Preproinsulin molecule and numbering of the twenty proteinogenic amino acids

Jean-Yves BOULAY

Jean-Yves BOULAY independent researcher – FRANCE – jean-yvesboulay@orange.fr ORCID: 0000-0001-5636-2375
https://www.researchgate.net/profile/Jean-Yves-Boulay

Abstract

The amino acid sequence of the 110-amino acid preproinsulin, the initial product of the translation of insulin mRNA, is in close dependence with the numbering of the twenty proteinogenic amino acids. Recalling here this new concept of amino acid numbering, classification deduced from their genetic translation, it is demonstrated that the orders of occurrence of the various preproinsulin amino acid, both direct and inverse sequence, are organized in numerous ratios of exact value 3/2. This, according to the new amino acid numbering concept. The degree of abundance of these amino acids in this initial single-chain molecule reveals same numerical rational phenomena.

1. Introduction

Today, it is now firmly established that living matter is organized via a so-called “universal” genetic code and that this genetic code encodes only, and very precisely, twenty proteinogenic amino acids. This number is not arbitrary, it is equal to 5x. More precisely and in a 3/2 ratio, this number of 20 entities is equal to 3x + 2x entities with a value of x equal to 4.

From a subtle numbering of the 64 codons of the universal genetic code, we propose a numbering (from 0 to 19) of the twenty amino acids. These two numbering systems, including the first proposed by Professor Sergey Petoukhov [1], are very directly dependent on the physico-chemical properties of the four nucleobases that make up DNA. They are therefore very legitimate to be used for the study of the genetic code mechanism. When we number the twenty amino acids, which are, very importantly, 5x in number, then we classify them into two symmetrical sets of 12 (or 3x) and 8 (or 2x) entities.

In preview published paper *"Numbering of the twenty proteinogenic amino acids"* [2], we have demonstrated that a large number of different amino acid attributes arrange themselves numerically in exact 3/2 value ratios according to this numbering system. Recalling here this new concept of amino acid numbering, we will also demonstrate that this numbering strongly influences the way in which all the amino acids of the 110-amino acid preproinsulin are organized. We therefore study here only he initial product of the translation of insulin mRNA.

We study in this paper the "human" version of preproinsulin. We chose this molecule because it constitutes a very essential protein placed very high in the evolutionary hierarchy of living matter.

In order to clarify and lighten the presentation of the phenomena described, some additional explanations are given in the appendix only.

2. Numbering of the twenty proteinogenic amino acids

In order to be able to number the twenty proteinogenic amino acids, we must first proceed to a numbering of the 64 codons of the universal genetic code. Also, this numbering of amino acids must depend on the physico-chemical character of the nucleobases constituting the codons. To this end, we use the very original numbering devised by Professor Sergey Petoukhov, which is based on the possible deamination and depurination of the four nucleobases. Additional explanations are available in the appendix to complement those in this chapter.

2.1. Petoukhov’s numbering of the 64 genetic code codons

In his investigations of the genetic code [1] Sergey Petoukhov assigns a number from 0 to 63 to each of the sixty-four codons. This Petoukhov numbering is directly dependent on the physico-chemical properties of the four DNA coding bases.

Using a very sophisticated method, Sergey Petoukhov manages to classify the full sixty-four codons set using a binary language (or alphabet, we invite the reader to consult the full article by Sergei Petoukhov [1]). Depending on whether each nucleobase can undergo deamination or not, Sergey Petoukhov assigns them either the value 1 or the value 0 (see table Figure A1 in appendix). Also, depending on whether each nucleobase can undergo depurination or not, Sergey Petoukhov assigns them either the value 0 or the value 1.

This double criterion makes it possible, for each codon, to create a six-digit binary number by juxtaposition of two three-digit numbers as described in Figure A2 in appendix. Sergey Petoukhov then classifies very subtly in superimposed squares of 4, 16 and 64 boxes the 64 codons and numbers them in the order of the bases G→T→A→C for the first, second and third bases. In this numbering imagined by Sergey Petoukhov, the GGG codon thus bears the number 0 (binary 000000) and the CCC codon the number 63 (binary 111111). Figure 1 illustrates this complete numbering of the 64 genetic code codons set.

	111	*110*	*101*	*100*	*011*	*010*	*001*	*000*
111	CCC Pro 63 111111	CCA Pro 62 111110	CAC His 61 111101	CAA Gln 60 111100	ACC Thr 59 111011	ACA Thr 58 111010	AAC Asn 57 111001	AAA Lys 56 111000
110	CCT Pro 55 110111	**CCG** **Pro** **54** **110110**	**CAT** **His** **53** **110101**	**CAG** **Gln** **52** **110100**	ACT Thr 51 110011	**ACG** **Thr** **50** **110010**	**AAT** **Asn** **49** **110001**	**AAG** **Lys** **48** **110000**
101	CTC Leu 47 101111	CTA Leu 46 101110	CGC Arg 45 101101	CGA Arg 44 101100	ATC Ile 43 101011	ATA Ile 42 101010	AGC Ser 41 101001	AGA Arg 40 101000
100	CTT Leu 39 100111	CTG Leu 38 100110	CGT Arg 37 100101	CGG Arg 36 100100	**ATT** **Ile** **35** **100011**	**ATG** **Met** **34** **100010**	AGT Ser 33 100001	**AGG** **Arg** **32** **100000**
011	TCC Ser 31 011111	TCA Ser 30 011110	TAC Tyr 29 011101	TAA Stop 28 011100	GCC Ala 27 011011	GCA Ala 26 011010	GAC Asp 25 011001	GAA Glu 24 011000
010	TCT Ser 23 010111	**TCG** **Ser** **22** **010110**	**TAT** **Tyr** **21** **010101**	TAG Stop 20 010100	GCT Ala 19 010011	**GCG** **Ala** **18** **010010**	**GAT** **Asp** **17** **010001**	**GAG** **Glu** **16** **010000**
001	TTC Phe 15 001111	TTA Leu 14 001110	TGC Cys 13 001101	TGA Stop 12 001100	GTC Val 11 001011	GTA Val 10 001010	GGC Gly 9 001001	GGA Gly 8 001000
000	**TTT** **Phe** **7** **000111**	**TTG** **Leu** **6** **000110**	**TGT** **Cys** **5** **000101**	**TGG** **Trp** **4** **000100**	GTT Val 3 000011	**GTG** **Val** **2** **000010**	GGT Gly 1 000001	**GGG** **Gly** **0** **000000**

Figure 1: Numbering of the 64 codons according to Sergey Petoukhov genetic code investigations [1] and distinction (grey areas) of the first appearance of each of the 20 coded amino acids. See Figure A1 and A2 also.

2.2. Numbering of the twenty proteinogenic amino acids

From this numbering system, in order to assign a number to each of the twenty proteinogenic amino acids, the most logical procedure is therefore proposed here, which is to follow the order of appearance of the amino acids according to this numbering of the codons (from 0 to 63) of the table by Sergey Petoukhov (Figure 1).

By this process, it is thus assigned (Figure 2) number 0 to Glycine, number 1 to Valine and to Proline, the last amino acid to appear according to this order of numbering of the sixty-four genetic code codons, 19 as number.

Only one number assigning to amino acids

Pro	Pro	His	Gln	Thr	Thr	Asn	Lys
Pro	**19** **Pro**	**18** **His**	**17** **Gln**	Thr	**16** **Thr**	**15** **Asn**	**14** **Lys**
Leu	Leu	Arg	Arg	Ile	Ile	Ser	Arg
Leu	Leu	Arg	Arg	**13** **Ile**	**12** **Met**	Ser	**11** **Arg**
Ser	Ser	Tyr	*Stop*	Ala	Ala	Asp	Glu
Ser	**10** **Ser**	**9** **Tyr**	*Stop*	Ala	**8** **Ala**	**7** **Asp**	**6** **Glu**
Phe	Leu	Cys	*Stop*	Val	Val	Gly	Gly
5 **Phe**	**4** **Leu**	**3** **Cys**	**2** **Trp**	Val	**1** **Val**	Gly	**0** **Gly**

Figure 2: Assigning a single only one number to each of 20 proteinogenic amino acids in the table of the complete genetic code. See Figure 1 also.

2.3. Symmetrical break-up of the 20 AAs in 3/2 ratio

Now that we have determined a numbering of amino acids by assigning them a unique and personal number, we propose to isolate these twenty entities in two sets of unequal size. We therefore distinguish, in Figure 3, a first set of 12 entities then a second set of 8 other entities. As illustrated in Figure 6, these two sets then oppose each other in a ratio of value 3/2.

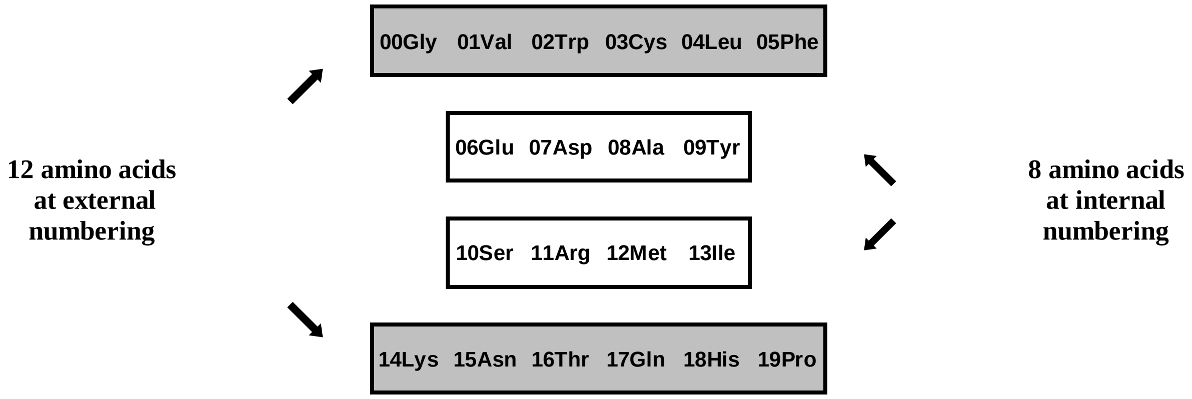

00Gly 01Val 02Trp 03Cys 04Leu 05Phe
06Glu 07Asp 08Ala 09Tyr
10Ser 11Arg 12Met 13Ile
14Lys 15Asn 16Thr 17Gln 18His 19Pro
12 amino acids at external numbering
8 amino acids at internal numbering

Figure 3: Conventional representation of 20 proteinogenic amino acids numbering in symmetry graphics. Since them numbering, symmetrical break-up of the 20 AAs into two sets of 2 times 6 versus 2 times 4 entities. See Figure 2 also.

Using symmetry graphics, many arithmetic phenomena presented in this paper will be presented in the way illustrated in Figure 3. Thereby, each of the 20 amino acids is symmetrically positioned to the one of opposite numbering in relation to the numbering order of these 20 AAs*: *00Gly* versus *19Pro*, *01Val* versus *18His*, etc.

Also, we therefore isolate two numbering zones:

- an area called “external” with inside the first six and last six numbered AAs
- an area called “internal” with inside the two times four centrally numbered AAs.

** To simplify, in some parts of text and tables, AA (or AAs) is used to replace amino acid appellation.*

2.4. Alphanumeric symbol of the 20 proteinogenic amino acids

In preview published paper *"Numbering of the twenty proteinogenic amino acids"* [2], we have proposed a new nomenclature of the twenty proteinogenic amino acids according to the numbering system that we have just presented above.

Now, we will therefore describe each of all these twenty amino acids in an 5 characters alphanumeric symbol: 2 digits and 3 letters. For example, Glycine is named *00Gly* and Proline *19Pro*. Table A3 in Appendix list this new nomenclature of the twenty proteinogenic amino acids.

3. The 110-amino acid preproinsulin molecule

3.1. Insulin biosynthesis

Insulin is a complex and essential protein managing in particular the energy needs of living organisms. The initial product of the translation of insulin mRNA is the insulin precursor preproinsulin, a single chain polypeptide, consisting by a sequence of 110 amino acids.

Here is studied this molecule in its initial configuration, i.e. in its complete structure as it is primarily coded. We therefore study here the human preproinsulin molecule [4]. The table in Figure 4 lists the complete sequence of the 110 AAs of preproinsulin in the order of their genetic encoding.

1st	*2nd*	*3rd*	*4th*	*5th*	*6th*	*7th*	*8th*	*9th*	*10th*
12Met	**08Ala**	**04Leu**	**02Trp**	**12Met**	**11Arg**	**04Leu**	**04Leu**	**19Pro**	**04Leu**
11th	*12th*	*13th*	*14th*	*15th*	*6th*	*17th*	*18th*	*19th*	*20th*
04Leu	**08Ala**	**04Leu**	**04Leu**	**08Ala**	**04Leu**	**02Trp**	**00Gly**	**19Pro**	**07Asp**
21st	*22nd*	*23rd*	*24th*	*25th*	*26th*	*27th*	*28th*	*29th*	*30th*
19Pro	**08Ala**	**08Ala**	**08Ala**	**05Phe**	**01Val**	**15Asn**	**17Gln**	**18His**	**04Leu**
31st	*32nd*	*33rd*	*34th*	*35th*	*36th*	*37th*	*38th*	*39th*	*40th*
03Cys	**00Gly**	**10Ser**	**18His**	**04Leu**	**01Val**	**06Glu**	**08Ala**	**04Leu**	**09Tyr**
41st	*42nd*	*43rd*	*44th*	*45th*	*46th*	*47th*	*48th*	*49th*	*50th*
04Leu	**01Val**	**03Cys**	**00Gly**	**06Glu**	**11Arg**	**00Gly**	**05Phe**	**05Phe**	**09Tyr**
51st	*52nd*	*53rd*	*54th*	*55th*	*56th*	*57th*	*58th*	*59th*	*60th*
16Thr	**19Pro**	**14Lys**	**16Thr**	**11Arg**	**11Arg**	**06Glu**	**08Ala**	**06Glu**	**07Asp**
61st	*62nd*	*63rd*	*64th*	*65th*	*66th*	*67th*	*68th*	*69th*	*70th*
04Leu	**17Gln**	**01Val**	**00Gly**	**17Gln**	**01Val**	**06Glu**	**04Leu**	**00Gly**	**00Gly**
71st	*72nd*	*73rd*	*74th*	*75th*	*76th*	*77th*	*78th*	*79th*	*80th*
00Gly	**19Pro**	**00Gly**	**08Ala**	**00Gly**	**10Ser**	**04Leu**	**17Gln**	**19Pro**	**04Leu**
81st	*82nd*	*83rd*	*84th*	*85th*	*86th*	*87th*	*88th*	*89th*	*90th*
08Ala	**04Leu**	**06Glu**	**00Gly**	**10Ser**	**04Leu**	**17Gln**	**14Lys**	**11Arg**	**00Gly**
91st	*92nd*	*93rd*	*94th*	*95th*	*96th*	*97th*	*98th*	*99th*	*100th*
13Ile	**01Val**	**06Glu**	**17Gln**	**03Cys**	**03Cys**	**16Thr**	**10Ser**	**13Ile**	**03Cys**
101st	*102nd*	*103rd*	*104th*	*105th*	*106th*	*107th*	*108th*	*109th*	*110th*
10Ser	**04Leu**	**09Tyr**	**17Gln**	**04Leu**	**06Glu**	**15Asn**	**09Tyr**	**03Cys**	**15Asn**

Figure 4: Listing of the 110 preproinsulin amino acids in order of their genetic coding.

3.2. Configuration of the 110-amino acid preproinsulin molecule

From the table in Figure 4, AAs primary attributes according to their apparition order and quantification are detailed in the following table in Figure 5. It is all these data that are subject of this item about numbering of twenty proteinogenic amino acid and their distribution in preproinsulin molecule.

AA	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*
00Gly	12	x		x			x	7	11
01Val	7	x		x		x		10	10
02Trp	2		x	x			x	4	19
03Cys	6	x			x	x		14	2
04Leu	20	x		x		x		3	5
05Phe	3		x	x			x	9	17
06Glu	8	x			x	x		16	4
07Asp	2		x	x			x	8	16
08Ala	10	x		x			x	2	14
09Tyr	4		x		x	x		17	3
10Ser	5	x			x	x		15	7
11Arg	5	x		x			x	5	12
12Met	2		x	x			x	1	20
13Ile	2		x		x	x		20	8
14Lys	2		x		x		x	19	13
15Asn	3		x		x	x		11	1
16Thr	3		x		x	x		18	9
17Gln	6	x			x	x		12	6
18His	2		x		x		x	13	18
19Pro	6	x		x			x	6	15

Cumulated values	10	10	10	10	10	10	210	210
12 external numbered AAs (from 0 to 5 and from 14 to 19)	6	6	6	6	6	6	126	126
8 internal numbered AAs (from 6 to 13)	4	4	4	4	4	4	84	84
ratio →	3/2	3/2	3/2	3/2	3/2	3/2	3/2	3/2

a total presence quantity

b 10 amino acids in largest number (out 20)

c 10 amino acids in smallest number (out 20)

d first 10 amino acids to appear (out 20) from first to last located AA (from 1st to 110th)

e last 10 amino acids to appear (out 20) from first to last located AA (from 1st to 110th)

f first 10 amino acids to appear (out 20) from last to first located AA (from 110th to 1st)

g last 10 amino acids to appear (out 20) from last to first located AA (from 110th to 1st)

h rank of appearance order from first to last located AA (from 1st to 110th)

i rank of appearance order from last to first located AA (from 110th to 1st)

Figure 5: According to their apparition order and quantification, some primary attributes of the preproinsulin amino acids. See Figure 4 also.

4. Order of the first AAs apparition in preproinsulin chain

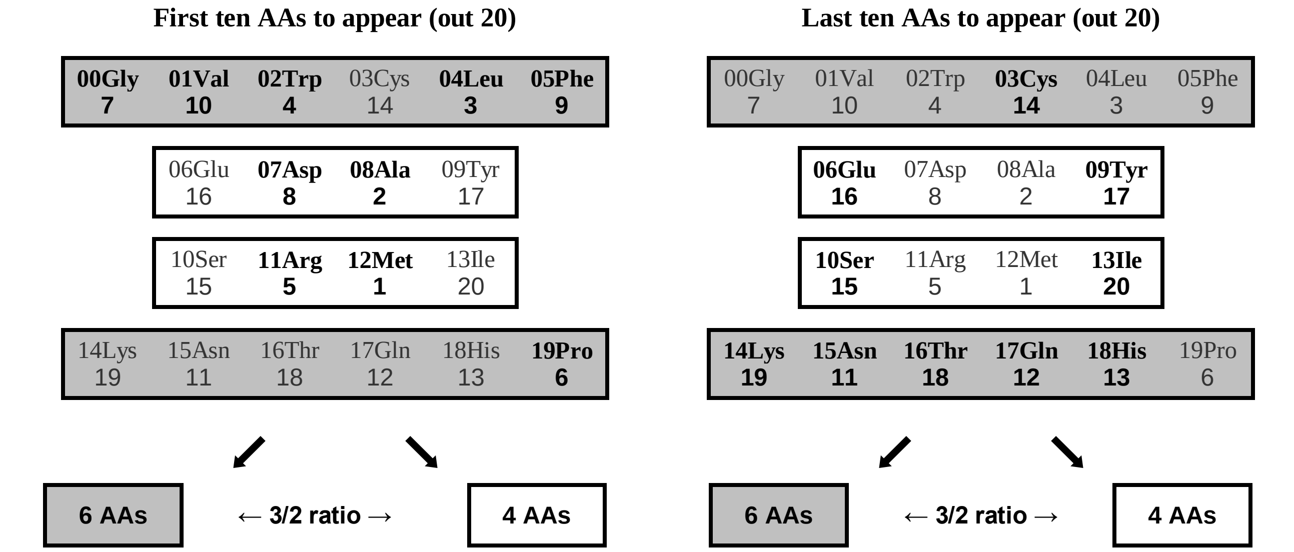
In Figure 6 are identified the first appearance of each of 20 different proteinogenic AAs inside of the 110-amino acid preproinsulin molecule. This, in order of their genetic coding.

1st	*2nd*	*3rd*	*4th*	*5th*	*6th*	*7th*	*8th*	*9th*	*10th*
12Met	08Ala	04Leu	02Trp	12Met	11Arg	04Leu	04Leu	19Pro	04Leu
11th	*12th*	*13th*	*14th*	*15th*	*6th*	*17th*	*18th*	*19th*	*20th*
04Leu	08Ala	04Leu	04Leu	08Ala	04Leu	02Trp	00Gly	19Pro	07Asp
21st	*22nd*	*23rd*	*24th*	*25th*	*26th*	*27th*	*28th*	*29th*	*30th*
19Pro	08Ala	08Ala	08Ala	05Phe	01Val	15Asn	17Gln	18His	04Leu
31st	*32nd*	*33rd*	*34th*	*35th*	*36th*	*37th*	*38th*	*39th*	*40th*
03Cys	00Gly	10Ser	18His	04Leu	01Val	06Glu	08Ala	04Leu	09Tyr
41st	*42nd*	*43rd*	*44th*	*45th*	*46th*	*47th*	*48th*	*49th*	*50th*
04Leu	01Val	03Cys	00Gly	06Glu	11Arg	00Gly	05Phe	05Phe	09Tyr
51st	*52nd*	*53rd*	*54th*	*55th*	*56th*	*57th*	*58th*	*59th*	*60th*
16Thr	19Pro	14Lys	16Thr	11Arg	11Arg	06Glu	08Ala	06Glu	07Asp
61st	*62nd*	*63rd*	*64th*	*65th*	*66th*	*67th*	*68th*	*69th*	*70th*
04Leu	17Gln	01Val	00Gly	17Gln	01Val	06Glu	04Leu	00Gly	00Gly
71st	*72nd*	*73rd*	*74th*	*75th*	*76th*	*77th*	*78th*	*79th*	*80th*
00Gly	19Pro	00Gly	08Ala	00Gly	10Ser	04Leu	17Gln	19Pro	04Leu
81st	*82nd*	*83rd*	*84th*	*85th*	*86th*	*87th*	*88th*	*89th*	*90th*
08Ala	04Leu	06Glu	00Gly	10Ser	04Leu	17Gln	14Lys	11Arg	00Gly
91st	*92nd*	*93rd*	*94th*	*95th*	*96th*	*97th*	*98th*	*99th*	*100th*
13Ile	01Val	06Glu	17Gln	03Cys	03Cys	16Thr	10Ser	13Ile	03Cys
101st	*102nd*	*103rd*	*104th*	*105th*	*106th*	*107th*	*108th*	*109th*	*110th*
10Ser	04Leu	09Tyr	17Gln	04Leu	06Glu	15Asn	09Tyr	03Cys	15Asn

Figure 6: Identification (grey area) of the first appearance of each of 20 different proteinogenic AAs inside of the 110-amino acid preproinsulin molecule in order of their genetic coding.

4.1. First ten and last ten AAs to appear in preproinsulin sequence.

According to data from table 5 and illustration in Figure 6, it is possible to distinguish the first ten amino acids (among the list of 20 proteinogens) and the last ten to appear in the sequence of the 110-amino acid preproinsulin.



First ten AAs to appear (out 20)

00Gly 7 | **01Val 10** | **02Trp 4** | 03Cys 14 | **04Leu 3** | **05Phe 9**

06Glu 16 | **07Asp 8** | **08Ala 2** | 09Tyr 17

10Ser 15 | **11Arg 5** | **12Met 1** | 13Ile 20

14Lys 19 | 15Asn 11 | 16Thr 18 | 17Gln 12 | 18His 13 | **19Pro 6**

6 AAs ← 3/2 ratio → 4 AAs

Last ten AAs to appear (out 20)

00Gly 7 | 01Val 10 | 02Trp 4 | **03Cys 14** | 04Leu 3 | 05Phe 9

06Glu 16 | 07Asp 8 | 08Ala 2 | **09Tyr 17**

10Ser 15 | 11Arg 5 | 12Met 1 | **13Ile 20**

14Lys 19 | **15Asn 11** | **16Thr 18** | **17Gln 12** | **18His 13** | 19Pro 6

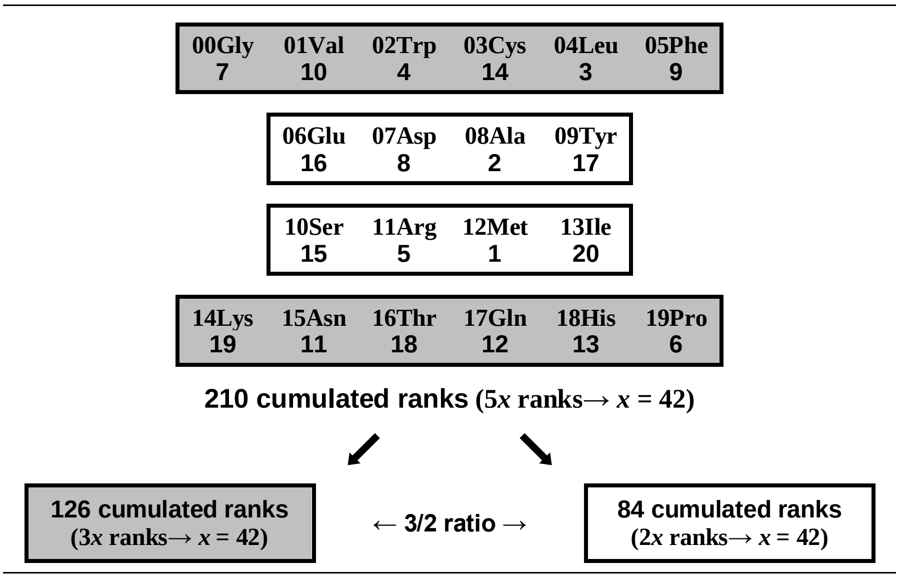
6 AAs ← 3/2 ratio → 4 AAs

Figure 7: According to their numbering, distribution in 3/2 ratio of two sets of first ten and last ten AAs to appear in preproinsulin sequence. Numbers are first occurrence rank. See Figure 5 and 6 also.

As shown in Figure 7, It therefore turns out that six of the first ten AAs to appear are externally numbered versus four internally numbered. This in an exact ratio of 3/2 value. The last ten AAs to appear in the preproinsulin sequence are distributed in this same ratio of value 3/2 in relation to their numbering.

4.2. Occurrence rank of the twenty AAs in preproinsulin sequence

In table Figure 5, in reference *h*, it is listed the rank of appearance (first occurrence) of each of the twenty proteinogenic amino acids in preproinsulin and in the order of the RNA translation sequence. For example *12Met* is the first AA to appear (rank 1) at the 1st position in preproinsulin chain and at the 91st position, *13Ile* appears last (rank 20).



00Gly	01Val	02Trp	03Cys	04Leu	05Phe
7	10	4	14	3	9

06Glu	07Asp	08Ala	09Tyr
16	8	2	17

10Ser	11Arg	12Met	13Ile
15	5	1	20

14Lys	15Asn	16Thr	17Gln	18His	19Pro
19	11	18	12	13	6

210 cumulated ranks (5*x* ranks→ *x* = 42)

126 cumulated ranks (3*x* ranks→ *x* = 42)

← 3/2 ratio →

84 cumulated ranks (2*x* ranks→ *x* = 42)

Figure 8: Occurrence ranks of the twenty AAs in preproinsulin chain (first occurrence in translation sequence order). See Figure 11 to comparison.

As illustrated Figure 8, it turns out that the cumulative value of these different occurrence ranks (from 1 to 20) oppose each other in ratio of an exact 3/2 value between the two numbering sets of amino acids. Indeed, the 12 AAs with external numbering accumulate 126 occurrence ranks (3*x* → *x* = 42) and the 8 with internal numbering accumulate 84 (2*x* → *x* = 42).

This is an primordial observation demonstrating how the preproinsulin components arrange themselves numerically according to the numbering of twenty proteinogenic amino acids. The next demonstrations will reinforce this point of view.

5. Order of the first AAs apparition in preproinsulin chain in reverse sequense order.

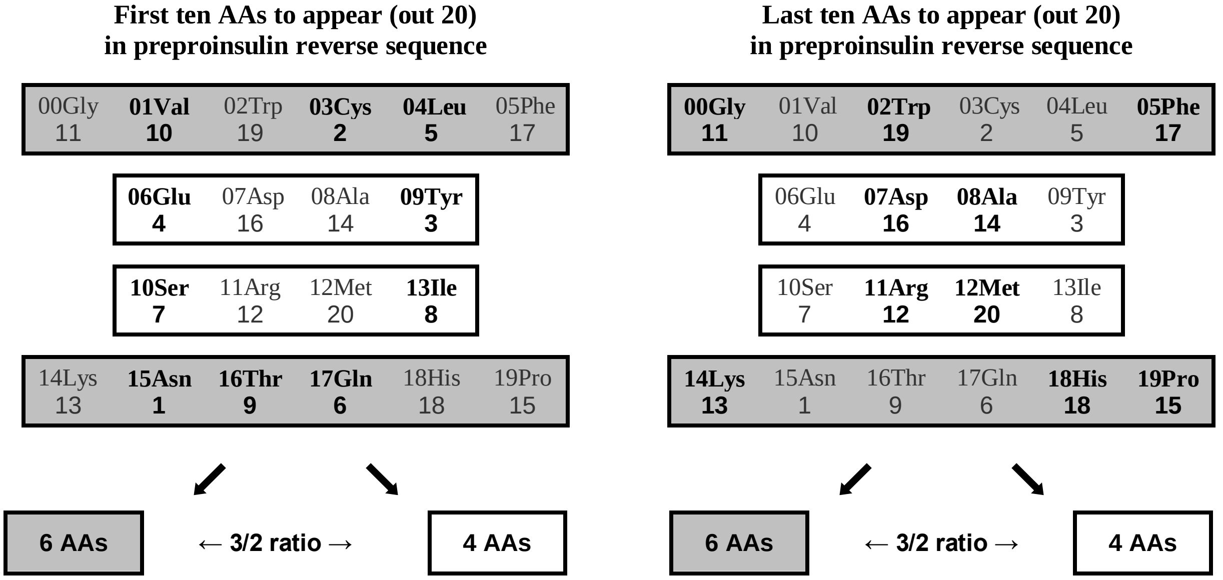
In Figure 9 are identified the first appearance of each of 20 different proteinogenic AAs inside of the 110-amino acid preproinsulin molecule. This, in reverse order of their genetic coding, so from 110th position to 1st position.

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
12Met	08Ala	04Leu	02Trp	**12Met**	11Arg	04Leu	04Leu	19Pro	04Leu
11th	*12th*	*13th*	*14th*	*15th*	*6th*	*17th*	*18th*	*19th*	*20th*
04Leu	08Ala	04Leu	04Leu	08Ala	04Leu	**02Trp**	00Gly	19Pro	07Asp
21st	*22nd*	*23rd*	*24th*	*25th*	*26th*	*27th*	*28th*	*29th*	*30th*
19Pro	08Ala	08Ala	08Ala	05Phe	01Val	15Asn	17Gln	**18His**	04Leu
31st	*32nd*	*33rd*	*34th*	*35th*	*36th*	*37th*	*38th*	*39th*	*40th*
03Cys	00Gly	10Ser	18His	04Leu	01Val	06Glu	08Ala	04Leu	09Tyr
41st	*42nd*	*43rd*	*44th*	*45th*	*46th*	*47th*	*48th*	*49th*	*50th*
04Leu	01Val	03Cys	00Gly	06Glu	11Arg	00Gly	05Phe	**05Phe**	09Tyr
51st	*52nd*	*53rd*	*54th*	*55th*	*56th*	*57th*	*58th*	*59th*	*60th*
16Thr	19Pro	14Lys	16Thr	11Arg	11Arg	06Glu	08Ala	06Glu	**07Asp**
61st	*62nd*	*63rd*	*64th*	*65th*	*66th*	*67th*	*68th*	*69th*	*70th*
04Leu	17Gln	01Val	00Gly	17Gln	01Val	06Glu	04Leu	00Gly	00Gly
71st	*72nd*	*73rd*	*74th*	*75th*	*76th*	*77th*	*78th*	*79th*	*80th*
00Gly	19Pro	00Gly	08Ala	00Gly	10Ser	04Leu	17Gln	**19Pro**	04Leu
81st	*82nd*	*83rd*	*84th*	*85th*	*86th*	*87th*	*88th*	*89th*	*90th*
08Ala	04Leu	06Glu	00Gly	10Ser	04Leu	17Gln	**14Lys**	**11Arg**	**00Gly**
91st	*92nd*	*93rd*	*94th*	*95th*	*96th*	*97th*	*98th*	*99th*	*100th*
13Ile	**01Val**	06Glu	17Gln	03Cys	03Cys	**16Thr**	10Ser	**13Ile**	03Cys
101st	*102nd*	*103rd*	*104th*	*105th*	*106th*	*107th*	*108th*	*109th*	*110th*
10Ser	04Leu	09Tyr	**17Gln**	**04Leu**	**06Glu**	15Asn	**09Tyr**	**03Cys**	**15Asn**

Figure 9: Identification (grey area) of the first appearance of each of 20 different proteinogenic AAs inside of the 110-amino acid preproinsulin molecule in reverse order of their genetic coding.

5.1. First ten and last ten AAs to appear in preproinsulin reverse sequence.

According to data from table 5 and illustration in Figure 9, it is possible to distinguish the first ten amino acids (among the list of 20 proteinogens) and the last ten to appear in the reverse sequence of the 110-amino acid preproinsulin (from 110th position to 1st position).



First ten AAs to appear (out 20) in preproinsulin reverse sequence

00Gly	**01Val**	02Trp	**03Cys**	**04Leu**	05Phe
11	**10**	19	**2**	**5**	17

06Glu	07Asp	08Ala	**09Tyr**
4	16	14	**3**

10Ser	11Arg	12Met	**13Ile**
7	12	20	**8**

14Lys	**15Asn**	**16Thr**	**17Gln**	18His	19Pro
13	**1**	**9**	**6**	18	15

6 AAs ← 3/2 ratio → 4 AAs

Last ten AAs to appear (out 20) in preproinsulin reverse sequence

00Gly	01Val	**02Trp**	03Cys	04Leu	**05Phe**
11	10	**19**	2	5	**17**

06Glu	**07Asp**	**08Ala**	09Tyr
4	**16**	**14**	3

10Ser	**11Arg**	**12Met**	13Ile
7	**12**	**20**	8

14Lys	15Asn	16Thr	17Gln	**18His**	**19Pro**
13	1	9	6	**18**	**15**

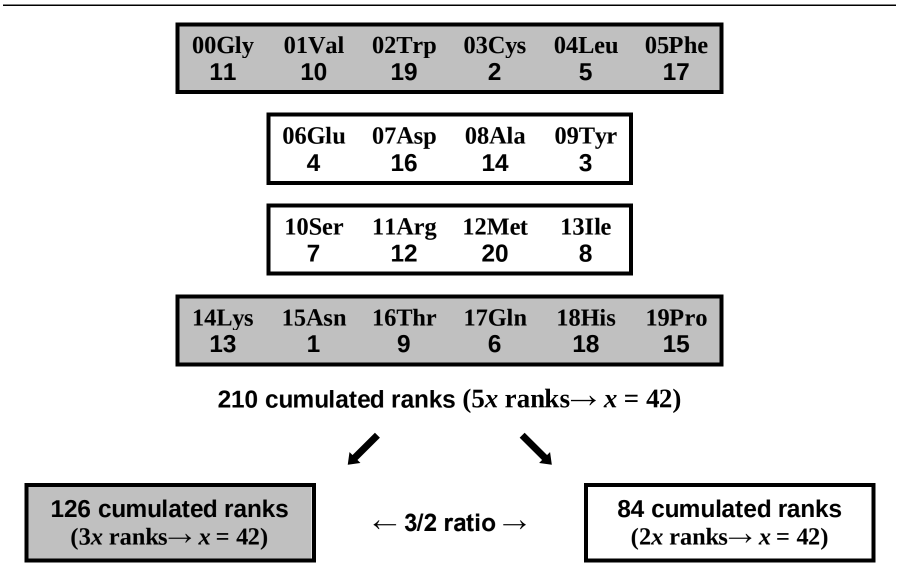
6 AAs ← 3/2 ratio → 4 AAs

Figure 10: According to their numbering, distribution in 3/2 ratio of two sets of first ten and last ten AAs to appear in preproinsulin reverse sequence. Numbers are first occurrence rank.

It therefore turns out that six of the first ten AAs to appear are externally numbered versus four internally numbered. This in an exact ratio of 3/2 value. The last ten AAs to appear in the reverse preproinsulin sequence are distributed in this same ratio of value 3/2 in relation to their numbering.

5.2. Occurrence rank of the twenty AAs in preproinsulin reverse sequence

In table Figure 5, in reference *i*, it is listed the rank of appearance (first occurrence) of each of the twenty proteinogenic amino acids in preproinsulin and in the reverse order of the RNA translation sequence. For example, in reverse sequence, at the 110th position, *15Asn* is the first AA to appear (rank 1) and at the 5th position, *20Met* appears last (rank 20).



00Gly	01Val	02Trp	03Cys	04Leu	05Phe
11	10	19	2	5	17

06Glu	07Asp	08Ala	09Tyr
4	16	14	3

10Ser	11Arg	12Met	13Ile
7	12	20	8

14Lys	15Asn	16Thr	17Gln	18His	19Pro
13	1	9	6	18	15

210 cumulated ranks (5*x* ranks→ *x* = 42)

126 cumulated ranks
(3*x* ranks→ *x* = 42)

← 3/2 ratio →

84 cumulated ranks
(2*x* ranks→ *x* = 42)

Figure 11: Occurrence rank of the twenty AAs in preproinsulin reverse sequence.
See Figure 8 to comparison.

As illustrated Figure 11, it turns out that the cumulative value of these different occurrence ranks (from 1 to 20) oppose each other in a perfect ratio of an exact 3/2 value between the two numbering sets of amino acids. Indeed, the 12 AAs with external numbering accumulate 126 occurrence ranks (3*x* → *x* = 42) and the 8 with internal numbering accumulate 84 (2*x* → *x* = 42).

This is exactly as for the ranks of occurrence in direct order of translation and although these ranks, according to this reverse sequence, are different for each of the twenty amino acids. This phenomenon, which operates both in direct sequence order and in reverse order, has very little chance of being the result of chance.

Once again, this is an primordial fact demonstrating how the preproinsulin components arrange themselves numerically according to the numbering of twenty proteinogenic amino acids.

6. Amino acid abundance in preproinsulin

We will now study the abundance of each of the twenty proteinogenic amino acids in the 110-amino acid preproinsulin and show a dependence between these respective abundances and the numbering of the twenty amino acids.

6.1. Amino acid abundance graph

Amino acid number in preproinsulin is equal to 110 so 3 times 37 – 1. As the AA abundance graph in Figure 12 illustrates, there is a strong imbalance between the ten amino acids of first numbering (from *00Gly* to *09Tyr*) and the following ten (from *10Ser* to *19Pro*).

Thus, to within one unit, there are exactly twice the number of amino acids present among the ten first numbered (74 so 2 times 37) than among the last ten numbered which are 36 so 37 -1.

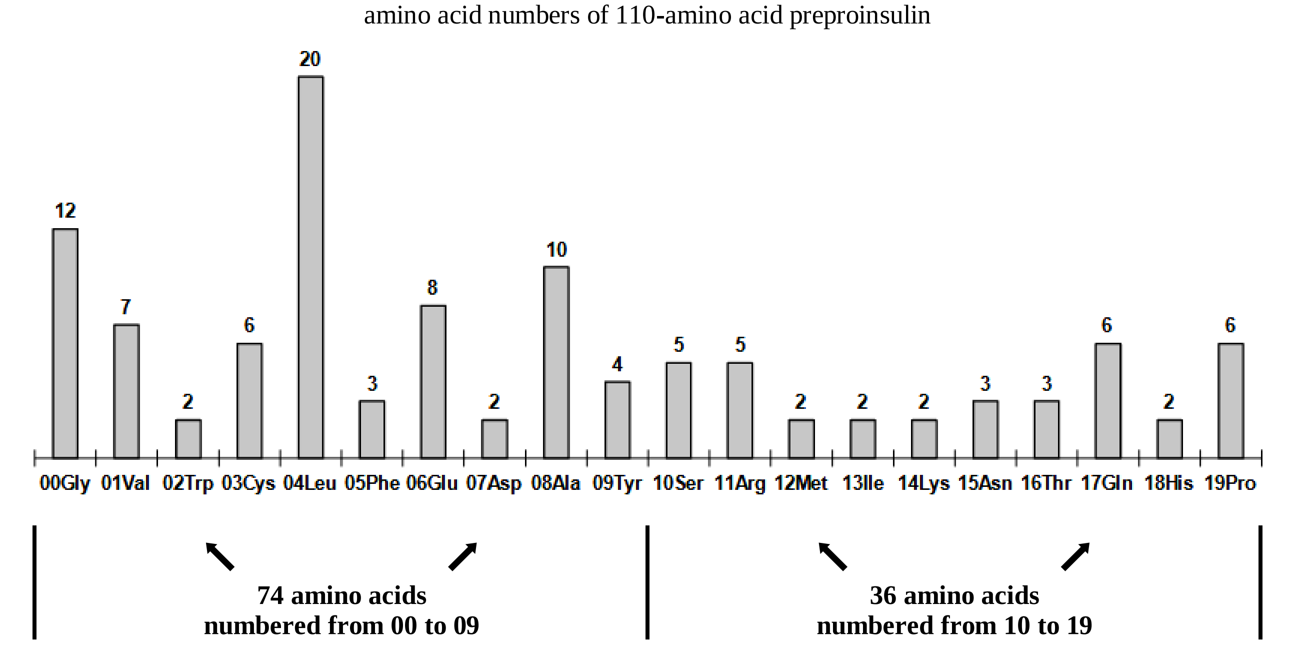
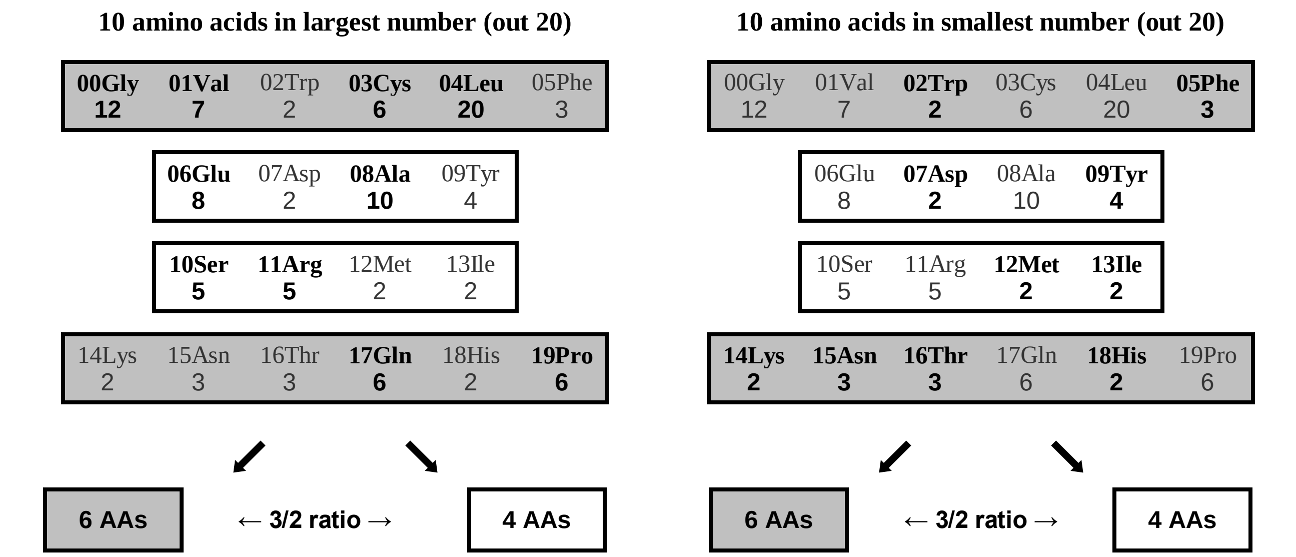


Figure 12: Amino acid abundance in 110-amino acid preproinsulin

It therefore seems very obvious that the numbering of the twenty proteinogenic amino acids greatly influences their rate of abundance in 110-amino acid preproinsulin. However, despite its imbalance operating in a near-perfect 2/1 ratio, same phenomena of 3/2 ratios as those presented above operates also about these abundance rates.

6.2. Ten largest numbers and ten smallest numbers of AAs

We have therefore just shown a strong tendency for the constituents of preproinsulin to be much more of the first half numbering amino acids than of the second half. Nevertheless, by distinguishing, in preproinsulin, the ten amino acids of greatest abundance from the ten of lowest abundance, we note that, for each of these two sets of ten AAs, six are of external numbering and four of internal numbering.



10 amino acids in largest number (out 20)

00Gly **12** | **01Val** **7** | 02Trp 2 | **03Cys** **6** | **04Leu** **20** | 05Phe 3

06Glu **8** | 07Asp 2 | **08Ala** **10** | 09Tyr 4

10Ser **5** | **11Arg** **5** | 12Met 2 | 13Ile 2

14Lys 2 | 15Asn 3 | 16Thr 3 | **17Gln** **6** | 18His 2 | **19Pro** **6**

6 AAs ← 3/2 ratio → 4 AAs

10 amino acids in smallest number (out 20)

00Gly 12 | 01Val 7 | **02Trp** **2** | 03Cys 6 | 04Leu 20 | **05Phe** **3**

06Glu 8 | **07Asp** **2** | 08Ala 10 | **09Tyr** **4**

10Ser 5 | 11Arg 5 | **12Met** **2** | **13Ile** **2**

14Lys **2** | **15Asn** **3** | **16Thr** **3** | 17Gln 6 | **18His** **2** | 19Pro 6

6 AAs ← 3/2 ratio → 4 AAs

Figure 13: According to their numbering, distribution in 3/2 ratio of two sets of ten AAs in largest number and ten AAs in smallest number in preproinsulin. Numbers are respective quantity for each AA. See Figure 12 also.

6.2.1. Ten different quantities of AAs

It turns out that the different amounts of amino acids are ten in number in the preproinsulin chain. So these ten different quantities are:

2 – 3 – 4 – 5 – 6 – 7 – 8 – 10 – 12 – 20

We therefore observe that this number is equal to 5*x*, a recurrent value in the organization of the genetic code.

Also, as it appears in Figure 14, 15 AAs, so 5*x* AAs (→ *x* = 3) are in quantities from 2 to 6, i.e. the five smallest quantities and 5 AAs (→ *x* = 1) are in quantities from 7 to 20, i.e. the five largest quantities.

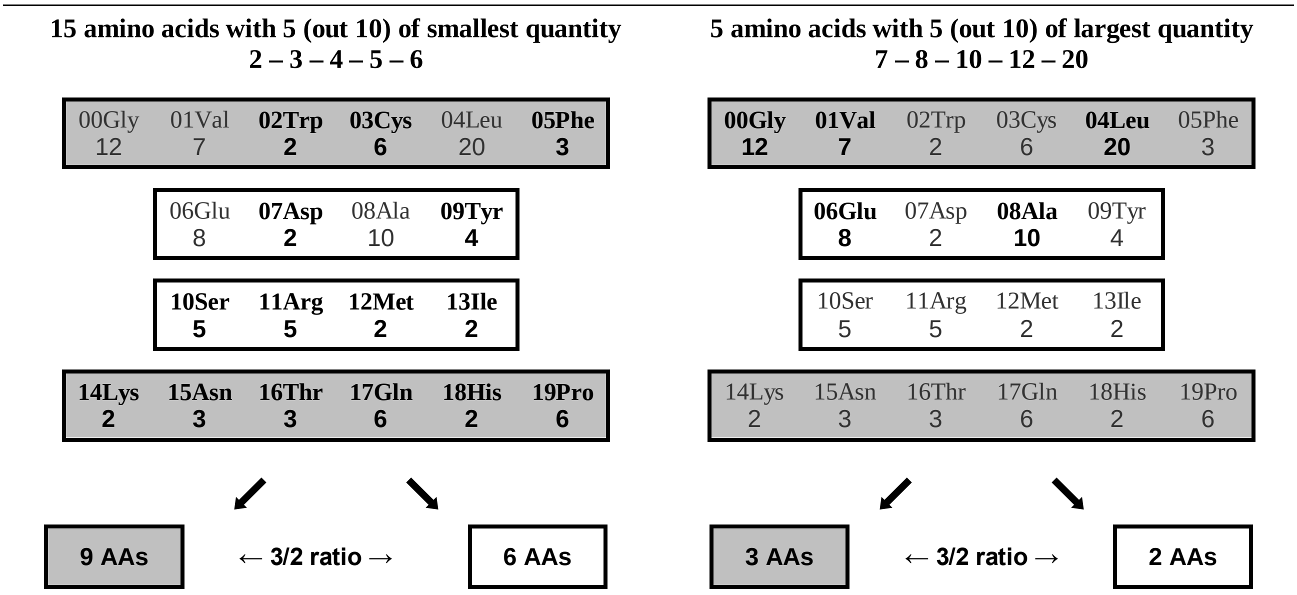


Figure 14: According to their numbering, distribution in 3/2 ratio of two sets of 15 AAs in smallest quantity and 5 AAs in largest quantity in preproinsulin. Numbers are respective quantity for each AA. See Figure 12 also.

As illustrated Figure 14, these two sets of 15 and 5 AAs, which are differentiated according to their degree of abundance, are both organized in the 3/2 ratio in accordance with the two numbering zones qualified as external and internal.

6.2.2. Five largest numbers of AAs

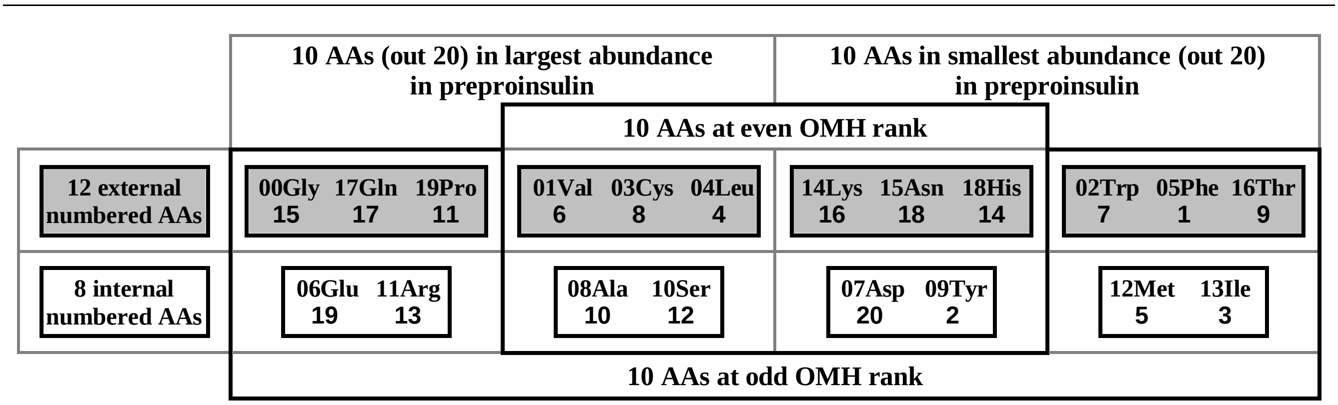
Concerning the set of five AAs present in the greatest quantity in preproinsulin (right part in Figure 14), it should be noted that all are among the ten of first numbering, that is to say among those numbered from 0 to 9. This appears more clearly in the graph of Figure 12. This reinforces even more greatly the idea that the phenomena presented about the distribution of AAs in preproinsulin are not due to chance.

6.3. AAs OMH rank

In the previous paper *"Numbering of the twenty amino acid"* [2], we demonstrated that the OMH index ranks [3] are organized in exact ratios of 3/2 values according to the two AA numbering zones. In this same article[2], we also demonstrated that the parity distinction of these OMH index ranks still generated the same phenomenon. In appendix, it is illustrated in detail these singular arrangements.

6.3.1. AA abundance and OMH index rank transcendence

As it is clearly visible and synthesized in Figure 15, it turns out that these two notions introduced here, that of AA abundance and that of OMH index rank parity transcend each other completely. All this, in accordance with the concept of numbering the twenty amino acids.



	10 AAs (out 20) in largest abundance in preproinsulin		10 AAs in smallest abundance (out 20) in preproinsulin	
		10 AAs at even OMH rank		
12 external numbered AAs	00Gly 17Gln 19Pro 15 17 11	01Val 03Cys 04Leu 6 8 4	14Lys 15Asn 18His 16 18 14	02Trp 05Phe 16Thr 7 1 9
8 internal numbered AAs	06Glu 11Arg 19 13	08Ala 10Ser 10 12	07Asp 09Tyr 20 2	12Met 13Ile 5 3
	10 AAs at odd OMH rank			

Figure 15: Distribution of four AAs subsets in perfect 3/2 ratios according to their numbering, their OMH rank parity and their abundance level in 110-amino acid preproinsulin molecule. See Figures 13 and 16 also. Numbers are OMH index rank, see Appendix.

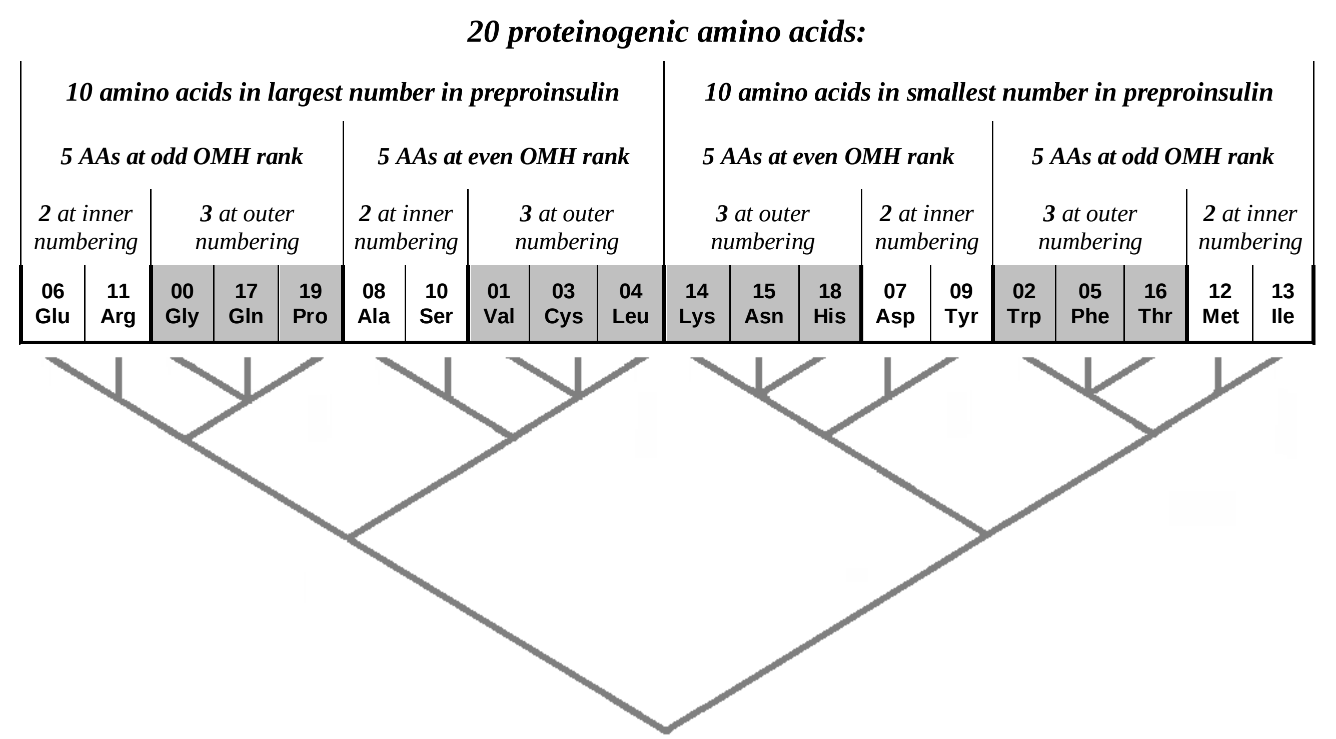
Thus, do we identify four subsets of five amino acids:

- 5 AAs among the 10 in largest number and at odd OMH rank,
- 5 AAs among the 10 in largest number and at even OMH rank,
- 5 AAs among the 10 in smallest number and at even OMH rank,
- 5 AAs among the 10 in smallest number and at odd OMH rank.

Also, systematically, in each of these four subsets, in exact 3/2 ratios, three amino acids are externally numbered and two AAs are internally numbered.

6.3.2. Amino acid symmetric fractal organization

Inside preproinsulin, this remarkable organization of the twenty proteinogenic amino acids turns out in reality to be fractal in nature, as illustrated by the graphic in Figure 16.



20 proteinogenic amino acids:

10 amino acids in largest number in preproinsulin										*10 amino acids in smallest number in preproinsulin*									
5 AAs at odd OMH rank					*5 AAs at even OMH rank*					*5 AAs at even OMH rank*					*5 AAs at odd OMH rank*				
***2** at inner numbering*		***3** at outer numbering*			***2** at inner numbering*		***3** at outer numbering*			***3** at outer numbering*			***2** at inner numbering*		***3** at outer numbering*			***2** at inner numbering*	
06 **Glu**	**11** **Arg**	**00** **Gly**	**17** **Gln**	**19** **Pro**	**08** **Ala**	**10** **Ser**	**01** **Val**	**03** **Cys**	**04** **Leu**	**14** **Lys**	**15** **Asn**	**18** **His**	**07** **Asp**	**09** **Tyr**	**02** **Trp**	**05** **Phe**	**16** **Thr**	**12** **Met**	**13** **Ile**

Figure 16: Symmetric fractal distribution of amino acids in the final 3/2 ratios according to three criteria: their numbering, their OMH rank parity and their abundance level in 110-amino acid preproinsulin molecule. See figure 15 also

This symmetric fractal representation makes better appear how we go from 20 entities to the final ratio 3/2. Indeed, from the twenty entities of the genetic code that are the proteinogenic amino acids, two sets of 10 entities can be isolated according to physico-chemical criteria. These two sets can each be split into two subsets of 5 AAs. Finally each of these subsets can be separated into sets with ultimate numbers of 3 and 2 entities.

We recall, as explained in the Chapter 2 and in appendix, that the numbering of the twenty amino acids is highly dependent on the physico-chemical properties of the four DNA bases.

7. The 3/2 ratio and genetic code organization

As a preliminary conclusion, it seems essential to us to speak about the importance of the arithmetic ratio of value 3/2 in the organization of the genetic code.

The numbering of the twenty proteinogenic amino acids is not the only concept to generate singular arithmetic phenomena opposing the entities of the genetic code in various ratios of value 3/2. In a preview *paper "Genetic code, quantum physics and the 3/2 ratio"* [5], we have revealed in great detail, a multitude of arithmetic arrangements of the components of the genetic code in this 3/2 ratio.

For example, we are drawing attention to the fact that Glycine, which is simply like an amino acid base, has all these various components at 5*x* in number (10 atoms, 40 protons, 75 nucleons, etc.) and that these can be opposed in 3*x* and 2*x* in number. The same phenomena are also observed in the composition of the five atoms constituting the twenty proteinogenic amino acids (Hydrogen, Carbon, Nitrogen, Oxygen and Sulphur) which can also be opposed in various ratios of 3/2 values. Finally, depending on whether or not they are organic, the first ten chemical elements also oppose their nuclear charge number (atomic number) in a ratio of value 3/2. These many observations confirm the main idea of this article that the genetic code, confused AAs and nucleobases, is arithmetically organized according with the ratio 3/2.

Also, various other genetic code investigations from many authors are in connections with the subject of this paper especially about ratio 3/2, symmetry, listing of proteinogenic amino acids or more generally connections between number theory and the genetic code. As example and not limited to, some of these investigations are listed in references [6 to 11].

8. Discusion and conclusion

We have just presented here the very first investigation on a possible connection between the structure of proteins and the concept of numbering of the twenty proteinogenic amino acids.

After recalling this concept of numbering which is dependent on the physico-chemical properties of the four coding nucleobases, we have analyzed the human insulin molecule in its initially translated version. So we studied the 110-amino-acid preproinsulin, the initial product of the translation of insulin mRNA. Investigations are so just on this single chain polypeptide, consisting by a no split sequence of 110 amino acids.

Focusing on the order of appearance of each of the twenty AAs in preproinsulin, we demonstrate that their configuration within the preproinsulin chain is not random but rather dependent on their numbering from 00 to 19. We demonstrated this for both forward and reverse translation sequence order.

We have also demonstrated this both by considering only two sets of first ten and last ten amino acids to appear (out of a total of twenty proteinogens), and individually in agreement with each occurrence rank of these twenty AAs.

We indeed demonstrate that the 110 amino acids constituting preproinsulin are arranged under the constraint of various ratios of value 3/2 in relation to the numbering system, from *00Gly* to *19Pro*, of the twenty proteinogenic AAs. These arithmetic arrangements in ratio 3/2 are exactly of the same nature as those, many numerous, presented in the preview published paper *"Numbering of the twenty proteinogenic amino acids"* [2].

Also, in addition to their order of appearance, the abundance of the different AAs in preproinsulin is still related to this numbering system since we show that 66% (so 2/3) of the 110 components of this molecule are made up of the first ten numbered AAs. The ten other AAs therefore constitute only 33% (so 1/3) of preproinsulin.

In view of all these demonstrations, it is therefore obvious that the new concept of numbering the twenty proteinogenic amino acids is of great use for the study of proteins and the connections that their configuration has with the physico-chemical structure of the four translator nucleobases.

The periodic table of the elements, support for the study of inert matter, is largely articulated with the domain of numbers with for example the numbering of atoms. It is therefore legitimate to also study the constituents of living matter, which are mainly the twenty proteinogenic amino acids and the DNA triplets encoding them, from a numerical angle. Thus it is legitimate the numbering of these twenty amino acids as is that of the chemical elements. This AAs numbering itself being deduced from the numbering of the 64 codons, the other primary constituents of the genetic code.

We therefore conclude by proposing to privilege the study of the constituents of living matter by making extensive use of the numbering system of the twenty proteinogenic amino acids.

Appendix

Here are presented additional explanations to the different chapters of this paper. These are mainly excerpts from the preview published paper *"Numbering of the twenty proteinogenic amino acids"* [2].

A1. About numbering of the 64 genetic code codons

Using a very sophisticated method, Sergey Petoukhov manages to classify the full sixty-four codons set using a binary language (or alphabet, we invite the reader to consult the full article by Sergei Petoukhov [1]). Depending on whether each nucleobase can undergo deamination or not, Sergey Petoukhov assigns them either the value 1 or the value 0 (table Figure A1). Also, depending on whether each nucleobase can undergo depurination or not, Sergey Petoukhov assigns them either the value 0 or the value 1.

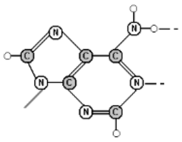
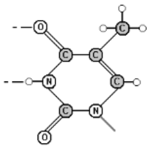
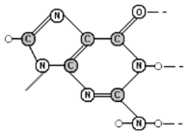
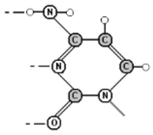
	Adenine	Thymine	Guanine	Cytosine
nucleobases				
Possible deamination: yes = 1 no = 0	1	0	0	1
Possible depurination: yes = 0 no = 1	0	1	0	1

Figure A1: Method of assigning a double binary value to the four DNA nucleobases according to Sergey Petoukhov [1].

This double criterion makes it possible, for each codon, to create a six-digit binary number by juxtaposition of two three-digit numbers as described in Figure A2.

physico-chemical criteria →	**possible deamination** *yes = 1 no = 0*			**possible depurination** *yes = 0 no = 1*		
codon →	A	T	G	A	T	G
binary convert →	*1*	*0*	*0*	*0*	*1*	*0*

ATG
Met
34
100010

ATG = *100010* = 34

Figure A2: Method of assigning a number to codons according to Sergey Petoukhov.
See Figures A1 and 1 also.

A2. About alphanumeric symbol of the 20 proteinogenic amino acids

The table in Figure A3 therefore lists all of the 20 proteinogenic amino acids involved in the mechanism of the universal genetic code. It is therefore described, from the conventional nomenclature, the trivial name, the symbol in 3 letters and the one letter symbol. To this is added, for each AA, its alphanumeric symbol of 5 characters that we propose as a new standardized and official nomenclature.

The 20 proteinogenic amino acids conventional nomenclature:			Alphanumeric symbol proposal
Trivial name	*symbol*	*one letter symbol*	
Glycine	Gly	G	**00Gly**
Valine	Val	V	**01Val**
Tryptophan	Trp	W	**02Trp**
Cysteine	Cys	C	**03Cys**
Leucine	Leu	L	**04Leu**
Phenylalanine	Phe	F	**05Phe**
Glutamic acid	Glu	E	**06Glu**
Aspartic acid	Asp	D	**07Asp**
Alanine	Ala	A	**08Ala**
Tyrosine	Tyr	Y	**09Tyr**
Serine	Ser	S	**10Ser**
Arginine	Arg	R	**11Arg**
Methionine	Met	M	**12Met**
Isoleucine	Ile	I	**13Ile**
Lysine	Lys	K	**14Lys**
Asparagine	Asn	N	**15Asn**
Threonine	Thr	T	**16Thr**
Glutamine	Gln	Q	**17Gln**
Histidine	His	H	**18His**
Proline	Pro	P	**19Pro**

Figure A3: Conventional nomenclature and alphanumeric symbol proposal to the twenty proteinogenic amino acids into 5 characters: 2 digits + 3 letters.

A3. About OMH hydrophobicity index

The OMH index [3] is universally recognized in the study of the twenty proteinogenic amino acids and it is highly unlikely that the next perfect arithmetic arrangements are so by pure chance.

A3.1. OMH hydrophobicity index ranks

According to the exact values of the OMH scale shown in the left part Figure A4, we created an index rank scale ranging from 1 (largest index) to 20 (lowest index) for the twenty amino acids.

The cumulative value of these ranks gives a amount of 126 for the external set of AAs and 84 for the internal one as this is illustrated in right part Figure A4.

Amino acid	OMH scale	rank of OMH hydrophobicity index*
00Gly	-0.67	**15**
01Val	0.91	**6**
02Trp	0.5	**7**
03Cys	0.17	**8**
04Leu	1.22	**4**
05Phe	1.92	**1**
06Glu	-1.22	**19**
07Asp	-1.31	**20**
08Ala	-0.4	**10**
09Tyr	1.67	**2**
10Ser	-0.55	**12**
11Arg	-0.59	**13**
12Met	1.02	**5**
13Ile	1.25	**3**
14Lys	-0.67	**16**
15Asn	-0.92	**18**
16Thr	-0.28	**9**
17Gln	-0.91	**17**
18His	-0.64	**14**
19Pro	-0.49	**11**

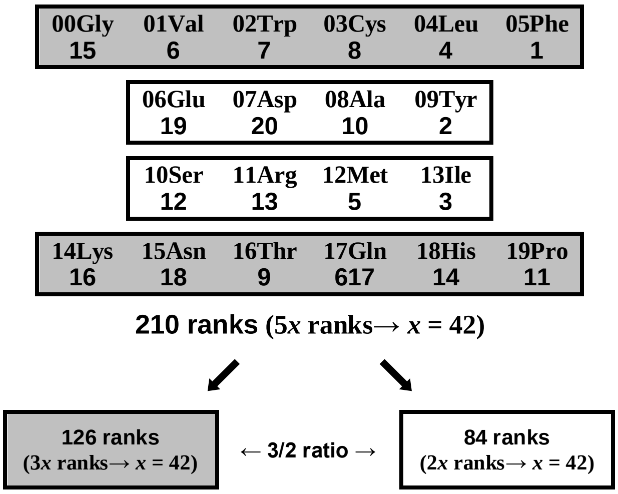


Figure A4: OMH index ranks distribution in exact 3/2 ratio into two external and internal sets of AAs. * rank from the highest index to the lowest index.

A3.2. OMH index ranks parity

Although the distribution of the different OMH index ranks (Figure A4) seems random within the two defined AAs sets of external and internal, the even and odd isolated values continue to generate (Figure A5) a perfect 3/2 ratio between these two sets.

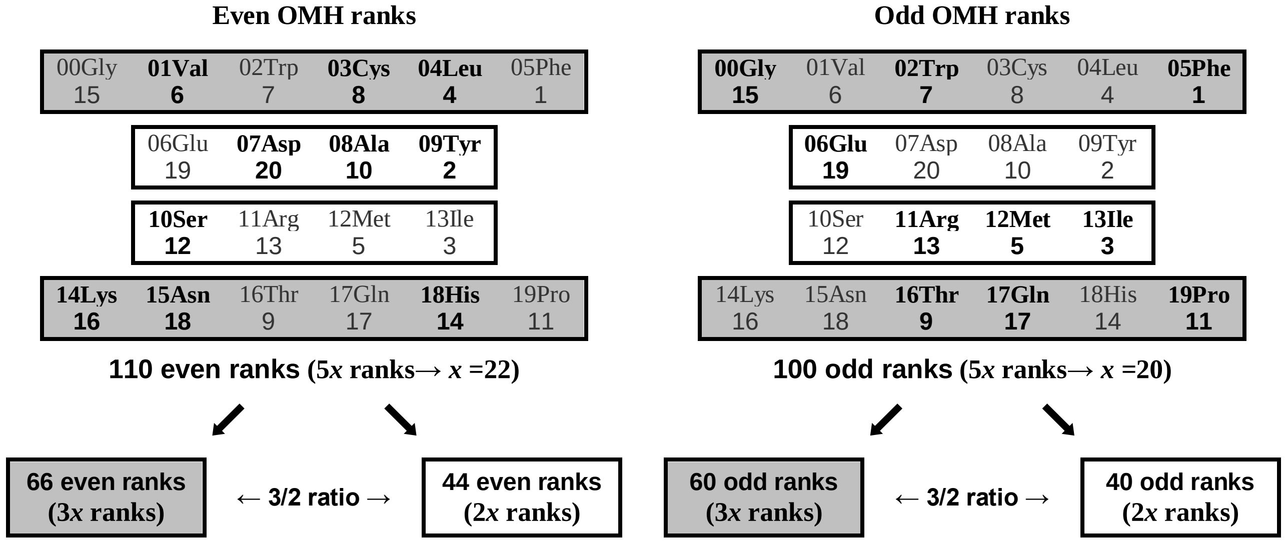


Figure A5: According of the rank parities: OMH index ranks distribution in exact 3/2 ratio into two external and internal sets of AAs.

References

1. S.V. Petoukhov. Genetic Code and the Ancient Chinese Book Of Changes. Symmetry: Culture and Science Vol. 10, Nos. 3-4, p. 211-226. 1999.

2. Jean-Yves Boulay. Numbering of the twenty proteinogenic amino acids: 3/2 ratios inside the genetic code. 2022. https://hal.science/hal-03793573

3. Sweet R.M., Eisenberg D. Optimized matching hydrophobicity (OMH). J. Mol. Biol. 171:479-488. 1983; https://doi.org/10.1016/0022-2836(83)90041-4.

4. Ming Liu. Biosynthesis, structure, and folding of the insulin precursor protein. Volume20, IssueS. 2Supplement: Update on Islet Hormone Production: A Tribute to Donald Steiner. Proceedings of the 19th Servier ‐ IGIS Symposium, St Jean Cap Ferrat, France, 22 ‐ 25 March 2018. https://doi.org/10.1111/dom.13378

5. Jean-Yves Boulay. Genetic code, quantum physics and the 3/2 ratio. 2020. https://hal.science/hal-02902700v4.

6. S.V. Petoukhov.The Bi-periodic Table of Genetic Code and Number of Protons, Foreword of K. V. Frolov, Moscow, 258. 2001.

7. Then, A., Mácha, K., Ibrahim, B. et al. A novel method for achieving an optimal classification of the proteinogenic amino acids. Sci Rep 10, 15321, 2020; https://doi.org/10.1038/s41598-020-72174-5.

8. Wohlin A. Numerical analysis of 3/2-relations in the genetic code and correlations with the basic series of integers 5-0. Biomed Genet Genomics 1, 2016; https://doi.org/10.15761/BGG.1000118.

9. Petoukhov S., He M. Symmetrical Analysis Techniques for Genetic Systems and Bioinformatics: Advanced Patterns and Applications. IGI Global, Hershey, USA, p. 271, 2009; https://doi.org/10.4018/978-1-60566-124-7.

10. Darvas G., Koblyakov A.A., Petoukhov S.V., Stepanyan I.V. Symmetries in molecular-genetic systems and musical harmony. Symmetry Culture and Science, vol. 23, №3-4, p. 343, 2012; http://symmetry.hu/scs_online/SCS_23_3-4.pdf.

11. Petoukhov S.V. Genetic code, musical harmony, stochastic resonance and the Ancient Chinese book of I-Ching. Editor-in-Chief Solar G. p. 160-180, 2022, . Published by: The First Clinic of Acupuncture and Natural Medicine of G.Solar, Ltd, Samorin, Slovak Republic. ISBN 978-80-974284-1-9. EAN 9788097428419. This publication was created thanks to support from the European Union Erasmus program, project number 2020-1-SK01-KA202-078222; https://www.acuclinic.eu/ecompendium/.

Jean-Yves BOULAY independent researcher (without affiliation) – FRANCE –
https://www.researchgate.net/profile/Jean-Yves-Boulay
jean-yvesboulay@orange.fr ORCID: 0000-0001-5636-2375