



# Evolution of the genetic code through progressive symmetry breaking



Reijer Lenstra\*

Route Cantonale 103, Saint Sulpice VD, Switzerland

## HIGHLIGHTS

- Evolution of the primitive protein synthesis machinery shapes the genetic code.
- Early ribosomes first determine a unique mRNA reading frame and subsequently enforce increasingly stringent basepairing to control codon-anticodon recognition.
- Aminoacyl-tRNA synthetases increasingly distinguish between tRNAs and amino acids.
- These symmetry breaking processes with selection for information generate the code.

## ARTICLE INFO

### Article history:

Received 12 December 2012

Received in revised form

18 December 2013

Accepted 1 January 2014

Available online 14 January 2014

### Keywords:

Aminoacyl-tRNA synthetase

Codon graph

Hamming distance


Ribosome

Shannon entropy

## ABSTRACT

Evolution of the genetic code in an early RNA world is dependent on the steadily improving specificity of the coevolving protein synthesis machinery for codons, anticodons, tRNAs and amino acids. In the beginning, there is RNA but the machinery does not distinguish yet between the codons, which therefore all encode the same information. Synonymous codons are equivalent under a symmetry group that exchanges (permutes) the codons without affecting the code. The initial group changes any codon into any other by permuting the order of the bases in the triplet as well as by replacing the four RNA bases with each other at every codon position. This group preserves the differences between codons, known as Hamming distances, with a 1-distance corresponding to a single point mutation. Stepwise breaking of the group into subgroups divides the 64 codons into progressively smaller subsets – blocks of equivalent codons under the smaller symmetry groups, with each block able to encode a different message. This formalism prescribes how the evolving machinery increasingly differentiates between codons. The model indicates that primitive ribosomes first identified a unique mRNA reading frame to break the group permuting the order of the bases and subsequently enforced increasingly stringent codon-anticodon basepairing rules to break the subgroups permuting the four bases at each codon position. The modern basepairing rules evolve in five steps and at each step the number of codon blocks doubles. The fourth step generates 16 codon blocks corresponding with the 16 family boxes of the standard code and the last step splits these boxes into 32 blocks of commonly two, but rarely one or three, synonymous codons. The evolving codes transmit at most one message per codon block and as the number of messages increases so does the specificity of the code and of protein synthesis. The selective advantage

metadata, citation and similar papers at [core.ac.uk](http://core.ac.uk)

brought to you by  CORE

provided by Elsevier - Publisher Connector

family of primitive aminoacyl-tRNA synthetases (aaRSs) divides the tRNA diversities into various different and overlapping subsets: each aaRS accepts some tRNAs but rejects all others and several aaRSs may accept the same tRNA species. Selection favoring less ambiguous codes eliminates these overlaps and also imposes the ribosomal anticodon block division as ambiguity arises when different aaRSs accept tRNAs of the same anticodon block. Only when the tRNAs of one or several anticodon blocks are accepted by a unique aaRS does the code become specific. This coding pattern is observed in the standard code and the evolution of amino acid assignments by primitive aaRSs onto tRNAs is traced back via tRNA trees that picture a gradual division of tRNA diversities into blocks with increasingly specific amino acid assignments. Symmetry breaking combined with continuous selection for codes carrying

\* Tel.: +41 21 534 5561.

E-mail address: [reijerlenstra@hispeed.ch](mailto:reijerlenstra@hispeed.ch)

more information evolves increasingly specific codes and efficiently traverses an immense space of all possible codes ( $> 10^{84}$ ) to give rise to the standard code.

© 2014 The Authors. Published by Elsevier Ltd. Open access under [CC BY license](#).

## 1. Introduction

The canonical genetic code maps 64 codons onto 21 messages – 20 amino-acids and a stop-signal. (Every codon encodes a single message and each message is encoded by at least one codon: a 64 → 21 onto mapping.) All extant living organisms use this code, or minor variations thereof (Knight et al., 2001; Koonin and Novozhilov, 2009), to synthesize the proteins encoded by their genomes and this strongly suggests that all modern life evolved from a last universal common ancestor (LUCA) of more than 3.5 billion years ago. The code would have evolved in a pre-LUCA RNA-world of life that, at least initially, was incapable of directed protein synthesis (Atkins et al., 2011; Deamer and Szostak, 2010). In modern cells RNA remains a key component of the protein synthesis machinery (PSM). There is no consensus about the early evolution of the code, which remains an area of active conjecturing and investigation ever since Crick's frozen accident theory (Crick, 1968), see recent reviews of a stereochemical code (Yarus et al., 2009), co-evolution (Di Giulio, 2004), error minimization (Freeland et al., 2003), a combination of these theories (Koonin and Novozhilov, 2009), early roles of amino acids (Szathmáry, 1999) and of aminoacyl-tRNA synthetases (aaRSs) (de Pouplana and Schimmel, 2001), broad reviews (Knight et al., 1999; Knight and Landweber, 2000; Trifonov, 2000; de Pouplana, 2004; Di Giulio, 2005) or a short introduction (Foltan, 2008).

In mathematical coding theory (Pretzel, 2000) the differences between codons (code words) are known as Hamming distances – the number of positions that differ between codons, i.e., 0, 1, 2 or 3. This metric measures how similar or dissimilar codons are and indicates how difficult distinguishing between codons would be. For example, in the standard code amino acids are commonly encoded by several codons at 1-distance (e.g., the four Gly-codons) as the PSM cannot differentiate between them. The standard code's characteristic Hamming distances define the basic symmetry group of the code in our model. Other models, in particular binary representations of the four common nucleotides by 2-bit codes {00, 01, 10, 11} (Sánchez et al., 2005; Jiménez-Montaño, 2009) do not preserve the code's Hamming distances. Group theoretic models based either on a quantum crystal basis (Sciarrino, 2003) or on 64 dimensional irreducible representations of Lie (Antoneli et al., 2010; Bashford et al., 1998) or finite groups (Antoneli and Forger, 2011) ignore Hamming metrics. Most models directly assign amino acids to codons without consideration of the intermediate biological machinery. They are only infrequently quoted by biologists as they are mathematically abstract and divorced from biological context (Freeland et al., 2003). In our model the biology and mathematics are two sides of the same coin: the increasing specificities of the PSM correspond mathematically with symmetry breaking and partitioning of codon and tRNA sets into subsets.

On a fundamental level, for a code to convey information, an Information Gathering and Using System (IGUS) must be able to distinguish between different code words (codons), i.e., discern differences (asymmetries) between them and this amounts to breaking symmetries that render them alike (Muller, 2007) – much like distinguishing left from right by human IGUS. Nucleotide strands were common in the early RNA-world but a cellular PSM (as an IGUS) using RNA strands as an information source for building proteins had yet to develop. Thus the code's evolution is understood as a direct consequence of the evolution of the PSM, which over time increasingly differentiates between (1) codons

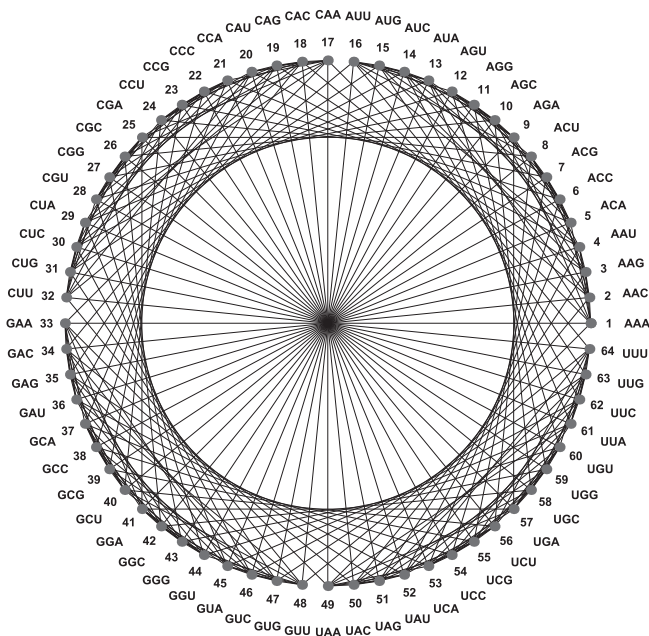
read by anticodons under ribosomal control – the code at the codon–anticodon level and (2) tRNAs and amino acids linked by primitive aaRSs – the code at the tRNA–amino acid level. Both decoding steps are interconnected by tRNAs only and, in our model, the evolution of the PSM is related with the number of tRNAs increasing over time. Evolution of precision of the translation process as key to understanding the code already was acknowledged by Woese (1965). The limited discrimination between certain codons by anticodons and between various isocoding tRNA species by aaRSs underlies the degeneracy of extant codes – the encoding of the same message by several codons. Early codes most likely were more degenerate. Modern proteinic aaRSs possess sophisticated editing mechanisms, such as double filtering for discrimination of nonpolar amino acids (Fukai et al., 2000), that distinguish their cognate from near-cognate amino acids (Ling et al., 2009). Primitive, initially RNA-based, aaRS precursors would lack such evolved specificities and differentiate less between amino acids. Similarly, modified tRNA nucleotides increase codon–anticodon recognition specificities (Agris et al., 2007), but their intricate, costly synthetic pathways and chemical differences between the three domains (Archaea, Bacteria, and Eukarya) suggest they were absent during pre-LUCA evolution (Grosjean et al., 2010). Both mechanisms illustrate that evolution towards greater molecular recognition specificities continued after the canonical code was established to ameliorate the efficiency (fidelity and speed) of translation.

The primary purpose of any code (such as the Morse code or English) is to transmit information, and the genetic code evolved to transmit instructions from an RNA genome to the PSM regarding the order of amino acids in proteins. Organisms that, early on and as fast as possible, evolved the capacity to transmit more information to build more sophisticated, better functioning proteins most likely gained an important selective advantage. In the model, selection for codes conveying more information drives progressive symmetry breaking and channels the code's evolution along relatively few, short paths through an immense space ( $> 10^{84}$ , Section 13) of all ways by which 64 codons could encode 21 messages. At each stage, evolution is constraint to explore by random variation only a limited number of relatively efficient codes, one of which equals the extant code. Thus the model provides an answer to a long standing riddle: why, did just one unique code evolve? Crick (1982) proposes directed panspermia with an organism possessing the standard code by an earlier civilization of a different planet. Instead of a LUCA bottleneck, primitive communal living organisms exchanging RNAs and evolving just one code is offered by Vetsigian et al., (2006).

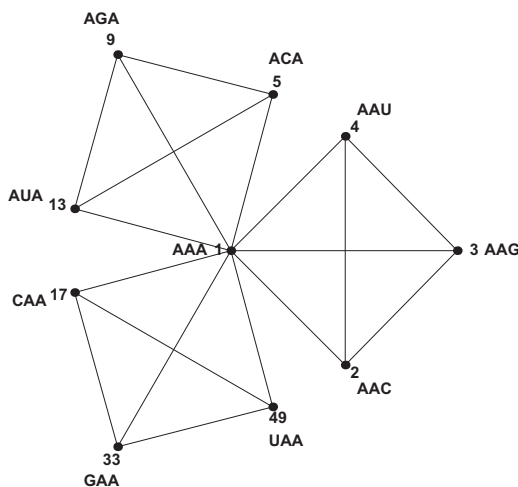
## 2. Intercodon Hamming distances define a graph representation of the genetic code

The genetic code is made up by all 64 length-3 code words, or codons, composed of a 4-letter alphabet, A, C, G, and U, representing the RNA bases, Adenine, Cytosine, Guanine, and Uracil. A single point mutation changes a codon to one of three other codons with a different base at the mutated position. All 64 codons are related via Hamming distances: the number of nucleotide differences between codons. Each codon is at one Hamming distance of nine other codons, at 2-distance of another 27 and at 3-distance of the remaining 27 codons. These relations define a 9-regular graph

comprising 64 vertices representing the codons and 288 edges connecting vertices at 1-distance: 9-regular as every vertex is incident on nine edges and 288 edges since every edge connects two vertices,  $288=64 \times 9/2$ . A circular embedding of this **Codon-Graph** is shown in Fig. 1. A subgraph of any vertex  $V_0$  and all nine vertices at 1-distance of  $V_0$  (a *closed neighborhood*) is the same for all 64 vertices and comprises three *K4-graphs* joined by *cutvertex*  $V_0$  as shown in Fig. 2. (Four vertices with six edges that link the vertices to each other form a *K4-graph* and a *cutvertex* is a unique connection between subgraphs.) As this figure shows, four codons that differ only in one position (e.g., AAA, AAC, AAG, and AAU) make up a *K4-graph* and each of the three codon positions corresponds with a different *K4-graph*. More details on Hamming distances and the CodonGraph are given in Appendix A.



**Fig. 1.** Circular embedding of labeled CodonGraph. The graph's 64 vertices are labeled with codons in lexicographical order and its 288 edges connect adjacent vertices at one Hamming distance.



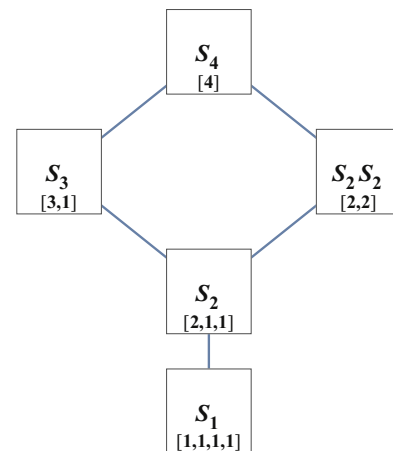
**Fig. 2.** Closed neighborhood of vertex-1 of the CodonGraph. The subgraph of the CodonGraph induced by vertex-1 and its nine adjacent vertices consists of three *K4* graphs connected by cut vertex-1 (Section 2). The vertices are labeled as in Fig. 1. Apart from the labeling, the closed neighborhoods of all 64 vertices of the CodonGraph are identical.

### 3. The symmetries of four codons differing only in one position as represented by a *K4* graph

The textbook genetic code table has 16 family boxes, sets of four codons that differ only at the third codon position, corresponding with 16 *K4*-graphs. The basic symmetry groups used in this article as illustrated by these boxes are detailed in Appendix B. When the *PSM* does not differentiate between codons, they encode the same message and such *synonymous* codons are *equivalent* for coding purposes. Sets of equivalent codons have *symmetries* that permute or exchange these codons without affecting the encoded message. For example, GAU and GAC both encode Asp and when these equivalent codons are exchanged  $GAU \leftrightarrow GAC$  the message remains Asp. The Asp 2-codon set can be ordered in two ways,  $(1, 2) = (GAC, GAU)$  and  $(1, 2) = (GAU, GAC)$ , with  $GAU \leftrightarrow GAC$  changing the order, as both correspond with  $(1, 2) = (\text{Asp}, \text{Asp})$ . The symmetry group of this 2-codon set is  $S_2$ , the *symmetric* or *permutation* group of two objects, with two orderings or permutations (including the identity permutation – not changing the order). Similarly (see Appendix B),  $S_4$  is the symmetry group of four codons at 1-distance represented by the vertices of a *K4*-graph.  $S_4$  contains *isotropy* subgroups  $S_3$ ,  $S_2$  and  $S_1$  that are the symmetry groups of subsets of the four codons:  $S_3$  of three codons or vertices of a triangle-subgraph of the *K4* graph,  $S_2$  of two codons or vertices of an edge, and  $S_1$  of one codon or single vertex. The isotropy lattice of Fig. 3 shows the stepwise breaking of the symmetries of  $S_4$ . The size or *order* of  $S_4$ ,  $S_3$ ,  $S_2$  and  $S_1$  equals respectively 24, 6, 2 and 1 – the number of permutations per group.

### 4. The symmetries of the CodonGraph are the symmetries of the stage-zero code

The three codon positions with four letters per position generate three *K4*-graphs, one for each position, as illustrated in Fig. 2, or, equivalently, three  $S_4$  permute the letters, one  $S_4$  per codon position. As the positions vary independently, a point mutation at the first position does not affect the other positions, the three  $S_4$  act independently, *i.e.*, as the direct product  $(S_4)_1 \times (S_4)_2 \times (S_4)_3$  – indexed by position 1, 2 and 3. For example, permutations  $A \leftrightarrow C$ ,  $A \leftrightarrow G$ , and  $A \leftrightarrow U$  of, respectively,  $(S_4)_1$ ,  $(S_4)_2$ , and  $(S_4)_3$ , exchange codons  $AAA \leftrightarrow CGU$ , with each individual permutation exchanging



**Fig. 3.** The isotropy lattice of  $S_4$ .  $S_4$ , the symmetry group of four codons at 1-Hamming distance or vertices of a *K4*-graph can break into subgroups,  $S_3$ ,  $S_2$  and  $S_1$ . These *isotropy* subgroups are the symmetry groups of subsets of the 4-set [4], *i.e.*, the [3], [2], and [1] subsets (Section 3 and Appendix B). The lattice is a partial ordering of the groups and subsets by inclusion, *e.g.*, an edge connects the node with  $S_3$  and subsets [3,1] with top-level node  $S_4$  and the [4] set because  $S_3$  is a subgroup of  $S_4$  and [3] and [1] are subsets of [4].  $S_1$  is omitted when possible, *e.g.*,  $S_3$  represents  $S_3 S_1$ .



codons at 1-distance – vertices of a K-graph. The direct product group is a symmetry group of the graph itself as the codon labels only serve to track the vertices – any correct labeling of the graph with codons is a symmetry of the graph, see [Appendix C](#). The three K4-graphs of [Fig. 2](#) are identical and inter-exchangeable – all assignments of three K-4 graphs to codon positions (1,2,3) are equivalent, *i.e.*, all permutations of (1,2,3) by  $S_3$  are symmetries of the graph. With  $S_3$  permuting the codon positions or K4-graphs, the full symmetry group of the CodonGraph is the wreath product  $(S_4)_1 \times (S_4)_2 \times (S_4)_3 \times_{\text{wreath}} S_3$  – all background group theory can be found in [Rotman \(1995\)](#). For example,  $S_3$  permutation  $1 \leftrightarrow 3$  exchanges the first and third codon positions, with  $\text{CGU} \leftrightarrow \text{UGC}$ , and using the direct product result  $\text{AAA} \leftrightarrow \text{CGU}$  from above, the wreath product exchanges  $\text{AAA} \leftrightarrow \text{UGC}$ . Importantly, both the direct product and the wreath product can permute any vertex into any other, *i.e.*, these symmetry groups are *transitive* on the graph and all vertices/codons are *equivalent* under these symmetries. These symmetry groups thus characterize an *initial, stage-zero code* with early RNA-world organisms not yet differentiating among any of the 64 codons. The *order* or number of symmetries of these groups equals the product of the orders of their subgroups, *i.e.*, the direct product  $(S_4)_1 \times (S_4)_2 \times (S_4)_3$  has order 13,824 ( $=24^3$ ) and the wreath product has order 82,944 ( $=24^3 \times 6$ ).

### 5. Codon–anticodon basepairing breaks $S_4$ into its isotropy subgroups

Modern ribosomes are ribozymes – RNA based machines, as their structure and function are dependent primarily on their RNA constituents and their primordial precursors of 3.5 billion years ago were likely all-RNA particles. The decoding center, located in the small ribosomal subunit, controls codon–anticodon pairing at the A-site, the acceptor site for aminoacyl-tRNAs. Three bases of the 16S RNA of this subunit measure the width of the minor groove of the short helix formed by the first two codon–anticodon nucleotides: A1493 at the first, A1492 and G530 at the middle codon position. In addition, G530 interacts with the third codon base but in a less specific manner. These interactions energetically and kinetically favor Watson–Crick base-pairing at the first two codon positions but permit wobble-pairing at the third position. The ribosomal control of codon–anticodon pairing is essential for the translation fidelity of protein synthesis with an error rate of one in 1000–100,000 amino acids. Error rates only based on energy differences between cognate and near cognate pairing are estimated at one in 10–100. (See reviews by Moore and Steitz; Noller; Ramakrishnan, in [Atkins et al., 2011](#).)

The evolution of the ribosomal decoding center in the RNA world is unknown, but we assume that the codon–anticodon base pairing stringencies evolved stepwise from pure physico-chemical interactions between RNA strands. In RNA double helices, WC pairing is observed most frequently, followed by GU-wobble pairing (a few percent) and then by rare other non-WC pairings, with the more common pairings having greater binding energies and forming more perfect helices ([Creighton, 2010](#)). As WC pairing differentiates among all four nucleotides, it partitions this 4-set into four one sets,  $[4] \rightarrow [1,1,1,1]$ , and breaks  $S_4$  into four  $S_1$  ([Appendix B](#) and [Fig. 3](#)). GU-wobble pairing differentiates pyrimidines  $Y=\{U, C\}$  from purines  $R=\{A, G\}$ , as G- and U-anticodon bases pair, respectively, with pyrimidine- and purine-codon bases. GU-wobble pairing partitions  $[4] \rightarrow [2,2]$  and breaks  $S_4$  into  $S_2S_2$ , with one  $S_2$  permuting the  $\{U, C\}$  and the other the  $\{A, G\}$  labels. The rare Inosine (I)-anticodon base, a deaminated A, wobble pairs with A-, C- and U-codon bases, partitions  $[4] \rightarrow [3,1]$  and breaks  $S_4$  into  $S_3S_1$ .  $S_4$  is realized through non-WC pairings, such as the U-superwobble, which pairs non-modified U with any codon base as observed in extant mitochondria and mycoplasma ([Agris et al.,](#)

[2007, Grosjean et al., 2010](#)). As mentioned in the introduction, modified bases present in modern cells played no role in pre-LUCA evolution of the genetic code and their later role is not considered here. Without ribosomal control, different anticodons compete for alignment with the same codon and through various base pairings realize different symmetry breakings. Thus the information transfer from mRNA codons to tRNA anticodons by pure physico-chemical interaction represents a *noisy channel* ([Cover and Thomas, 2006](#)). The noise is caused by various energetically less favored, less frequent alignments competing with the most frequent alignment. In particular, free energy differences between cognate and near cognate codon–anticodon alignments are insufficient for accurate decoding ([Ogle and Ramakrishnan, 2005](#)). Thus, ribosomal control of codon–anticodon interactions most likely evolved to suppress this channel noise.

### 6. Ribosomal identification of a unique mRNA reading frame breaks the $S_3$ -wreath product

The  $S_3$ -wreath product exchanges the three codon positions ([Section 4](#)), which renders them equivalent and their order irrelevant for coding.  $S_3$  can break into isotropy subgroups  $S_2$  and  $S_1$  ([Section 3, Fig. 3](#)). Under  $S_2$  two codon positions are equivalent and distinguished from the third, while under  $S_1$  all positions are unique. The six permutations of  $S_3$  correspond with the six ways length-3 anticodons can align on a linear RNA strand: three ways differing by one nucleotide frameshifts in both the sense and antisense direction. These alignments effectively change the order of the codon positions on the mRNA as read by anticodons. Consider an RNA strand composed of ACG repeats: due to frameshifts these align as ACG, CGA and GAC in one, and GCA, CAG and AGC in the other direction. Modern ribosomes process mRNA only in the  $5' \rightarrow 3'$  direction and, commencing at a unique start codon, permit only antisense length-3 codon–anticodon alignments. These ribosomal controls select one out of six possible reading frames and break  $S_3$  to  $S_1$ . We argue in [Section 7](#) that primitive ribosomes evolved these functions very early on. Antisense pairing is commonly observed in RNA double strands ([Creighton, 2010](#)) and likely adopted as such by primitive ribosomes. New RNA strands are transcribed  $5' \rightarrow 3'$  so that first their  $5'$ -ends become available to primitive ribosomes, which possibly therefore adopted processing mRNA  $5' \rightarrow 3'$ . Co-transcriptional mRNA translation with initiation of translation while mRNA is being transcribed is observed in modern prokaryotes ([Elliott and Ladomery, 2011](#)). Selection against frameshifting favors the *densest packing* of tRNA anticodon loops on the linear mRNA that primitive ribosomes could enforce as it leaves no wiggle room for frameshifts. We conjecture here that this packing imposed length-3 codons. Why triplet codons evolved, rather than longer or shorter ones, is a hitherto unresolved issue. The polymerization activity of ribosomes, which is entirely due to entropy effects, *i.e.*, on bringing two amino-acylated tRNAs close together, is dependent on dense tRNA packing and primitive ribosomes initially might have adopted template RNA-strands to tightly pack tRNAs rather than for coding purposes. Early on, before start codons evolved, a general run-on preference, such as for the first physical triplet of the RNA strand, could fix the reading frame, and without stop codons, run-off from RNA strands would end transcription.

### 7. Primitive ribosomes broke the $S_3$ -wreath product before the onset of the code's evolution

Symmetry breaking of  $S_4$  while  $S_3$  permutes the three codon positions generates codes that are very different from the

canonical code. Using a generic set {A, B, C, D} as four letter alphabet, and with  $S_4$  reduced to  $S_1$  so that all letters are different, such codes comprise 20 distinct code words: ABC, ABD, ACD, BCD, AAB, AAC, AAD, BBA, BBC, BBD, CCA, CCB, CCD, DDA, DDB, DDC, AAA, BBB, CCC and DDD – as letter order is irrelevant due to  $S_3$ . Breaking the  $S_3$ -wreath product to  $S_2$  generates three different codes, each with a different fixed position, of up to 40 distinct code words. For example, with  $S_2$  permuting the first two positions and X representing any of the four letters in the fixed third position: AAX, ABX, ACX, ADX, BBX, BCX, BDX, CCX, CDX, and DDX. None of these codes, using any substitution of the generic letters {A, B, C, D} by the four bases {A, C, G, U}, corresponds with the canonical code, not even approximately. Here the symmetry breaking analysis identifies the consequences of the wreath product persisting during the code's evolution and permits the rejection of this thesis. So instead, primitive ribosomes must have broken the wreath product through mechanisms discussed in Section 6 before the onset of the code's evolution. This fixed the letter order of the code, *i.e.*, distinguished codon positions one, two and three, which in turn permitted the evolution of the code as described in the sections below. Our model lends support to the hypothesis that primitive ribosomes polymerized amino acids in a random, non-directed manner before a code evolved (Fox, 2010).

## 8. Stepwise breaking of $S_4S_4S_4$ partitions the codon set into synonymous codon blocks

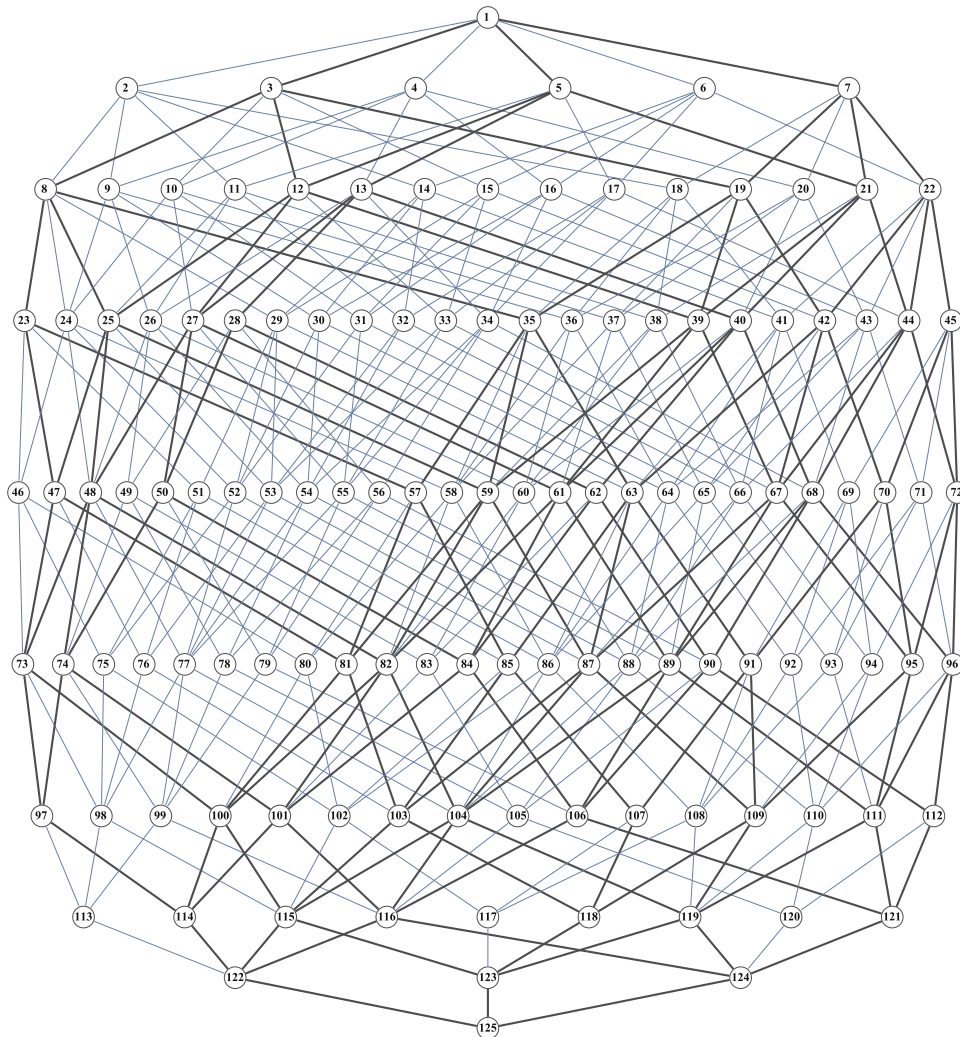
With the  $S_3$ -wreath product broken, the codon positions (1,2,3) are fixed and the direct product  $(S_4)_1 \times (S_4)_2 \times (S_4)_3$ , abbreviated  $S_4S_4S_4$ , characterizes the departure stage for the code's evolution. Under this symmetry group, all codons are equivalent and synonymous – conveying the same message (Section 4). Breaking the three  $S_4$  (Sections 3 and 5) generates the isotropy subgroups of the direct product that are symmetry groups for subsets of the 64 codons. In other words, the codon set is *partitioned* into non-overlapping subsets or *blocks* (no codon belongs to two blocks) of equivalent, synonymous codons. Each codon block can convey a different message as no isotropy subgroup permutes codons belonging to different blocks. Stepwise symmetry breaking progressively *refines* the codon set partition into smaller and smaller blocks, a process comparable to dividing a set of playing cards first into two stacks and subsequently splitting these stacks into smaller stacks and so on, but the isotropy groups permit just a few specific splits. There are just six possible first symmetry breakings: one of the three  $S_4$  breaks into either  $S_2S_2$  or  $S_3$  (Fig. 3). If, for example, the middle  $S_4$  breaks to  $S_2S_2$ , as per GU-wobble pairing, the two daughter symmetry groups  $S_4S_2S_4$  each permute 32 ( $=4 \times 2 \times 4$ ) codons, and these NYN and NRN codon blocks correspond with the left and right halves of the textbook codon table (N standing for any base). Alternatively, Inosine wobble pairing (Section 5) breaks this  $S_4$  into  $S_3$  resulting in  $S_4S_3S_4$  and  $S_4S_1S_4$ , that permute, respectively, 48 ( $=4 \times 3 \times 4$ ) and 16 ( $=4 \times 1 \times 4$ ) codons – blocks comprising, respectively, the left three columns and right column of the codon table. Thus the first symmetry breaking generates a code able to convey two messages, one for each codon block. Subsequent breakings refine these blocks and produce codes able to transmit more messages. Theoretically this process can proceed until  $S_4S_4S_4$  is reduced to 64  $S_1S_1S_1$  with 64 one-codon blocks capable of transmitting 64 messages.

Stepwise symmetry breaking traverses an enormous space ( $> 1.72 \times 10^{65}$ , Appendix D) of all possible codon set partitions via a limited number of well prescribed pathways. For example, as described above, the first breaking can generate only six out of

$\approx 9.2 \times 10^{18}$  possible two block partitions (the number of ways to divide 64 cards into two stacks). The biologically most relevant (Sections 9 and 10) pathways are depicted in Fig. 4 as an isotropy lattice with top lattice vertex-1 (LV1)  $S_4S_4S_4$  linked, one level down, via six edges to the six vertices corresponding with the six different sets of two smaller isotropy subgroups discussed above, *etc.*, all the way to bottom LV125, which represents the 64  $S_1S_1S_1$  groups. (This isotropy lattice is the Cartesian product of three  $S_4$  isotropy lattices of Fig. 3: an ordered triple XYZ with {X, Y, Z} any of  $\{S_4, S_3, S_2S_2, S_2, S_1\}$ ) The isotropy lattice comprises 125 vertices, 375 edges, and 13,440 different paths (9-step chains composed of vertices and edges) connecting top to bottom vertex. Excluding rare Inosine  $S_3$ , the lattice encompasses only 64 vertices, 144 edges and 1680 paths – highlighted in black in Fig. 4 – running from LV1 to LV125. Evolution constraint by symmetry breaking combined with selection for codes conveying more information will explore these 1680 paths preferentially over any others (Section 9). The lattice of Fig. 4 does not show all theoretically possible symmetry breakings. On the lattice, LV5 is linked to LV28, which represent, respectively, two  $S_4S_2S_4$  generated by GU-wobble pairing at the 2nd position and four  $S_4S_1S_4$  resulting from subsequent WC pairing at the same position. But LV5 also links via breaking just one of the two  $S_4S_2S_4$  to an off-lattice vertex corresponding with one  $S_4S_2S_4$  plus two  $S_4S_1S_4$  due to GU-wobble pairing at the 2nd position with a missing G- or U-anticodon causing a  $S_1$ -stop codon (see also Section 10). In general, off-lattice vertices represent isotropy groups generated by breaking similar symmetry groups at the same codon position in different ways, such as breaking one  $S_2$  in  $S_1S_1$  but not the other  $S_2$  as in the example above; or breaking the  $S_4$  of the 16  $S_1S_1S_4$  (LV112) alternatively in  $S_2S_2$ ,  $S_3S_1$ , or  $S_2S_1S_1$  as discussed in Section 3 for the 16 family boxes. Thus, in principle, the code space accessible by symmetry breaking comprises more than the 125 vertices of Fig. 4. The first four symmetry breakings under ribosomal control (Section 10) break the first two  $S_4$  but leave the third one intact and all such symmetry breakings generate 311 vertices, still a rather small space for evolution to explore. The bottom vertex of this subspace, LV112, 16  $S_1S_1S_4$ , corresponds with the 16 family boxes of the standard code (Section 3). All theoretically possible symmetry breakings of the three  $S_4$  in any order generate 40,193,906 vertices (as determined by computational enumeration). This space is possibly still searchable by random variation, but selection for codes conveying most information (Section 9) and stepwise evolution of ribosomal control of basepairing (Section 10) keeps evolution close to the paths on the isotropy lattice of Fig. 4.

## 9. Selection for encoded information drives symmetry breaking and channels evolution

Proteins are built to specifications by the genetic code and the more information a code conveys, the more protein synthesis can be fine tuned and optimized to produce better functioning proteins. Thus, during evolution of the code, when various codes potentially coexist and compete, codes transferring more information confer a selective advantage. Information can be objectively measured by Shannon entropy, a weighted average of the frequencies (probabilities) of the code's messages:  $SE = -\sum_m f(m) \log_2 f(m)$ , with  $f(m)$  the frequency of a message (Cover and Thomas, 2006) – explanatory examples are given below. Assuming that in an RNA-world all codons are equally likely, the frequency of a message corresponds with the relative number of codons transmitting the same message. Even in modern genes, amino acid frequencies (Creighton, 2010) are highly correlated with the number of codons encoding them (correlation coefficient  $\approx 0.75$ ). The relative sizes of the codon blocks of the

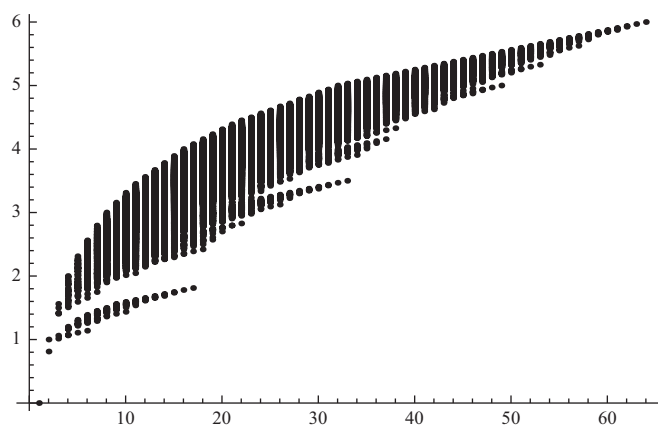


**Fig. 4.** Pathways for evolution of the codon–anticodon code: an isotropy lattice of  $S_4S_4S_4$ . This lattice is the product of three  $S_4$  isotropy lattices (Fig. 3) and shows the principal paths by which  $S_4S_4S_4$  can be broken into smaller symmetry groups (Section 8). The 125 vertices represent isotropy subgroups of  $S_4S_4S_4$  and the 375 edges various symmetry breakings. Below,  $S_n$  is abbreviated as  $n$  and indices are omitted so that  $(S_4)_1 \times (S_4)_2 \times (S_4)_3$  becomes 444,  $S_1=1$  is omitted when possible, and parentheses are used to separate groups of different index for clarity or when necessary. The 144 black edges represent symmetry breaking to equal subgroups (i.e.,  $4 \rightarrow 22$  and  $2 \rightarrow 1$ ), while the 231 gray edges represent  $4 \rightarrow 3$  and  $3 \rightarrow 2$  breaking (Section 3). The 1680 black chains represent the main evolutionary pathways for the code (Sections 9 and 10). The vertices (in bold numbers) and groups are: **1.** 444, **2.** 443, **3.** 44(22), **4.** 434, **5.** 4(22)4, **6.** 344, **7.** (22)44, **8.** 442, **9.** 433, **10.** 43(22), **11.** 4(22)3, **12.** 4(22)(22), **13.** 424, **14.** 343, **15.** 34(22), **16.** 334, **17.** 3(22)4, **18.** (22)43, **19.** (22)4(22), **20.** (22)34, **21.** (22)(22)4, **22.** 244, **23.** 441, **24.** 432, **25.** 4(22)2, **26.** 423, **27.** 42(22), **28.** 414, **29.** 342, **30.** 333, **31.** 33(22), **32.** 3(22)3, **33.** 3(22)(22), **34.** 324, **35.** (22)42, **36.** (22)33, **37.** (22)3(22), **38.** (22)(22)3, **39.** (22)(22)(22), **40.** (22)24, **41.** 243, **42.** 24(22), **43.** 234, **44.** 2(22)4, **45.** 144, **46.** 431, **47.** 4(22)1, **48.** 4(2)2, **49.** 413, **50.** 41(22), **51.** 342, **52.** 332, **53.** 3(22)2, **54.** 323, **55.** 32(22), **56.** 314, **57.** (22)41, **58.** (22)32, **59.** (22)(22)2, **60.** (22)23, **61.** (22)2(22), **62.** (22)14, **63.** 242, **64.** 233, **65.** 23(22), **66.** 2(22)3, **67.** 2(22)(22), **68.** 2(2)4, **69.** 143, **70.** 14(22), **71.** 134, **72.** 1(22)4, **73.** 421, **74.** 412, **75.** 331, **76.** 3(22)1, **77.** 3(2)2, **78.** 313, **79.** 31(22), **80.** (22)31, **81.** (22)(22)1, **82.** (22)(2)2, **83.** (22)13, **84.** (22)1(22), **85.** 241, **86.** 232, **87.** 2(22)2, **88.** 2(2)3, **89.** 2(2)(22), **90.** 214, **91.** 142, **92.** 133, **93.** 13(22), **94.** 1(22)3, **95.** 1(22)(22), **96.** 124, **97.** 411, **98.** 321, **99.** 312, **100.** (22)21, **101.** (22)12, **102.** 231, **103.** 2(22)1, **104.** 2(2)2, **105.** 213, **106.** 21(22), **107.** 141, **108.** 131, **109.** 1(22)2, **110.** 123, **111.** 12(22), **112.** 114, **113.** 311, **114.** (22)11, **115.** 2(2)1, **116.** 212, **117.** 131, **118.** 1(22)1, **119.** 1(2)2, **120.** 113, **121.** 11(22), **122.** 211, **123.** 121, **124.** 112, **125.** 111.

evolving codes (Section 8) correspond with the frequencies of their messages and the sizes are summarized by an *integer-64 partition*, commonly written between square brackets, that lists the number of codons per block. For example, LV1,  $S_4S_4S_4$ , equals [64], and the two codon blocks resulting from first symmetry breaking correspond either with [32, 32]=[32<sup>2</sup>] or [48, 16], as discussed in Section 8. The entropies of these codes are, respectively, 0 bits ( $-1 \log_2 1$ ), 1 bit ( $-2(1/2 \log_2 1/2)$ ) and 0.81 bit ( $-1(3/4 \log_2 2/4 + 1/4 \log_2 1/4)$ ). Entropies increase with the number of blocks and are greater for partitions into blocks of more equal sizes. Thus, selection for more information drives symmetry breaking with progressive refinement of the codon partitions from top LV1 representing the 0 bit [64]-code down to bottom LVX125 with the 6 bit [1<sup>64</sup>]-code.  $S_4$  breaking to  $S_2S_2$  generates equal sized blocks with greater entropies than  $S_4$  breaking

to  $S_3S_1$  and this favors the 1680 paths highlighted in black in Fig. 4. Indeed, these paths comprise the maximum entropy codes for the fewest symmetry breaking steps: 1 bit [32<sup>2</sup>], 2 bit [16<sup>4</sup>], 3-bit [8<sup>8</sup>], 4 bit [4<sup>16</sup>], and 5-bit [2<sup>32</sup>] partitions as well as the 6 bit [1<sup>64</sup>] partition. Off-lattice vertices represent unequal symmetry breakings (Section 8) and thus codes with smaller entropies for the same number of symmetry breaking steps. All 40,193,906 vertices generated by symmetry breaking correspond with 66,700 different integer partitions (as per computation), a small subset,  $\approx 3.8\%$ , of all 1,741,630 different 64-partitions (all ways for positive integers to add up to 64). The other  $\approx 96.2\%$  not generated is inaccessible to evolution by symmetry breaking. The 66,700 integer partitions define an entropy hill, shown in Fig. 5, on the accessible code space, analogous to McGhee's (2007) adaptive landscape dimension on a theoretical





**Fig. 5.** The entropy hill of the code space. Information (Shannon entropy) transmitted by codes is measured in bits (Section 9). The codes vary from 1 to 64 messages (X-axis), and the entropies from 0 to 6 bits (Y-axis). The dots, which represent the entropies of all possible codes generated by symmetry breaking, outline the entropy hill. Different codon partitions are generated by the 40,193,906 ways to break  $S_4S_4S_4$  (Section 8) and these partitions correspond with 66,700 integer-64 partitions that measure the number of codons per block (Section 9). The Shannon entropy of the codes, with one message per block, is calculated from the integer partitions by the formula given in Section 9.

morphospace. Selection drives the code's evolution up this hill from the initial 0-bit [64]-code, represented by the lower left dot of Fig. 5, to the 6-bit [1<sup>64</sup>]-code upper right dot. The pathways on the isotropy lattice of Fig. 4 represent steep climbs of this hill, and the stronger the selection for entropy, the less likely evolution strays from these climbs into the wider, flatter code space.

## 10. Ribosomal control of basepairing evolves in five steps

As argued in Section 5, modern ribosomes suppress noise in the codon-anticodon channel through the enforcement of specific basepairing rules. Intricate kinetic and structural studies of ribosomes have elucidated that this function depends entirely on RNA and might have been exercised by primordial ribosomes (Ogle and Ramakrishnan, 2005; Schmeing and Ramakrishnan, 2009). Assuming that evolution of the ribosomal decoding center gave rise to the modern basepairing rules, early codes would be characterized by less stringent rules. In principle this hypothesis could be tested: engineered mutations of the decoding center should relax the extant rules. We model a stepwise increase of ribosomal basepairing restrictions from initially none – allowing all physico-chemical pairings at all codon positions (Section 5), to finally permitting only WC-pairings. Ribosomes can only enforce more stringent rules if these do not lead to stop codons – codons not recognized by any anticodons, as such codons stunt polypeptide synthesis, especially in early codes. For example, enforcement of wobble pairing at the 2nd position, LV1 → LV5, makes all  $A_2$ -codons – 25% of all codons – stop codons if  $U_2$ -anticodons are not available, and similarly for  $C_2$ -codons if  $G_2$ -anticodons are lacking. The next step, WC-pairing at position-2, LV5 → LV28, demands the presence not only of  $\{C_2, U_2\}$  but also of  $\{A_2, G_2\}$  anticodons and subsequent steps demand larger anticodon repertoires. Over time paralogous tRNA evolution, gene duplication followed by mutation, expands tRNA diversities and paces the evolution of the basepairing rules. Our stochastic modeling of this process (Appendix E) shows that, with great probability, basepairing rules evolved from none to wobble to WC pairing initially at one codon position, then at another, and finally at the remaining position. In other words, the three  $S_4$  break sequentially via  $S_2S_2$  to  $S_1$  along

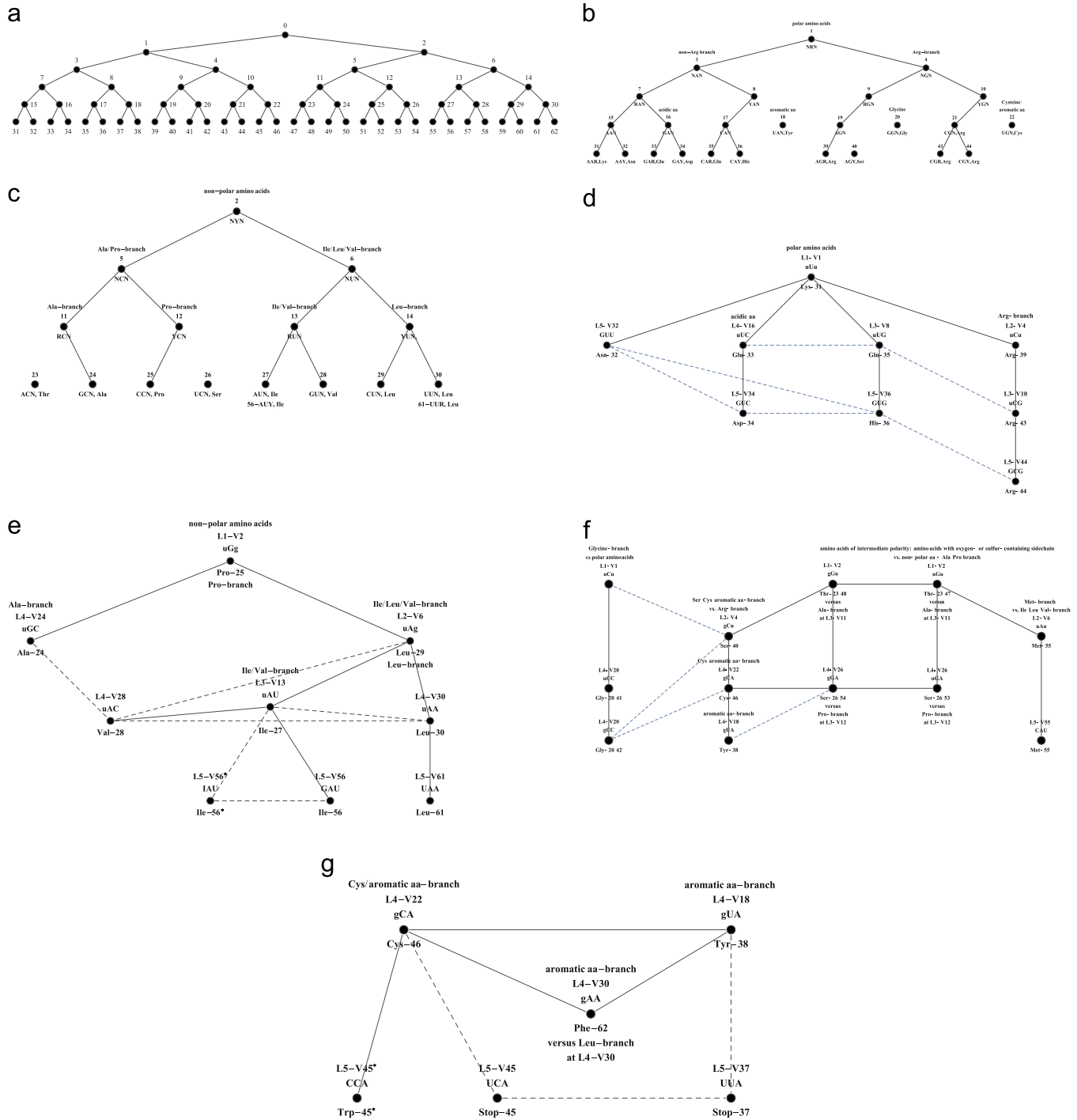
any of just six paths on the isotropy lattice of Fig. 4 – three choices for breaking the first  $S_4$  and two for the second  $S_4$ . The vertices on these paths correspond with maximal entropy codes (Section 9). With reference to the literature, Crick (1968) remarks that early codes should have few stop codons. To avoid stop codons, Van der Gulik and Hoff (2011) propose a rapid expansion of the tRNA diversity, followed by codon reassignments and non-sense suppression of unassigned codons in the presence of minimally 20 tRNAs. Barricelli (1977) suggests wobbling at all three positions in early codes and that persisting U-wobble pairing of anticodon 5'C<sub>3</sub>A<sub>2</sub>U<sub>1</sub>3' permits translation initiation at codons AUG and GUG in extant E.coli.

In our stochastic tRNA expansion model (Appendix E), the three codon positions are mathematically equivalent and chance alone (spontaneous symmetry breaking) determines the order in which the three  $S_4$  are broken. For example, if by chance, initially the  $\{U_2, G_2\}$  anticodons evolve before  $\{U_1, G_1\}$  or  $\{U_3, G_3\}$  anticodons,  $S_4$  breaks first at the 2nd position. However, for stereo-chemical reasons, pairing at the middle position is most relevant for codon-anticodon recognition (Ogle and Ramakrishnan, 2005) and initial breaking at this position might have conferred a selective advantage. Codon amino acid assignments of the standard code (Sections 11 and 12) indicate that, historically pre-LUCA ribosomes broke  $S_4$  symmetries initially at the 2nd, then at the 1st, and finally at the 3rd codon position. The third  $S_4$  broke only partially and most frequently to  $S_2S_2$  when ribosomes imposed wobble pairing at the 3rd position, but did not subsequently enforce WC-pairing. Because of these historical contingencies, symmetry breaking proceeds along a unique 5-step path on the isotropy lattice of Fig. 4: LV1,  $S_4S_4S_4$  → LV5,  $S_4(S_2S_2)S_4$  → LV28,  $S_4S_1S_4$  → LV62,  $(S_2S_2)S_1S_4$  → LV112,  $S_1S_1S_4$  → LV121,  $S_1S_1(S_2S_2)$ . (The path skips vertices LV-13 and LV-90 with  $(S_2S_1S_1)$  caused by stop codons – GU-wobbles without G- or U-anticodons.) This path progressively partitions the codon set along a binary tree from root vertex-0 representing [64] down to 32 vertices at the 5th-level corresponding with the [2<sup>32</sup>] partition (Fig. 6a). The blocks of these partitions correspond with divisions of the textbook codon table: [64] is the table, [32<sup>2</sup>] are the left and right halves, [16<sup>4</sup>] the four columns, [8<sup>8</sup>] the eight upper and lower halves of these columns, [4<sup>16</sup>] the 16 family boxes and [2<sup>32</sup>] the 32 upper and lower halves of these boxes. The 32 synonymous 2-codon blocks are observed in extant codes of vertebrate and insect mitochondria in which both Met and Trp are encoded by two codons (Grosjean et al., 2010). The canonical code displays two small differences: one  $S_4$  family box is split into  $S_2S_1S_1$  as a missing  $U_3$ -anticodon generates a stop codon in the presence of the Trp  $C_3$ -anticodon, and another box is broken to  $S_3S_1$  compatible with a rare  $I_3$ -anticodon, as observed in Eukarya, encoding Ile and a  $C_3$ -anticodon for Met. In Bacteria and Archaea chemically modified anticodons cause  $S_3S_1$ , but as the bacterial tRNA-modifying enzyme is not found in Archaea (Soma et al., 2003), these modifications are a post-LUCA development (Grosjean et al., 2010). Perhaps convergent evolution seeking to minimize the number of start (=Met) codons selected the rare  $I_3$ -anticodon and modified anticodons in the different domains. Post-LUCA the LV121 [2<sup>32</sup>] code might have evolved to the standard code, which is represented by an off-isotropy lattice vertex near LV121. In extant organisms, some chemically modified  $U_3$ -anticodons permit super wobbling (Agris et al., 2007) and restore  $S_4$  for some family boxes, such as for Ala. Lehman and Libchaber (2008) propose that modern ribosomes permit such super wobbling only when the first two codon positions are GC pairings, or if only one GC pair is present, when a purine at the middle anticodon position stabilizes the tRNA anticodon loop. This post-LUCA evolved stringency corresponds with eight  $S_4$  plus eight  $S_2S_2$  at the 3rd position or [4<sup>82</sup>16], a vertex just off the isotropy lattice between LV112 and LV121, and was probably selected for greater translation speeds (as super wobbling anticodons match four codons rather than two, they are less often rejected by ribosomes) with preservation of the canonical codon assignments.

### 11. Basepairing rules partition the codon and anticodons in matching blocks

As discussed in Section 10, ribosomal enforcement of increasingly stringent basepairing rules evolves along a 5-step path that partitions the 64 codons along a binary tree (Fig. 6a). The expanding tRNA repertoire is partitioned in parallel as, at each stage, each codon block is recognized by one or several anticodons but no anticodon recognizes more than one codon block under the basepairing rules. For example, the 1st step, GU-wobble pairing at position-2, partitions the codon set into 32 Y<sub>2</sub>- and 32 R<sub>2</sub>-codons, and, at the same time, partitions a small number of early tRNAs

(≈ 5, Appendix E) into, respectively, matching R<sub>2</sub>- and Y<sub>2</sub>-anticodon blocks, and the 2nd step, WC pairing at position-2, partitions the tRNA diversity (of ≈ 7 anticodons) into four blocks {A<sub>2</sub>, C<sub>2</sub>, G<sub>2</sub>, U<sub>2</sub>}. Especially early on the tRNA blocks are much smaller than the matching codon blocks but the expanding tRNA repertoire is partitioned along a tree as well. Through this one-codon-block to one-tRNA-block decoding, codon and tRNA blocks convey the same message at every stage: *the code at the codon-anticodon level*. The evolution of amino acid assignments of this code is studied by tracing the known assignments of the standard code back up the codon and tRNA trees. The codon trees for polar and non-polar amino acids are shown in Fig. 6b and c, and the





matching tRNA-anticodon trees in Fig. 6d and e. The parsimonious tRNA trees represent each tRNA-block with one anticodon that aligns with all codons of the matching codon block under the stage-specific basepairing rules. Since the trees for polar and non-polar amino acid assignments do not overlap after the 1st step, the code distinguishes between these amino acid classes at this stage but not between various amino acids within the same class. Early on amino acid encoding is thus necessarily *ambiguous* and while this ambiguity diminishes with each step, it fully resolves only after the 5th step. Parsimonious tRNA trees for amino acids of intermediate polarity (glycine and the amino acids with oxygen- or sulfur-group containing sidechains) and aromatic amino acids are shown in Fig. 6f and g. The topology of these trees is not self-evident, but irrespective of topology, anticodons of these trees recognize various codon blocks that are also recognized by anticodons of the polar and non-polar amino acid tRNA trees and this increases the ambiguity of early codes. For example, after the 1st step, the NYN codon block is recognized by all G<sub>2</sub>-anticodons of the early tRNA diversity but these tRNAs potentially carry non-polar amino acids (Fig. 6e) as well as amino acids of intermediate polarity (Fig. 6f), so that the NYN codons encode a mix of both amino acid classes.

## 12. Evolution of amino acid assignments by aminoacyl-tRNA synthetases

Modern aaRSs charge one cognate amino acid on one or more isocoding tRNA species – *the code at the tRNA-amino acid level*. Although proteinic aaRS probably evolved early in the RNA world, they are thought to have replaced more primitive ribozymal aaRS (de Pouplana and Schimmel, 2001) and RNA-based aaRS activities have been observed in *in vitro* experiments as reviewed by Knight and Landweber (2000). We assume that early ribozymal aaRSs evolve as a paralogous gene family with gradually increasing molecular recognition specificities and will not explicitly consider their later replacement by proteinic aaRS here. Primitive aaRS initially only distinguish between amino acid classes with similar physicochemical characteristics, in line with Woese (1965), with specificities gradually evolving to just one cognate amino acid. Presumably few abiotic amino acids existed in the early RNA world but their variety increased over time by organic synthesis (Trifonov, 2000), which likely stimulated evolution of additional

aaRS varieties. The aaRSs also increasingly differentiate, based on anticodon and other recognition elements, between tRNAs of the expanding tRNA repertoire, *i.e.*, at each stage, *each* aaRS partitions the tRNA diversity into two blocks: accepted isocoding tRNAs and rejected tRNAs. Early on these multiple partitions most likely overlap when the same tRNA species is accepted by several primitive aaRSs, which causes ambiguity at the tRNA-amino acid level. Acceptance of different tRNAs belonging to the same anticodon block by different aaRS causes ambiguity at the codon-anticodon level (Section 11). Selection for more specific codes (which convey more information or *entropy*) gradually eliminates both sources of ambiguity and imposes that, in the end, different aaRSs recognize different blocks of a single, unique tRNA partition that resembles the codon-anticodon partition enforced by ribosomes. In this final partition, a unique aaRS accepts all tRNAs of one or several anticodon blocks as observed for the standard code (16 family boxes split into two codon blocks, Sections 3 and 10): nine amino acids are encoded by one block of two codons, five amino acids by two, and three amino acids by three such blocks, while Ile, Met, and Trp are encoded by single blocks of different sizes. Selection for coding specificity favors that, at each stage, evolving aaRS accept the anticodons of the tRNA trees of Fig. 6d–g, as these match the stage-specific codon blocks, and reject all tRNAs not matching these codon blocks. For example, after the 1st step, the code only discriminates between polar and non-polar amino acids at the tRNA and codon level, if a primitive aaRS for polar amino acids accepts Y<sub>2</sub>-anticodons and rejects R<sub>2</sub>-anticodons (Fig. 6d), with opposite preferences for a non-polar amino acid aaRS (Fig. 6e). The tRNA tree of Fig. 6d suggests that, after the 2nd step, two different aaRS for polar amino acids evolved: one preferring Arg and recognizing C<sub>2</sub>-anticodons and another one for non-Arg polar amino acids and recognizing U<sub>2</sub>-anticodons. Similarly (Fig. 6e), the non-polar amino acid aaRS split into an Ala/Pro branch for G<sub>2</sub>- and an Ile/Leu/Val- branch for A<sub>2</sub>-anticodons. In this manner, the tRNA trees indicate pathways for the evolution of all extant aaRS.

Because primitive aaRS differentiate between tRNAs based on anticodons *and* non-anticodon tRNA recognition elements, they can accept different anticodons of tRNAs sharing other features. For example, after the 2nd step, the single Arg-preferring aaRS recognizes all C<sub>2</sub>-anticodons (above), but subsequently evolves to accept C<sub>2</sub>-anticodons of three different anticodon blocks {5'YCU3', YCG, RCG} as shown in the tRNA tree in Fig. 6d (no basepairing

**Fig. 6.** Tree-graphs of the codon and corresponding tRNA anticodon partitions. (a) *Binary codon partition tree*. Vertices represent codon blocks and are numbered identically in (a–c). Five symmetry breakings split the codon set in binary fashion from root vertex-0 (corresponding with the block of all 64 codons) at level-0 down to level-5 comprising 32 vertices (31–62) representing blocks of two codons. At each level, the vertices are arranged in alphabetical order: **N**=any of the four regular bases, **R**=purine, **Y**=pyrimidine. The binary partition splits N in {R, Y}, and subsequently, R in {A, G} and Y in {C, U}. Codon labels are shown in (b and c), but not in (a). (b) *Codon tree of codons encoding polar amino acids, a subtree of the codon tree*. Labeling as for (a), amino acids is abbreviated as **aa**. The standard code assignments for polar amino acids to the codons at the bottom of the tree are traced back up to the root of this tree, located at vertex 1 of the codon tree of (a). This subtree comprises, among others, branches for Arg, non-Arg polar amino acids, and the acidic amino acids. (c) *Codon tree of codons encoding non-polar amino acids, a subtree of the codon tree*. Labeling as for (b) The standard code assignments for non-polar amino acids to the codons at the bottom of the tree are traced back up to the root of this tree, located at vertex 2 of the codon tree of (a). This subtree comprises, among others, branches for Ala/Pro and Ile/Leu/Val. Vertices 56 and 61 (at level 5 of the codon tree of (a)) encoding, respectively Ile and Leu are listed below level-4 vertices 27 and 30. (d) *Parsimonious tree of tRNA anticodons for polar amino acids*. The vertices are labeled with anticodons 5' → 3' (as in G<sub>3</sub>U<sub>2</sub>U<sub>1</sub> – the order of the codon positions is 3–2–1) with nucleotides abbreviated as in (a) and amino acids as **aa**. The tree branches match those of the codon tree of (b): bases in large cap are essential for pairing with codon blocks at the indicated level and vertex of the codon tree, as in L1–V1 standing for level-1, vertex-1 of (a). Bases in small cap are chosen to parsimoniously minimize the number of vertices and edges of the tRNA tree. The amino acid encoded by the anticodon is indicated below the vertex, followed by the vertex number of the codon tree for this amino acid. For example, Lys-31 indicates that the AAR-codon block of vertex 31 (b) encodes Lys. Edges connect anticodons at one Hamming distance: solid edges parallel the branching of the codon tree; dashed edges indicate other anticodon network relations. (e) *Parsimonious tree of tRNA anticodons for non-polar amino acids*. Labeling as for (d). The tree branches match those of the codon tree of (c). Ile-anticodon uAU evolves at level 5 to either anticodon GAU or IAU, aligning, respectively, with codons AUU of vertex 56 of the codon tree (c) or codons {AUU, AUA} of vertex 56\*, not shown, of the standard code. (f) *Parsimonious tree of tRNA anticodons for amino acids of intermediate polarity*. Labeling as for (d). Many anticodons of this tree compete for codon alignment with anticodons of the trees of (d and e) as indicated by versus (vs.) followed by a general indication (such as Arg-branch) or by a specific amino acid with the lowest level and vertex of the codon tree at which the coding ambiguity persists. For example, the Thr-anticodons GGU and UGU compete with Ala-anticodon UGC (e) for codon binding down to the L3–V11 RCN-codon block (c). This coding ambiguity resolves at level-4. The Met-branch anticodons UAU and CAU match, respectively, codon block AUR of vertex 55 of the codon tree (Fig. 6a) or the single codon AUG of vertex 55\* (not shown) of the standard code. (g) *Parsimonious tree of tRNA anticodons for aromatic amino acids*. Labeling as for (f). Section 12 conjectures why the three anticodons {GAA, GCA, GUA} encode aromatic amino acids at level 4. The aromatic amino acid-anticodon GCA competes with the Cys-anticodon GCA (f) for binding to codon block UGN of level-4 vertex 22 (b). This coding ambiguity resolves at level-5 with anticodon GCA encoding Cys and anticodon UCA encoding Trp (as in mitochondria) or, alternatively, with UCA missing, CCA encoding Trp as in the standard code. Trp-anticodon UCA aligns with codon block UGR of vertex 45 (a); Trp-anticodon CCA only with codon UGG (vertex 45\*, not shown) so that the missing UCA-anticodon generates the stop codon UGA of the standard code. Missing anticodons UUA and CUA correspond with the stop codon block UAR of vertex 37 (a) and the Tyr-anticodon GUA with codon block UAY of vertex 38.

rules can select these three anticodons so other tRNA-recognition elements are required). The parallel evolving specificity for just one cognate amino-acid can lead to the exclusion of other amino acids from peptide synthesis, e.g., the Arg-preferring aaRS might accept similar amino acids, such as citrulline and ornithine, at early but not at later stages. The tRNA tree for amino acids containing sulfur/oxygen groups (Fig. 6f) suggests that the aaRS family with a preference for this class of amino acids evolves a C<sub>2</sub>-anticodon Ser/Cys/Tyr branch that, by the 4th step evolves an aaRS for aromatic acids that incurs a rare mutation to I<sub>2</sub> (from G<sub>2</sub>) at its anticodon recognition site. As a result, this aaRS accepts 5'G<sub>3</sub>A<sub>2</sub>A<sub>1</sub>3', GU<sub>2</sub>A, and GC<sub>2</sub>A anticodons, but NAA and NCA anticodons are also recognized by, respectively, Leu- and Cys-aaRSs (Fig. 6g). The resulting coding conflicts are resolved by the 5th step. That these events occurred in the final steps is in agreement with the hypothesis that Cys and the aromatic amino acids arose relatively late (Trifonov, 2000) and the close relationship still observed for Phe- and Tyr-tRNAs in slowly evolving Archaea (Xue et al., 2003). Also at the final, 5th step, stop codons evolve within the aromatic amino acid tRNA tree (Fig. 6g), possibly due to the rarity of charged Tyr- and Trp-tRNAs causing prolonged pausing and pairing with uncharged tRNAs. Extant peptide release factors resemble non-acylated tRNAs and probably evolved post-LUCA as they differ between the three domains (Atkins et al., 2011). Modern proteinous aaRS, which likely evolved from early aaRS via ribozyme-peptide complexes (de Poupplana and Schimmel, 2001; de Poupplana, 2004), also offer support for the above scenarios. For example, the first branching of the tRNA trees for non-polar amino acids (Fig. 6e) coincides with the specificities of different modern aaRS classes: Class-1a are specific for {Leu, Ile, Val} and Class-2a for {Pro, Ala}. Similarly, an aaRS Class 1a enzyme is specific for Arg, the first branch of the polar amino acid tree (Fig. 6d), while non-Arg polar amino acids are substrates of other classes. Taken together the codon and tRNA trees illustrate how the code's initial ambiguity gradually resolves through symmetry breaking and why, in general, physico-chemically similar amino acids are encoded by neighboring codon blocks. After the 5th step, ribosomes differentiate between 32 anticodon blocks that can convey up to 32 messages, but the LUCA code comprises only 21 messages. Possibly ribosomal symmetry breaking and aaRS specificities evolved so fast that no additional aaRS, such as for selenocysteine or pyrrolysine, both incorporated in extant proteins, were able to evolve within the allotted time span. The theoretically possible 6th symmetry breaking step, WC pairing at the 3rd codon position, only slows down translation (63 out of 64 tRNAs are rejected at each codon) without conveying more than 21 messages, and thus faces negative but no positive selection and is not observed.

### 13. Discussion

As detailed in Sections 8–12, the model constrains evolution to sample relatively few paths through an immense code space of more than 10<sup>84</sup> different codes. (Appendix D provides numerical information.) Relatively few historical contingencies influenced the final outcome, among which the 2–1–3 order of the codon positions at which primitive ribosomes imposed base pairing rules (Section 10), and a rare Inosine anticodon for Ile that broke one S<sub>4</sub> into S<sub>3</sub> instead of S<sub>2</sub>S<sub>2</sub> as well as a missing anticodon that broke one S<sub>2</sub> into S<sub>1</sub> with a stop codon but these were possibly post-LUCA events (Section 10). The assignment of amino acids to tRNAs was deduced from those of the standard code (an extraneous input for the model) as illustrated by the tRNA trees of Fig. 6. Therefore the mathematical model prescribes how random variation evolves relatively few competing codes in pre-LUCA organisms, one of

which passed on via the LUCA bottleneck to extant organisms as standard code.

Our model is compatible with various conjectures about the role of the early ribosomes, tRNAs and aaRSs, as highlighted in Sections 10–12, but it presents in particular an alternative to the *error minimization conjecture* (Freeland et al., 2003) which holds that selection for reduced mutation loads (as opposed to selection for entropy in our model) determines the code structure. The impact of some point mutations at the 3rd, a few at the 1st and hardly any at the 2nd codon position is muted in the standard code, in agreement with the sequential acquisition of coding at positions 2–1–3 (Massey, 2006), which, unrelated to error minimization, is the order of symmetry breaking of S<sub>4</sub> in our model (Section 10). Freeland et al. (2003) cite statistical evidence for their thesis, in particular, that the genetic code was less sensitive to mutations than a million random codes (Freeland and Hurst, 1998), while Sella and Ardell (2006) provide support with code-gene co-evolution computer simulations. These findings are *not* shared by others: computational searches found less mutation-sensitive codes than the standard one (Goldman, 1993; Buhrman et al., 2011) and it was argued that random natural selection should be able to find such codes (Di Giulio, 2000). Frequently, in up to 22% of the simulations, simple models that sequentially add similar amino acids to similar codons generate codes that minimize errors as well as or better than the standard code (Massey, 2008). The standard code is only a relatively low, unstable peak of a rugged fitness landscape of error minimization (Koonin and Novozhilov, 2009) and genetic algorithm simulations readily find better codes (Santos and Monteagudo, 2010). We note that the value of any *supportive* evidence from computer simulations is rather limited as even modern computers cannot randomly explore the immense theoretical code space, e.g., one million is an insignificant sample of 10<sup>84</sup>. No biological evolution, especially acting on early ambiguous codes, of error minimized codes has yet been proposed. In mathematical coding theory (Pretzel, 2000) Hamming distances determine error sensitivities. Codes only detect a single error when the minimum Hamming distance between codons equals two – so that a single point mutation creates a codon not encoding an amino acid or stop signal, but such a code, for example {AAA, CCA, GGA, UUA, CAC, GAG, UAU, ACC, AGG, AUU}, could transmit only 10 different messages. For correction of a single error a minimal 3-distance is required, but then only four messages can be conveyed, for example by {AAA, CCC, GGG, UUU}. Since the standard code conveys 21 messages, single errors are neither detected nor corrected.

The code's Hamming distances also are *not* optimized to reduce the impact of errors by other means: (1) As a single error changes a codon to any of nine codons at 1-distance, error minimization dictates that these nine neighbors encode, as much as possible, the same or similar amino acids, but, on average, 6.84 (76%) of the 1-distance neighbors encode different messages, among which always some physico-chemically very different amino acids or stop codons. (2) Optimally, synonymous codons should be at 1-distance, as in the K4-graphs of Fig. 2, so that as many as possible single errors do not change coding. With only 16 K4-graphs (related to variation at the third position, Section 3) but 21 messages, none of the amino acids should be encoded by more than four codons but three amino acids are encoded by six codons in the standard code. Remarkably, the three stop codons are not within 1-distance of each other, and (3) one stop codon instead of the canonical three would limit the greatest impact of a single error – chain termination. (4) Hamming distances between codons encoding similar amino acids should be minimized but they are not, for example, among the codons encoding the most hydrophilic amino acids, four Arg-codons are at maximum 3-distance from the two Asp-codons. (5) If the most extreme hydrophobic or hydrophilic amino acids were encoded by single codons with

intermediate physico-chemical characteristics at 1-distance then point mutations would change these characteristics only gradually, but this is not observed in extant codes. Therefore, as the canonical code actually is very sensitive to point mutations, life had no choice but to evolve a communication channel comprised of high-fidelity components. And it did: the ribosomal selection of the correct anticodon (Sections 5 and 10) combined with the accurate amino-acylation of cognate tRNAs by aaRSs (Section 12) limit the incorporation of wrong amino acids to one in  $10^3$ – $10^4$  (Ogle and Ramakrishnan, 2005). In our model, the code is shaped by the evolution of the specificities of these components in pre-LUCA organisms.

## Acknowledgments

The author thanks his spouse emeritus professor Toni Claudio for fruitful discussions and critical proofreading of the manuscript and professor Veit Witzeman at the Max Planck Institute for Medical Research in Heidelberg as well as the Wellcome Foundation in London for library access. The revision of the article greatly benefitted from helpful comments by the editor and reviewers of the journal.

## Appendix A. Hamming distances and the CodonGraph

The InterCodon Distance Matrix of Supplemental material 1 shows the Hamming distances between the 64 codons (Section 2) as a  $64 \times 64$  matrix with the codons indexing the columns and rows and with the distance between them as matrix entries. Since the CodonGraph is 9-regular (Section 2) certain subgraphs are readily enumerated. Each vertex is incident on three K4 graphs (Fig. 2), and as each K4 graph comprises four vertices, there are 48 K4 subgraphs ( $48 = 64 \times 3/4$ ). Similarly, each vertex is connected with 27 codons at 2-distance via 432 square subgraphs ( $64 \times 27/4$ ) and with 27 codons at 3-distance via 216 cube subgraphs ( $64 \times 27/8$ ), see Supplemental material 2. As it takes only three point mutations to change a codon into any other codon, the graph has 3-width, and these subgraphs depict all Hamming relations.

## Appendix B. Permutation groups and symmetries of codon sets of the canonical code

The textbook genetic code table is partitioned into 16 sets of four codons that differ only at the third codon position, *i.e.*, the four codons are at Hamming 1-distance from each other. Eight of these 16 sets or family boxes encode single amino-acids. For example, {GGA, GGC, GGG, GGU} all encode Gly, as the PSM does not distinguish between them. These four Gly-codons are equivalent for coding purposes as they map to the same message, *e.g.*, GGA → Gly. By contrast, the codons GGA and GAA are not equivalent as GAA → Glu. Because of the equivalency of the four Gly codons, they can be listed in any order in the Gly-family box, or assigned in any order to the vertices of a K4-graph as in Fig. 2. (The four vertices of a K4 are neighbors and represent four codons differing at only one codon position, Section 2.) All orders of the Gly-codon set → (1,2,3,4) also map Gly to all positions, *e.g.*, both the alphabetical order, (1, 2, 3, 4)=(GGA, GGC, GGG, GGU) and its reverse (GGU, GGG, GGC, GGA) do not change the box or the coding of the vertices, (1, 2, 3, 4)=(Gly, Gly, Gly, Gly). All rearrangements of the order of the codons that do not affect the coding are symmetries of the codon set. The Gly-codon set can be ordered in 24 ways: four codon choices for the first position/vertex 1, three for vertex 2, two for vertex 3 and one for vertex 4:  $4 \times 3 \times 2 \times 1 = 4!$  (4-factorial) = 24 choices. These 24 rearrangements or permutations constitute

the symmetric group-4 or  $S_4$ , the mathematical group of all permutations of four objects, which is thus the symmetry group of the eight family boxes encoding single amino-acids. Not coincidentally,  $S_4$  is the symmetry group of a K4 graph: its four vertices are equivalent, adjacent vertices (connected via one edge) to all three other vertices – exchanging vertices does not alter the graph. Similarly, the three codons, {AUA, AUC, AUU}, encoding Ile are equivalent and their symmetry group is  $S_3$ , comprised of all six ( $3 \times 2 \times 1 = 3!$ ) permutations of three objects, the symmetry group of a 3-vertex triangle-graph. Sets of two codons encoding the same amino acid, such as {GAA, GAG} for Glu, are comprised of two equivalent codons, with symmetry group  $S_2$ , the symmetry group of a 2-vertex edge. The unique {AUG} Met-codon has the  $S_1$  symmetry of a single vertex. These codon sets are subsets of 4-codon family box sets: [4] (the number of codons per set) can be partitioned into [3,1], [2,2], [2,1,1] and [1,1,1,1]. The corresponding graphs are sub-graphs of the K4 graph: [3,1] corresponds with a triangle and single vertex, [2,2] with two edges, [2,1,1] with an edge and two vertices and [1,1,1,1] with four vertices. The smaller symmetric groups are subgroups of  $S_4$ , which comprises all permutations of its subsets. For example, the four codons of the Ile+Met [3, 1] box, (1, 2, 3, 4)=(AUA, AUC, AUG, AUU)=(Ile, Ile, Met, Ile) has  $S_3$  as symmetry group for the triangle formed by vertices {1,2,4} encoding Leu, and  $S_1$  for the single vertex {3} encoding Met. All six rearrangements of the three Leu-codons on triangle vertices {1,2,4} are symmetries of the Ile+Met codon set, but all other rearrangements of  $S_4$  that assign AUG to any of these three vertices are not; for example, the exchange AUA ↔ AUG would change the vertex coding to (1, 2, 3, 4)=(Met, Ile, Ile, Ile), unequal to the original ordering (Ile, Ile, Met, Ile). In this [3,1] box,  $S_4$  breaks into  $S_3 + S_1$  subgroups, which are isotropy or stabilizer subgroups – symmetry groups of subsets. The [2,2] partition corresponds with the symmetry breaking of  $S_4$  into  $S_2 + S_2$ , [2,1,1] with  $S_2 + S_1 + S_1$  and [1,1,1,1] with four  $S_1$ . Similarly,  $S_3$  breaks into  $S_2 + S_1$  and  $S_2$  into two  $S_1$ . The progressive, stepwise symmetry breaking of  $S_4$  into its smaller isotropy subgroups is shown in Fig. 3 as an  $S_4$  isotropy lattice. In conclusion, symmetry breaking of  $S_4$  corresponds with the partitioning a set of four codons represented by a K4 graph into codon subsets encoding different messages. Symmetry breaking is the mathematical formalism for differentiation between codons (which creates codon subsets) by the protein synthesis machinery and progressive, stepwise symmetry breaking models a gradual evolution of the capacities of the PSM to distinguish between codons.

## Appendix C. Symmetries of the CodonGraph

Section 4 identifies the wreath product  $(S_4)_1 \times (S_4)_2 \times (S_4)_3 \times_{\text{wreath}} S_3$  with order 82,944 as the symmetry group of the CodonGraph. As a check, the order should, and does, equal the number of ways the vertices of the unlabeled graph can be labeled with codons: using Fig. 2, there are 64 ways to label cutvertex  $V_0$  with any codon, then 9 ways for the first vertex at 1-distance one, then 2 and 1 ways for the two vertices belonging to the same K4 graph, subsequently 6 ways for the first vertex of a second K4 graph, etc. –  $64 \times (9 \times 2 \times 1) \times (6 \times 2 \times 1) \times (3 \times 2 \times 1) = 82,944$ . These first 10 labels fix the codon assignments for all remaining vertices via square- and cube-subgraph adjacency relations described in Appendix A.

## Appendix D. The theoretical code space comprises more than $10^{84}$ different codes

The math in this appendix describes the theoretical code space but is not used for the model. The genetic code maps 64 codons



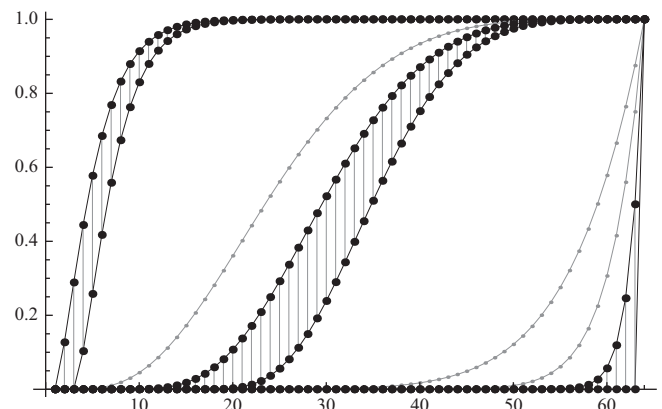
onto 21 messages 21 messages – 20 amino-acids and a stop-signal. Mathematically there are  $\approx 1.51 \times 10^{84}$  ways for 64 codons to encode 21 messages so that every message is encoded at least by one codon but no codon encodes more than one message. The number of these  $64 \rightarrow 21$  onto mappings or surjections is given by  $\sum_{j=0}^{21} C(21, j)(-1)^j(21-j)^{64}$ , Mazur (2010) gives an elementary derivation, – the binomial coefficient or combinations  $C(21, j)$  equals the number of ways to take  $j$ -messages from the total of 21 (formula in Appendix F). These surjections are 36% of all  $21^{64} \approx 4.2 \times 10^{84}$  possible  $64 \rightarrow 21$  mappings with 21 choices for each codon as most of these mappings do not reach one or more of the 21 messages. For example, all  $20^{64} \approx 1.8 \times 10^{83}$  mappings of  $64 \rightarrow 20$  miss at least one message and there are  $C(21, 20)=21$  ways to pick these 20 messages among 21. The formula above corrects for these deficient mappings: the first term ( $j=0$ ) of the sum equals all  $21^{64}$  mappings, the second term ( $j=1$ ) deducts all mappings that miss one or more messages  $=21 \times 20^{64}$ , then the third term adds back mappings that miss two or more messages as these were double counted by the second term, etc. In our model, the number of encoded messages gradually increases from one to 21, so that the total size of the theoretical code space traversed by evolution equals  $\approx 1.60 \times 10^{84}$  (the number of messages is varied from 1 to 21 in the above formula and results are summed.)

The number of partitions of the codon set (Section 8) equals the number of onto mappings divided by the number of permutations of the messages (as each message can be assigned to any block of the partition, but no two messages to the same block.) For example, for an onto mapping of 64 cards to 21 numbers, one first divides the cards into 21 different stacks – the partition, and then one can assign a number to these stacks in  $21!$  ways (with  $21! = 21$ -factorial the number of permutations of 21 numbers  $=21 \times 20 \times \dots \times 1$ ). The number of partitions of the codon set into 21 blocks thus equals  $(1/21!) \times$  the number of 21-message codes  $\approx (5.11 \times 10^{-19}) \times (1.51 \times 10^{84}) = 2.96 \times 10^{64}$ . (Enumerations of set partitions are known as Stirling numbers of the second kind (Mazur, 2010).) All possible partitions of the codon set (from one to 64 different blocks) number  $\approx 1.72 \times 10^{65}$  (the Bell number for 64). Symmetry breaking under ribosomal control evolves along a single pathway through this enormous space to a final partition of the 64 codons in 32 codon blocks (Section 10). Anticodon and tRNA diversities are partitioned at the same time (Section 11), and with selection favoring less ambiguous codes, coevolving aaRSs adopt the anticodon partitions imposed by the ribosomes (Section 12). The aaRSs do not randomly assign amino acids to the tRNA blocks, but again gradual symmetry breaking, i.e., the progressive differentiation among tRNAs and amino acid classes by aaRSs, channels evolution along a restricted number of possible paths as illustrated by the tRNA trees of Fig. 6 (Section 12). These paths map the 32-block codon partition onto the 21 messages in just one way out of  $\approx 2.42 \times 10^{39}$  possible  $32 \rightarrow 21$  onto mappings (substitute 32 for 64 in the formula counting onto mappings).

#### Appendix E. Ribosomal enforcement of the basepairing rules depends on tRNA-diversity

Ribosomes can only enforce basepairing rules without causing stop codons if all codons can pair with anticodons under these rules. As mentioned in Section 10, GU-wobbling at one codon position demands a set of at least two anticodons with {G,U} at that position and WC-pairing four anticodons {A,C,G,U}. The combination of two GU-wobbles needs anticodons with four {GG,UG,GU,UU}, a GU-wobble and a WC pairing with eight {GA, GC,GG,GU,UA,UC,UG,UU} and two WC pairings with all 16 dinucleotides. And so on for combinations of three basepairing rules: 8 trinucleotide anticodons for three wobbles, 32 for wobble plus

two WC and 64 for three WC pairings. Therefore primitive ribosomes can enforce increasingly stringent basepairing rules as the number of tRNA paralogs increases from initially just one to 64 anticodons made up by the common four RNA bases – we ignore rare Inosine-anticodons here. The likelihood that a random sample of the 64 tRNAs comprises the required anticodons can be calculated exactly using the generalized hypergeometric distribution, an elementary probability distribution also used for card counting as detailed in Appendix F. The probability that the minimal set of two {U,G} anticodons required for the first step are present increases from zero for one tRNA to virtually hundred percent for a diversity of 15 tRNAs and is shown as the leftmost curve of Fig. 7. As this curve shows, when the number of tRNAs increases to five tRNAs, the probability that ribosomes can enforce the GU-wobble pairing without causing stopcodons increases to over 50%, so that this first step (Section 10) becomes likely after just a few tRNA paralogs evolve. As the tRNA diversity increases further to seven different anticodons, the probability of WC pairing (replacing the previous GU-wobble) increases to over 50%, as shown by the second leftmost curve. Importantly, the likelihood of WC pairing as second step (breaking  $S_2S_2$  to  $S_1$ ) is much greater than the probabilities for a second GU-wobble (i.e., breaking a second  $S_4$  to  $S_2S_2$ ) shown by the third curve from the left – colored gray as this two GU-wobble configuration can no longer evolve after the first WC-pairing. Similarly, the likelihoods for WC+wobble pairing and two WC pairings, which correspond, respectively, with the third and fourth black curves from the left, are much greater than the probabilities for alternative basepairing rules shown by the far right gray curves. The probabilities for wobble+two WC and three WC pairings are given by the two rightmost black curves. These rules demand, respectively, 63 and 64 tRNAs to



**Fig. 7.** Probabilities for base pairing stringencies at one, two or three codon positions. The probabilities that ribosomes can enforce more stringent base pairing rules without causing stop codons increases as the tRNA diversity expands over time (Section 10 and Appendix E). The Y-axis shows probabilities (0–1); the X-axis the number of tRNAs (1–64). The probability that a set of tRNAs of a certain size (as shown on X-axis) contains anticodons required for various basepairing stringencies without causing stop codons is calculated using the generalized hypergeometric distribution (Appendix F). From left to right, the six black curves represent the probabilities for: 1. GU-wobble pairing at one codon position, 2. WC pairing at one position, 3. WC pairing at one position plus wobble pairing at a second position, 4. WC pairing at two positions, 5. WC pairing at two positions plus wobble pairing at the third position, 6. WC pairing at all three positions – note the edge connecting (63, 0) with (64, 1). To facilitate visual comparison, vertical fill lines connect probabilities for wobble and WC pairings at same codon position for the same number of tRNAs. The three gray curves, from left to right, correspond with wobble pairing at two positions, at three positions, and, with wobble pairing at two positions plus WC pairing at the third position. As the anticodon diversity expands, the base pairing stringencies corresponding with the black curves are enforced by the ribosomes in six successive steps (1–6, corresponding with the curve numbers) so that thereafter alternative stringencies, shown by the gray curves, can no longer evolve (Section 10 and Appendix E). The extant codon–anticodon pairing rules evolve with the first five steps out of the theoretically possible six steps (Section 10).



avoid stop codons with  $\geq 50\%$  probability, but missing anticodons at these last stages cause few stop codons and interfere less with protein synthesis. In summary, according to this stochastic model of tRNA evolution, ribosomal control of codon–anticodon recognition broke the three  $S_4$  sequentially to  $S_1$  (and historically, the 2nd  $S_4$  broke first to  $S_1$ , then the 1st  $S_4$ , while lastly the 3rd  $S_4$  broke to  $S_2S_2$ , Section 10), rather than breaking them in any other order, such as for example, one  $S_4$  to  $S_2S_2$  in step 1 and another  $S_4$  to  $S_2S_2$  in step 2.

#### Appendix F. Hypergeometric distribution for tRNA anticodon samples

The number of ways a sample of size  $N$  can be taken from the 64 different anticodons composed of the four regular RNA bases is given by the binomial coefficient or combinations  $C(64,N)=64!/((64-N)! \times N!)$ , with  $N!$  or  $N$ -factorial  $=N \times (N-1) \times \dots \times 2 \times 1$ , e.g., two tRNAs can be taken in  $C(64,2)=64!/((62! \times 2!))=64 \times 63/2=2016$  ways. The set of anticodons can be divided into different classes – similar to dividing a pack of cards into four different suits: hearts, diamonds, spades, and clubs. Looking at a single codon position, the 64-anticodon set has 4 different classes {A, C, G, U} of 16 anticodons, while two positions define 16 classes of four anticodons {AA, AC, etc.}, and three positions define 64 classes of just one codon each {AAA, AAC, etc.}. A random sample of tRNAs may or may not have representatives of all of these classes – as a hand of cards may or may not have any hearts. One can take two tRNAs from the 16 {A} class tRNAs in 120 ways:  $C(16,2)=16!/((14! \times 2!))=16 \times 15/2=120$ . The probability that two tRNAs taken from the 64 set are both of the {A} class equals the number of ways to take the two {A} tRNAs divided by the total number of ways to take two tRNAs from the total of 64 = *positive outcomes/total outcomes* =  $120/2106=5.7\%$ . In general, the probabilities that a sample of  $N$  tRNAs comprise  $n_j$  representatives of class  $j$  with class size  $K_j$  is given by the generalized hypergeometric distribution  $p(N)=C(K_1,n_1) \times C(K_2,n_2) \times \dots \times C(K_j,n_j)/C(64,N)$ . This distribution exactly describes the probabilities related to sampling (without replacement) of small populations, such as the 64-codon set or a 52-card pack (Epstein, 1977). As mentioned in Section 10 and Appendix E, different representatives of certain tRNA classes are required to avoid stop codons under various ribosomal basepairing rules. For example, for GU-wobble pairing at one position at least one representative of both the G- and the U-class is required. The probabilities that samples varying from 1 to 64 tRNAs contain representatives of both these classes are plotted as dots for the left-most curve in Fig. 7, with similar calculations for the other curves.

#### Appendix G. Supporting information

Supplementary data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.jtbi.2014.01.002>.

#### References

- Agris, P.F., Vendeix, F.A.P., Graham, W.D., 2007. Review. tRNA's wobble decoding of the genome: 40 years of modification. *J. Mol. Biol.* 366, 1–13.
- Antoneli, F., Forger, M., 2011. Symmetry breaking in the genetic code: finite groups. *Math. Comput. Model.* 53, 1469–1488.
- Antoneli, F., Forger, M., Gaviria, P.A., Hornos, J.E.M., 2010. On amino acid and codon assignment in algebraic models for the genetic code. *Int. J. Mod. Phys. B* 24, 435–463.
- Atkins, J.F., Gesteland, R.F., Cech, R. (Eds.), 2011. *RNA Worlds*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Bashford, J.D., Tsohantjis, I., Jarvis, P.D., 1998. A supersymmetric model for the evolution of the genetic code. *Proc. Natl. Acad. Sci. USA* 95, 987–992.
- Barricelli, N.A., 1977. On the origin and evolution of the genetic code. I. Wobbling and its potential significance. *J. Theor. Biol.* 67, 85–109.
- Buhrman, H., van der Gulik, P.T.S., Kelk, S.M., Koolen, W.M., Stougie, L., 2011. Some mathematical refinements concerning error minimization in the genetic code. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8, 1358–1372.
- Cover, T.M., Thomas, J.A., 2006. *Elements of Information Theory*. Wiley-Interscience, John Wiley & Sons, Inc., Hoboken, New Jersey.
- Creighton, T.E., 2010. *The Biophysical Chemistry of Nucleic Acids & Proteins*. Helvetian Press. ([www.HelvetianPress.com](http://www.HelvetianPress.com)).
- Crick, F.H.C., 1968. The origin of the genetic code. *J. Mol. Biol.* 38, 367–379.
- Crick, F., 1982. *Life Itself*. Futura Publications, London.
- Deamer, D., Szostak, J.W. (Eds.), 2010. *The Origins of Life*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- de Pouplana, L.R., Schimmel, P., 2001. Aminoacyl-tRNA synthetases: potential markers of genetic code development. *Trends Biochem. Sci.* 26, 591–596.
- de Pouplana, L.R. (Ed.), 2004. *The Genetic Code and the Origin of Life*. Kluwer Academic/Plenum Publishers, New York, NY.
- Di Giulio, M., 2000. Genetic code origin and the strength of natural selection. *J. Theor. Biol.* 205, 659–661.
- Di Giulio, M., 2004. The coevolution theory of the origin of the genetic code. *Phys. Life Rev.* 1, 128–137.
- Di Giulio, M., 2005. The origin of the genetic code: theories and their relationships, a review. *BioSystems* 80, 175–184.
- Elliott, D., Ladomery, M., 2011. *Molecular Biology of RNA*. Oxford University Press, Oxford.
- Epstein, R.A., 1977. *The Theory of Gambling and Statistical Logic*. Academic Press, San Diego, California.
- Foltan, J.S., 2008. tRNA genes and the genetic code. *J. Theor. Biol.* 253, 469–482.
- Fox, G.E., 2010. Origin and evolution of the ribosome. In: Deamer, D., Szostak, J.W. (Eds.), *The Origins of Life*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, pp. 261–278.
- Freeland, S.J., Hurst, L.D., 1998. The genetic code is one in a million. *J. Mol. Evol.* 47, 238–248.
- Freeland, S.J., Wu, T., Keulmann, N., 2003. The case for an error minimizing standard genetic code. *Orig. Life Evol. Biosph.* 33, 457–477.
- Fukai, S., Nureki, O., Sekine, S., Shimada, A., Tao, J., Vassilyev, D.G., Yokoyama, S., 2000. Structural basis for double-sieve discrimination of L-Valine from L-Isoleucine and L-Threonine by the complex of tRNA<sup>Val</sup> and Valyl-tRNA Synthetase. *Cell* 103, 793–803.
- Goldman, N., 1993. Further results on error minimization in the genetic code. *J. Mol. Evol.* 37, 662–664.
- Grosjean, H., de Grécy-Lagard, V., Marck, C., 2010. Review. Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. *Febs Lett.* 584, 252–264.
- Jiménez-Montaño, M.A., 2009. The fourfold way of the genetic code. *BioSystems* 98, 105–114.
- Knight, R.D., Freeland, S.J., Landweber, L.F., 1999. Review. Selection, history and chemistry: the three faces of the genetic code. *Trends Biochem. Sci.* 24, 241–247.
- Knight, R.D., Landweber, L.F., 2000. Minireview. The early evolution of the genetic code. *Cell* 101, 569–572.
- Knight, R.D., Freeland, S.J., Landweber, L.F., 2001. Rewiring the keyboard: evolvability of the genetic code. *Nat. Rev. Genet.* 2, 49–58.
- Koonin, E.V., Novozhilov, A.S., 2009. Origin and evolution of the genetic code: the universal enigma. *Life* 61, 99–111.
- Lehman, 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.
- Ling, J., Reynolds, N., Ibba, M., 2009. Aminoacyl-tRNA synthesis and translational quality control. *Annu. Rev. Microbiol.* 63, 61–78.
- Massey, S.E., 2006. A sequential “2–1–3” model of the genetic code evolution that explains codon constraints. *J. Mol. Evol.* 62, 809–810.
- Massey, S.E., 2008. A neutral origin for error minimization in the genetic code. *J. Mol. Evol.* 67, 510–516.
- Mazur, D.R., 2010. *Combinatorics, A Guided Tour*. The Mathematical Association of America Inc, Washington, DC.
- McGhee, G.R., 2007. *The Geometry of Evolution, Adaptive Landscapes and Theoretical Morphospaces*. Cambridge University Press, Cambridge.
- Muller, S.J., 2007. *Asymmetry: The Foundation of Information*. Springer-Verlag, Berlin, Heidelberg.
- Ogle, J.M., Ramakrishnan, V., 2005. Structural insights into translational fidelity. *Annu. Rev. Biochem.* 74, 129–177.
- Pretzel, O., 2000. *Error-correcting Codes and Finite Fields*. Oxford University Press, Oxford.
- Rotman, J.J., 1995. *An introduction to the theory of groups*, 4th ed. Graduate Texts in Mathematics, vol. 148. Springer-Verlag, New York.
- Sánchez, R., Morgado, E., Grau, R., 2005. A genetic code boolean structure. I. The meaning of boolean deductions. *Bull. Math. Biol.* 67, 1–14.
- Schmeing, T.M., Ramakrishnan, V., 2009. Review. What recent ribosome structures have revealed about the mechanism of translation. *Nature* 461, 1234–1242.
- Sciarrino, A., 2003. A mathematical model accounting for the organization in multiplets of the genetic code. *BioSystems* 69, 1–13.
- Sella, G., Ardell, D.H., 2006. The coevolution of genes and genetic codes: Crick's frozen accident revisited. *J. Mol. Evol.* 63, 297–313.
- Soma, A., Ikeuchi, Y., Kanemasa, S., Kobayashi, K., Ogasawara, N., Ote, T., Kato, J., Watanabe, K., Sekine, Y., Suzuki, T., 2003. An RNA-modifying enzyme that governs both the codon and amino acid specificities of isoleucine tRNA. *Mol. Cell* 12, 689–698.

- Santos, J., Monteagudo, A., 2010. Study of the genetic code adaptability by means of a genetic algorithm. *J. Theor. Biol.* 264, 854–865.
- Szathmáry, E., 1999. The origin of the genetic code. Amino acids as cofactors in an RNA world. *Trends Genet.* 15, 223–229.
- Trifonov, E.N., 2000. Consensus temporal order of amino acids and evolution of the triplet code. *Gene* 261, 139–151.
- Van der Gulik, P.T.S., Hoff, W.D., 2011. Unassigned codons, nonsense suppression, and anticodon modifications in the evolution of the genetic code. *J. Mol. Evol.* 73, 59–69.
- Vetsigian, K., Woese, C., Goldenfeld, N., 2006. Collective evolution and the genetic code. *Proc. Natl. Acad. Sci. USA* 103, 10696–10701.
- Woese, C.R., 1965. On the evolution of the genetic code. *Proc. Natl. Acad. Sci. USA* 54, 1546–1552.
- Xue, H., Tong, K., Marck, C., Grosjean, H., Wong, J.T., 2003. Transfer RNA paralogs: evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life. *Gene* 310, 59–66.
- Yarus, M., Widmann, J.J., Knight, R., 2009. RNA-amino acid binding: a stereochemical era for the genetic code. *J. Mol. Evol.* 69, 406–429.