine	ILE2	011001
26	UGA	Stop	STOP2	100001

Legend

A=01,G=00, C=11,
U=10
*=A,G,C,or U
#=C, or U (Pyrimidine)
x= 0 or 1
?=A or G (Purine)