On the origin of the genetic code: a 27-codon hypothetical precursor of an intricate 64-codon intermediate shaped the modern code.

Bernard Dujon *

Institut Pasteur, Dept. Genomes and Genetics, CNRS (UMR3525) and Sorbonne Université (UFR927), Paris, France

Supplementary material: 2 tables and 3 figures

Supplementary Table S1: Analysis of the codon table in the standard and variant forms used in this work.

Nucleotide at position **1**	**A**	**C**	**G**	**U**
Unsplit codon families	ACN > Thr	CCN > Pro CGN > Arg CUN > Leu	GCN > Ala GGN > Gly GUN > Val	UCN > Ser
Split codon families	AGN * > Ser + Arg (*Gly*) AUN > Ile + Met AAN > Asn + Lys	CAN > His + Gln	GAN > Asp + Glu	UGN > Cys + Trp UUN > Phe + Leu UAN *> Tyr + (*Gln*)

Nucleotide at position **2**	**A**	**C**	**G**	**U**
Unsplit codon families		GCN > Ala CCN > Pro ACN > Thr UCN > Ser	GGN > Gly CGN > Arg	GUN > Val CUN > Leu
Split codon families	GAN > Asp + Glu CAN > His + Gln AAN > Asn + Lys UAN * > Tyr + (*Gln*)		AGN * > Ser + Arg (*Gly*) UGN > Cys + Trp	AUN > Ile + Met UUN > Phe + Leu

The table groups in columns the codon families of the modern code according to nucleotides at first or second positions of codons, with indication of their coding significance in the standard or variant (brackets) forms. * Gln replaces stop for UAR codons in the UAN family; Gly replaces Arg for AGR codons in the AGN family. Color code of a.a.s as in Table 1.

Note that all codons with **C** at the second position belong to unsplit families (fully degenerated boxes of 4 codons) whereas all codons with **A** at the same position belong to split families (partially degenerated boxes of 2, exceptionally 3, codons). Codons with **G** or **U** at the second position are equally distributed between split and unsplit families. No similar bias exists for the first position of codons. Due to anticodon interactions with the ribosomal grip, codons of unsplit families always have either a **C-G** pair at the central position of codon-anticodon duplexes or a **C-G** / **G-C** pair at their 1st position, whereas codons of split families always have either a **A-U** pair at the central position or a **U-A** / **A-U** pair at position 1 [39].

Note that all codons with **C** at the second position encode amino acids charged by class II aaRS [43, 44] whereas almost all codons with **U** at this position encode amino acids charged by class I aaRS. Codons with **U** at second positions encode hydrophobic amino acids, those with **C** encode small neutral amino acids and those with **A** encode hydrophilic and larger neutral amino acids [42].

Finally, note that the second position of codon is also the most sensitive to potential mutational errors [70].

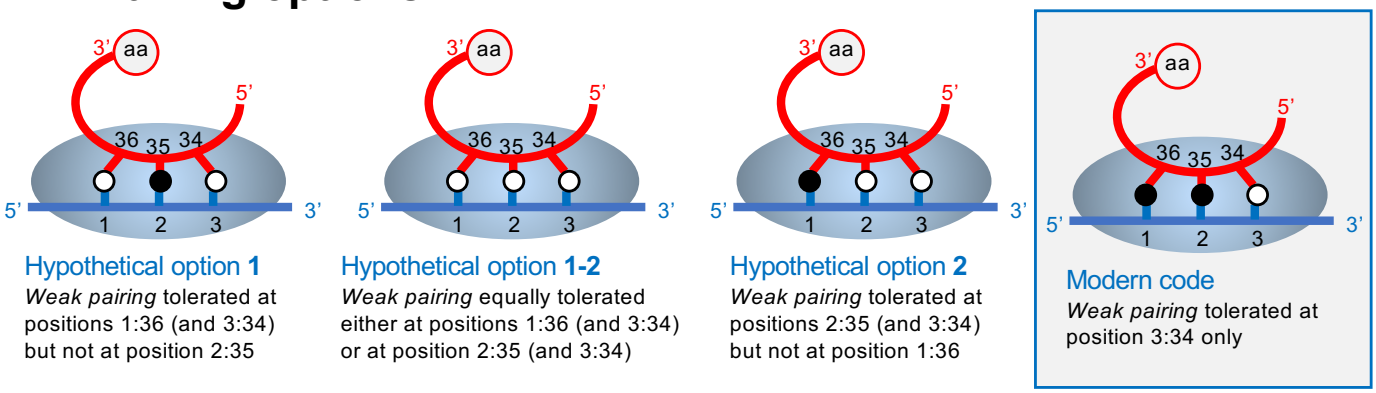
Supplementary Table S2: Predicted codon *intricacy* under pairing option 1 and its traces in the modern code.

Shared anticodon type	*Intricate* families		Coding significance		Observable coincidence	Remarks
	Family n°1	Family n°2	Family n°1	Family n°2		
NGG	**CCN**	**UCN**	Pro (*IIA*)	Ser (*IIA*)	Same aaRS subclass (IIA)	Possible Pro/Ser early decoding ambiguity (see Figure 1).
NGU	**GCN**	**ACN**	Ala (*IIC*)	Thr (*IIA*)	Same aaRS class (II)	Likely Ala/Thr early decoding ambiguity (see Figure 2).
NAG	**CUN**	*UUN*	Leu (*IA*)	Leu (*IA*) Phe (*IIC*)	Same a.a. (Leu)	CUB and UUB families were non-coding in the 27-codon precursor code.
NAU	**GUN**	*AUN*	Val (*IA*)	Ile (*IA*) Met (*IA*)	Same aaRS subclass (IA)	Val and Ile are the most similar of all proteinogenic a.a.s. The GUB family was non-coding in the 27-codon precursor code. Likely Val / Ile decoding ambiguity at intermediate period.
NCG	**CGN**	*UGN*	Arg (*ID*)	Cys (*IB*) Trp (*IC*)	Same aaRS class (I)	For Arg and Cys, see Discussion.
NCU	**GGN**	*AGN*	Gly (IIA IIC)	Ser (*IIA*) Arg (*ID*) *Gly* [2] (*IIA IIC*)	Same aaRS subclass (IIA)	Likely accidental coincidence for Gly in a variant mitochondrial code.
NUG [1]	*CAN*	*UAN*	His (*IIA*) Gln (*IB*)	Tyr (*IC*) *Gln* [3] (*IB*)	Same a.a. (Gln)	Likely significant coincidence for Gln in variant nuclear codes.
NUU [1]	*GAN*	*AAN*	Asp (*IIB*) Glu (*IB*)	Asn (*IIB*) *Lys* (*IIB IE*)	Same aaRS subclass (IIB)	Possible Asp/Asn early decoding ambiguity or late replacement of Asp by Asn (see transamidation process).

The table lists all intricate pairs of codon families (columns 2 and 3, **bold type** unsplit families, *italics*, split families) predicted under hypothetical pairing option **1** with their common anticodon type (first column) and their coding significance (columns 4 and 5) in the modern code (a.a. color refer to aaRS class as in Table I, subclasses under brackets are according to [64]. For each predicted pair of intricate codon family, observed coincidence in the modern code is noted in column 6 with remarks in column 7. Notes: [1] Inactive anticodon type in the 27-codon precursor code (see Figure 1). [2] Gly is found in place of Arg in the mitochondrial code of urochordates (see text). [3] Gln encoded by UAR codons in the nuclear code of many unicellular eukaryotes (see text).

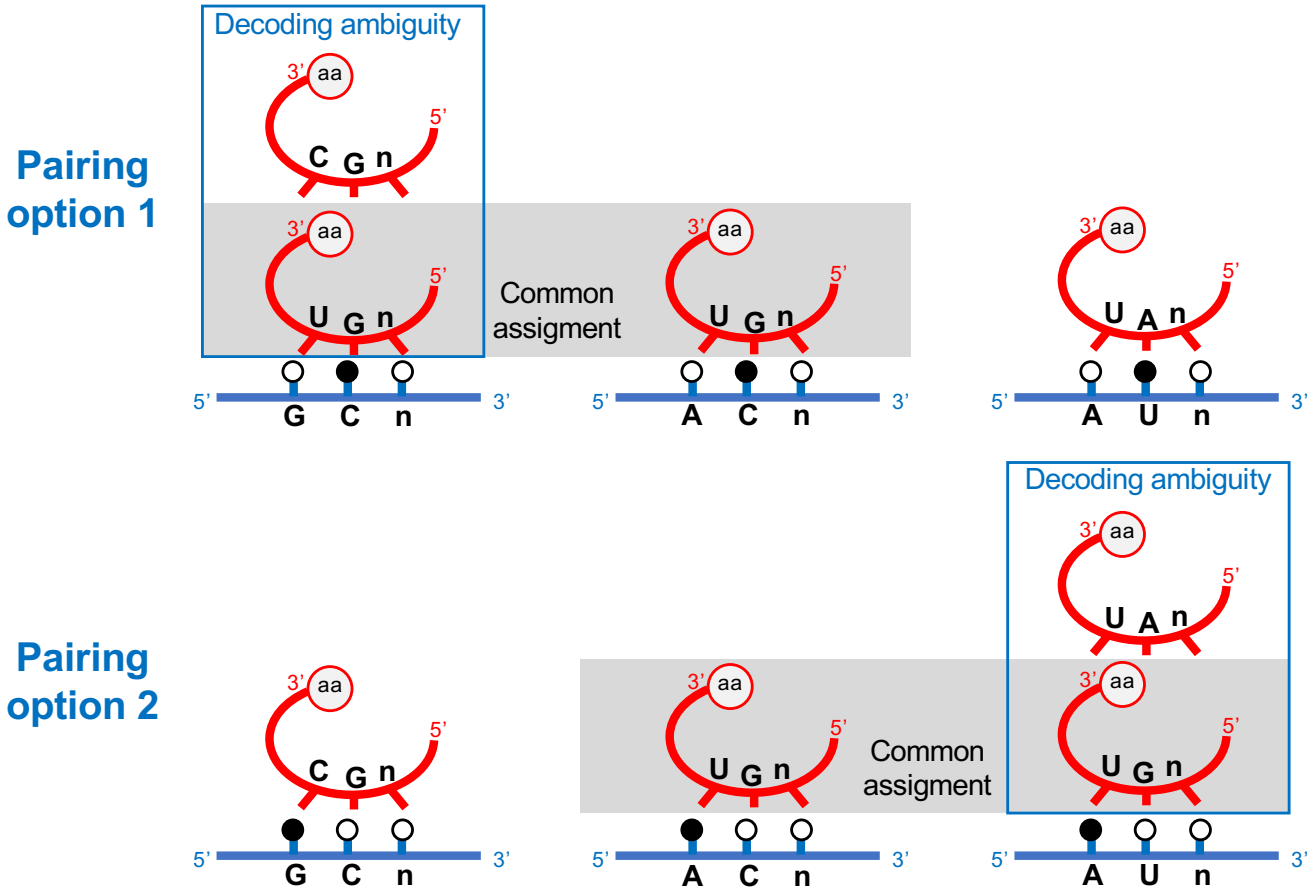
Suppl. Figure S1

A: Pairing options



The figure illustrates the 3 hypothetical codon-anticodon pairing options examined in this work, compared to the pairing rules observed in the modern code (right insert). Grey ovals symbolize the decoding machinery; blue lines codons; red lines: anticodons; nucleotide numbering as in [39]. Circles: purine-pyrimidine interactions tolerated at each position of the codon - anticodon duplex: black: Watson-Crick pairs only; white: same plus weaker pairs (*e.g.* **G•U** or **U•G**). Note that under the option 1-2, weak pairing is not simultaneously tolerated at positions 1:36 and 2:35.

B: Examples of decoding ambiguity and codon *intricacy* predicted by pairing options 1 and 2



The figure illustrates the consequences of pairing options **1** or **2** on the hypothetical codes using the codon families GCN, ACN and AUN as an example. Under each option, a given codon family can be recognized by more than one anticodon types, creating a potential decoding ambiguity (box), and distinct codon families can share a same anticodon type (grey background), creating a common assignment (designated here as *intricacy*). Both properties are intrinsic characteristics of all codon-anticodon matrices explored in this work. Affected codons depend on the pairing option selected, as illustrated.

Suppl. Figure S2: Codon-anticodon interaction matrix in a hypothetical 27-codon G, A, U precursor code and consequences.

A

Anticodon types (5' – 3')			Codon families (5' – 3')								
			RRn				RYn		YRn		YYn
			GGD	*GAD*	*AGD*	*AAD*	**GUD**	*AUD*	*UGD*	*UAD*	*UUD*
nRR	5a	DGG	-	-	-	-	-	-	-	-	-/-/-
	5b	DGA	-	-	-	-	-	-	-	-	-/+/+
	5c	DAG	-	-	-	-	-	-	-	-	+/+/-
	5d	DAA	-	-	-	-	-	-	-	-	**++**
nRY	6a	DGU	-	-	-	-	-/-/-	-/+/+	-	-	-
	6b	DAU	-	-	-	-	+/+/-	**++**	-	-	-
nYR	7a	DUG	-	-	-	-	-	-	-/-/-	+/+/-	-
	7b	DUA	-	-	-	-	-	-	-/+/+	++	-
nYY	8a	DUU	-/-/-	+/+/-	-/+/+	++	-	-	-	-	-

C

Anti - codons	Corresponding amino acids		
	Pairing option **1**	Pairing option **1-2**	Pairing option **2**
DGG	-	-	-
DGA	-	?	?
DAG	?	?	-
DAA	?	?	?
DGU	-	?	?
DAU	Val , ?	Val , ?	?
DUG	?	?	-
DUA	?	?	?
DUU	?	?	?

B

Coding ambiguity and codon intricacy	**GGD**	*GAD*	*AGD*	*AAD*	**GUD**	*AUD*	*UGD*	*UAD*	*UUD*
Pairing option **1**	-	8a	-	8a	6b	6b	-	7a 7b	5c 5d
Pairing option **1-2**	-	8a	8a	8a	6b	6a 6b	7b	7a 7b	5c 5c 5d
Pairing option **2**	-	-	8a	8a	-	6a 6b	7b	7b	5b 5d

Codon-anticodon interaction matrix for primitive RNA molecules composed of two hypothetical purines (**G** and **A)** and a single pyrimidine (**U**). Compare to Figure 1. Same legend except for symbols: **++** active pairing independent of chosen option (two **A-U** or **U-A** pairs at positions 1 and 2 of codons): **+** active pairing dependent of chosen option (one **A-U** or **U-A** pair plus one **G•U** or **U•G** pair at positions 1 and 2 of codons); **-** no pairing (any other combination). **D** = not **C**.

Suppl. Figure S3: Codon-anticodon interaction matrix in an intermediate 64-codon G', C, U, A' code and consequences.

A

Ancient and novel anticodon types (5'-3')			Ancient codon families (5' – 3')									Novel codon families (5' – 3')						
			RRn	RYn		YRn		YYn				RRn			RYn		YRn	
			G'G'N	G'CN	G'UN	CG'N	UG'N	CCN	CUN	UCN	UUN	A'A'N	G'A'N	A'G'N	A'CN	A'UN	CA'N	UA'N
nRR	1'a	NG'G'	-	-	-	-	-	++	-/+/+	+/+/-	- /-/-	-	-	-	-	-	-	-
nRY	2'a	NG'C	-	++	-/+/+	-	-	-	-	-	-	-	-	-	+/+/-	-/-/-	-	-
	2'b	NG'U	-	+/+/-	-/-/-	-	-	-	-	-	-	-	-	-	++	-/+/+	-	-
nYR	3'a	NCG'	-	-	-	++	+/+/-	-	-	-	-	-	-	-	-	-	-/+/+	-/-/-
	3'b	NUG'	-	-	-	-/+/+	-/-/-	-	-	-	-	-	-	-	-	-	++	+/+/-
nYY	4'a	NCC	++	-	-	-	-	-	-	-	-	-/-/-	-/+/+	+/+/-	-	-	-	-
	4'b	NCU	+/+/-	-	-	-	-	-	-	-	-	-/+/+	-/-/-	++	-	-	-	-
	4'c	NUC	-/+/+	-	-	-	-	-	-	-	-	+/+/-	++	-/-/-	-	-	-	-
	4'd	NUU	-/-/-	-	-	-	-	-	-	-	-	++	+/+/-	-/+/+	-	-	-	-
nRY	2'c	NA'C	-	-/+/+	++	-	-	-	-	-	-	-	-	-	-/-/-	+/+/-	-	-
	2'd	NA'U	-	-/-/-	+/+/-	-	-	-	-	-	-	-	-	-	-/+/+	++	-	-
nYR	3'c	NCA'	-	-	-	+/+/-	++	-	-	-	-	-	-	-	-	-	-/-/-	-/+/+
	3'd	NUA'	-	-	-	-/-/-	-/+/+	-	-	-	-	-	-	-	-	-	+/+/-	++
nRR	1'b	NA'G'	-	-	-	-	-	-/+/+	++	-/-/-	+/+/-	-	-	-	-	-	-	-
	1'c	NG'A'	-	-	-	-	-	+/+/-	-/-/-	++	-/+/+	-	-	-	-	-	-	-
	1'd	NA'A'	-	-	-	-	-	-/-/-	+/+/-	-/+/+	++	-	-	-	-	-	-	-

C

Anti-codons	Corresponding amino-acids		
	Pairing option **1**	Pairing option 1-2	Pairing option 2
NG'G'	*Pro, Ser*	*Pro, Ser, Leu*	*Pro, Leu*
NG'C	*Ala*, **Thr**	*Ala, Val*	*Ala, Val*
NG'U	*Ala*, **Thr**	*Ala*, **Thr**, ?	**Thr**, ?
NCG'	*Arg*, ?	*Arg*, ?	*Arg*, ?
NUG'	?	*Arg*, ?	*Arg*, ?
NCC	*Gly*, ?	*Gly*, ?	*Gly*, ?
NCU	*Gly*, ?	*Gly*, ?	?
NUC	?	*Gly*, ?	*Gly*, ?
NUU	?	?	?
NA'C	**Val**	Val, Ala, ?	Val, Ala, ?
NA'U	**Val**, ?	Val, **Thr**, ?	**Thr**, ?
NCA'	Arg, ?	Arg, ?	?
NUA'	?	?	?
NA'G'	**Leu**, ?	Leu, Pro, ?	Leu, Pro
NG'A'	Ser, Pro	Ser, Pro, ?	**Ser**, ?
NA'A'	**Leu**, ?	Leu, Ser, ?	Ser, ?

B

Coding ambiguity and codon intricacy	G'G'N	G'CN	G'UN	CG'N	UG'N	CCN	CUN	UCN	UUN	A'A'N	G'A'N	A'G'N	A'CN	A'UN	CA'N	UA'N
Pairing option 1	4'a 4'b	2'a 2'b	2'c 2'd	3'a 3'c	3'a 3'c	1'a 1'c	1'b 1'd	1'a 1'c	1'b 1'd	4'c 4'd	4'c 4'd	4'a 4'b	2'a 2'b	2'c 2'd	3'b 3'd	3'b 3'd
Pairing option 1-2	4'a 4'b 4'c	2'a 2'b 2'c	2'a 2'c 2'd	3'a 3'b 3'c	3'a 3'c 3'd	1'a 1'b 1'c	1'a 1'b 1'd	1'a 1'c 1'd	1'b 1'c 1'd	4'b 4'c 4'd	4'a 4'c 4'd	4'a 4'b 4'd	2'a 2'b 2'd	2'b 2'c 2'd	3'a 3'b 3'd	3'b 3'c 3'd
Pairing option 2	4'a 4'c	2'a 2'c	2'a 2'c	3'a 3'b	3'c 3'd	1'a 1'b	1'a 1'b	1'c 1'd	1'c 1'd	4'b 4'd	4'a 4'c	4'b 4'd	2'b 2'd	2'b 2'd	3'a 3'b	3'c 3'd

Codon-anticodon interaction matrix for hypothetical RNA molecules composed of two hypothetical purines (symbolized as **G**' and **A')** with opposite pairing preferences for the two pyrimidines (**C** and **U**) *i.e.* **G'-C** or **C-G'** > **G'•U** or **U•G'** and **A'-U** or **U-A'** > **A'•C** or **C•A'.** Compare to Figure 2. Same legend except for symbols: **++** active pairing independent of chosen option (two **G'-C** or **C-G'** pairs, two **A'-U** or **U-A'** pairs, or one **G'-C** or **C-G'** pair plus one **A'-U** or **U-A**' pair at positions 1 and 2 of codons); **+** active pairing dependent of chosen option (one **G'-C**, **C-G'**, **A'-U** or **U-A**' pair plus one **G'•U**, **U•G'**, **A'•C** or **C•A**' pair at positions 1 and 2 of codons); **-** no pairing (any other combination). Note that part **C** is only presented here to facilitate the comparison with Figure 2, assuming that the two purines, **G'** and **A',** are equivalent to **G** and **A,** respectively.