Exceptional error minimization in putative primordial genetic codes

Artem Novozhilov

Applied mathematics department at
Moscow State University of Communication Means

MATHEMATICS. COMPUTING. EDUCATION

Puschino, January 24–29

The standard genetic code:

The standard genetic code table has a distinctly non-random structure, with similar amino acids are often encoded by codon series that differ by a single nucleotide substitution, typically in the third or the first position of the codon.

UUU [F] Phe	UCU [S] Ser	UAU [Y] Tyr	UGU [C] Cys
UUC [F] Phe	UCC [S] Ser	UAC [Y] Tyr	UGC [C] Cys
UUA [L] Leu	UCA [S] Ser	UAA [ ] Ter	UGA [ ] Ter
UUG [L] Leu	UCG [S] Ser	UAG [ ] Ter	UGG [W] Trp
CUU [L] Leu	CCU [P] Pro	CAU [H] His	CGU [R] Arg
CUC [L] Leu	CCC [P] Pro	CAC [H] His	CGC [R] Arg
CUA [L] Leu	CCA [P] Pro	CAA [Q] Gln	CGA [R] Arg
CUG [L] Leu	CCG [P] Pro	CAG [Q] Gln	CGG [R] Arg
AUU [I] Ile	ACU [T] Thr	AAU [N] Asn	AGU [S] Ser
AUC [I] Ile	ACC [T] Thr	AAC [N] Asn	AGC [S] Ser
AUA [I] Ile	ACA [T] Thr	AAA [K] Lys	AGA [R] Arg
AUG [M] Met	ACG [T] Thr	AAG [K] Lys	AGG [R] Arg
GUU [V] Val	GCU [A] Ala	GAU [D] Asp	GGU [G] Gly
GUC [V] Val	GCC [A] Ala	GAC [D] Asp	GGC [G] Gly
GUA [V] Val	GCA [A] Ala	GAA [E] Glu	GGA [G] Gly
GUG [V] Val	GCG [A] Ala	GAG [E] Glu	GGG [G] Gly

The standard genetic code
The codon series are shaded in accordance with the Polar Requirement Scale values (Woese, Dugre et al. 1966).

Three basic theories of the code nature, origin, and evolution:

- **Stereochemical theory:** codon assignments for particular amino acids are determined by physicochemical affinities;
- **Coevolution theory:** the structure of the standard code reflects the pathways of amino acid biosynthesis;
- **Adaptive theory:** the structure of the genetic code was shaped under selective forces that minimize the effect of errors (point mutations and translational misreadings).

Knight RD, Freeland SJ, Landweber LF: *J Biol Chem*, 1999, **273**:23019–23025
Koonin EV, Novozhilov AS: *IUBMB Life*, 2009, 61(2):99–111

Three basic theories of the code nature, origin, and evolution:

- **Stereochemical theory:** codon assignments for particular amino acids are determined by physicochemical affinities;
- **Coevolution theory:** the structure of the standard code reflects the pathways of amino acid biosynthesis;
- **Adaptive theory:** the structure of the genetic code was shaped under selective forces that minimize the effect of errors (point mutations and translational misreadings).

Knight RD, Freeland SJ, Landweber LF: *J Biol Chem*, 1999, **273**:23019–23025
Koonin EV, Novozhilov AS: *IUBMB Life*, 2009, 61(2):99–111

- **The frozen accident:** the allocation of amino acids is mainly accidental.

The stereochemical theory:

'...Thus the question arises about the way four-digital numbers can be translated into such 'words'.
It seems to me that such translation procedure can be easily established by considering the 'key-and-lock' relation between various amino acids and the rhomb-shaped 'holes' formed by various nucleotides in the deoxyribonucleic acid chain.'

Gamov G: *Nature*, 1954, **173**:318

George Gamow (1904–1968)

The frozen accident and the coevolution theory:

'...The evolution of the code has the property that it could produce a code in which the actual allocation of amino acid to codons is mainly accidental and yet related amino acids would be expected to have related codons.'

Crick F: *J Mol Biol*, 1968, **38**:367–379

'...The structure of the codon system is primarily an imprint of the prebiotic pathways of amino-acid formation, which remain recognizable in the enzymic pathways of amino-acid biosynthesis.'

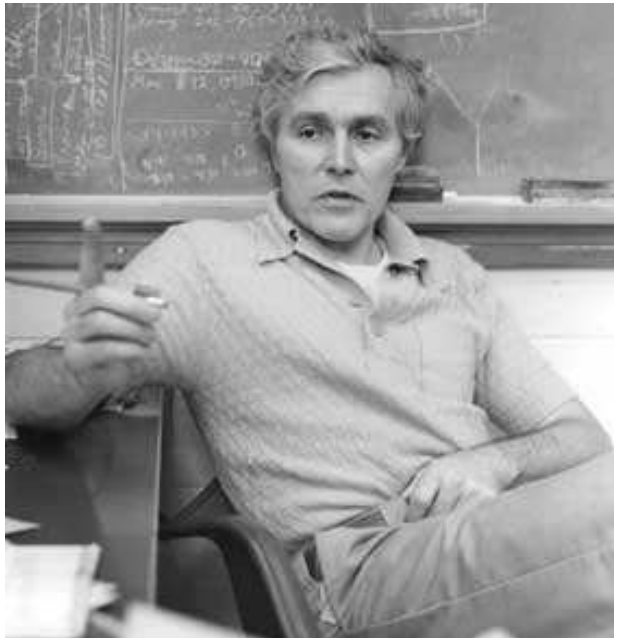
Wong J: *Proc Nat Acad Sci*, 1975, **72**(5):1909–1912

Francis Crick (1916-2004)

The adaptive theory:

'In brief, the codon catalogue which we observe today is considered to have arisen through a series of evolutionary steps which served gradually to reduce an initial inherent high error rate in the translation processes of the primitive cell.'

Woese C.: *Proc Nat Acad Sci*, 1965, **54**:1546–1552



Carl Woese (born 1928)

The modern state of the basic theories:

- **Stereochemical theory:** 'the escaped triplet theory', there is experimental evidence that short RNA molecules selected from random sequence mixtures by amino acid-binding were significantly enriched with cognate triples for the respective amino acids (M. Yarus, R. Knight);
- **Coevolution theory:** elaborated scenarios of the genetic code evolution by M. Di Giulio;
- **Adaptive theory:** extensive statistical support from comparison of the standard code with random alternatives ('the genetic code is one in a million', L. Hurst, S. Freeland, D. Ardell).

Synthesis?

'At first, in the RNA world, stereochemical interactions would have largely determined the correspondence between certain RNA-sequence tags and amino acids. [...] As amino acid and peptide cofactors, and eventually catalysts, became more prevalent at the onset of the RNA–protein world, coevolution of the code and the amino acid set might have led to expansion of the code on the basis of metabolic relatedness. This expansion would also have preserved the rules initially established by stereochemical interactions in order to continue making the original templated protein or proteins. Finally, after the evolution of the mRNA–tRNA–aminoacyl-tRNA-synthetase system removed direct interaction between amino acids and codons, codon swapping in different lineages would have permitted some degree of code optimization by codon reassignment.'

Knight R: *PhD Thesis*, 2001

Technical details:

Code is a mapping

$$a\colon C \to A,$$

where C is the set of codons, and A is the set of amino acids. The cost function for a given code can be written as

$$\varphi(a(c)) = \sum_{c'} \sum_{c} p(c'|c) d(a(c), a(c')),$$

where $p(c'|c)$ gives the probability of misreading codon c as codon c', and $d(a(c), a(c'))$ defines the cost of replacing amino acid $a(c)$ with amino acid $a(c')$ (I use Polar Requirement Scale).

Technical details:

Random code algorithms: Total number $\approx 1.5 \times 10^{84}$

Classical algorithm: permutations of amino acid assignments keeping the block structure of the standard code intact ($20! \approx 2.4 \times 10^{18}$ codes; changes the number of synonymous codons).

New algorithm: assignments of 8 amino acids that are encoded by 4-codon series are distributed randomly among 14 blocks; assignments of the remaining amino acids that are encoded by 2-codon series are distributed randomly among the remaining half-blocks ($\approx 10^{19}$; retains the degeneracy pattern of codons).

UUU [F] Phe	UCU [S] Ser	UAU [Y] Tyr	UGU [C] Cys
UUC [F] Phe	UCC [S] Ser	UAC [Y] Tyr	UGC [C] Cys
UUA [L] Leu	UCA [S] Ser	UAA [∏] Ter	UGA [∏] Ter
UUG [L] Leu	UCG [S] Ser	UAG [∏] Ter	UGG [W] Trp
CUU [L] Leu	CCU [P] Pro	CAU [H] His	CGU [R] Arg
CUC [L] Leu	CCC [P] Pro	CAC [H] His	CGC [R] Arg
CUA [L] Leu	CCA [P] Pro	CAA [Q] Gln	CGA [R] Arg
CUG [L] Leu	CCG [P] Pro	CAG [Q] Gln	CGG [R] Arg
AUU [I] Ile	ACU [T] Thr	AAU [N] Asn	AGU [S] Ser
AUC [I] Ile	ACC [T] Thr	AAC [N] Asn	AGC [S] Ser
AUA [I] Ile	ACA [T] Thr	AAA [K] Lys	AGA [R] Arg
AUG [M] Met	ACG [T] Thr	AAG [K] Lys	AGG [R] Arg
GUU [V] Val	GCU [A] Ala	GAU [D] Asp	GGU [G] Gly
GUC [V] Val	GCC [A] Ala	GAC [D] Asp	GGC [G] Gly
GUA [V] Val	GCA [A] Ala	GAA [E] Glu	GGA [G] Gly
GUG [V] Val	GCG [A] Ala	GAG [E] Glu	GGG [G] Gly

Technical details:

Minimization percentage is calculated as follows:

$$MP = \frac{E[\varphi] - \varphi_{code}}{E[\varphi] - \varphi_{opt}},$$

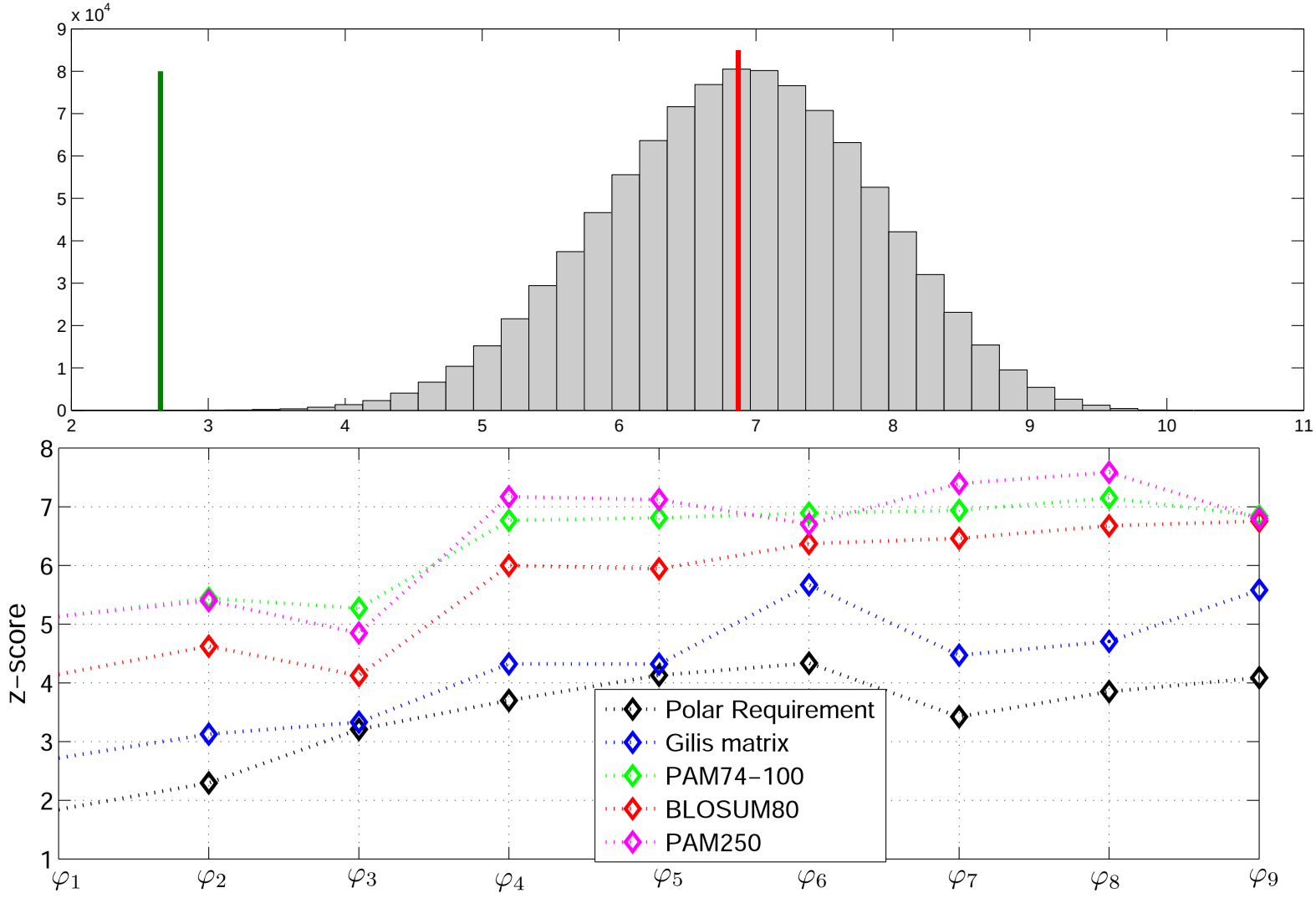
where $E[\varphi]$ is the mean value of the distribution of code costs, φ_{code} is the cost of the given code, φ_{opt} is the cost of the optimal code.

UUU [F] Phe	UCU [S] Ser	UAU [Y] Tyr	UGU [C] Cys
UUC [F] Phe	UCC [S] Ser	UAC [Y] Tyr	UGC [C] Cys
UUA [L] Leu	UCA [S] Ser	UAA [ ] Ter	UGA [ ] Ter
UUG [L] Leu	UCG [S] Ser	UAG [ ] Ter	UGG [W] Trp
CUU [L] Leu	CCU [P] Pro	CAU [H] His	CGU [R] Arg
CUC [L] Leu	CCC [P] Pro	CAC [H] His	CGC [R] Arg
CUA [L] Leu	CCA [P] Pro	CAA [Q] Gln	CGA [R] Arg
CUG [L] Leu	CCG [P] Pro	CAG [Q] Gln	CGG [R] Arg
AUU [I] Ile	ACU [T] Thr	AAU [N] Asn	AGU [S] Ser
AUC [I] Ile	ACC [T] Thr	AAC [N] Asn	AGC [S] Ser
AUA [I] Ile	ACA [T] Thr	AAA [K] Lys	AGA [R] Arg
AUG [M] Met	ACG [T] Thr	AAG [K] Lys	AGG [R] Arg
GUU [V] Val	GCU [A] Ala	GAU [D] Asp	GGU [G] Gly
GUC [V] Val	GCC [A] Ala	GAC [D] Asp	GGC [G] Gly
GUA [V] Val	GCA [A] Ala	GAA [E] Glu	GGA [G] Gly
GUG [V] Val	GCG [A] Ala	GAG [E] Glu	GGG [G] Gly

UUU [E] Glu	UCU [S] Ser	UAU [Y] Tyr	UGU [F] Phe
UUC [E] Glu	UCC [S] Ser	UAC [Y] Tyr	UGC [F] Phe
UUA [D] Asp	UCA [H] His	UAA [ ] Ter	UGA [ ] Ter
UUG [D] Asp	UCG [H] His	UAG [ ] Ter	UGG [W] Trp
CUU [N] Asn	CCU [G] Gly	CAU [V] Val	CGU [C] Cys
CUC [N] Asn	CCC [G] Gly	CAC [V] Val	CGC [C] Cys
CUA [K] Lys	CCA [G] Gly	CAA [V] Val	CGA [L] Leu
CUG [K] Lys	CCG [G] Gly	CAG [V] Val	CGG [L] Leu
AUU [Q] Gln	ACU [A] Ala	AAU [T] Thr	AGU [L] Leu
AUC [Q] Gln	ACC [A] Ala	AAC [T] Thr	AGC [L] Leu
AUA [R] Arg	ACA [A] Ala	AAA [T] Thr	AGA [L] Leu
AUG [R] Arg	ACG [A] Ala	AAG [T] Thr	AGG [L] Leu
GUU [R] Arg	GCU [S] Ser	GAU [P] Pro	GGU [I] Ile
GUC [R] Arg	GCC [S] Ser	GAC [P] Pro	GGC [I] Ile
GUA [R] Arg	GCA [S] Ser	GAA [P] Pro	GGA [I] Ile
GUG [R] Arg	GCG [S] Ser	GAG [P] Pro	GGG [M] Met

$$MP = 0.78$$

Some results:



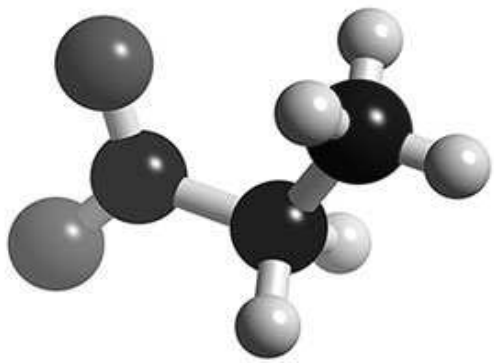
Novozhilov, A., Wolf, Yu., Koonin E.: *Biol Direct*, 2007, 2:24

The origin of the genetic code:

'...it seems likely that only a few amino acids were involved. [...] Again, it seems unlikely that the primitive code could code specifically for more than a few amino acids, since this would make the origin of the system terribly complicated.'

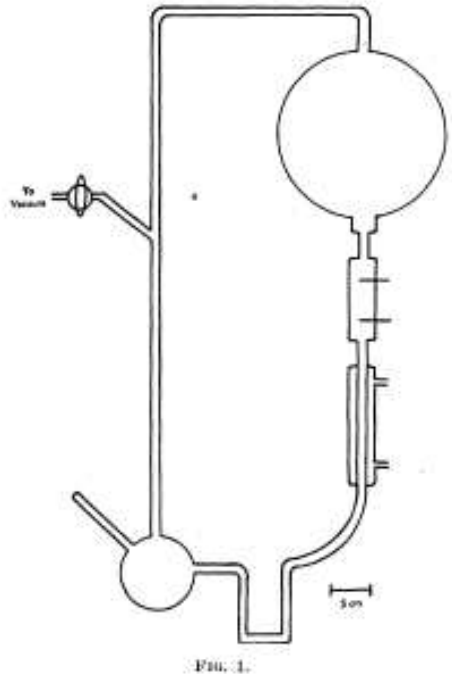
Crick F: *J Mol Biol*, 1968, **38**:367–379

Question: Which amino acids were the first?



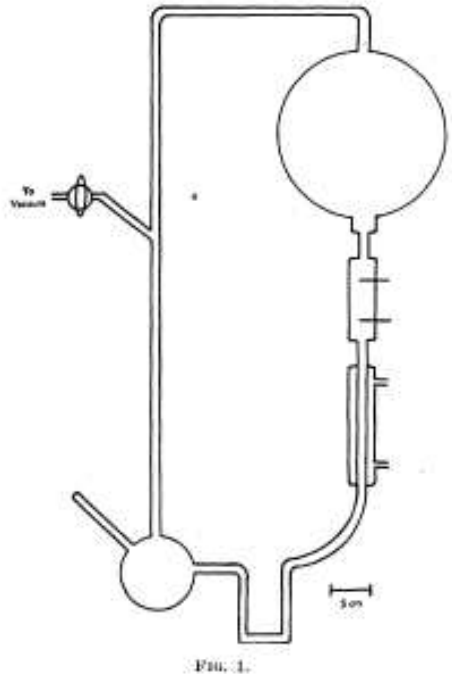
Prebiotic amino acid synthesis:

- Miller SL: *Science*, 1953, **117**:528–529
- Miller SL, Urey CH: *Science*, 1959, **130**:245–251
- Kobayashi K, et al.: *Org Life Evol Biosph*, 1990, **20**:99-109
- Cleaves HJ, et al: *Org Life Evol Biosph*, 2008, **38**:105–115
- many others...



Prebiotic amino acid synthesis:

- Miller SL: *Science*, 1953, **117**:528–529
- Miller SL, Urey CH: *Science*, 1959, **130**:245–251
- Kobayashi K, et al.: *Org Life Evol Biosph*, 1990, **20**:99-109
- Cleaves HJ, et al: *Org Life Evol Biosph*, 2008, **38**:105–115
- many others...



Observable amino acids:

Gly, **Ala**, **Asp**, **Glu**, **Val**, **Ser**, **Ile**, **Leu**, **Pro**, **Thr**

Consensus temporal order of amino acid formation:

Table V
Consensus temporal order of amino acids (final).

Amino acids of Miller		Average rank (± 0.7)	Order	Codon capture cases
+	G	3.5	1	
+	A	4.0	2	
+	D	6.0	3	
+	V	6.3	4	
+	P	7.3	5	
+	S	7.6	6	
+	E	8.1	7	
+	T	9.4	8	
+	L	9.9	9	(+)
	R	11.0	10	
	N	11.3	11	
+	I	11.4	12	(+)
	Q	11.4	13	(+)
	H	13.0	14	+
	K	13.3	15	
	C	13.8	16	+
	F	14.2	17	+
	Y	15.2	18	+
	M	15.4	19	+
	W	16.5	20	+

Gly
Ala
Asp
Glu
Val
Ser
Ile
Leu
Pro
Thr

Trifonov EN: *J Biomol Struct Dynam*, 2004, **22**:1–11

Thermodynamics of amino acid formation

Free energies of formation

	R_{obs}	R_{code}	C_{rel}	ΔG_{surf}
G - Gly	1.1	3.5	1.0000	80.49
A - Ala	2.8	4.0	0.4970	113.66
D - Asp	4.3	6.0	0.1633	146.74
E - Glu	6.8	8.1	0.1153	172.13
V - Val	8.5	6.3	0.0724	178.00
S - Ser	8.6	7.6	0.0286	173.73
I - Ile	9.1	11.4	0.0226	213.93
L - Leu	9.4	9.9	0.0116	205.03
P - Pro	10.0	7.3	0.0437	192.83
T - Thr	11.7	9.4	0.0008	216.50
K - Lys	12.6	13.3	0	258.56
F - Phe	13.2	14.4	0	303.64
R - Arg	13.3	11.0	0	409.46
H - His	13.3	13.0	0	350.52
N - Asn	14.2	11.3	0	201.56
Q - Gln	14.2	11.4	0	223.36
C - Cys	14.2	13.8	0	224.67
Y - Tyr	14.2	15.2	0	334.20
M - Met	14.2	15.4	0	113.22
W - Trp	14.2	16.5	0	431.17

Higgs PG, Pudritz RE: *ArXiV*, 2009

Stability of base step interactions

XZ(N)	ΔG/base-step (B-*D*NA)		Aminoacids in cc
	Melting (kcal/mol)	Stacking (kcal/mol)	
GC	−2.70	−2.17	Ala[b]
GT	−2.04	−1.81	Val[b]
AC			Thr[b]
GG	−1.97	−1.44	Gly[b]
CC			Pro[b]
GA	−1.66	−1.43	Asp, Glu
TC			Ser[b]
CG	−1.44	−0.91	Arg[b]
CT	−1.29	−1.06	Leu[b]
AG			Arg, Ser[a]
AT	−1.27	−1.34	Ileu, Met
AA	−1.04	−1.11	Phe, Leu[a]
TT			Asn, Lys
CA	−0.78	−0.55	His, Gln
TG			Cys, Trp, Term[c]
TA	−0.12	−0.19	Tyr, Term[c]

Travers A: *Orig Life Evol Biosph*, 2006, **36**:549–555

‘Early’ and ‘late’ amino acids:

‘Early’ amino acids:

Gly, **Ala**, **Asp**, **Glu**, **Val**, **Ser**, **Ile**, **Leu**, **Pro**, **Thr**

‘Late’ amino acids:

Arg(?), **Asn**, **Gln**, **His**, **Lys**, **Cys**, **Phe**, **Tyr**, **Met**, **Trp**

Stereochemical theory and 'early' amino acids:

- Testes amino acids: **Phe**, **Ile**, **His**, **Leu**, **Arg**, **Trp**, **Tyr**, **Gln**;
- Only **Arg** showed strong statistical support;
- **Gln** showed no correlation between the codon and selected aptamers;
- The stereochemical association and error minimization properties are independent;
- The hypothesis: prior fixation of a stereochemical core and an effective later minimization of error (Caporaso et al: *J Mol Evol*, 2005, **61**:597–607).

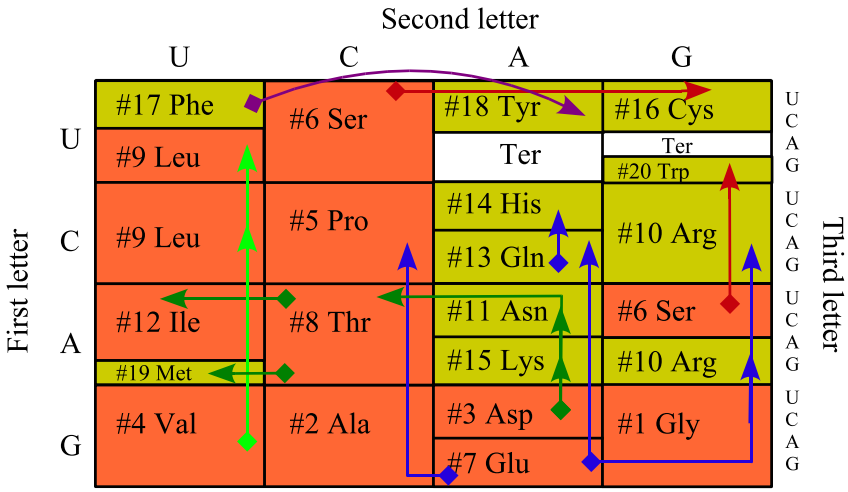
Stereochemical theory and 'early' amino acids:

- Testes amino acids: **Phe**, **Ile**, **His**, **Leu**, **Arg**, **Trp**, **Tyr**, **Gln**;
- Only **Arg** showed strong statistical support;
- **Gln** showed no correlation between the codon and selected aptamers;
- The stereochemical association and error minimization properties are independent;
- The hypothesis: prior fixation of a stereochemical core and an effective later minimization of error (Caporaso et al: *J Mol Evol*, 2005, **61**:597–607).
- Only **Ile** and **Leu** are 'old' amino acids;

Coevolution theory and ‘early’ amino acids:

’Coevolution theory suggests that there are three phases of amino acid entry into proteins. Phase 1 amino acids came from prebiotic synthesis, and phase 2 ones from biosynthesis’

Wong J: *BioEssays*, 2005, **27**:416–425



According to Wong J, 2005

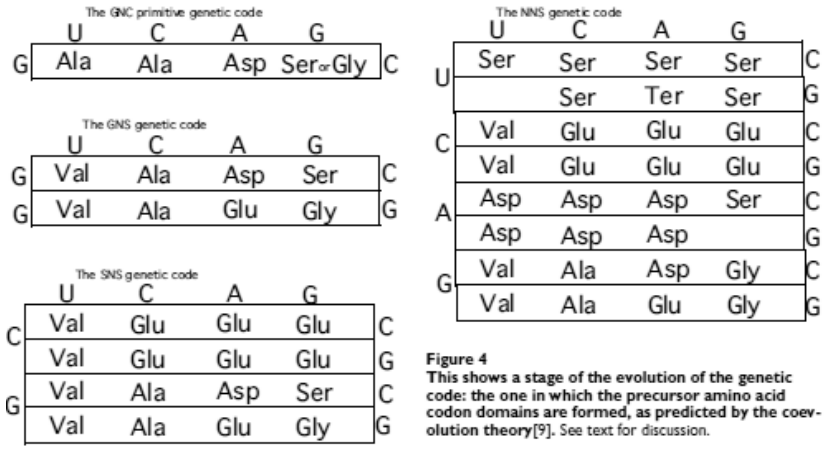


Figure 4
This shows a stage of the evolution of the genetic code: the one in which the precursor amino acid codon domains are formed, as predicted by the coevolution theory[9]. See text for discussion.

Di Giulio M: *Biol Dir*, 2008, **3**:37

Origin and evolutionary process of the genetic code:

- 'The primitive code was a triplet code (in the sense that the reading mechanism moved along three bases at each step) but that only, say, the first two bases were read. This is not at all implausible'.

  Crick F: *J Mol Biol*, 1968, **38**:367–379

Origin and evolutionary process of the genetic code:

- 'The primitive code was a triplet code (in the sense that the reading mechanism moved along three bases at each step) but that only, say, the first two bases were read. This is not at all implausible'.
  Crick F: *J Mol Biol*, 1968, **38**:367–379
- Glycine code: Hartman H: *Orig Life*, 1975, **6**:423–427;
- GNS code: Ikehara K, Niihara Y: *Curr Med Chem*, 2007, **14**:3221–3231;
- GNN code: Higgs PG: *Biol Dir*, 2009, **4**:16
- many others...

Why 2-letter triplet codons?

- Wobble rule: the base at the 5' end of the anticodon does not have as strict base-pairing requirements as the other two base pases, allowing it to form hydrogen bonds with several bases at the 3' end of the codon);
- Thermodynamics of codon-anticodon interactions: the codon–anticodon pairs for the codes in non-plant mitochondria on the one hand and prokaryotic and eukaryotic organisms on the other can be unequivocally divided into two classes — the most stable base steps define a common code specified by the first two bases in a codon while the less stable base steps correlate with divergent usage and the adoption of a 3-letter code. (Travers A: 2006).

The parsimony principle:

if the primordial code encoded and amino acid, then this amino acid was encoded by the same four-codon series (a supercodon) that encodes the same amino acid in the standard genetic code (or, at least, a subset of the series encodes the same amino acid)

UUU [F] Phe	UCU [S] Ser	UAU [Y] Tyr	UGU [C] Cys
UUC [F] Phe	UCC [S] Ser	UAC [Y] Tyr	UGC [C] Cys
UUA [L] Leu	UCA [S] Ser	UAA [ ] Ter	UGA [ ] Ter
UUG [L] Leu	UCG [S] Ser	UAG [ ] Ter	UGG [W] Trp
CUU [L] Leu	CCU [P] Pro	CAU [H] His	CGU [R] Arg
CUC [L] Leu	CCC [P] Pro	CAC [H] His	CGC [R] Arg
CUA [L] Leu	CCA [P] Pro	CAA [Q] Gln	CGA [R] Arg
CUG [L] Leu	CCG [P] Pro	CAG [Q] Gln	CGG [R] Arg
AUU [I] Ile	ACU [T] Thr	AAU [N] Asn	AGU [S] Ser
AUC [I] Ile	ACC [T] Thr	AAC [N] Asn	AGC [S] Ser
AUA [I] Ile	ACA [T] Thr	AAA [K] Lys	AGA [R] Arg
AUG [M] Met	ACG [T] Thr	AAG [K] Lys	AGG [R] Arg
GUU [V] Val	GCU [A] Ala	GAU [D] Asp	GGU [G] Gly
GUC [V] Val	GCC [A] Ala	GAC [D] Asp	GGC [G] Gly
GUA [V] Val	GCA [A] Ala	GAA [E] Glu	GGA [G] Gly
GUG [V] Val	GCG [A] Ala	GAG [E] Glu	GGG [G] Gly

The parsimony principle:

if the primordial code encoded and amino acid, then this amino acid was encoded by the same four-codon series (a supercodon) that encodes the same amino acid in the standard genetic code (or, at least, a subset of the series encodes the same amino acid)

Phe/Leu	**Ser**	**Tyr/Ter**	**Cys/Trp**
Leu	**Pro**	**His/Gln**	**Arg**
Iso/Met	**Thr**	**Asn/Lys**	**Ser/Arg**
Val	**Ala**	**Asp/Glu**	**Gly**

Two-letter triplet code and 'early' amino acids:

	U	C	A	G
U	?/Leu	Ser	?	?
C	Leu	Pro	?	?
A	Ile	Thr	?	?/Ser
G	Val	Ala	Asp	Gly

Question: What is the level of error minimization of doublet genetic codes having the core shown in the figure?

The arrangement of 'early' amino acids is almost perfect:

If I ignore the question marks (i.e., put $d(a_1, a_2) = 0$ if a_1 or a_2 are question marks):

(a)

	U	C	A	G
U	?	Ser	?	?
C	Leu	Pro	?	?
A	Ile	Thr	?	?
G	Val	Ala	Asp	Gly

(b)

	U	C	A	G
U	Leu	Thr	?	?
C	Leu	Pro	?	?
A	Ile	Ala	?	?
G	Val	Ser	Asp	Gly

(c)

	U	C	A	G
U	Leu	Thr	?	?
C	Leu	Pro	?	?
A	Ile	Ala	?	Ser
G	Val	Ser	Asp	Gly

(d)

	U	C	A	G
U	?	Ser	?	?
C	Leu	Pro	?	?
A	Ile	Thr	?	Ser
G	Val	Ala	Asp	Gly

$$MP > 0.98$$

What about unknown assignments?

We can assume that the genetic code table is filled columnwise:

(a)

	U	C	A	G
U	Val	Ser	Asp	Gly
C	Val	Ala	Asp	Gly
A	Ile	Thr	Asp	Gly
G	Leu	Pro	Asp	Gly

(b)

	U	C	A	G
U	Leu	Thr	Asp	Gly
C	Leu	Pro	Asp	Gly
A	Val	Ser	Asp	Gly
G	Ile	Ala	Asp	Gly

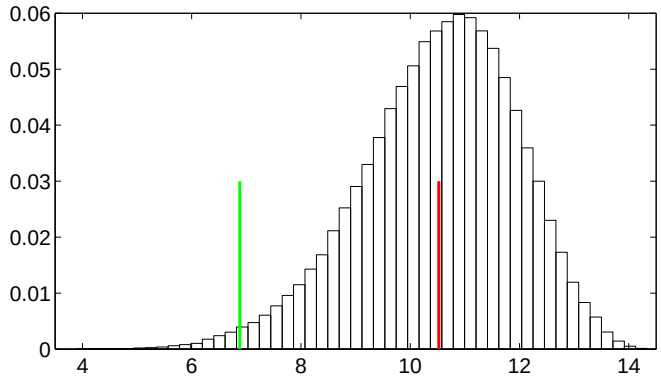
$$MP > 0.94$$

Two-letter code with 16 amino acids:

(a)

	U	C	A	G
U	Leu	Ser	Tyr	Cys
C	Leu	Pro	His	Arg
A	Ile	Thr	Asn	Ser
G	Val	Ala	Asp	Gly

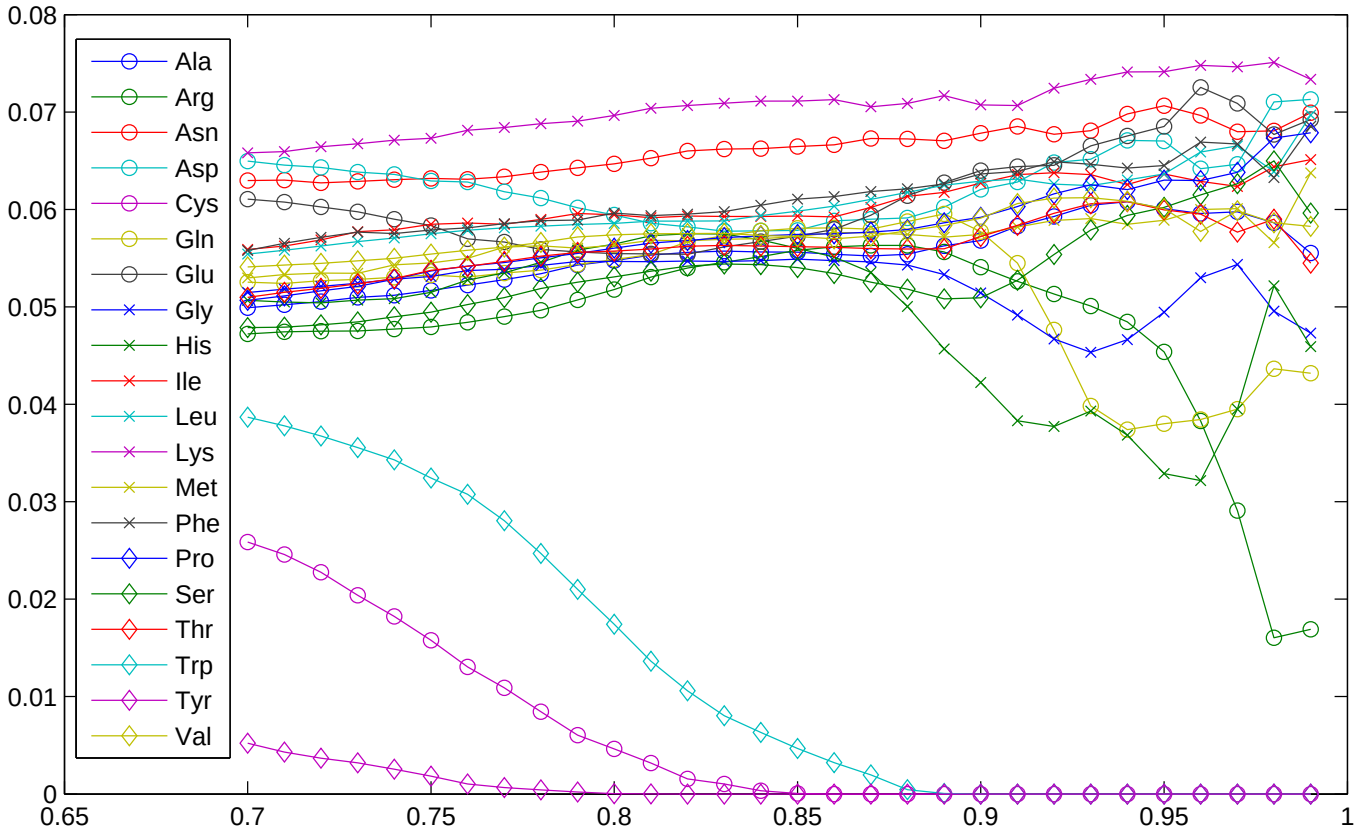
(b)



Error minimization level of 2-letter codes. (*a*) A 2-letter code obtained using the parsimony principle. For the cells with an ambiguous assignment, one random amino acid is chosen; (*b*) The distribution of the costs of the random 2-letter codes obtained by permutation of amino acid assignments in (*a*), the green line shows the cost of the code from (*a*) and the red line shows the mean; $MP = 0.51$, the distance from the mean is 2.6 standard deviations.

Which amino acid's position is the worst?

- The codons **UAN** and **UGN** are the least stable according to Travers, 2006;
- amino acids **Cys**, **Trp**, **Tyr** are 'the worst' amino acids with respect to error minimization of the doublet genetic code:

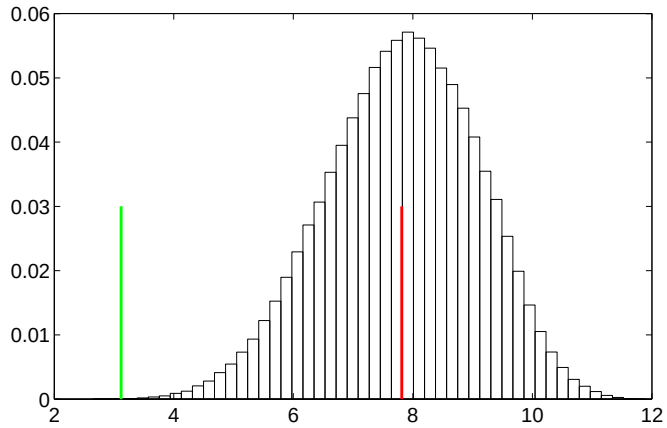


Two-letter code with disregarded stop codons:

(a)

	U	C	A	G
U	Leu	Ser	?	?
C	Leu	Pro	His	Arg
A	Ile	Thr	Asn	Ser
G	Val	Ala	Asp	Gly

(b)



Error minimization levels of 2-letter codes. (*a*) A 2-letter genetic code obtained using the parsimony principle. For the cells with ambiguous assignment, one random amino acid is chosen; two supercodons, UAN and UGN, are disregarded; (*b*) The distribution of the costs of random 2-letter codes obtained by permutation of amino acid assignments in (*a*), the green line shows the cost of the code in (*a*), and the red line shows the mean; $MP = 0.88$, the distance from the mean is 3.7 standard deviations.

Fixing the position of Arginine:

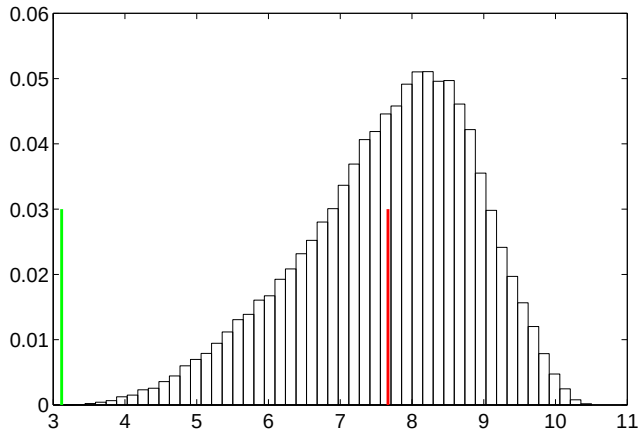
(a)

	U	C	A	G
U	Leu	Ser	?	?
C	Leu	Pro	His	Arg
A	Ile	Thr	Asn	Ser
G	Val	Ala	Asp	Gly

(b)

	U	C	A	G
U	Ile	Pro	?	?
C	Leu	Thr	His	Arg
A	Leu	Ala	Asn	Ser
G	Val	Ser	Asp	Gly

(c)



$$MP = 0.983$$

Conclusions:

- Given the list of 'early' amino acids, stereochemical theory and coevolution theory cannot be taken as a reasonable explanation for a primordial genetic code;
- Taking into account the parsimony rule and likely doublet genetic code we can infer that the assignments of the 'early' amino acids is nearly 'ideal' with respect to each other;
- If we fix the assignments of only two particular amino acids, **Asn** and **Arg**, in the doublet genetic code table, then the selective force of the error minimization yields the code extremely close to the standard one.

Hypothesis: the primordial code was shaped almost exclusively by the selective forces to minimize the impact of translational mistakes.

Collaborators:

- Eugene Koonin, NCBI/NIH 
- Yuri Wolf, NCBI/NIH 
- Special thanks to the members of Koonin's group at NCBI

Thank you for your attention!

Questions?

- Novozhilov AS, Wolf Y, Koonin E: *Biol Direct*, 2007, 2:24
- Koonin EV, Novozhilov AS: *IUBMB Life*, 2009, 61(2):99–111
- Novozhilov AS, Koonin EV: *Biol Direct*, 2009, 4:44