Davis, J.[1,2], Khan, S.[3], Cromer, K.[4], Church, G. M.[1]

Multiplex, cascading DNA-encoding for making angels

AFFILIATIONS

[1] Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston MA 02115
[2] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139
[3] Peshawar 2.0, Arbab Tower, Nasir Bagh Road, Peshawar, Pakistan
[4] Department of Pediatrics, Stanford University, Stanford, CA 94305

ABSTRACT

DNA molecules having three base-pairs or more can simultaneously hold three unique numbers. We show how a coding strategy based on these three numbers can be used to encode a single molecule with multiple, independent data sets and furthermore, that many layers of information can be encoded in this way. Our example is a molecule holding multiple encodings of "Subhan Allah," an Arabic phrase that is said to have been repeated for more than a thousand years as an invocation associated with creating angels.

[KEYWORDS: DNA Manifolds; DNA Silent Code; DNA Amino Code; Subhan Allah; Angels]

INTRODUCTION

A centuries-old tradition involves the practice of making angels by uttering a certain phrase in Arabic language. According to this tradition, whenever a particular phrase is pronounced, an angel is automatically created. Moreover, any number of angels can be generated in this way. Like mantras on Tibetan prayer flags, it makes no difference whether the phrase is spoken, or written, or caused to be printed. Anytime the phrase is repeated, in any form or iteration, the result is believed to be an angel.

The phrase is "Subhan Allah" (سبحان الله), which roughly translates to the word, "Hallelujah" in English. We assume that repeating "Hallelujah" in any language could rarely be construed to offend anyone, and in the midst of so many COVID-19 victims, the symbolism about creating many angels may be a comfort to many.

19th century Islamic scholars reported that connections between repeating "Subhan Allah," and the subsequent proliferations of angels, date at least as far back as the 9th century CE.1 The practice is also referred to in hadith collections, accounts from verbal and physical teachings and traditions dating from the early Islamic era. Although these accounts have been contested, and not uniformly endorsed by many religious scholars, persistent narratives about pronouncements of "Subhan Allah" and the appearance of angels have endured for hundreds of years. Symbolism about changing the demographic of heaven can be elegantly aligned with the objectives and capabilities of information-keeping in DNA.

DISCUSSION

ASCII encoding

A preliminary "Subhan Allah" coding strategy is based on representing corresponding characters of Arabic text as hexadecimal ASCII (American Standard Code for Information Interchange) numbers, and their equivalent expressions in mathematical bases 2 (binary), and 4 (DNA).

Arabic Text ("Subhan Allah"): سبحان الله

"Subhan Allah" Arabic text to hexadecimal code (ASCII):

d8 b3 d8 a8 d8 ad d8 a7 d9 86 20 d8 a7 d9 84 d9 84 d9 87

"Subhan Allah" hexadecimal ASCII conversion to binary code:

11011000 10110011 11011000 10101000 11011000 10101101 11011000 10100111 11011001 01101000 00100000 11011000 10100111 11011001 10000100 11011001 10000100 11011001 10000111

"Subhan Allah" binary code conversion to 76-mer DNA where C=00 T=01 A=10 G=11 (increments based on molecular weight) :

GTACAGCGGTACAAACGTACAAGTGTACAATGGTATTAACCACCGTACAATGGTATACTCGTATACTCGTATACTG

= سبحان الله

Abjad encoding

Another option for encoding "Subhan Allah" entails the use of an alphabetic numeral system of notation related to gematria, an ancient practice using the Hebrew alphabet, and the ancient alphabetical number system practices of many other ancient cultures.2 Before Arabic numerals were promoted in Western Europe in the 13th century, most forms of European mathematics were also predominantly written with alphabetical numerals. In Arabic, abjad numerals are a decimal alphanumeric code in which the 28 letters of the Arabic alphabet are assigned numerical values. Alif, the first letter of the Arabic alphabet is used to represent the number 1; the second latter, bā, is used to represent the number 2, and so on. Individual Arabic alphabetical characters appearing after the 9th letter in the Arabic alphabet are used to represent 10s and 100s. The letter, yā represents 10. The number 20 is assigned to the letter, kāff. The letter, qāf represents 100, etc. Abjad numbers are also traditionally used to assign numerical values to whole Arabic words for purposes of numerology, belief in the divine or mystical relationship between numbers and one or more coinciding events. Ilm al-Hurūf or, "Science of Letters" is the practice of Arabic numerology whereby numerical values assigned to Arabic alphabetical characters are added up to provide total values for words in the Quran (though most Islamic scholars and religious authorities do not recommend its use for interpreting Quran for purposes of divination or prediction). [3]

The abjad number of سبحان الله is 187 = (5+30+30+1) + (50+1+8+2+60)

= 11 10 11 10 [binary]

= GAGA [DNA] in reading right to left

or, AGAG [DNA] reading left to right ("ه ل ل ا ن ا ح ب س")

"Subhan Allah" spoken one hundred times per day1 x 10 billion Muslims since 632 AD x 60 person-yrs each = 20 quadrillion angels

50 quadrillion copies of the 4-mer DNA = 100 micrograms

Taking all of this into account, and inspired in part by repeating geometry of Islamic tiling, we have implemented a DNA coding technique that combines several simultaneous levels of informational symmetry:

Silent Code

Here we use the term, "Silent Code" to describe a method for DNA-encoding using "silent mutations" to hold binary information in redundant codons with values incremented by molecular weight. That is, values are assigned to individual codons according to the respective incremental mass of all codons translating for a particular amino acid. [see Figures 1a, 1b,] By itself, Silent Code is not a very efficient coding technique in terms of bits-per-nucleotide, but it can be written into highly conserved genes.

Amino Code

If amino acids are given values too (in this case, mathematical base-20 values are assigned), then "Subhan Allah" can be coded for in a molecule that simultaneously codes for something else. [see: Figure 2] A message can be independently written into a number assigned to the sequence of amino acids (Amino Code) irrespective of information written into the number that corresponds with the sequence of redundant codons (Silent Code). This is a very flexible coding technique, since even in the case of relatively small genes, astronomical numbers of distinctly different DNA sequences can code for the same sequence of amino acids.

Nature has built functional redundancy into the genetic code, but non-functional redundancy is up for grabs. In addition to having information of its own, values assigned to amino acids (peptide sequences), may also be used as a check for copying errors. A predictable peptide sequence can hold core information while its triplet variants can encode separate data sets. As a given sequence of amino acids is repeated many times, the probability for error increases. So, if the core peptide sequence is "xyz," and there is a region with an erroneous peptide sequence, then there will be a high likelihood of errors appearing in corresponding Silent Code. Methods for such over encoding of information are common aspects of electronic and broadcast communications where multiple layers of information are added to guarantee the integrity of information sent or received.

Three numbers

There is a third number too, and this third number is one that corresponds with the DNA sequence itself, where C=00, T=01, A=10, and G=11. [See: Figure 3] In this way, every DNA molecule larger than a 2-mer can hold three arbitrary numbers or, three "pages" of information, and it seems nature uses only two of them.

DNA Manifolds

These three pages of information (inherent to almost all DNA molecules) are key to a coding method we have termed, "DNA Manifolds." Using DNA Manifolds, "Subhan Allah" can be written over itself again and again in the same DNA molecule. In the example given here, the Amino Code number codes for binary values of the 76-mer "Subhan Allah" DNA in "Line one" and the corresponding Silent Code number holds identical "Subhan Allah" binary values in "Line four":

Line one:	1101	1000	1011	00	1111	01	1000	1010	1000	110	1100	01	01	01	1011	...
Line two:	CYS	SER	ALA	ILE	GLY	VAL	SER	THR	SER	LYS	GLU	VAL	VAL	VAL	ALA	...
Line three:	TGT	TCG	GCT	ATA	GGC	GTT	TCT	ACA	TCT	AAG	GAG	GTA	GTG	GTC	GCT	...
Line four:	1	1	01	1	00	01	01	10	01	1	1	10	11	00	01	...

Both of these numbers are automatically contained in the DNA sequence, TGTTCGGCTATAGGCGTTTC-TACATCTAAGGAGGTAGTGGTCGCT... (Line three), which becomes the "third" number of the molecule:

0111010100111100011001111100110101010001100010010001101000001000000110110111
110100110001....

1 see: Sahih al-Bukhari 6405, Book 80, Hadith 100

This last number holds all of the information coded into the other two numbers, including the specific sequence of the initially encoded 76-mer DNA molecule.

This "third" number can be subsequently encoded into the Amino Code and Silent Code numbers of yet another molecule, and so on, cascading input data (in this case, "Subhan Allah") into many layers of encoded information. A multi-layer "manifold" can be systematically unpacked into a set of imaginary, but precisely described DNA molecules. Just as in this case, the initial 76-mer coding for "Subhan Allah" exists only as a mathematical construct that is decoded from the sequence of another molecule. Only the final sequence is synthesized as a real DNA molecule, one that can ideally be encoded with the maps of many other "virtual" DNA molecules – and all of the information they contain.

There is a very large number of possibilities to select from when searching for the most efficient simultaneous encoding of input data into amino acid sequences and redundant codons (Lines two and three above). As the number of coded 'virtual' molecules increase in a DNA Manifold, so does the number of corresponding values sets that become available to code for them. After a few steps of manifold encoding, huge numbers of alternatively coded value sets can be composed to hold the same input data. One way to select a value set from many possible value sets is to determine the load of Silent Coded bits that can be contained in respective sets of Amino Code values. Ideal value sets maximize the number of Silent Code bits that can be contained in fewest possible Amino Code values. Computational search engines may obviously be applied to this problem. Otherwise, the process of encoding many such layers of information is a tedious one, and prone to human error.

RESULTS

Angel Manifold

The Subhan Allah DNA Manifold given in Table 1 is shown as 86 respective DNA triplet codons annotated with "Subhan Allah" Amino Code. Corresponding Silent Code is also shown having identical "Subhan Allah" binary data, as well as a set of abjad "GAGA' repeats. The complete sequence is a 258-mer DNA.

Subhan Allah Manifold DNA:

TGTTCGGCTATAGGCGTTTCTACATCTAAGGAG-
GTAGTGGTCGCTGCTCTTTCCGTTGATTGCATA-
AATACCCTTGTTCTTATATGCAGTACAGTGCAC-
CACGTTCCTTCCGTGATACATGTACCGTCCGTA-
ATATGTCCTTCCGTCCACGATGTCAAACGCAGG-
CGTAGACGCAGAAGACGTAGGAGGCGTAGAC-
GTAGACGCCGTCGCAGGCGTAGACGTAGACGC-
CGTCGCAGGCGTAGACGTAGACGCCGTAGG

This 258-mer sequence is roughly 3 x larger than the 76-mer encoded as a single "Subhan Allah," but the Subhan Allah DNA Manifold sequence contains a total of 19.5 "Subhan Allah" repeats: one as Amino Code, one as Silent Code, and 17.5 abjad "GAGA" (11 10 11 10) repeats (abjad numbers encoded as Amino Code).

Coding efficiency

Computer-assisted encoding may be useful to select shorter, doubly-encoded sequences from inherently very large sets of possible solutions, and so, more highly efficient applications of this method can be anticipated. Nevertheless, this example is reasonably efficient in terms of maximizing the number of bits that can be stored per DNA base. In this case, 2 x 152-bit "Subhan Allah" texts (binary Arabic ASCII) are encoded, as well as 17.5 x 8-bit abjan "GAGA" encodings (152 bits), totaling 444 bits in 258 DNA bases, or 1.72 bits/DNA base. If the two, encoded 76-mer DNA sequences and the 70-mer DNA encoding the 17.5 "GAGA" repeats are also counted as input data, then input data total 748 bits in 258 DNA bases or, 2.89 bits/DNA base. To date, information density of 2 bits/DNA base has been considered theoretically possible. But when taking into account inevitable DNA reading and writing errors, a maximum of 1.8 bits of data per nucleotide of DNA has been cited as the practical limit.[4] For perspective, information density achieved with "DNA Fountain" encoding, one of the most efficient DNA data-encoding methods to date, was 1.57 bits/base.[5]

BioBricks

One option to increase potential iterations of "Subhan Allah" would be to clone into a plasmid vector using restriction sites for EcoR1 & XbaI on one end, and SpeI & PstI on the other (the basic biobrick prefix and suffix). "BioBricks" comprise a kind of warehouse of resources for the International Genetically Engineered Machine (iGEM) community, and their foundation maintains an 'open source' supply.

Adding the "BioBrick" prefix, GAATTCGCGGCCGCTTCTAGAG and suffix, TACTAGTAGCGGCCGCTGCAG, yields a 301-mer DNA:

GAATTCGCGGCCGCTTCTAGAGTGTTCGGC-TATAGGCGTTTCTACATCTAAGGAGGTAGTG-GTCGCTGCTCTTTCCGTTGATTGCATAAATAC-CCTTGTTCTTATATGCAGTACAGTGCACCAC-GTTCCTTCCGTGATACATGTACCGTCCGTAATAT-GTCCTTCCGTCCACGATGTCAAACGCAGGCG-TAGACGCAGAAGACGTAGGAGGCGTAGACG-TAGACGCCGTCGCAGGCGTAGACGTAGACG-CCGTCGCAGGCGTAGACGTAGACGCCGTAGG-TACTAGTAGCGGCCGCTGCAG

Assembly and cloning

The 301-mer BioBrick-compatible "Subhan Allah" Manifold DNA was synthesized as a gene block by GeneUniversal, Inc. (Newark, Delaware, USA) and cloned into a pUC57 bacterial expression plasmid. The gene block sequence was confirmed by Sanger sequencing using the following primers — Forward: TGTTCGGCTATAGGCGTTTC and Reverse: CGTGGACGGAAGGACATATT.

CONCLUSION

A succinct explanation of the DNA Manifolds idea to general audiences is expected to be challenging. In this case, since we don't re-encode the 258-mer into yet another ("virtual") molecule, the level of complexity will be that of only a single "manifold" and so, the "Subhan Allah" example may be easier to communicate.

Regardless of coding efficiency and potential practical applications, in this example, DNA Manifolds also becomes a tool of art and culture. It seems especially beautiful to compose this particular text as a message that folds into itself. It recalls profoundly mathematical traditions and intricate repeating calligraphy in Islamic art, and so, seems particularly appropriate.

A 1mm layer of our 258-mer on a 1.5mm pinhead would correspond with approximately 2.417 quintillion angels (6E23 x 1 x pi x 0.752 x 1E-3/(330 x 258/19.5) = 2.4E+17 angels).

6.02214076 × 1023 × 1 x 3.14 x 0.752 x 10-3/(330 x 258/19.5) = 2.417 x 1018 = 2.417 quintillion

---

REFERENCES

1.) Ibn Nur al-Din, al-Abbas. Nuzhat al-Jalis wa Munyat al-Adib al-Anis [two volumes] Al-Matba a al-Wahbiyya, Publishers, Cairo V. 1 pp. 287 (1876)

2.) https://en.wikipedia.org/wiki/Alphabetic_numeral_system

3.) http://www.oxfordislamicstudies.com/article/opr/t125/e1005

4.) Service, R. F. DNA could store all of the world's data in one room. Science; https://www.sciencemag.org/news/2017/03/dna-could-store-all-worlds-data-one-room (02 Mar. 2017)

5.) Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. Science 355, 950–954 (2017)

[See Table and Appendix below]

APPENDIX

"Subhan Allah" pUC57 Plasmid:

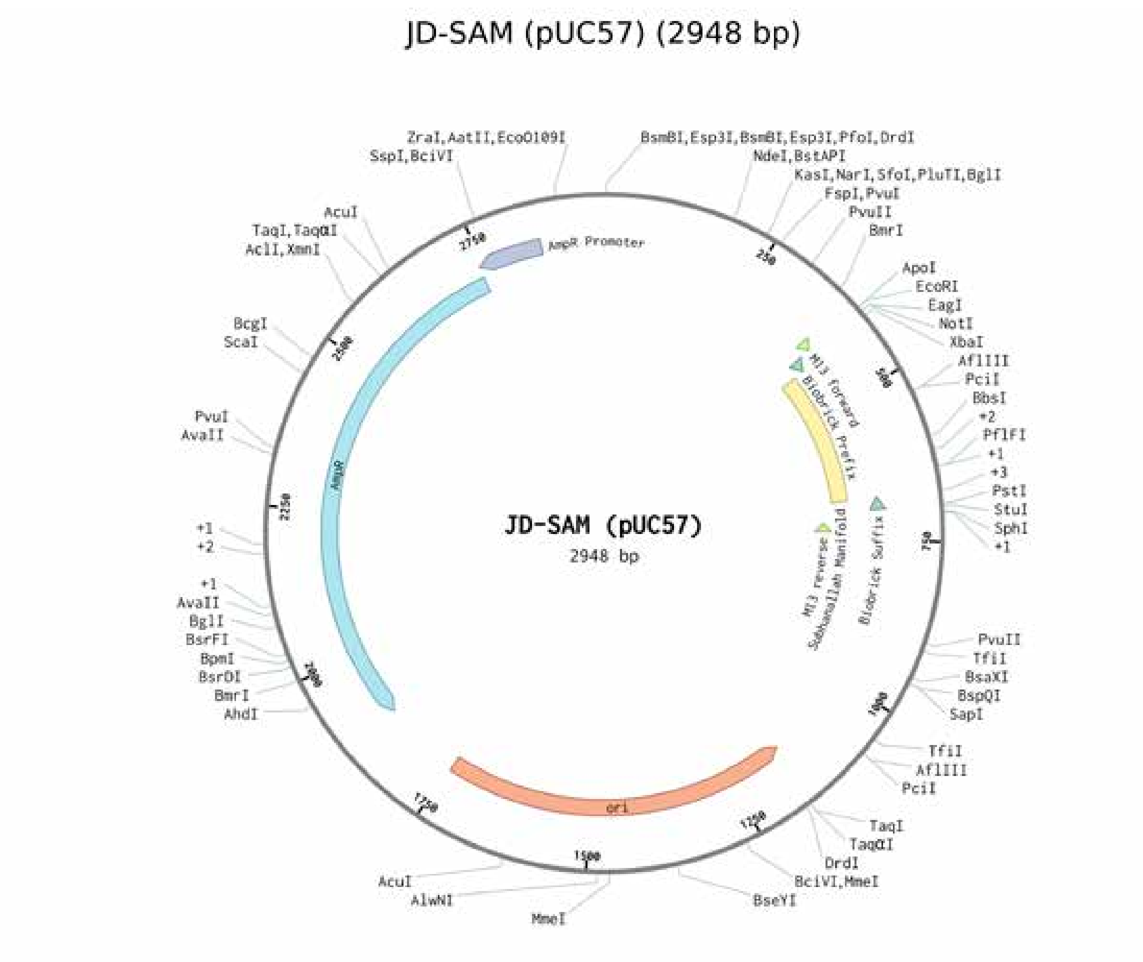

Table 1

"SUBHAN ALLAH" ARABIC ASCII BINARY:

11011000 10110011 11011000 10101000 11011000 10101101 11011000 10100111 11011001 01101000 00100000 11011000 10100111 11011001 10000100 11011001 10000100 11011001 10000111

SINGLE "SUBHAN ALLAH" ENCODING AS A 76-MER DNA:

GTACAGCGGTACAAACGTACAAGTGTACAATGGTATTAACCACCGTACAATGGTATACTCGTATACTCGT ATACTG

SUBHAN ALLAH MANIFOLD ENCODING:

Amino Code:	1101	1000	1011	00	1111	01	1000	1010	1000	110	1100	01	01	01	1011
Amino acid:	CYS	SER	ALA	ILE	GLY	VAL	SER	THR	SER	LYS	GLU	VAL	VAL	VAL	ALA
DNA:	TGT	TCG	GCT	ATA	GGC	GTT	TCT	ACA	TCT	AAG	GAG	GTA	GTG	GTC	GCT
Silent Code:	1	1	01	1	00	01	01	10	01	1	1	10	11	00	01

Amino Code:	1011	000	1010	01	111	1011	00	101	1010	000	01	000	00	1101	1000	1010
Amino acid:	ALA	LEU	SER	VAL	ASP	CYS	ILE	ASN	THR	LEU	VAL	LEU	ILE	CYS	SER	THR
DNA:	GCT	CTT	TCC	GTT	GAT	TGC	ATA	AAT	ACC	CTT	GTT	CTT	ATA	TGC	AGT	ACA
Silent Code:	01	01	00	01	1	0	1	1	00	01	01	01	1	0	11	10

Amino Code:	01	11	11	01	1001	1000	01	00	11	01	1001	1000	01	00	1101	1001
Amino acid:	VAL	HIS	HIS	VAL	PRO	SER	VAL	ILE	HIS	VAL	PRO	SER	VAL	ILE	CYS	PRO
DNA:	GTG	CAC	CAC	GTT	CCT	TCC	GTG	ATA	CAT	GTA	CCG	TCC	GTA	ATA	TGT	CCT
Silent Code:	11	0	0	01	01	00	11	1	1	10	11	00	10	1	1	01

Amino Code:	1000	01	11	–	111	01	110	1110	1110	1110	1110	1110	1110	1110	1110
Amino acid:	SER	VAL	HIS	–	ASP	VAL	LYS	ARG	ARG	ARG	ARG	ARG	ARG	ARG	ARG
DNA:	TCC	GTC	CAC	–	GAT	GTC	AAA	CGC	AGG	CGT	AGA	CGC	AGA	AGA	CGT
Silent Code:	00	00	0	–	1	00	0	00	11	01	10	00	10	10	01

Amino Code:	1110	1110	1110	1110	1110	1110	1110	1110	1110	1110	1110	1110	1110	1110
Amino acid:	ARG	ARG	ARG	ARG	ARG	ARG	ARG	ARG	ARG	ARG	ARG	ARG	ARG	ARG
DNA:	AGG	AGG	CGT	AGA	CGT	AGA	CGC	CGT	CGC	AGG	CGT	AGA	CGT	AGA
Silent Code:	11	11	01	10	01	10	00	01	00	11	01	10	01	10

Amino Code:	1110	1110	1110	1110	1110	1110	1110	1110	1110	1110	1110
Amino acid:	ARG	ARG	ARG	ARG	ARG	ARG	ARG	ARG	ARG	ARG	ARG
DNA:	CGC	CGT	CGC	AGG	CGT	AGA	CGT	AGA	CGC	CGT	AGG
Silent Code:	00	01	00	11	01	10	01	10	00	01	11

SUBHAN ALLAH MANIFOLD DNA:

TGTTCGGCTATAGGCGTTTCTACATCTAAGGAGGTAGTGGTCGCTGCTCTTTCCGTTGATTGCATAAATA
CCCTTGTTCTTATATGCAGTACAGTGCACCACGTTCCTTCCGTGATACATGTACCGTCCGTAATATGTCCT
TCCGTCCACGATGTCAAACGCAGGCGTAGACGCAGAAGACGTAGGAGGCGTAGACGTAGACGCCGTCG
CAGGCGTAGACGTAGACGCCGTCGCAGGCGTAGACGTAGACGCCGTAGG

This 258-mer "Subhan Allah DNA Manifold" sequence is roughly 3 x larger than our 76-mer encoded as a single "Subhan Allah" but this sequence contains 19.5 "Subhan Allah" repeats: one as Amino Code, one as Silent Code, and 17.5 abjad "GAGA" (11 10 11 10) repeats.

Fig. 1A

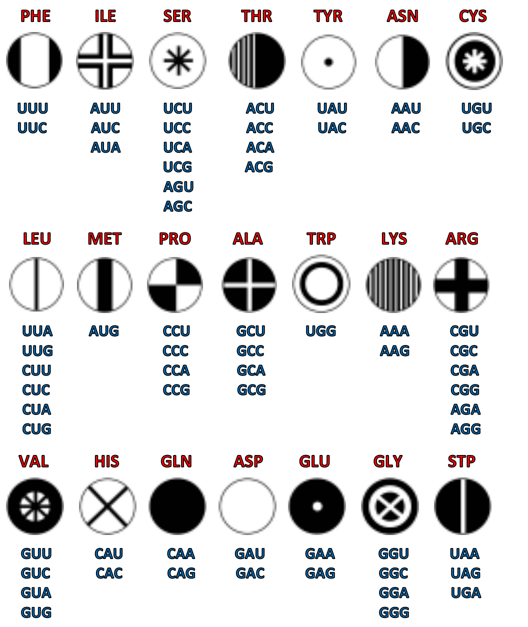

Fig. 1B

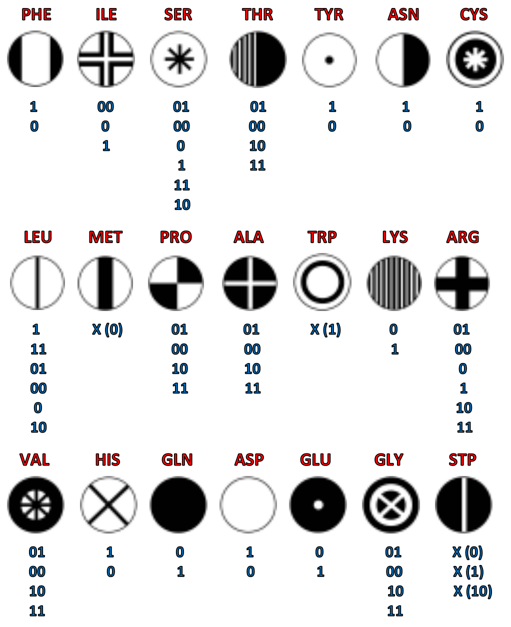

Fig. 2

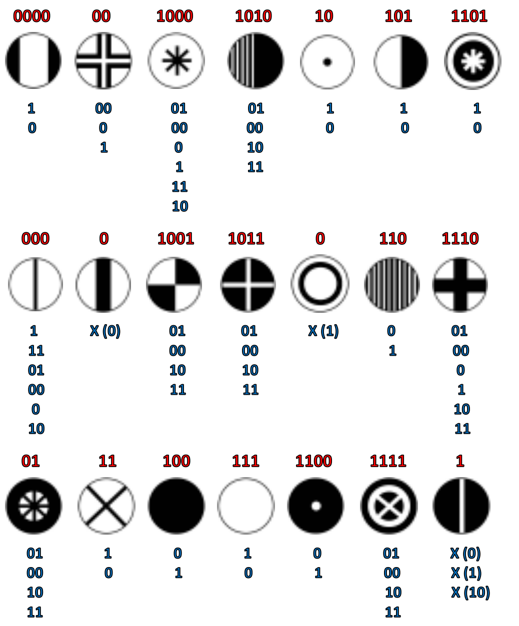

Fig. 3