Journal of
Proteomics & Bioinformatics

Nahalka, J Proteomics Bioinform 2011, 4:8
DOI: 10.4172/jpb.1000184

Research Article Open Access

Quantification of Peptide Bond Types in Human Proteome Indicates How DNA Codons were Assembled at Prebiotic Conditions

Jozef Nahalka[1,2]*

[1]Institute of Chemistry, Center for Glycomics, Slovak Academy of Sciences, Dúbravská cesta 9, SK-84538 Bratislava, Slovak Republic
[2]Institute of Chemistry, Center of excellence for white-green biotechnology, Slovak Academy of Sciences, Trieda Andreja Hlinku 2, SK-94976 Nitra, Slovak Republic

Abstract

[GADV]-protein world hypothesis [1] leaded me to quantification of decapeptides assembled from G, A, V, D in human proteome. The G, A, V, and D amino acids were related to the nucleotides Guanine (g), Cystosine (c), Uracil (u), and Adenine (a). The search revealed agreement with the genetic code. The types of prebiotic peptide bonds represent probably the first selection power that established the base order in the codons. The genetic code underwent three phases of formation, which explain why modern codons have their particular order of nucleotides: the monobase, dibase and the modern phase (tribase). Sequence alignments and 3D structures of aminoacyl-tRNA synthetases confirm the depicted picture of "relatedness" and the picture indicates how "relatedness" is used by aminoacyl-tRNA synthetases for navigation into and within of the C-terminal anticodon-binding domain. The findings presented here illustrate the novel concept of possible translation of the amino acid sequence into a nucleotide sequence that can be in interactive or contrary mode regarding to desired protein-RNA interactions. Hopefully, it could be used in synthetic biology.

Keywords: Origin of genetic code; Peptide bond types; Human proteome; Nucleotide order of codons; Protein-RNA interactions; Aminoacyl-tRNA synthetases

Introduction

It seems that determining the origin of the genetic code is a much more challenging issue then deciphering the code. The state-of-the-art for the origin of the code still remains one of the fundamental unsolved problems. Generally, five basic alternatives exist:

I. The protein world in the beginning - protein coacervates or protein microspheres have developed proto-cells and the genetic code [2-4].

II. The RNA world in the beginning – polymerization of amino acids by poly-RNA [5,6].

III. A parallel evolution of the genetic code and protein synthesis [7].

IV. Transformation of other life systems based on the conditions on the Earth [8].

V. The genetic code and life has been designed by intelligence.

Alternatives IV and V are not within the scope of this manuscript; therefore we will only address the first three. Proteins (I) represented the first target for scientific exploration; however, because RNA (II) was shown to act simultaneously as an informational and catalytic molecule [9], it became the most accepted theory. The scenario in which the genetic code and protein synthesis evolved in parallel (III) resulted as a combination of the two preceding theories. The RNA world theory postulates that self-replicable RNAs (ribozymes) evolved from primordial soup, independently of proteins, and their cellularisation created the first proto-cell, from which followed the evolution of transcription and then translation [6]. Therefore, phylogenetic rooting of transfer-RNAs is a powerful tool that is used to study the early evolution of life and the emergence of the genetic code. For example, Sun and Caetano-Anollés have built phylogenies derived from the sequence and structure of tRNAs and generated timelines of amino acid charging and codon discovery [10]. This rooting showed that charging of Sec, T, S, and L appeared to be ancient, whereas specificities for Q, M, and R amino acids were derived. However, codons for A and P were identified as the most ancient according to their detection. Their study indicates the separate discoveries of amino acid encoding and charging (the genetic and operational code of tRNA). An analysis of a model situation where early tRNAs were not selectively charged with amino acids revealed that it is possible to observe a coded polymerization [11]. Considering this system, a coding regime could have naturally occurred primarily under prebiotic conditions and the operational code could have evolved secondarily [11].

RNA world theory seems be correct; however, the development of amino acids-DNA affinity chromatography [12] over the last decade has revealed that amino acid-nucleotide biomolecular recognition is realistically observed, and it is difficult to imagine any stage of evolution without the influence of peptides. Shimizu showed that single amino acids are able to act as catalysts and anticodonic tri-ribonucleotides corresponding to the amino acid (for example, "uuu" to lysine) act similarly in specific metabolic reaction [13-15]. The idea that amino acids are specifically "related" to nucleotides is not new. Woese probably first suggested the relationship between all amino acids and their codons [16]. He proposed the existence of a "codon-amino acid" logic at some earlier stage in evolution, where he describes amino acid-nucleotide interactions as follows:

"It will not do to refer to the usual "picture" of a molecule garnered from 2-dimensional formulas or their 3-dimensional equivalents, for this purpose, because these give a picture of only one type of interaction

***Corresponding author:** Joaef Nahalka, Institute of Chemistry, Center for Glycomics, Slovak Academy of Sciences, Dúbravská cesta 9, SK-84538 Bratislava, Slovak Republic, E-mail: nahalka@savba.sk

Received July 23, 2011; **Accepted** August 23, 2011; **Published** August 25, 2011

Citation: Nahalka J (2011) Quantification of Peptide Bond Types in Human Proteome Indicates How DNA Codons were Assembled at Prebiotic Conditions. J Proteomics Bioinform 4: 153-159. doi:10.4172/jpb.1000184

of which a molecule is capable (i.e., van der Waals repulsion). "Relatedness" should be defined in terms of a composite of all the interactions of which a molecule is capable and these in a proportion defined by the context in question." [16].

A BLAST search of human taxid: 9606 is used here as the ideal marker of "relatedness", which was not possible at that time because bioinformatics technology had not yet been developed. The search is based on the new idea that the RNA bond formation was driven by peptide bond between two amino acids with each interaction being specific for a nucleotide. The human proteome is chosen for search because it is only one organism and the genome is completely sequenced. Sequences of most concentrated peptides are used for the nucleotide pairing.

Aliberti thought that the affinity of different amino acids to bind to specific nucleic acid triplets is insufficient and he alternatively proposed the idea of repetitive interactions between specific amino acid side chains and specific bases and between polypeptides and polynucleotides [7], which supports the theory of the parallel evolution of the genetic code and protein synthesis (III). CLUSTAL W Multiple Sequence Alignments of aminoacyl-tRNA synthetase (aaRSs) anticodon loops, and the crystal structures of tRNAs that bind to specific aaRSs are used here to prove the definition of "anticodon-amino acid" "relatedness".

Miller [17] experiments with the abiotic synthesis of amino acids under conditions similar to those of the primitive Earth [17,18] and the amino acid composition of carbonaceous meteorites [19] support [GADV]-protein world hypothesis, which is based on pseudo-replication of [GADV]-proteins [1]. Generally, GAVD amino acids are accepted as the first amino acids to be created and they occupy the most thermodynamically stable complementary codons [20]. The second base from these codons is taken here for first nucleotide-amino acid paring (g-G, c-A, a-D and u-V), which resulted in the "monobase codon-amino acid" phase of the genetic code evolution. The intermediate "relatedness" was designed based on above mentionned idea that the RNA bond formation was driven by peptide bond between two amino acids with each interaction being specific for a nucleotide. For example, I found in the human proteome 985 hits for GGGGGGGGGG, 658 hits for GVGVGVGVGV, 496 hits for GAGAGAGAGA and 285 hits for GDGDGDGDGD. It means that glycin preferentially reacted with other glycin in prebiotic GAVD soup; GG peptide bond was the most accessible for guanine nucleotides, so gg intermediate code was reserved for G. The human proteome served as a sample of the proteome, giving order of the peptide sequence hits. The hits depict how amino acids reacted at prebiotic pool. Interestingly, it is in accordance with the amino acid hydrophaty and modern codon table. Finally, the final phase was illustrated analogically, for example, the highest number among GXGXGXGXGX (X=all 20AA) peptides was obtained for GGGGGGGGGG, so glycine connected to di-guanine nucleic acid preferentially reacted with other glycine connected to mono or diguanine nucleotide (gg-G + g-G = ggg-GG) - "tribase codon-amino acid" phase of creation. These three phases were named according to Crick [21] who postulated that the evolution of the genetic code involved three phases:

I. The primitive code

II. The intermediate code

III. The final code, as we have it today [21].

Materials and Methods

The Basic Local Alignment Search Tool (BLAST), the program which compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches, was used to infer evolutionary relationships between RNA and peptide sequences (http://blast.ncbi.nlm.nih.gov/). The peptide sequence has been entered into "blast query sequence" and "non-redundant protein sequences (nr)" and "human (taxid:9606)" has been selected at "choose search set" and program was started using blastp algorithm. Searching was done for decapeptides which included five same peptide bonds for each amino acid including itself, with exception of selenocysteine (U) and pyrrolysine (O), (Table 1). For example, tendency to create stable peptide bond between glycine and alanine was explored by number of

	R	K	D	E	N	Q	H	P	Y	W	S	T	G	A	M	C	F	L	V	I
R	319	248	893	1606	213	321	202	349	282	215	1236	376	471	326	265	260	267	343	250	300
K	260	534	371	878	234	300	350	372	891	298	451	292	276	338	342	256	270	343	290	361
D	929	342	923	572	194	299	315	291	194	356	4619	234	288	226	274	488	290	534	568	231
E	1627	834	571	1305	301	581	269	610	341	311	479	479	482	458	215	413	257	423	273	300
N	177	206	203	297	271	243	255	219	211	294	998	268	388	275	366	311	226	252	278	225
Q	305	244	275	539	267	2064	245	640	310	327	333	321	299	977	230	364	281	472	453	271
H	238	322	333	254	223	258	419	351	266	274	307	419	336	242	238	267	297	271	294	340
P	369	433	323	569	212	594	357	853	292	291	428	1723	788	746	186	321	265	415	753	33
Y	240	738	181	327	225	298	223	299	173	310	318	209	257	277	277	303	275	307	289	245
W	208	278	291	269	251	326	245	293	284	507	252	260	222	301	265	245	270	323	377	270
S	1210	500	5816	493	666	284	281	423	322	232	1723	345	475	818	336	479	222	305	252	358
T	374	301	215	567	256	336	401	1113	244	253	280	1528	373	453	203	277	443	474	364	269
G	428	283	258	511	384	314	304	729	379	246	422	392	950	458	327	369	251	292	733	223
A	302	277	203	505	275	1078	247	649	261	319	532	487	496	985	113	250	245	231	343	264
M	264	351	266	201	339	264	257	200	323	272	346	203	269	212	336	306	312	262	308	291
C	287	295	448	418	282	370	254	328	303	258	472	302	407	253	286	303	320	324	489	464
F	249	274	310	233	218	298	304	286	288	279	256	389	271	246	338	319	334	289	347	344
L	313	341	536	407	211	540	297	442	319	297	393	371	266	286	258	323	276	359	211	196
V	256	232	535	170	227	483	261	806	312	384	306	490	658	358	297	472	292	261	449	238
I	311	255	264	319	231	246	343	383	278	287	392	344	238	257	285	448	452	236	195	278

* Amino acids are ordered by Doolittle's hydropathy index. Blue – hydrophilic, green – neutral and brown – hydrophobic.

Table 1: Number of hits for $(A_{row}A_{column})_5$ peptides in blastp human (taxid:9606).

matches for "GAGAGAGAGA" peptide and for "AGAGAGAGAG" peptide (496 and 458 hits, Table 1).

Aminoacyl-tRNA synthetases complexed with cognate tRNAs were visualized using RCSB Protein Data Bank, view in 3D (Jmol), (http://www.rcsb.org/)

Results and Discussion

Analysis of the genetic code

Glycine, alanine, aspartic acid and valine are accepted as the first amino acids to be created. The order of the nucleotide triphosphates according to their solubility is as follows: ATP, CTP, UTP, and GTP. The most stable RNA triplets are ggc and gcc (28.3 Kcal/mole), which are followed by ggg and ccc (26.8 Kcal/mole) [20]. These characteristics are the basis of the primitive "monobase codon-amino acid" code, where the most abundant amino acid residues (AA), glycine and alanine, are related to g and c. Alanine is first hydrophobic amino acid residue in Kyte-Doolittle Hydrophaty index, so it was evolutionary convenience to be paired with better soluble pyrimidine base to produce "c-A" and the neutral glycine residue was paired with the least soluble purine base to produce "g-G". According to above mentioned theory, the most abundant amino acid residues, glycine and alanine, created the first peptide bonds at prebiotic pool, which catalyzed RNA bonds formation. The AA abundance and RNA stability are reasons for the evolution of "c-A" and "g-G" "relatedness". According to modern codon table, the most hydrophilic residue, aspartic acid, is related to most soluble adenine "a-D", and the most hydrophobic amino acid residue, valine, is most closely related to uracil "u-V". Looking over modern codon table, alanine is related to neutral AA and histidine to hydrophilic AA, so all AA in Kyte-Doolittle Hydrophaty index are shifted to left, and for better understanding, the "g-G", "c-A", "u-V", "a-D" relatedness are colored, blue – hydrophilic, green – neutral and brown – hydrophobic (Table 1). VG and GV are the most conserved of the peptide bonds between V and the other amino acids from the group of G, A, V, and D, so the "dinucleotide" relatedness has been created according to the

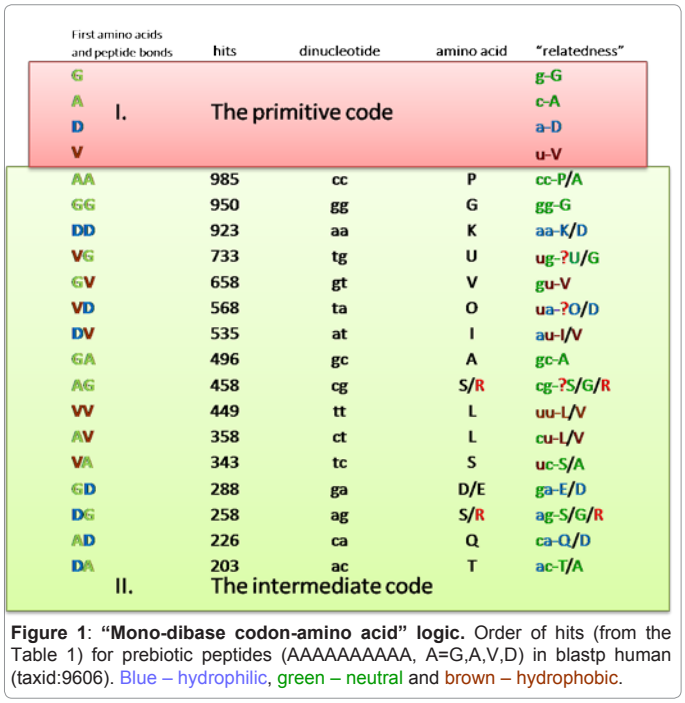

Figure 1: **"Mono-dibase codon-amino acid" logic.** Order of hits (from the Table 1) for prebiotic peptides (AAAAAAAAAA, A=G,A,V,D) in blastp human (taxid:9606). Blue – hydrophilic, green – neutral and brown – hydrophobic.

following: g-G + u-V = gu-VG (chemical equation summary) or gu-V (the relatedness writing), and it is in accordance with modern codons (Figure 1). Based on the valine "relatedness" to "u", valine has been replaced by the more hydrophobic isoleucine in the case of the "au" dinucleotide (au-I), and by the less hydrophobic leucine in the case of uu-L and cu-L(Figure 1). Glycine, which is likely the most abundant amino acid in the prebiotic soup, preferentially forms peptide bonds with itself in among G, A, V, and D, according to the following: 2 x g-G = gg-GG or gg-G, and it is in accordance with modern codons (Figure 1). Alanine and aspartic acid formed the most evolutionary conserved peptide bonds with itself too, but cc-A and aa-D "relatedness" were adopted later by proline (cc-P) and lysine (aa-K) (Figure 1). Lysine is the second most hydrophilic neighbor of aspartic acid in Doolittle's hydropathy index; the substitution is expected despite the opposite charge of K. In contrast, proline is quite far from alanine in Doolittle's hydropathy index, which indicates that the amino acid group "Y, W, S, T", placed between P and A in Doolittle's hydropathy index, occurred later. Generally, P is accepted as the fifth appeared amino acid. So one can hypothesize about [GAVDP] protein world. The other most conserved amino acid peptide bonds created in the G, A, V, and D prebiotic pool for alanine and aspartic acid are GA and GD, so that g-G + c-A = gc-AG and g-G + a-D = ga-DG or gc-A and ga-D formed as we recognize them today in tribase codons. In the case of aspartic acid, more hits are observed for VD bonds and "relatedness" u-V + a-D = ua-DV, so probably ua-O/D (O = Pyrrolysine) was present at "dibase codon-amino acid" code, until it was reserved for tyrosine and stop codons. Aspartic acid, glutamic acid and glutamine have the same hydropathy index (HI = -3.5); there is a high probability that the intermediate "dibase" code did not distinguish between the negatively charged amino acids and shared one codon for aspartic and glutamic acid, ga-E/D, and polar uncharged glutamine assumed a different codon, ca-Q/D. Interestingly, serine and threonine replaced both alanine and glycine in the "dibase-codon" world and uc-S/A, ac-T/A, ag-S/G as we recognize today. Serine is a little more hydrophilic than threonine, so it adopts uc-S and threonine ac-T. Arginine is placed at the hydrophilic end, far from glycine, so it seems that arginine displaced serine or threonine later when the other forces were more important than the hydropathic similarity. Our first analytical step shows that the "dinucleotide-amino acid relatedness" regime naturally occurred primarily under prebiotic conditions, and hydropathic similarity played the main role at this stage of the evolution of the genetic code. The synthesis of prebiotic RNA bonds was influenced by the creation of prebiotic peptides, and everything originated from the "primitive relatedness" g-G, c-A, a-D and u-V. A number of hits of conserved fossil peptide sequences indicated the tendency of the peptide bonds created in the G, A, D, and V pool and when we accept "relatedness" between the creation of peptide and the nucleotide bonds, it is possible to depict the structure of the RNA oligomers and their concentration in the prebiotic pool. One can hypothesize that prebiotic peptides and oligonucleotides participated on the same reaction [13-15], and despite the fact that evolutionary latter they split, proteins were specialized for catalysis and nucleic acids for coding, the relatedness is still conserved. The "dinucleotide-A" relatedness and the prebiotic peptide concentration (number of hits) estimate which and how many amino acids were involved in the reactions. According to quantity of prebiotic peptides and the "dinucleotide-amino acid relatedness", it is possible to order the first fifteen amino acids into the chronology of their appearance as the first coded amino acids. For example the most concentrated cc-A was substituted by cc-P relatedness, that prove first proline appearance from next neutral AA (P,Y,W,S,T). The resulting

order is as follows: "G,A,D,V,P,K,U,O,I,S,L,E,Q,T,R" (see Figure 1). Arginine does not follow hydropathy similarity, it was placed to the end. The final "tribase" codon evolved from the "dibase" codon and, again, peptide bond formation has been involved in its creation. For example, the creation of RD peptide bonds influenced RNA oligonucleotide formation as follows: cg-R + a-D = cga-RD, cg-R + ga-D = cgga-RD, ag-R + a-D = aga-RD, ag-R + ga-D = agga-RD. The process is illustrated in Figure 2, where it is possible to see that aspartic acid, serine, glutamine, threonine, proline, arginine, and glutamic acid have been preferentially involved. It is possible to estimate from the reactivity between amino acids that "gau" is the first tribase codon for aspartic acid, "uca, ucg" are the first tribase codons for serine, "acc" is the first tribase codon for threonine, "cgg, agg" are the first tribase codons for arginine and "gaa, gag" are the tribase codons for glutamic acid. With careful attention, one basic difference between Figures 1 and 2 can be detected. C-terminal amino acids are related to the intermediate code formation; this is illustrated by the GV peptide bond in which g-G + u-V = gu-VG or gu-V (Figure 1), which is contrary to the final code where the N-terminal amino acid is involved, and by the DS peptide bond, in which ga-D + uc-S = gauc-DS. This fact

Peptide bonds	hits	"tribase codon-amino acid relatedness"			
DS	5816	gauc-DS	auc-DS/I	gaag-DS/E	
SD	4619	uca-SD	ucga-SD	aga-SD/R	agga-SD/R
QQ	2064	caca-QQ			
SS	1723	ucuc-SS	agag-SS/R		
TP	1723	accc-TP			
RE	1627	cgga-RE	agga-RE		
ER	1606	gaag-ER	gacg-ER/D		
TT	1528	acac-TT			
EE	1305	gaga-EE			
SR	1236	uccg-SR	ucag-SR	agcg-SR	agag-SR/R
RS	1210	cguc-RS	cgag-RS	agag-RS	aguc-RS/S
PT	1113	ccac-PT			
QA	1078	cagc-QA	cac-QA/H		
SN	998	ucaa-SN			
AA	985	gcc-AA	cgc-AA/R		
AQ	977	gcca-AQ	cca-AQ/P		
GG	950	gggg-GG			
RD	929	cga-RD	cgga-RD	aga-RD	agga-RD
DD	923	gaa-DD/E	gaga-DD/E	aga-DD/R	
DR	893	gacg-DR	gaag-DR/E	acg-DR/T	aag-DR/K
YK	891	Y			
EK	878	gaaa-EK			
PP	853	cccc-PP			
KE	834	aaga-KE			
AS	818	gcuc-AS	gcag-AS	cuc-AS/L	cag-AS/Q
PV	806	ccu-PV	ccgU-PV		
GP	788	ggcc-GP	gcc-GP/A		
VP	753	gucc-VP	ucc-VP/S		
AP	746	gccc-AP	ccc-AP/P		
KY	738	Y			
VG	733	gug-VG	gugg-VG	ugg-VG/W	
PG	729	ccg-PG	ccgg-PG		

III. The final code, as we have it today.

Figure 2: **"Tribase codon-amino acid" logic.** Order of hits (from the Table 1) for prebiotic peptides (AAAAAAAAAA, A=amino acid) in blastp human (taxid:9606). Blue – hydrophilic, green – neutral and brown – hydrophobic; red – "tribase codon-amino acid".

Figure 3: ***T. thermophilus* EaRS•tRNA(E) complex.** L_{442} and opposite L_{447} recognize "$c_{534}u_{535}$". RCSB Protein Data Bank, 1N78.

simply shows that it was created in the latter stage of the evolution by different mechanism and the triplet code has two letters related to the N-terminal amino acid and only one letter for the C-terminal amino acid, therefore gau is related to D and not to S or V.

Our third analytical step will include a study of the anticodon reading by aminoacyl-tRNA synthetases. Figure 3 shows a solved crystal structure of glutamyl-tRNA synthetase complexed with tRNA(E). The glutamyl-tRNA synthetase is a class-I enzyme that can catalyze the initial amino acid activation reaction only in the presence of the cognate tRNA [22]. The loop of complexed tRNA(E) contains $c_{534}u_{535}c_{536}$ anticodon nucleotides, and the anticodon-binding domains of EaRS contains "A_{431}QPLRAALTG SL_{442}ETPGL$_{447}$FEI LALLGKERAL RRLERALA$_{468}$" sequence. Application of the "dibase codon-glutamic acid" logic explains how the synthetase recognizes the anticodon ("cuc","uuc"): SL_{442} recognizes uc cu/uu, and L_{447}F recognizes cu/uu uu; serine provides the read-through for the second and third anticodon letters (uc), and leucine, together with phenylalanine, provides the read-through for the first and second anticodone letters (cu/uu-L, uu-F). SL_{442} and L_{447}F reading sequences are followed by E, which is a charged amino acid that has opposite "relatedness" ga-E. Figure 3 clearly shows that the $c_{534}u_{535}c_{536}$ anticodon bases are inserted into a level between the SL_{442} loop and the L_{447}F α-helix. Figure 4 shows the crystal structure of cysteinyl-tRNA synthetase bound to tRNA(C). Two currently known codons for cysteine are "ugc" and "ugu", so the first two letters of the anticodon are "gc" and "ac" and "dibase codon-amino acid relatedness" translation will lead us to look for A and T. As we know from above results, the triplet code has been evolved from the duplet code, so searching for the last two letters of the anticodon "ca-Q" will be more important. The anticodon-binding domain of CaRS contains "Q_{421}QRLDARKAK DWAAADAARD RLNEMGIVLE DGPQGTTWRRK$_{461}$" sequence. The crystal structure of the complex revealed that the anticodon "$g_{34}c_{35}a_{36}$" is identified by the Q_{454}GTT sequence. Using the "relatedness" translation, g-G and ca-Q are responsible for the "gca" anticodon read-through and ac-T with ca-Q are responsible for the "aca" anticodon read-through. A sequence alignment of the most diverse CaRS group members (see

Figure 5) shows that the translation process will be more variable. The "relatedness" g-G - ac-T is much more conserved then g-G - ca-Q or ac-T - ca-Q, so the GT dipeptide can read "gca", and TT/TQ can read "aca". Interestingly, the substitution of TT/TQ by TS/SK is still functional, which indicates the importance of the second letter in the "dibase codon-amino acid relatedness", and confirms a preexistence of "monobase codon-amino acid relatedness" (see Figure 5). The GSK amino acid sequence translated based on "monobase codon-amino acid relatedness" means g-G, (t)c-S, (a)a-K or so "gca" anticodon. Another interesting point is that tryptophan is conserved next to the "reading sequence" and it has the same "double letter codon ug" as cysteine, which is the charged amino acid to Ca-RS. When we return to the expected results from the "dibase codon-amino acid relatedness" translation, the expected alanine residue was not identified in the short "reading" sequence, but the CaRS-tRNA complex structure shows that alanine is used to pull down the U-turn of the tRNA anticodon into the cavity of the CaRS C-terminal α/β domain (Figure 4). The "dibase codon-amino acid" translation for A, R and Q is "gc", "cg" and "ca", the "monobase codon-amino acid" translation for A, R, K and D is "c", "g", "a" and "a", and the cavity is full of these amino acids: "Q_{421}QRLDARKAKDWAAADAARD". It seems that the indirect orientation of "$g_{34}c_{35}$" is controlled mainly by alanine and "$c_{35}a_{36}$" mainly by glutamine (Figure 4).

Prediction of anticodon reading sequences inside the aminoacyl-tRNA synthetases by "codon-amino acid" relatedness

As described above, the one "di-base relatedness" can be substituted with two "mono-base relatedness" what complicates the potential modelling of the protein ↔ nucleic acid interaction process, but it can still be helpful for the identification of the protein sequences for nucleic acid "reading". The *Thermotoga maritima* tryptophanyl-tRNA synthetase (WaRS) has only one codon, and the enzyme is relatively small (328AA). The C-terminal domain follows the sequence: "E_{205}ISEKELEQTILRMMTDP$_{222}$ARVRRSD-P_{230}GNP$_{233}$ENCP_{237}VW$_{239}$KYH_{242}QAFDISEEESKWVWEGCTTA-

Figure 4: ***Escherichia coli*** **CaRS•tRNA(C) complex.** Using the "relatedness" translation, ac-T and ca-Q are responsible for the "$c_{35}a_{36}$" read-through and gc-A and cg-R are responsible for the "$g_{34}c_{35}$" read-through. RCSB Protein Data Bank, 1U0B.

```
1U0B          453 PQGTTWRR 460
gi 89093601   452 REGTSWFR 459
gi 152996132  452 REGTTWTR 459
gi 227372957  444 PEGSKWRL 451
gi 94499488   452 REGTTWVK 459
gi 83644974   459 REGTSWQR 466
gi 149376716  428 REGTSWRR 435
gi 146328731  448 ATGTQWYY 455
gi 30248074   457 PQGTTWRR 464
gi 91774855   447 PQGTSWRR 454
```

Figure 5: **Part sequence of anticodon-binding domain of cysteinyl tRNA synthetases (CaRSs).** Sequence Alignment of the most diverse group members.

SIGCVDCKKLLLKNMKRKLAP$_{283}$IWENFRKIDEDP$_{295}$HYVDD-VIMEGTKKAREVAAKTMEEVRRAMNLMF$_{328}$". The "dibase codon-amino acid relatedness" translation of the "cca" anticodon will lead us to look for ca-Q, ca-H, ac-T and cc-P, or the "monobase codon-amino acid relatedness" translation of "cca" provides c-A, c-S, c-T and aa-K, aa-N, a-D, a-E letters. This is quite a lot of combinations; however, it helps that the "reading sequence" is usually close to the charging amino acid because it has opposite relatedness as its anticodon, which is tryptophan or the cysteine with the same "double-codon" can be considered. The potential amino acids are coloured in the above C-terminal domain sequence and W_{239} with P_{233}, P_{237} on one side and H_{242}QA sequence on other side appear to be potential "cca" reading amino acids. The whole C-terminal domain has only two glutamines and two histidines, so H_{242}QA and P_{237} appear to be the most important. Determining the sequence alignment (see Figure 6) indicates that P_{233} and H_{242} are the most important anticodon-reading amino acids. This can be considered to be a result of "relatedness" analysis of the primary sequence. The crystal structure of WaRS is accessible (Figure 7), so we can illustrate anticodon reading, which can be accomplished without complex crystallization with cognate tRNA. The "cca" anticodon nucleotides are navigated by oppositely oriented proline residues, cc-P_{233} and cc-P_{237}, and by aa-N_{235} placed up the prolines, into the anticodon domain cavity, where it is oriented by ca-H_{242} on one side and by (t)c-S_{254} plus aa-K_{253} on the opposite side. The opposite oriented tryptophan residues, W_{239} and W_{256} are also internally exposed into the cavity to provide smooth movement within; they have the exact contrary "relatedness", ugg-W (see Figure 7). C_{236}, C_{259}, C_{266}, C_{239} cysteine residues with contrary "dinucleotide-AA relatedness" do not interrupt the anticodon reading but hold four iron atoms inside the cavity by their S atoms. Perhaps it only stabilises the cavity structure, because they showed WaRS is hyperthermophilic, or the irons fix the anticodon position [23].

Summary

Various possible scenarios, which lead to the emergence of the genetic code, have been indicated in the introduction. However the presented results imply that the strongly supported RNA world theory is not clear. The results support researchers that suggest that a full-blown RNA world in fact never existed [24]. Instead, very early coevolution of nucleotides and amino acids evolved to produce a "ribonucleopeptide" world where ribonucleotides were larger, because the RNA nucleotides were more stable than peptides (for example Prokaryotic ribosomes are composed of 65% ribosomal RNA and 35% ribosomal proteins), but all rules have been already designed based on the GAVD time period when interactions of emerged amino acids with nucleotides leaded to short RNA synthesis. The bioinformatic evidence presented here, that the first nucleotide bonds were connected to the

```
rs_YP_003146066_[YP_003146066]      IKTMPTDPARVRRTDPGTPEKCPVWEFHKIYSSD---EVKDWVQKGCTTA
rs_YP_343076_[YP_343076]_trypt      LRTMPTDPARVRRTDPGDPEKCPVWQFHRVYSDD---EVKEWVQKGCRTA
tr_D1KE93_9GAMM_[D1KE93]_SubNa      IKRMPTDPARVKLTDSGNPEKCPVWQLHKVYSDE---QTQDWVVDGCTKA
rs_ZP_01451501_[ZP_01451501]_t      VKTMPTDPARVRRDDPGTPEACPVWDFHKVYSTE---AEREWVQDGCTTA
gpu_CP002416_1499_[CP002416]_t      VSSMITDPARIRKDDPGHPEVCTVFSFHKVFNEN---EVPEIEQH-CRGG
query                               ILRMMTDPARVRRSDPGNPENCPVWKYHQAFDIS--EEESKWVWEGCTTA
sp_SYW_THEMA_[Q9WYW2]_RecName_      ILRMMTDPARVRRSDPGNPENCPVWKYHQAFDIS--EEESKWVWEGCTTA
rs_YP_001410396_[YP_001410396]      VLPMVTDPARKRRTDPGNPENCPVWDYHKAFGTADNEEEKQWVFEGCTQA
rs_YP_001567417_[YP_001567417]      ILPMMTDPARIRRTDPGNPEKCPVWDYHKAFTKS--QDEKDWVWNGCTTA
rs_YP_003264676_[YP_003264676]      LAQAVTDPKRETREDPGNPDDCNLYTLHTFFSSE---DEQQWVRQGCTTA
                                                *: *     *  :         .   . *  .
```

Figure 6: **Part sequence of anticodon-binding domain of tryptophanyl-tRNA synthetases (WatRS).**

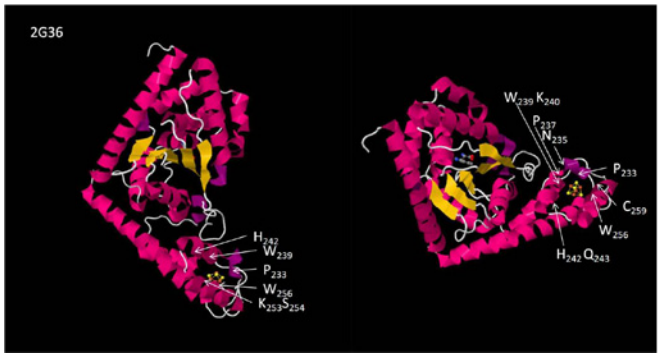

Figure 7: ***Thermotoga maritima*** **tryptophanyl-tRNA synthetase**. Using the "relatedness" translation of "cca" anticodon, cc-P_{233} and cc-P_{237} are responsible for the "cc" read-through; ca-H_{242} and (t)c-S_{254} plus aa-K_{253} are responsible for the "ca" read-through; aa-N_{235} completes the prolines to "cca" read-through; and the codon relatedness ugg-W does not interrupt the read-through. RCSB Protein Data Bank, 2G36.

peptide bonds and that distribution of prebiotic peptides still correlates with DNA code, confirms the "relatedness" picture that has been designed by Woese [16] and other previous theories [24]. Lehmann and coworkers [11] simulated the polymerization of amino acids along RNA templates and concluded that a system of four codons (gnc, n= g, c, u, a) and four amino acids (G, A, V, D) could be a plausible original genetic code. Independently, Higgs [25] proposed four column triplet code (gnn, n= g, c, u, a) and four amino acids (G, A, V, D) as the earliest genetic code. The presented work changes the scenario, four amino acids (G, A, V, D) "related" to four nucleotides (g, c, u, a) created the first peptides and the first RNA oligomers were synthesized along the peptide templates, [GADV]-protein world [1] or better [g-G,c-A,u-V,a-D]-ribonucleopeptide world. It is probable that ribonucleopetides catalyzed first metabolic reactions and latter they split, proteins were specialized for catalysis (20 AA - better variability) and nucleic acids for coding (DNA - better stability).

As have already Woese, [26] pointed out, similar polar properties are required for amino acids coded by the same base in the second codon position, hydropathic similarity played the main role at the evolution of the intermediate (dibase) code (Figure 1, Table 1). The third letter adition to triplet code creation, in difference, is independent on hydropathic similarity, but still follow the peptide bonds formation in the enviroment. Lehmann and Libchaber, [27] analyzed the anticodon-codon association within the ribosome decoding center. Interestingly, they pointed out stability of the base pair at the second position of the anticodon and fact that the canonical U-turn of the tRNA anticodon loop contributes to the degeneracy at the third position of the codon, what is in accordance with the presented theory of the genetic code evolution.

Despite many structural, kinetic and thermodynamic studies that have been published [28,29], all the forces that relate nucleic and amino acids are unknown, so it is still not sufficient to consider all of the components together, such as energy, frequency and propensities of amino acid-nucleotide interactions, water-mediated interaction, or other physicochemical and stereochemical forces. Based on Woese's work [16], when we do not obtain information from "all the interactions of which a molecule is capable", we cannot predict or calculate universal recognition rules. A complete equation is not achievable today, but it is possible search a correlation by bioinformatical tools and efficiently evaluate empirical data from structures in PDB [30] or from amino acid-nucleotide affinity chromatography [12], chip-seq data [31], or data from other sources.

Acknowledgments

This contribution is the result of the project implementation: Centre of excellence for white-green biotechnology, ITMS 26220120054, supported by the Research & Development Operational Programme funded by the ERDF.

References

1. Ikehara K (2005) Possible steps to the emergence of life: The [GADV]-protein world hypothesis. Chem Rec 5: 107-118.
2. Oparin AI, Gladilin KL (1980) Evolution of self-assembly of probionts. BioSystems 12: 133-145.
3. Orlovskiĭ AF, Gladilin KL, Vorontsova VI, Kirpotin DB, Oparin AI (1977) Stabilization of coacervate drops by orthophosphate and nucleotides. Dokl Akad Nauk SSSR 232: 236-239.
4. Muller-Herold U, Nickel G (1994) The stability of proteinoid microspheres. BioSystems 33: 215-220.
5. Gilbert W (1986) The RNA world. Nature 319: 618.
6. Hirao I, Ellington AD (1995) Re-creating the RNA world. Current Biology 5: 1017-1022.
7. Aliberti S (1997) The origin of the genetic code and protein synthesis. J Mol Evol 45: 352-358.
8. Cairns-Smith AG, Hall AJ, Russell MJ (1992) Chapter 9 mineral theories of the origin of life and an iron sulfide example. Orig Lif Evol Biosp 22: 161-180.
9. Jaeger L (1997) The new world of ribozymes. Curr Opin Struct Biol 7: 324-335.
10. Sun FJ, Caetano-Anollés G (2008) Evolutionary patterns in the sequence and structure of transfer RNA: A window into early translation and the genetic code. PLoS ONE 3: e2799.
11. Lehmann J, Cibils M, Libchaber A (2009) Emergence of a code in the polymerization of amino acids along RNA templates. PLoS ONE 4: e5773.
12. Sousa F, Cruz C, Queiroz JA (2010) Amino acids-nucleotides biomolecular recognition: From biological occurrence to affinity chromatography. J Mol Recognit 23: 505-518.

13. Shimizu M (2007) Amino acid and anticodon enhance metabolic reaction rates weakly but specifically: Genetic code world. J Phys Soc Japan 76: 053801.

14. Shimizu M (1995) Specific aminoacylation of C4N hairpin RNAs with the cognate aminoacyl-adenylates in the presence of a dipeptide: Origin of the genetic code. J Biochem 117: 23-26.

15. Shimizu M (2004) Histidine and its anticodon GpUpG are similar metabolic reaction rate enhancers: Molecular origin of the genetic code. J Phys Soc Japan 73: 323-326.

16. Woese CR (1965) Order in the genetic code. Proc Natl Acad Sci U S A 54: 71-75.

17. Miller SL (1953) A production of amino acids under possible primitive earth conditions. Science 117: 528-529.

18. Johnson AP, Cleaves HJ, Dworkin JP, Glavin DP, Lazcano A, et al. (2008) The miller volcanic spark discharge experiment. Science 322: 404.

19. Pizzarello S, Shock E (2010) The organic composition of carbonaceous meteorites: The evolutionary story ahead of biochemistry. Cold Spring Harbor Perspectives in Biology 2: a002105.

20. Trifonov EN (2004) The triplet code from first principles. J Biomol Struct Dyn 22: 1-11.

21. Crick FHC (1967) Origin of the genetic code. Nature 213: 119.

22. Deutscher MP (1967) Rat liver glutamyl ribonucleic acid synthetase. II. further properties and anomalous pyrophosphate exchange. J Bio Chem 242: 1132-1139.

23. Han GW, Yang X, McMullan D, Chong YE, Krishna SS, et al. (2010) Structure of a tryptophanyl-tRNA synthetase containing an iron-sulfur cluster. Acta Crystallogr Sec F: Struct Biol Cryst Commun 66: 1326-1334.

24. Di Giulio M (2005) The origin of the genetic code: Theories and their relationships, a review. BioSystems 80: 175-184.

25. Higgs PG, (2009) A four-column theory for the origin of the genetic code: Tracing the evolutionary pathways that gave rise to an optimized code. Biol Direct 4: 16.

26. Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger WC (1966) On the fundamental nature and evolution of the genetic code. Cold Spring Harbor Symp Quant Biol 31: 723-736.

27. Lehmann J, Libchaber A (2008) Degeneracy of the genetic code and stability of the base pair at the second position of the anticodon. RNA 14: 1264-1269.

28. Morozov AV, Havranek JJ, Baker D, Siggia ED (2005) Protein-DNA binding specificity predictions with structural models. Nucleic Acids Res 33: 5781-5798.

29. Luscombe NM, Laskowski RA, Thornton JM (2001) Amino acid-base interactions: A three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res 29: 2860-2874.

30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235-242.

31. Mokry M, Hatzis P, de Bruijn E, Koster J, Versteeg R, et al. (2010) Efficient double fragmentation ChiP-seq provides nucleotide resolution protein-DNA binding profiles. PLoS ONE 5: e15092.