

University of Tasmania Open Access Repository

Cover sheet

Title

Spectroscopy of the genetic code

Author

Peter Jarvis, Bashford, JD

Bibliographic citation

Jarvis, Peter; Bashford, JD (2008). Spectroscopy of the genetic code. University Of Tasmania. Chapter.
https://figshare.utas.edu.au/articles/chapter/Spectroscopy_of_the_genetic_code/23056406

Is published in: [10.1142/9781848162556_0009](https://doi.org/10.1142/9781848162556_0009)

Copyright information

This version of work is made accessible in the repository with the permission of the copyright holder/s under the following,

Licence.

If you believe that this work infringes copyright, please email details to: oa.repository@utas.edu.au

Downloaded from [University of Tasmania Open Access Repository](#)

Please do not remove this coversheet as it contains citation and copyright information.

University of Tasmania Open Access Repository

Library and Cultural Collections

University of Tasmania

Private Bag 3

Hobart, TAS 7005 Australia

E oa.repository@utas.edu.au

CRICOS Provider Code 00586B | ABN 30 764 374 782

utas.edu.au

Chapter 1

Spectroscopy of the Genetic Code

Discussions of the nature of the genetic code cannot be divorced from the biological context of its origin and evolution. We briefly review some of the main arguments that have been put forward for the evolution of the genetic code, together with the salient biological background. Longstanding observations of genetic code regularities have led to combinatorially-based assertions about its structure. However, it is also possible to extend such ‘symmetry’ descriptions to continuous symmetries or supersymmetries, especially in relation to the pattern of redundancy (degeneracy) of the genetic code. We give an account of some recent work along these lines. This is supported by graphical presentations, and some data fits, of samples of measured physico-chemical properties of codons and amino acids across the genetic code table. Finally, we review codon-anticodon recognition in terms of conformational degrees of freedom, and structural, stereochemical and kinetic considerations. Based on this, we suggest a possible role for quantum processes at important stages of codon reading and translation.

J.D. Bashford and P.D. Jarvis

1.1 Background: systematics of the genetic code

Discussion of the nature, and organisation, of the genetic code dates from almost before the early work on its detailed elucidation, and has spawned a great variety of ingenious suggestions and insights¹.

The dual aims of this chapter are firstly, to review physics-inspired approaches for describing and analysing the patterns of codon-amino acid

¹One of the first such speculations was the so-called ‘diamond code’ proposed by the physicist Gamow (1954).

assignments which characterise the nature of the genetic code, almost universally across all life forms, and secondly, to discuss possible roles for quantum processes within the genetic code recognition system.

The simplest abstraction of the “genetic code” is as a mapping of genetic information, encoded by one type of biological macromolecule, the nucleic acids, into another family, the amino acids, which constitute the building blocks of proteins². The scheme is simple enough to state: a dictionary of 64 possible code-words (codons), is associated with 20 amino acids plus a “stop” signal. The vast bulk of built-in redundancy in this mapping is conserved within all living organisms, and this strongly suggests that the code in its present form is the result of an evolutionary process at the molecular level, whereby the code derived from some more primitive form. A common first step in attempting to describe the evolution of the genetic code is simply to explain the two numbers, ‘64’ and ‘20’.

The remainder of this section is devoted to a rapid survey of the salient biological and biochemical facts relating to the biomolecules and processes involved in the genetic coding system, followed by a brief commentary on the traditional ‘explanations’ of the origins of the code. More recent, information-theory perspectives, of potential relevance to a role for quantum processes, are also mentioned. In §1.2 the ‘systematics’ of the genetic code in terms of the observed patterns and regularities of the genetic code Table 1.1, are developed in the group-theoretical language of dynamical Lie algebras and superalgebras. In particular, we show that an $sl(6/1)$ supersymmetry model proposed by us (Bashford *et al.*, 1998), is able to give an interpretation in this language, of two main models of code evolution developed quite independently in the biological literature, [Jiménez Sanchez (1995); Jiménez-Montaña (1999)]. The most subtle stage of code evolution relates to the third codon base (see below), and this is taken up in §1.4, with a detailed discussion of codon-anticodon recognition taking into account ribonucleotide conformational degrees of freedom. On this basis we suggest a possible role for quantum processes at important stages of codon-anticodon reading. Meanwhile in §1.3, the dynamical symmetry description is corroborated by giving some simple numerical fits between various measured biological and physico-chemical codon and amino acid properties, and appropriate polynomials in the group labels, or ‘quantum’ numbers of the codons.

²A recent review of the origin of the genetic code is, for example, (Szathmáry, 1999); see also articles in the special issue volume 33, 2003 of *Origins of Life and Evolution of the Biosphere*.

1.1.1 RNA translation

The machinery of gene translation is dependent upon four major kinds of biomolecule [Woese *et al.* (1966)]: *mRNA* - the gene transcript, which contains sequences of codons; amino acids (*a.a.*)- the building blocks of polypeptides; *tRNA* - the intermediary molecules, which carry amino acids and recognise specific codons on *mRNA* via base-pairing with the *anticodons* which they present, and the amino-acyl (*-tRNA*-) synthetases, *aaRS* - enzymes which bind specific amino acids to their cognate *tRNA*. The basic reactions connecting these biomolecules are sketched in Figure 1.1. Other kinds of molecule also participate: the ribosome complex, at which translation occurs, and elongation factor *EF-Tu* : a protein which transports amino-acylated *tRNAs* to the ribosome. However these latter pathway components have more generic roles, independent of codon or amino acid properties, and will not be discussed in great detail. The basic unit of the

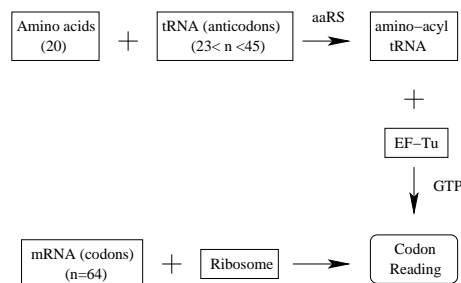


Fig. 1.1 Organisational chart of key steps in codon reading.

genetic code is the *codon*: a triplet of nucleotide bases . Four such bases, Guanine, Cytosine, Adenosine and Uracil (Thymine for *DNA*) occur naturally in *mRNA*. Therefore there exist $64 = 4^3$ possible triplet combinations which are distributed, unequally, amongst 20 amino acids as shown in Table 1.1. Even a cursory inspection of this, mitochondrial, genetic code reveals systematic patterns in the observed degeneracy of amino acids to codons. Specifically, changes in the first and second codon letters always change the coded amino acid, whilst changes in the third position often do not. Furthermore, codons with the purine (*R*)-derivative bases (*A* and *G*) in the third position always code for the same amino acid, as do those with pyrimidine-derivatives (*U* and *C*)³. The overall effect seen in Table

³The code for eukaryotic organisms is more involved, as will be discussed below.

Table 1.1 Mitochondrial genetic code.

a. a. ^a	codon	a.c. ^b	a. a. ^a	codon	a.c. ^b	a. a. ^a	codon	a.c. ^b	a. a. ^a	codon	a.c. ^b
Phe	UUU	GAA	Ser	UCU	UGA	Tyr	UAU	GUA	Cys	UGU	GCA
Phe	UUC	GAA	Ser	UCC	UGA	Tyr	UAG	GUA	Cys	UGC	GCA
Leu	UUA	UAA	Ser	UCA	UGA	Ter	UAA	-	Trp	UGA	UCA
Leu	UUG	UAA	Ser	UCG	UGA	Ter	UAG	-	Trp	UGG	UCA
Leu	CUU	UAG	Pro	CCU	UGG	His	CAU	GUG	Arg	CGU	UCG
Leu	CUC	UAG	Pro	CCC	UGG	His	CAC	GUG	Arg	CGC	UCG
Leu	CUA	UAG	Pro	CCA	UGG	Gln	CAA	UUG	Arg	CGA	UCG
Leu	CUG	UAG	Pro	CCG	UGG	Gln	CAG	UUG	Arg	CGC	UCG
Ile	AUU	GAU	Thr	ACU	UGU	Asn	AAU	GUU	Ser	AGU	GCC
Ile	AUC	GAU	Thr	ACC	UGU	Asn	AAC	GUU	Ser	AGC	GCC
Met	AUA	UAU	Thr	ACA	UGU	Lys	AAA	UUU	Ter	AGA	-
Met	AUG	UAU	Thr	ACG	UGU	Lys	AAG	UUU	Ter	AGG	-
Val	GUU	UAC	Ala	GCU	UGC	Asp	GAU	GUC	Gly	GGU	UCC
Val	GUC	UAC	Ala	GCC	UGC	Asp	GAC	GUC	Gly	GGC	UCC
Val	GUA	UAC	Ala	GCA	UGC	Glu	GAA	UUC	Gly	GGA	UCC
Val	GUG	UAC	Ala	GCG	UGC	Glu	GAG	UUC	Gly	GGG	UCC

^a a.a. = amino acid.^b a.c. = anticodon. Anticodon base modifications not shown.

1.1 is that codons cluster in families, with either 4-fold (“family boxes”) or 2-fold degeneracies (“mixed boxes”), which all code for the same amino acid; moreover there are two cases of ‘hexanumerate’ codons in which both a family box and a mixed box contribute. This structure is a direct consequence of the *tRNA-mRNA* pairing: bases in the first⁴ two positions of the codon always bind to their anticodon *complement* (*G* with *C*, *A* with *U*), as shown in Figure 1.2. The binding in the third position is less precise. The nature of this ambiguous or “wobble” pairing, first postulated by Crick (1966), is still not completely understood, and we will review current knowledge in a subsequent section (see §1.4), as it forms the cornerstone of our suggestions for quantum processes in codon reading.

Finally let us mention the fourth class of molecule, the *aaRS*. Each enzyme contains a receptor for a specific amino acid, and binds to the anticodon-containing region of the cognate *tRNA*. Detailed comparison of *aaRS* structures [Eriani *et al.* (1990)] led to the discovery that two structurally-distinct families of molecule exist. Furthermore, these structural motifs are strongly conserved amongst organisms with only one, prim-

⁴Relative to the standard (5' carbon→3' carbon) orientation of the sugar-phosphate backbone; see section 1.4.

itive, exception amongst archaebacteria, discovered to date [Fabrega *et al.* (2001)]. Remarkably, each *aaRS* class contains species cognate for 10 amino acids, with the resulting families being “complete” in the sense that each contains physicochemically-distinct (in terms of hydrophobicity and acidity) amino acids [Cavalcanti *et al.* (2004)], capable of producing key protein structural motifs. Moreover the so-called class II *aaRS* are associated with smaller, polar *a.a.*'s, commonly believed to have been incorporated in the genetic code earlier than the bulkier residues of class I. This observation has led to speculation that the modern genetic code formed via a ‘doublet’ predecessor.

1.1.2 The nature of the code

During the 1960s and 1970s organisms as diverse as certain bacteria, viruses and vertebrates were all found to have the same genetic code, leading to the concept of a “universal” genetic code, present at least since the Last Universal Common Ancestor (LUCA). This observed universality was the motivation for the ‘frozen accident’ hypothesis [Crick (1968)], which stated that as evolution progressed, the increased complexity of proteins made incorporation of new amino acids unlikely to be beneficial.

Although the “universal” genetic code incorporates 20 amino acids, recognised by the procedure in Figure 1.1, several recently-discovered exceptions exist, whereby “new” amino acids are encoded by *tRNA* simultaneously binding to a “stop” codon and recognising a secondary structural motif. Examples include selenocysteine [Bock *et al.* (1991)] in eukaryotes and pyrrolysine [Srinivasan *et al.* (2002)] in archaebacteria. Differences in amino acid-codon assignments have also been discovered (for a review see the paper by Osawa *et al.* (1992)) and currently 16 variants on the “universal” code are catalogued on the NCBI Taxonomy webpage <http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi>

In this paper we shall be concerned with two codes: the (vertebrate) mitochondrial Code (VMC), posited to be related to an ancestor of the universal or eukaryotic Code (EC). The EC has qualitatively similar 2- and 4-fold degeneracies and codon assignments to amino acids as are observed for the VMC. As mentioned, these degeneracies are due to the ambiguous or “wobble” nature of pairing between third codon, and first anticodon, bases [Crick (1966); Agris (1991)]. Here it suffices to state the wobble rules (Table 1.2); we shall discuss them in greater detail in §1.4. As seen from Table 1.1, in the VMC only first *a.c.* position *U* and *G* are present, leading

Table 1.2 Wobble pairing rules^a

First a.c. position ^b	Third codon position
<i>U</i>	<i>U, C, A, G</i>
<i>G</i>	<i>U, C</i>
<i>C</i>	<i>G</i>
<i>I</i>	<i>U, C, A</i>

^a Adapted from [Osawa *et al.* (1992)].^b Nucleoside modifications not shown.

to the characteristic “4” and “2+2” box degeneracies. While *a.a.* – codon assignments are very similar in the EC, the *anticodon* usage is different. Firstly the purine-derived base Inosine (*I*) replaces *A* (in all but a few exceptions) in the first *a.c.* position, while *C* may compete with *I* and *U* for codons ending with *G*. We shall comment further on this competition in §1.2.

Regularities inherent in the nucleobase “alphabet” allow discussion of codon-amino acid relationships to be abstracted from a biochemical setting to a mathematical/logical one. Nucleobases are commonly classified in terms of three dichotomous indices [Saenger (1984)]: Strong (*G, C*) versus Weak (*A, U/T*) pertaining to the number of H-bonds formed in canonical pairs (Figure 1.2); puRine-derived bases (*A, G*) contain two heterocyclic rings, while pYrimidine bases (*C, U/T*) have one. Thirdly one can distinguish the proton acceptor/donor nature of the functional group attached to the C1 atom: aMino (*A, C*) versus Keto (*G, U/T*). Of course any two of these indices are sufficient to establish the identity of any given nucleobase. Finally there exists the common notation aNy base, that is $N = U/T, C, A, G$. In terms of this language, regularities in the code are easily expressed. Arguably the best known is Rumer’s rule [Rumer (1966)]: replacement of the bases in codon positions *I* and *II* by their *M/K* counterparts changes the nature of the 4-codon box. For example, the box *UUN* is split, with codons *UUY* and *UUR* coding for Phe(nylalanine) and Leu(cine) respectively. Replacing *U(I)* and *U(II)* by the other *K*-type base (*G*) changes the structure to *GGN*, which is a family (unsplit) box, coding for Gly(cine). Another, more recent example is the observed correlation [Biro *et al.* (2003); Chechetkin (2006)] between class I and class II *aaRS* and anticodon families of the forms

$$(WWW, WWS, SWW, WWS), (SSS, SSW, WSS, WSW), \\ (MMM, MMK, KMM, KMK), (KKK, KKM, MMK, MKM).$$

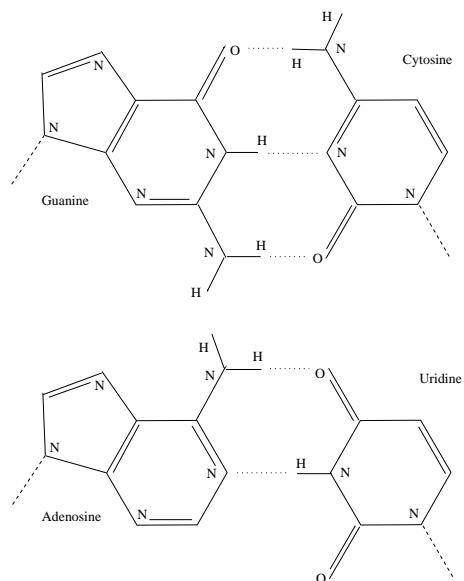


Fig. 1.2 Watson-Crick pairing of RNA bases.

Other regularities apparent in the code, in particular relating amino acid and codon physico-chemical properties will be discussed in more detail in subsequent sections.

Theories on the evolution of the code fall into one of three broad categories. The *co-evolution* theory posits that the genetic code evolved in parallel with the emergence of increasingly complex amino acids [Wong (1976); Weberndorfer *et al.* (2003)]. Thus similar amino acids would be coded by similar codons because more recent *a.a.*'s “captured” codons from their precursors. The *physico-chemical* hypothesis [Di Giulio (2003)] suggests that, at an early stage of evolution, direct contacts between amino acids and codons/anticodons facilitated translation, dictating patterns of physico-chemical regularities observed within the modern, universal code. Finally the *selection* theory suggests that the code evolved to minimise phenotypic errors [Freeland *et al.* (2003); Ronneberg *et al.* (2000)] and, indeed, secondary structure of *mRNA* transcripts [Shabalina *et al.* (2006)]. The three streams of thought are not, however, mutually exclusive and each mechanism may have influenced different stages of evolution.

1.1.3 Information processing and the code

On an abstract level, the flow of genetic information from DNA to polypeptide can readily be viewed in terms of a digital code. Generally the Y/R and S/W characteristics of each nucleobase are represented as “bits” and regularities within patterns of codon and amino acid assignments are investigated [Freeland *et al.* (2003); Mac Dónaill and Manktelow (2004)]. For example in a 2-bit scheme (1 each for Y/R and S/W) Rumer’s conjugate rule ($G \leftrightarrow U$, $A \leftrightarrow C$) can be implemented as the negation operation [Négadi (2003)]. Informational aspects of coding evolution, including adaptor enzymes (*aaRS*) have been discussed by Nieselt Struwe and Wills (1997). A fuller discussion of such coding labelling is described in §1.2 and §1.3.

From the evolutionary point of view however Gray, or error-checking codes [Freeland *et al.* (2003)] are of particular interest. For example Mac Dónaill (2003) proposed a 4-bit scheme, where 3 bits indicated proton donor/acceptor sidegroups on the bases and 1 labelled the base Y/R nature, viz.

$$G = (0, 1, 1; 0), \quad C = (1, 0, 0; 1) \\ aA = (1, 0, 1; 0), \quad U/T = (0, 1, 0; 1).$$

Here “ aA ” (amino-Adenosine) was considered on theoretical grounds. With base parity defined as the sum of all 1’s appearing in the corresponding vector, it is clear that the fourth, Y/R bit acts as a parity check upon permissible H-bonding patterns (*i.e.* those which do not disrupt the regular helical geometries of DNA or RNA)⁵. Using this scheme an extended set of 2^4 candidate bases was considered for inclusion in the nucleotide alphabet, whereupon it was argued that same-parity alphabets have high recognition fidelity, in contrast to those of mixed parity. Further alphabets with even parity, such as the natural one, are likely to be favoured over odd ones as the occurrence of tautomers is generally less likely.

A different kind of information-processing hypothesis was proposed by Patel (2001a,b,c) whereby DNA or protein assembly is reduced to a computational task. The problem is to determine a maximally-sized “database” of items (nucleobases), which are readily distinguishable, with a minimal number of search queries. For a database containing N randomly-ordered items denote the number of binary (“yes/no”) queries, Q , required to locate the desired item. In a classical ensemble, where rejected items are returned to the database, the expected mean number of questions, $\langle Q \rangle = N/2$.

⁵Changing from aA to A (deletion of the third “1” of aA) does not affect the even parity of the resulting alphabet.

However in a quantum-mechanical system, superpositions of items are permitted, and Grover's algorithm [Grover (1997)] exploits this feature. Starting with the symmetric initial state $N^{-1/2}(|1\rangle + |2\rangle + \dots |N\rangle)$, the number of queries needed is determined by

$$(2Q + 1)\text{Arcsin}N^{-1/2} = \pi/2. \quad (1.1)$$

The potential significance of several solutions of this equation for DNA replication and expression have been pointed out by Patel (2001a,b,c). For $Q = 1$ the database has $N = 4$ items which, if the "items" are nucleobases, and (Watson-Crick) base-pairing the "quantum oracle", has implications for DNA replication. For $Q = 3$ (one question per base pair in a codon) $N \simeq 20.2$, while the number of "letters" in the genetic code is 21 (20 amino acids plus a stop signal). Lastly when $Q = 2$ one finds $N \simeq 10.5$, which is potentially of interest in regard to the two classes of aaRS and amino acids [Patel (2005)].

In order for quantum genetic information-processing to occur, as outlined above, there are two important considerations. Firstly quantum decoherence, which occurs far more quickly than proton tunneling, at room temperature needs to be mitigated. Indeed particular enzymes have been suggested to facilitate proton tunneling (for a recent review see the paper by Knapp and Klinman (2002)) in other reactions, via the exclusion of H_2O , although whether DNA polymerases have such capabilities is unknown. Secondly there is the inference that, somehow a quantum superposition of base molecules is set up in the vicinity of the assembly site. It is unclear precisely how an enzyme might achieve this. However, as has recently been pointed out [Shapira *et al.* (2005)], the Grover algorithm can produce advantageous searches under a variety of initial conditions including mixed states.

1.2 Symmetries and supersymmetries in the genetic code

Attempts to understand the non-random nature of the genetic code invite a description in a more abstract setting. Combinatorial symmetry considerations amount to statements about certain transformations amongst the basic ingredients, like bases and codons, or groupings thereof, which display or predict regularities⁶. A natural extension of this language is to formal

⁶There have been many studies attempting to unlock the "secret of the genetic code" by careful examination of code patterns. See §1.1 for historical remarks, and (Szathmáry, 1999) for a modern account. A recent orthodox study attempting to establish objective

continuous linear transformations amongst the physical objects themselves. This mathematical viewpoint thereby allows access to the rich theory of Lie groups and their representations. In group theoretical language the existence of the genetic code itself entails a simple counting problem – find the semisimple Lie groups⁷ which have irreducible representations of dimension 64, the number of codons in the genetic code. Secondly, such a genetic code group should have a “reading subgroup”, for which the dimension and multiplicity of its representations in the decomposition of the 64-dimensional codon representation, coincides with, or is a refinement of, the known pattern of codon redundancies in the amino acid assignments. Such groups and subgroups are candidates for ‘symmetries of the genetic code’.

The first work along these lines was carried out in a pioneering paper by Hornos and Hornos (1993), and elaborated in (Forger *et al.*, 1997) (for a review see (Hornos *et al.*, 1999)). A comprehensive sort by rank and dimension of representations led to the almost unique identification of the 64-dimensional representation of the group $Sp(6)$. It was shown that standard symmetry-breaking scenarios using eigenvalues of Casimir operators in group-subgroup branching chains, for plausible assignments of codons to basis vectors, could provide a quantitative and statistically significant match to a certain major composite index of amino acid functionality, the so-called Grantham polarity index. The group-subgroup branching chain also gave support to the interpretation of the code as a ‘frozen accident’: generically unequal polarity eigenvalues of certain distinct representations of the ‘reading subgroup’ are constrained by the numerical fit to be degenerate, because in practice they still code for the same amino acid.

Subsequent work has exploited the connection between *algebraic* structures and allied transformation groups. In (Bashford *et al.*, 1997, 1998), and (Forger and Sachse, 2000a,b), the notion of symmetry transformations is generalised to that of supersymmetries, which make allowance for the fact that objects in the underlying space may possess a grading, that is an assigned ‘even’, ‘odd’, or in physics language ‘bosonic’, ‘fermionic’, character. In (Bashford *et al.*, 1997, 1998) a specific type of scheme based on the Lie superalgebra $sl(6/1)$ is developed, and in (Forger and Sachse, 2000a,a) more general possibilities are identified. The work (Frappat *et al.*, 1998),

support for code trends is (Biro *et al.*, 2003). Intriguing geometrical insights have been developed by Yang (2005) (also see references therein).

⁷A restriction to simple groups would proscribe the reasonable candidate $SL(4) \times SL(4) \times SL(4)$ for example (see below); on the other hand, weakening the search criteria (for example to non-classical groups or even reducible representations) would considerably complicate the discussion, and so is avoided just on technical grounds.

reviewed by Frappat *et al.* (2001) on the other hand, deviates further from the Lie group–Lie algebra connection in exploiting a specific type of ‘quantum’, or q -deformed algebra, here $sl_q(2) \times sl_q(2)$, in which tensor products of representations reduce in a simple way, as desirable if unique assignments of abstract state vectors to objects in the genetic code system are to be maintained. An important paper which critically reviews *all* group-based and related attempts at a description of the genetic code, especially the $Sp(6)$ models and an alternative $SO(13)$ version, is the article (Kent *et al.*, 1998).

As reviewed in §1.1 above, current thinking is that the origin of the genetic code is distinct from the origin of life itself, and that its present-day, near-universal structure is the result of some evolutionary process of refinement from earlier, primitive versions. The potential of group-theory based accounts is that, in contrast to combinatorial schemes, which merely serve to express genetic code regularities via succinct statements about various discrete transformations, the code structure as a whole can be described in terms of a succession of group-subgroup steps, in a chain starting with the initial codon group, and ending with the final ‘reading’ group. In the remainder of this section we describe the $sl(6/1)$ model in some detail, in this context. The focus is not so much on quantitative predictions (but see §1.3 below), but rather to demonstrate that a group-theory based account can indeed be broadly compatible with established biochemically- and biologically-based understandings about the origin and evolution of the code. The claim would be that an eventual, ‘dynamical code model’ will bear out the group-theoretical steps in detail.

A useful starting point for code degeneracy, compatible with both physico-chemical and coevolution views of the code origin, is to regard the outputs of the early translation system as ‘stochastic proteins’. Possibly, early proto-amino acid/nucleic acid associations were useful in the context of optimising replication, and only incrementally acquired sufficient specificity for the synthesis of functional oligopeptides to emerge as an end in itself. It is reasonable to suggest that early coding for such primitive enzymes was quite non-specific and error-prone, but also, that the system as a whole was error-tolerant. The group-theoretical counterpart of this is that the degeneracies of the early code should be associated with the decomposition of the 64-dimensional representation of the codon group into irreducible representations (irreps) of intermediate subgroup(s), such that codon assignments within, or between, such subgroup irreps, respectively minimize, or maximise, variations in amino acid properties at the largest

possible levels of functional synonymy.

It was pointed out in §1.1 above, a major determinant of amino acid type is the character of the second codon base. Specifically *weak* bases $W = \{A, U\}$ are associated with hydrophilicity/hydrophobicity extremes respectively (see (Weber and Lacey, 1978)). Also, there is an argument from biosynthetic complexity [Jiménez Sanchez (1995)] that the earliest utilised bases should be the simplest chemically, namely A, U again (there are of course other arguments, for example thermodynamic stability of codon-anticodon pairs, [Baumann and Oro (1993)] for the strong bases $S = \{C, G\}$ to have been the earliest coding bases). Finally, it can be argued [Woese *et al.* (1966)] that a minimal requirement for useful oligopeptides, should be the existence of tunable hydrophobic/hydrophilic regions in the primary structure, so as to allow the possibility of folding and the presentation of stereochemically specific, enzymatically active contact regions.

1.2.1 $sl(6/1)$ model: $UA + S$ scheme

The representation-theoretical equivalent is thus that there should be an assignment of codons to a basis for the 64-dimensional representation of the genetic code algebra, which is adapted to a subalgebra decomposition which distinguishes the second base letter and assigns codons NAN and NUN to different representations (necessarily of dimension 16). It turns out that the class of superalgebras $sl(n/1)$ possesses a family of so-called typical irreducible representations of dimension 2^n , which moreover branch to members of the corresponding family under restriction to a smaller subalgebra $sl(n'/1)$, $sl(n''/1)$ with $n'' < n' < n$. This property, shared by spinor representations of the orthogonal groups, singles out for attention in the genetic code context the superalgebra $sl(6/1)$, and its typical irreducible representations of dimension $2^6 = 64$, with Dynkin label $(0, 0, 0, 0, 0; b)$ for appropriate values of $b > 5$ (denoted hereafter by $\mathbf{64}_b$). The branching rule (for $n' = n - 2 = 4$) reads

$$sl(6/1) \rightarrow sl(2)^{(2)} \times sl(4/1)^{(13)} \times gl(1),$$

$$\mathbf{64}_b \rightarrow \mathbf{1} \times \mathbf{16}_{b+2} + \mathbf{2} \times \mathbf{16}_{b+1} + \mathbf{1} \times \mathbf{16}_b,$$

where the Dynkin label of the $2^4 = 16$ dimensional typical irreducible representation $\mathbf{16}$ of $sl(4/1)$ is given as a subscript⁸, and the superscripts on superalgebra labels (or multiplets as needed) refer to the codon positions

⁸This label is closely related to the weight of the irreducible representations of $gl(1)$ which occur in the decomposition; however these are not given explicitly.

on which the subalgebra factors act. Making the natural identification of the $\mathbf{1}$'s with the A and U codons as suggested by the above discussion, and assigning the strong bases S to the doublet $\mathbf{2}$, a more descriptive form of the branching rule is thus

$$\mathbf{64}_b \rightarrow \mathbf{1}_A \times \mathbf{16}_{b+2} + \mathbf{2}_S \times \mathbf{16}_{b+1} + \mathbf{1}_U \times \mathbf{16}_b$$

with the understanding that the codon groups being assigned to the symmetry adapted bases for the subalgebra representations are NAN , NSN , and NUN , respectively.

From the standpoint that redundancy in codon reading and amino acid translation, equates with degeneracy in codon assignments to irreducible representations in the group theoretical schemes, it can be suggested that this stage of code evolution would have corresponded to the existence of three proto-amino acids, or possibly three groups of amino acids with shared functional uses within each group. Alternatively, in the earliest stages the middle NSN group could have simply been unassigned to a definite amino acid coding role. Code elaboration became possible once the developing translation system had achieved a requisite degree of accuracy and reliability. Further major determinants of amino acid assignments to codons are once again the precise identity of the second codon base (thus, not just a coding role for the NSN group, but perhaps separately for NCN and NGN), but also the modulation of codon assignments afforded by the identity of the *first* codon base. Both options are plausible, and lead to different group branching scenarios.

Consider, for example, the second option. It is natural to repeat, at the level of the first base letter, the previous branching pattern, this time at the level of $sl(4/1)^{(13)} \rightarrow sl(2)^{(1)} \times sl(2/1)^{(3)} \times gl(1)$,

$$\mathbf{16}_{b'} \rightarrow \mathbf{1} \times \mathbf{4}_{b'+2} + \mathbf{2} \times \mathbf{4}_{b'+1} + \mathbf{1} \times \mathbf{4}_{b'},$$

where again the $gl(1)$ label has been omitted in favour of the related nonzero Dynkin index of the $sl(4/1)^{(3)}$ typical irreps $\mathbf{4}$ (given as a subscript). At this stage the full list of $sl(2)^{(1)} \times sl(2)^{(2)} \times sl(2/1)^{(3)}$ irreps (again omitting $gl(1)$ factors but including the nonzero Dynkin label of the $sl(2/1)$ third base letter quartets) in the decomposition of the codon representation is

$$\begin{aligned} \mathbf{64}_b \rightarrow & (\mathbf{1} \times \mathbf{1} \times \mathbf{4}_{b+2} + \mathbf{1} \times \mathbf{2} \times \mathbf{4}_{b+3} + \mathbf{1} \times \mathbf{1} \times \mathbf{4}_{b+4}) + \\ & (\mathbf{2} \times \mathbf{1} \times \mathbf{4}_{b+3} + \mathbf{2} \times \mathbf{2} \times \mathbf{4}_{b+2} + \mathbf{2} \times \mathbf{1} \times \mathbf{4}_{b+1}) + \\ & (\mathbf{1} \times \mathbf{1} \times \mathbf{4}_b + \mathbf{1} \times \mathbf{2} \times \mathbf{4}_{b+1} + \mathbf{1} \times \mathbf{1} \times \mathbf{4}_{b+2}), \end{aligned}$$

corresponding to the codon groups

$$\begin{aligned}
 &(AUN+ASN + AAN)+ \\
 &(SUN+SSN + SAN)+ \\
 &(UUN+USN + UAN)
 \end{aligned} \tag{1.2}$$

respectively. Once again, depending on whether the codons with middle letter S are translated or unassigned (or ambiguous), this code stage suggests 5 or 6, or possibly as many as 8 or 9, active groups of mutually exchangeable proto-amino acids. This group-theoretical description closely matches the scheme for code evolution proposed by Jiménez Sanchez (1995) where the weak U, A bases are argued to be the first informative parts of primordial (three-letter) codons (with the strong bases merely providing stability for the codon-anticodon association). Van den Elsen and coworkers have argued for an intermediate expansion stage of evolution of the genetic triplet code via two types of doublet codons, namely both ‘prefix codons’ in which both the middle and first bases are read (as in the above scenario), but also ‘suffix codons’ involving reading of the middle and third codon bases [Elsen *et al.* (2005)]. Conflicts are resolved by allowing certain amino acids to possess both prefix and suffix codons, which are still visible in the present eukaryotic code in the form of the six-fold codon degeneracies for *Arg*, *Leu* and *Ser*. It should be noted that the above scheme involving A, U and S in both first and second bases, which has been introduced as a partial doublet prefix codon genetic code, could also develop suffix codon reading; for example $SWN \rightarrow SWW$, and $SSN \rightarrow SSW$ wherein the third base position is read and the remaining S base positions confer stability.

The final step in this scheme is the breaking of the strong base $sl(2)$ symmetries which hold C, G bases in the first and second codon positions degenerate (or unassigned). If the second codon position is the major determinant of amino acid differentiation, then the $sl(2)^{(2)}$ breaking step proceeds first, yielding (referring to (1.2) above)

$$\begin{aligned}
 &\text{four } WWN \text{ quartets;} \\
 &ASN \text{ and } USN \rightarrow \text{four quartets } ACN, AGN, UCN, UGN; \\
 &\text{two unbroken octets } SUN, SAN; \\
 &SSN \rightarrow \text{two octets } SCN, SGN.
 \end{aligned}$$

The code has thus expanded to eight degenerate quartets and four octets, for up to twelve readable amino acids⁹. The final step is first base position

⁹In the account of Jiménez Sanchez (1995), the original code used triplet codons entirely of the WWW form, which later acquired S codons in all positions.

symmetry breaking leading to 16 family boxes, with both first and second base positions being read. This proposal for code evolution, with or without the variation of prefix and suffix doublet codons, can be referred to as the 'UA+S' scheme, to distinguish it from the following alternative model.

1.2.2 $sl(6,1)$ model: 3CH scheme

A somewhat different proposal for the origin and organisation of the genetic code has been developed by Jiménez-Montaña *et al.* (1996); Jiménez-Montaña (1999) under the motto 'protein evolution drives the evolution of the genetic code, and vice-versa'. According to this scenario, the code has evolved by sequential *full* elaboration of the second codon base letter, followed by the first (and lastly the third); however at each stage pyrimidine-purine reading occurs, before further strong-weak base reading within the Y, R types. This is argued by strictly applying a systematic criterion of code evolution via incremental, minimum change coding pathways, whereby the hierarchical order of codon-anticodon Gibbs free energy of interaction, $\mathbb{C}_2 > \mathbb{H}_2 > \mathbb{C}_1 > \mathbb{H}_1$ (followed by $\dots > \mathbb{C}_3 > \mathbb{H}_3$), which can be established *in vitro*, is adopted to infer a temporal sequence of code expansion. This means that the *chemical type* $\mathbb{C} = \{Y, R\}$ (that is, whether bases are pyrimidines Y or purines R), and then the *hydrogen bonding type* $\mathbb{H} = \{S, W\}$ (that is, whether bases are strong, S , or weak, W) of the second, first (and finally third) codon base are successively able to be read by the evolving translation system.

A group-theoretical branching scheme reflecting this scenario would entail symmetry breaking of transformations on successively the second, first (and lastly the third) base letters, in contrast to the hierarchical UA+S scheme's adoption of the partial A, U and S breaking scheme on the second and first base letters before differentiation of the S bases into C, G . Thus corresponding to the chemical type Y, R , within each of which is in turn a strong and a weak base (Y includes C, U and R includes G, A respectively), each base quartet is assigned two dichotomic labels, which serve to distinguish subgroup transformations. Remarkably group-theoretical branching rules incorporating these steps are once again natural within the class of 64-dimensional typical representations of the $sl(6/1)$ superalgebra. In this scheme the required labels are eigenvalues of $gl(1)$ generators, and the relevant $sl(n'/1)$, $sl(n''/1)$ subalgebras are now successively $n' = n - 1 = 5$, $n'' = n' - 1 = 4$, rather than $n' = n - 2 = 4$, $n'' = n - 4 = 2$ as in the UA+S scheme. The branching rules read finally (with the $gl(1)$'s being

tagged according to whether they refer to chemical or hydrogen bonding type),

$$\begin{aligned}
 sl(6/1) &\rightarrow sl(5/1) \times gl(1)^{\mathbb{C}_2}, \\
 \mathbf{64} &\rightarrow \mathbf{32}_Y + \mathbf{32}_R \cong NYN + NRN; \\
 sl(5/1) \times gl(1)^{\mathbb{C}_2} &\rightarrow sl(4/1)^{(1,3)} \times gl(1)^{\mathbb{H}_2} \times gl(1)^{\mathbb{C}_2}, \\
 \mathbf{32}_Y &\rightarrow \mathbf{16}_{Y_S} + \mathbf{16}_{Y_W}, \mathbf{32}_R \rightarrow \mathbf{16}_{R_S} + \mathbf{16}_{R_W}, \\
 NYN &\rightarrow NCN + NUN, NRN \rightarrow NGN + NAN.
 \end{aligned}$$

This pattern is repeated for the first codon base letter giving eventually

$$sl(6/1) \rightarrow sl(2/1)^{(3)} \times gl(1)^{\mathbb{H}_1} \times gl(1)^{\mathbb{C}_1} \times gl(1)^{\mathbb{H}_2} \times gl(1)^{\mathbb{C}_2}$$

with 16 codon quartets in which the first two bases are read. This scenario can be referred to as the '3CH' scheme.

A variant of this picture was in fact proposed earlier by Swanson (1984). That version considered code elaboration based on a $\mathbb{C}_2 > \mathbb{C}_1 > \mathbb{H}_2 > \mathbb{H}_1$ hierarchy. In the $sl(6/1)$ model, the corresponding subalgebra branching pattern would be

$$\begin{aligned}
 sl(6/1) &\rightarrow sl(4/1)^{(1,3)} \times sl(2)^{(3)} \times gl'(1) \\
 &\rightarrow sl(4/1)^{(1,3)} \times gl(1)^{(3)} \times gl(1)' \\
 &\rightarrow sl(2/1)^{(3)} \times sl(2)^{(1)} \times gl''(1) \times gl(1)^{(3)} \times gl(1)' \\
 &\rightarrow sl(2/1)^{(3)} \times gl(1)^{(1)} \times gl''(1) \times gl(1)^{(3)} \times gl(1)'.
 \end{aligned}$$

However, in a quantitative study of the effect of base changes on amino acid similarity across the code (using amino acid correlation matrices from alignment methodologies), it was shown by Mac Dónaill and Manktelow (2004) that this scheme is less supported than the standard 3CH version above.

Thus far both group-theoretical branching scenarios based on the $sl(6/1)$ scheme have arrived at the 16 codon boxes (quartets) of the standard genetic code, regarded as 4-dimensional irreps of the residual third letter $sl(2/1)^{(3)}$ dynamical symmetry to which both branching chains reduce. These are shown in Figures 1.3 and 1.4 for the $UA+S$ and 3CH schemes respectively (compare Figure 1.3 with (Jiménez Sanchez, 1995) Table 2 and Fig. 2, and Figure 1.4 with (Jiménez-Montaña, 1999), Fig. 1a, 1b and 1c). The $SO(13)$ model mentioned above, would have similar counterparts because of the intimate relation between the typical irreducible representations of the $gl(n/1)$ superalgebras and the spinor representations of $SO(2n)$ or $SO(2n+1)$ groups¹⁰.

¹⁰In the first scenario the $UA+S$ split is natural within spinor reductions $SO(n) \rightarrow$

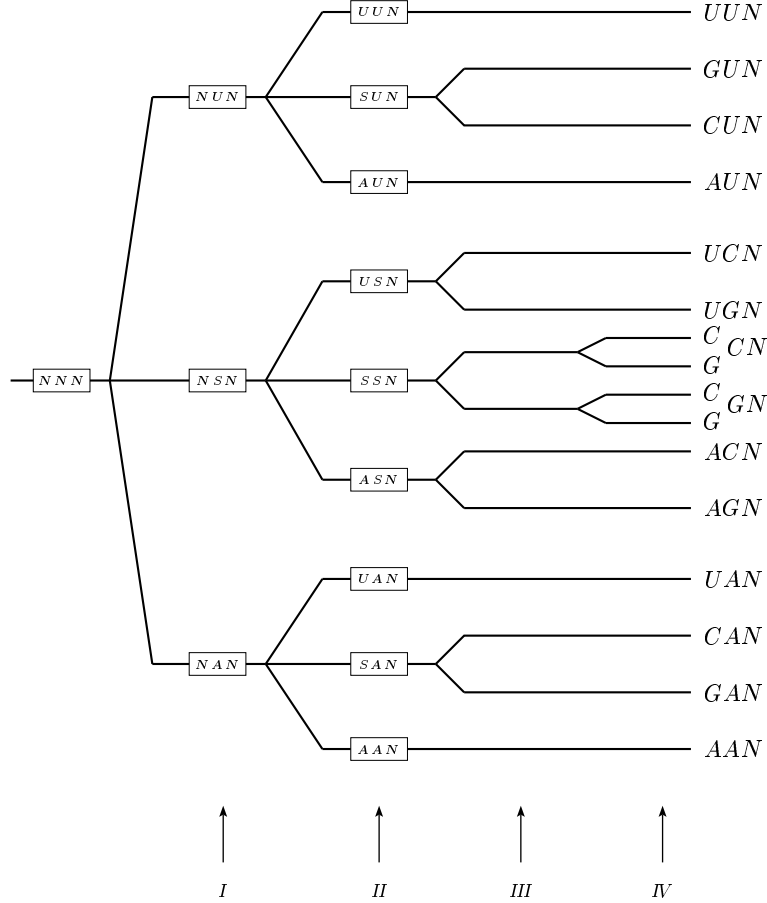


Fig. 1.3 Code evolution according to Jiménez-Sánchez 1995 [Jiménez Sanchez (1995)] transcribed into a group branching scheme in the $sl(6/1)$ chain. Dynamical symmetry breaking stages: *I*: $sl^{(1,3)}(4/1) \times gl^{(2)}(2)_S \times gl^{(2)}(1)$. *II*: $sl(2/1)^{(3)} \times gl^{(1)}(2)_S \times gl^{(1)}(1) \times gl^{(2)}(2)_S \times gl^{(2)}(1)$. *III*: $sl(2/1)^{(3)} \times gl^{(1)}(2)_S \times gl^{(1)}(1) \times gl^{(2)}(1)_m \times gl^{(2)}(1)_d$. *IV*: $sl(2/1)^{(3)} \times gl^{(1)}(1)_m \times gl^{(1)}(1)_d \times gl^{(2)}(1)_m \times gl^{(2)}(1)_d$.

$SO(n-4) \times SO(4)$, wherein a four dimensional spinor of $SO(4) \simeq SU(2) \times SU(2)$ decomposes into a direct sum $\mathbf{1} + \mathbf{1} + \mathbf{2}$ with respect to one of the $SU(2)$ factors. The second scenario is compatible with successive branchings of the form $SO(n) \rightarrow SO(n-2) \times SO(2)$ wherein a spinor representation of a certain dimension reduces to a pair of spinors of the subgroup.

1.2.3 *Dynamical symmetry breaking and the third base wobble*

The final stage in code evolution is the expansion of the amino acid repertoire via reading of the third codon letter. The relevant feature of the genetic code in this respect, is that the canonical Crick-Watson pairing between the codon base letters on the *mRNA* strand, and the *tRNA* base anticodon (recognition) letters, breaks down. Namely, the 'first', 5'-3' base of the anticodon triplet (which structurally occurs at base position 34 in the so-called anticodon loop of 7 bases in each *tRNA*), admits the so-called Crick wobble pairing with respect to the third codon base, which is more flexible than canonical pairing¹¹. The 'degeneracy of the genetic code' as a whole, is in fact a convolution of the association between the 45 or so used anticodons and amino acids (and the 20-strong *aaRS* enzyme system), and the wobble pairing. Indeed, pairing at the third codon position determines almost all of the degeneracy in the genetic code and, as such, is least correlated with amino acid properties. On the other hand, as will be seen in detail below, the pattern of such pairing depends upon genomic *G+C* content, and also post-transcriptional modification of *tRNA* bases; usually at bases 34 (*a.c.* position 1) and also at base 37 (downstream of *a.c.* position 3). It is reasonable to contend then that reading at this codon position is associated with the latter stages of evolution of the genetic code, the basic translation apparatus necessarily already having been established.

The viewpoint adopted here is that the dynamical symmetry description must relate to codon-anticodon binding and amino acid recognition as a whole; in the *UA + S* scheme stereochemical or other considerations are dominant in organising coding according to hydrophobicity; in the 3CH scheme the free energy of formation is the major determinant of coding. For the third codon base, attention is naturally focussed on the patterns not only of amino acid assignments within codon boxes, but also, in view of the wobble pairing, of anticodon useage.

We now take up in detail these issues of the codon-anticodon degeneracy and the wobble rules. The mitochondrial codes, believed to show similar structural simplifications to a hypothetical ancestral code¹², are especially simple: in each *family* box, *one tRNA* codes; Uridine *U*(34) in the first *a.c.* position can pair with *U, C, A, or G* – in the case of *U · A* by canonical

¹¹The explicit notation $N(34) \cdot N'(III)$ can be used to denote this anticodon-codon base (wobble) pairing, or simply $N \cdot N'$ where no confusion arises.

¹²Code variations across biota have been intensively studied in recent years; see for example (Osawa *et al.*, 1992) for a comprehensive early review.

pairing, and in $U \cdot U$, $U \cdot C$ and $U \cdot G$ by wobble pairing. *Mixed* boxes on the other hand have *two* *tRNA* species; Guanine $G(34)$ recognises U and C – via wobble pairing in $G \cdot U$ and canonical pairing in $G \cdot C$, while $U(34)$ binds A and G – again via canonical pairing in $U \cdot A$, and in $U \cdot G$ by wobble pairing. Uridine $U(34)$ in these mixed boxes is commonly prevented from misreading U and C by chemical modification (and in the family boxes, may indeed undergo different modification to *facilitate* the $U \cdot N$ wobbles)¹³.

The dynamical symmetry description of code elaboration for the third codon position in the mitochondrial codes is therefore rather straightforward. Codon boxes (quartets) either lead to *family* boxes ('intact' $sl(2/1)^{(3)}$ irreps), or reduce for *mixed* boxes to two doublets of some subalgebra, $N_1 N_2 N \rightarrow N_1 N_2 Y + N_1 N_2 R$. It follows that the symmetry breaking at the third position is *partial* (8 family, and 8 mixed, boxes); moreover the pattern of breaking turns out to be specified completely by the identity of bases occupying the first two codon positions. We shall return to these points presently, along with a specific choice for the unbroken subalgebra.

In higher organisms, which use the 'universal' genetic code (for example the eukaryotic code), codon usage in higher organisms is strongly linked with genomic base content. In particular, within the open reading frames in a genome, the $G+C$ content of the third codon position correlates well with genomic $G+C$ content, in contrast to the other positions. In genomes with high $A+T$ content, $G+C$ -rich codons are seldom used, and can even be deleted from the code (for example CGG in *Mycoplasma capricolum* [Andachi *et al.* (1989)]). Conversely, if $G+C$ content is high, $A+U$ -rich codons become rare, and may also disappear (for example NNA within family boxes in *Micrococcus luteus* [Kano *et al.* (1991)].) In between these extremes, different *tRNA* species can 'compete' for the same set of synonymous codons. Representative patterns of codon degeneracy within a box are shown in Table 1.3. The $2+2$ and $2+1+1$ patterns on the left-hand side of Table 1.3 are utilised in seven of the eight mixed boxes, while the $2+1+1$ patterns on the right (the triplet of anticodons involving Inosine is actually $2+1$ arising from $NNY+NNA$) determine seven of the family boxes. The exceptions are boxes AUN (in which *Met* has the single codon AUG , and *Ile* is coded for by three codons AUY and AUC) and GGN (the *Gly* family box) in which these patterns are reversed. However, as emphasised already, we also need to consider how codon usage shifts with genomic base content. For $A+T$ -rich genomes, codons NNG are se-

¹³We shall discuss further the effects of post-transcriptional modification in a quantal model of codon recognition in §1.4

Table 1.3 Anticodon usage patterns in eukaryotes.

mixed boxes		family boxes	
codon	a.c	codon	a.c.
NNU	GNN	NNU	INN
NNC	GNN	NNC	INN
NNA	UNN	NNA	(INN, UNN)
NNG	(CNN, UNN)	NNG	(CNN, UNN)

lected against, resulting in the probable disappearance of anticodons CNN . In this instance, the codon/anticodon box pattern coincides with that of the mitochondrial code $\mathbf{2} + \mathbf{2}$ mixed boxes, $N_1N_2N \rightarrow N_1N_2Y + N_1N_2R$. Conversely, in $G+C$ -rich organisms codon NNA is relatively rare, and consequently there is little need for $tRNA$ species UNN : recognition of codon NNG is predominantly due to anticodon CNN . Thus, in this case there tends to be a further ‘breaking’ of the NNR codon doublet assignment leading to $N_1N_2N \rightarrow N_1N_2Y + N_1N_2R \rightarrow N_1N_2Y + N_1N_2A + N_1N_2G$.

It remains to tie the third base reading patterns to a dynamical symmetry breaking account relating to the proposed $sl(2/1)^{(3)}$ codon quartet superalgebra. We suggest [Bashford *et al.* (1998)] that the basic pattern of $sl(2/1)^{(3)}$ breaking for the primitive and mitochondrial codes is to a $gl(1/1)^{(3)}$ superalgebra, which is further reduced to an appropriate $gl(1)^{(3)}$ label for the eukaryotic code. $gl(1/1)$ is the well-known superalgebra of supersymmetric quantum mechanics, with supercharge generators Q^\pm satisfying $\{Q^-, Q^+\} = H$, $[F, Q^\pm] = \pm Q^\pm$ where H is the ‘Hamiltonian’ and F (with eigenvalues $= 0, 1$) labels fermion number. Irreducible representations are generically two-dimensional, so that quartets of $sl(2/1)^{(3)}$ decompose under $gl(1/1)^{(3)}$ to two degenerate doublets as $\mathbf{4} \rightarrow \mathbf{2} + \mathbf{2}$ (the mixed box codon-anticodon pattern). *Partial* symmetry breaking – the fact that 8 family boxes remain intact, and do not show this codon-anticodon splitting – must be attributed to the varying *strength* of this breaking across the code. The same applies to the final $\mathbf{2} + \mathbf{2} \rightarrow \mathbf{2} + \mathbf{1} + \mathbf{1}$ decomposition manifested in the eukaryotic code, which can be attributed in turn to partial $gl(1/1)^{(3)}$ beaking, this time to the generator F of its Abelian $gl(1)^{(3)}$ subalgebra. The codon-anticodon doublet degeneracy is potentially completely lifted by the additional fermion number-dependent shift; however again this is realised only in certain of the mixed boxes for N_1N_2R , and never for N_1N_2Y .

As mentioned already, the existence of partial symmetry breaking was

argued by Hornos and Hornos (1993) to support the ‘frozen accident’ account of the structure of the genetic code (see also §1.1). In terms of dynamical symmetry breaking, appropriate breaking parameters are to be fine-tuned, so that otherwise non-degenerate codons remain degenerate, and can consistently be assigned the same amino acids. The mechanism operates similarly in our present $sl(6/1)$ scheme, except that we have been discussing codon-anticodon pairings rather than codon-amino acid assignments (which are not the same if different *tRNA*’s can be charged with the same amino acid). Moreover, we have linked the emergence of $\mathbf{2} + \mathbf{1} + \mathbf{1}$ pairing to genomic $G+C$ content, and so effectively injected an organism-dependence into the breaking patterns. Further numerical aspects of the partial breaking, and of the related issue of codon-amino acid assignments, are given in §1.3. A more refined discussion of the codon-anticodon recognition process is given in §1.4.

1.3 Visualising the genetic code

The discussion of the genetic code so far has centred on qualitative aspects of its systematics. These include both longstanding trends, noticed almost as soon as the code was fully elucidated in the sixties (§1.1), as well as more elaborate Lie symmetry and supersymmetry-based schemes (§1.2), which served to transcribe selected accounts grounded in biological understanding, into a mathematical language.

The utility of symmetry schemes in physics is specifically, in helping to quantify the hierarchy of symmetry and symmetry-breaking in the spectroscopy of complex quantum systems such as atoms, molecules and nuclei. Schematically, suppose that a Hamiltonian operator can be constructed as a series of the form

$$H = H_0 + H_1 + H_2 + \cdots$$

where the terms are successively ‘smaller’ in the appropriate sense. Also, on the Hilbert space of quantum-mechanical states of the system, suppose there are operators representing the transformations of a hierarchy of symmetry groups $G_0 \supset G_1 \supset G_2 \supset \cdots$ such that G_0 is a symmetry of (commutes with) H_0 , G_1 commutes with H_1 , \cdots , and so on. Then, by general theorems, the energy eigenfunctions of the system (the energy levels of physical states) are organised into unitary irreducible representations of the successive subgroups. The spectra of the partial Hamiltonians H_0 ,

$H_0 + H_1$, $H_0 + H_1 + H_2$ can be labelled by these irreducible representations, each of which corresponds to states with degenerate energy levels. Moreover, as the corrections introduced become smaller, this labelling thus provides a hierarchical ‘symmetry breaking’ scheme for understanding the structure of the system. In ideal cases the contributions to H are moreover appropriate combinations of so-called Casimir operators of the various subgroups, such that when the states are accorded their correct ancestry in terms of the descending hierarchy of subgroups and respective irreducible representations, their energies (eigenvalues of H) are the corresponding sum of Casimir eigenvalues (polynomials in the labels, or quantum numbers, of the respective representations, for example the highest weight labels).

This methodological approach has indeed been taken, with some success, for the genetic code problem in the work of Hornos and Hornos (1993). As mentioned in §1.2, the symmetry groups were taken to be a chain of subgroups of $Sp(6)$, with codons assigned to its 64-dimensional representation, with the role of the energy being played by a composite measure of codon and amino acid organisation, the Grantham polarity index. An attractive feature of the argument was that although the symmetry breaking chain taken implied complete degeneracy in the generic case, the ‘frozen accident’ visible in the instances of synonymous codon assignments in the real genetic code could be explained by particular parameter constraints between the strengths with which Casimirs belonging to the partial Hamiltonians appeared in the total Hamiltonian H .

In our work we have taken a somewhat weaker approach to quantifying the structure of the genetic code. Within the group branching scenario, it is often sufficient¹⁴ to distinguish states within an irreducible representation of a starting group G by their so-called weights, which are labels for (one dimensional) representations of the smallest available continuous subgroup, the Cartan (maximal Abelian) subgroup. Thus a parametrisation of physical properties via *fitted* polynomials in these labels, can be regarded as a kind of general proxy for the more specific approach sketched above, where a definite subgroup chain is declared, and specific Casimir operators are included at the outset. It is this more flexible method that we have used as an attempted confirmation of the $sl(6/1)$ -based supersymmetric schemes for the structure of the genetic code in §1.2 above. It is apparent from the discussion in §1.2 that the subgroup and state labelling required is closely matched to the four base letter alphabet, three letter word lexicon of the

¹⁴Technically the irreducible representation of G must have no *weight multiplicities*.

genetic code. Mention has already been made of the fact that the nucleic acid bases stand in very symmetrical relationships with respect to each other, and it is natural to reflect this in the state labelling appropriate to the 64-dimensional codon 'space' (see the introductory discussion in §1.2 above). Indeed, any *bipartite* labelling system which identifies each of the four bases A, C, G, U , extends naturally to a *composite* labelling for codons, and hence amino acids. We choose for bases two coordinates $d, m = 0, \pm 1$ as $A = (-1, 0)$, $C = (0, -1)$, $G = (0, 1)$, $U = (1, 0)$, so that codons are labelled as ordered sextuplets, $NNN = (d_1, m_1, d_2, m_2, d_3, m_3)$; for example $ACG = (-1, 0, 0, -1, 0, 1)$. Our choice of dichotomic base labels is of course equivalent to 0, 1 binary labelling and the geometrical picture of the code as a 6 dimensional hypercube¹⁵ as has been noted by several authors (see the discussion above). Fitting polynomial functions in these labels to code properties is furthermore compatible with *any* group labelling scheme for which the 64-dimensional codon representation is equivalent to a hypercube in weight space; as discussed earlier, candidate groups and algebras include $sl(6/1)$ but also $so(13)$, and non-simple groups such as $so(4) \times so(4) \times so(4)$. With these preliminaries it only remains to present sample genetic code (codon and or amino acid), physico-chemical or biological, data, and compare this data to polynomials in the codon labels.

Figure 1.5 gives a two-dimensional presentation of the genetic code whereby each of the m, d paired labels for each base letter are plotted or projected onto the plane. In the case of the first two base letters this occurs by showing four d_2, m_2 diamonds separated by their different d_1, m_1 coordinates; for the third base letter instead a linear rank ordering U, C, A, G of the bases is used (corresponding to a one-dimensional projection of the diamond to a line skew to the sides of the basic diamond). Remarkably, essentially this organisation of the code was discussed some time ago by Siemion (1994) in connection with so-called 'mutation rings', designed to present a rank ordering of codons reflecting their relative interconvertability or functional similarity (so that near neighbours in the mutation ring are also likely to be correlated in their occurrences in nucleic acid coding). Figure 1.6 shows Siemion's rings, with the linear rank ordering (related by Siemion to a 'mutation number' $0 \leq k \leq 63$), labelling a looping closed path around the main second base letter rings with excursions into and out of the G ring starting with GAU (*Asp*) (compare fig 1.5 which has had the $G_2 = \{d_2 = 0, m_2 = 1\}$ ring shifted downwards, with Figure 1.6).

¹⁵The m, d labels give a *diamond* rather than a square orientation to the fundamental base quartet.

Important quantitative indicators of coding functions are the so-called Chou-Fasman parameters, which give *log* frequency measures of the presence of each amino acid in protein tertiary structures such as β sheets and α helices and associated turns. We have fitted several of these parameters to the codon weight labels as described above, and we present here representative fits (taken from (Bashford and Jarvis, 2000)). In Figures 1.8 and 1.9 are plotted histograms of the P^α and P^β parameters against the Siemion number, together with least-squares polynomial fits to functions $F^\alpha(d_1, m_1, d_2, m_2, d_3, m_3)$, $F^\beta(d_1, m_1, d_2, m_2, d_3, m_3)$ given by

$$\begin{aligned} F^\alpha &= 0.86 + .24d_2^2 + .21m_1m_2(m_2 - 1) - .02(d_3 - m_3) - .075d_2^2(d_3 - m_3), \\ F^\beta &= 1.02 + .26d_2 + .09d_1^2 - .19d_2(d_1 - m_1) - .1d_1m_2(m_2 - 1) \\ &\quad - .16m_1^2m_2(m_2 - 1), \end{aligned}$$

respectively.

In Figures 1.7(a) to 1.7(f) measured values for a selection of further experimental parameters are plotted (without any fitting), not against Siemion number, but as histograms over the rings themselves¹⁶. It is clear from the fitted plots Figures 1.8 and 1.9 and from these further plots that simple numerical fitting of the type given in (1.3) can capture the major trends in such genetic code data. For example in [Bashford and Jarvis (2000)], it was found that for the Grantham polarity itself, the important terms were simply d_2 (second codon base hydrophobicity) and $d_3 - m_3$ (third base chemical *Y/R* type) with appropriate coefficients, modulo some first base dependence. Similar numerical considerations also support the ‘partial symmetry breaking’ scenarios. For example, the polynomial

$$(d_1d_2)^2 + \frac{1}{2} (d_1^2m_2^2(1 + m_2) + m_1^2d_2^2(1 - d_2)) \quad (1.3)$$

takes the value 1 on *WWN*, *WGN* and *SAN* and 0 on *SSN*, *WCN* and *SUN* and so serves to ‘turn on’ the partial codon-anticodon $sl(2/1)^{(3)}$ breaking leading to mixed versus family boxes (the key parameter underlying Rumer’s rule [Rumer (1966)]). Finally it can be noted that the difference between codon-anticodon pairing degeneracy and codon-amino acid assignment synonymy also has numerical support: periodicity or symmetry patterns of codon-amino acid properties over the Siemion rings is consistent with repeated amino acid assignments (belonging to different codon boxes) occurring on certain symmetrical ring locations (see (Bashford and Jarvis, 2000; Siemion, 1994)).

¹⁶ Recently entire databases of physico-chemical and biological codon and amino acid properties have become available; see for example (Kawashima and Kanehisa, 2000).

1.4 Quantum aspects of codon recognition

In this section we present the proposition that the codon-anticodon recognition process has an initial, quantum-mechanical step. Previously Patel (2001a,b) discussed the genetic code in terms of quantum information processing, however despite the striking numerical predictions stemming from Grover's search algorithm, the model required some unlikely properties of enzymes.

Our basic assertion rests on the observation that the first anticodon base (labelled henceforth as $N(34)$) is conformationally flexible, whereas *a.c.* sites 35, 36 are constrained by the geometry of the *tRNA* anticodon loop (in addition to modifications to base 37). In an unpaired *tRNA*, $N(34)$ could therefore be expected to be in a superposition of conformational states. In proximity to the complementary codon base, one such state becomes increasingly favoured, facilitating the "collapse" to the classical, paired state. Thus, in contrast to the Patel picture, the superposition of nucleobase states occurs at a structural, rather than chemical level. There is still the issue of thermal effects; however in this regard, we note that aminoacyl-*tRNA* is transported to the ribosome by elongation factor (*EF-Tu*). There are thus two distinct *tRNA*-protein environments, in either of which quantum coherence could be maintained.

1.4.1 $N(34)$ conformational symmetry

In order to develop this quantal hypothesis it is necessary to first discuss nucleobase conformational states. The RNA oligomer is formed of repeated ribonucleotide-phosphate units, one of which is sketched in Figure 1.10(a). The conformer degrees of freedom fall into three broad categories (for a full discussion see (Yokoyama and Nishimura, 1995)). First are the torsion angles between ribonucleotide and phosphate groups: there are respectively three C4'-C5' (*gg*, *gt* and *tg*) and two C3'-O3' (G^\pm) bond rotamers. Secondly there is a twofold degree of freedom (*anti/syn*) describing the relative orientation of the base to the ribose ring. Only *R*-type bases can form two H-bonds (commonly argued to be the minimum required for recognition) in the *syn* conformation, however such *R · R* pairings are not observed *in vivo* [Yokoyama and Nishimura (1995)]. Finally there is a nonplanar deformation of the ribose ring (Figure 1.10(b)), commonly described by the pseudorotation parameter τ . Typically one of two conformers: C2'-*endo* ($\tau \simeq 180^\circ$) or C3'-*endo* ($\tau \simeq 0^\circ$) is favoured, as sketched in Figure

1.10(b). Note however that other states, such as $O4'-exo$, may also become favourable under special circumstances. By correlating these degrees of freedom with stereochemical considerations arising from linking nucleobase units, it is possible to identify likely, low-energy conformer states. For example, according to Altona and Sundralingam (1972), for mononucleosides in the solid state favoured low-energy conformers of R and Y bases can be summarised as in Table 1.4. Any extrapolation to duplex

Table 1.4 Mononucleoside conformations

Base	Conformer	Rotamer ^a	States
R	$C2'-endo$	$\begin{pmatrix} \text{syn} \\ \text{anti} \end{pmatrix} \times \begin{pmatrix} gg \\ gt \end{pmatrix}$	4
R	$C3'-endo$	$(\text{anti}) \times \begin{pmatrix} gg \\ gt \end{pmatrix}$	2
Y	$C3'-endo$	$(\text{anti}) \times (gg)$	1
Y	$C2'-endo$	$(\text{anti}) \times \begin{pmatrix} gg \\ gt \\ tg \end{pmatrix}$	3

^a Rotamer states G^\pm have been neglected.

RNA is likely to restrict the number of favourable states even further. For example, the above classification is *modulo* G^\pm rotamer states. Within a duplex the combination of G^- and $C2'-endo$ ribose places an oxygen ($O2'$) group in close proximity to a (backbone) P unit, with the resulting repulsion making such conformers highly unfavourable. In fact ($C3'-endo$, G^-) and ($C2'-endo$, G^+) are the stable combinations [Yokoyama and Nishimura (1995)] of these degrees of freedom. Additional constraints upon allowable states may arise since the binding occurs with the codon as a ribosomal substrate, rather than in solution. To date only R_{C3} and $Y_{C2,C3}$ conformers have been observed *in vivo* [Takai (2006)] and on these grounds we may neglect the R_{C2} states in a first approximation.

From the rules in Table 1.4 it is easy to see how anticodon GNN might accommodate codons NNC and NNU : $G(34)$ is predominantly in the $C3'-endo$ form. The WC pairing geometry ($G \cdot C$) requires the gg rotamer, while $G \cdot U$ requires a Guanine deformation towards the major groove, possibly facilitated by transition to the gt form. In the present picture the flexible $G(34)$ base would be in a superposition of conformer states, until it encounters the third codon base, whereupon it is required to collapse to either an optimal or suboptimal state (in the contexts of $G \cdot C$ and $G \cdot U$ respectively).

The case of $U(34)$ is more complex. From the rules above, one identifies the $U_{C3}(34)$ *gg* singlet, which participates in WC pairing, and a triplet of U_{C2} rotamers. Empirical evidence strongly suggests $U \cdot G$ and $U \cdot U$ wobble pairs occur in the C2'-*endo* form, hence can be placed in the triplet. However little is known about the $U \cdot C$ pair. Such $U \cdot Y$ mismatches are physically impossible in the C3'-*endo* form; on the other hand the C2'-*endo* conformer theoretically suffers from steric hindrance. Other proposals for the $U \cdot C$ pairing geometry include water-mediated H-bonds [Agris (2004)] and protonation of C or, possibly a different ribose conformer. Based upon current knowledge, the $U \cdot C$ pair is not inconsistent as the third member of the U_{C2} triplet. Uridine is unique amongst the bases, in that the C2'-*endo* and C3'-*endo* forms are almost equally favoured: ΔG^* as defined in Figure 1.11(b) is of the order of $-0.1 \text{ kcal mol}^{-1}$ [Yokoyama *et al.* (1985)] and it therefore readily forms wobble pairs. However $U(34)$ is almost invariably modified post-transcriptionally, presumably to enhance recognition fidelity in one of several ways. The 5-hydroxyuridine derivatives¹⁷ (xo^5U^*) almost always participate in 4-way wobbles [Takai (2006)]. This modification shifts the pseudorotation double well in 1.11(a) in favour of the C2'-*endo* form ($\Delta G^* = 0.7 \text{ kcal mol}^{-1}$), thereby enhancing recognition of the $U \cdot U$, $U \cdot G$ (and presumably $U \cdot C$) wobble pairs. Conversely, 5-methyl-2-thio-uridine derivatives (m^5s^2U^*) strongly stabilise the C3' form ($\Delta G^* = -1.1 \text{ kcal mol}^{-1}$) [Takai and Yokoyama (2003)]. These modifications appear in the split boxes, where misreading of U - and C -ending codons would be potentially lethal. Note that such misreadings still occur, albeit several orders of magnitude less frequently than the “correct” $G \cdot C$ and $G \cdot U$ pairings [Inagaki *et al.* (1995)].

1.4.2 Dynamical symmetry breaking and the third base wobble

In the “modified wobble hypothesis”, [Agris (1991, 2004)] patterns of nucleotide modification are proposed to modify anticodon loop dynamics so as to be compatible with the codon-ribosome complex. In addition to the effects of post-transcriptional modification upon $N(34)$ conformations, as discussed above, bases 32 and 38 (which demarcate the anticodon loop) are commonly modified to enhance H-bonding, thereby facilitating an “open” loop. Further, modifications to $R(37)$ (just downstream of *a.c.* position 3), the so-called “universal purine”, generally correlate with the base content

¹⁷The ‘*’ superscript denotes possible further modification.

of position 36).

Structural studies lend support to kinetic models of codon reading [Takai (2006); Ninio (2006)] describing multiple-stage processes. Initial contacts between ribosome and canonical A-form duplex RNA (for pairs $N(35) \cdot N(II)$ and $N(36) \cdot N(I)$) have been observed to promote conformational changes [Ogle *et al.* (2003)] in the ribosome which facilitate the release of the amino acid from *tRNA*. In fact the conservation of A-form *structure* is more important than stability conferred by base-pairing. For example, thermodynamically, the difference between contributions of canonical, $U(36) \cdot A(I)$, and a potential first codon position wobble, $U(36) \cdot G(I)$, pairs is of the order of 10%. Yet when bound to the ribosome, reading of the “correct” Watson-Crick pair proceeds 3-4 orders of magnitude faster [Kurland *et al.* (1996)].

With the codon bound at the ribosomal A-site, it is reasonable to assume that these three bases are intrinsically rigid, and amenable to A-form duplex formation. Moreover, it can be argued [Lim and Curran (2001)] that anticodon nucleotides 35, 36 are inflexible, whether by way of proximity to the 5' end of the anticodon loop, or due to $R(37)$ modifications. With these considerations it is straightforward to envisage a simple, lattice Hamiltonian containing three sites with one, corresponding to position (34), carrying an internal conformer degree of freedom which feels a double-well potential of the sort sketched in Figure 1.11(a) above.

Using the “low energy” conformer states of Table 1.4 (and tentatively assuming the $xo^5U^*(34)$ base adopts the $C2'$ -*endo tg* form in the context of $U(34) \cdot C(III)$) we can write down correlated anticodon states associated with third letter codon recognition. In the mitochondrial code, for example, in the mixed boxes the possible states for the first anticodon letter are

$$\begin{aligned} G(34) &\rightarrow (|C3' \rangle \otimes |G^- \rangle \otimes (\alpha_1 |gg \rangle + \alpha_2 |gt \rangle), \\ m^5s^2U^*(34) &\rightarrow \beta_1 |C3' \rangle \otimes |G^- \rangle \otimes |gg \rangle + \beta_2 |C2' \rangle \otimes |G^+ \rangle \otimes |gg \rangle, \end{aligned}$$

while the family boxes have

$$\begin{aligned} xo^5U^*(34) &\rightarrow \gamma_1 |C3' \rangle \otimes |G^- \rangle \otimes |gg \rangle \\ &\quad + |C2' \rangle \otimes |G^+ \rangle \otimes (\gamma_2 |gg \rangle + \gamma_3 |gt \rangle + \gamma_4 |tg \rangle). \end{aligned}$$

In their model of permitted wobble-rules, Lim and Curran (2001) predicted that, given certain wobble pairs $U(34) \cdot N(III)$, a second position wobble,

$U(35) \cdot G(II)$ was not forbidden, but prevented from occurring by the use of anticodons GNN . This kind of observation naturally connects the above, quantum picture of codon reading with our original discussion of broken dynamical symmetries. Second position wobbles are likely to be suppressed by the local rigidity of the 7-member anticodon loop, in addition to ribosomal contacts with the bound codon-anticodon complex. It is therefore plausible that higher, dynamically-broken symmetries could describe some ancestral translation system with simpler structural features (and lower fidelity). In this manner the pattern of dynamical symmetry breaking may be indicative of the evolution of the translation apparatus (see, for example, (Seligmann and Amzallag, 2002; Poole *et al.*, 1996; Beuning and Musier Forsyth, 1999)).

The possibility of a graded symmetry underlying the scenario just described is left open. Whenever a 4-way wobble ($xo^5U(34)^*$) is present, the bound states described in Figure 1.11(d) are in one-to-one correspondence with the “physical” codon-anticodon reading complexes. In a mixed box (or indeed in any eukaryotic box) multiple *tRNA* species exist, and several analogues of Figure 1.11(d) are required, with the lowest-lying states of each comprising the set of reading complexes. One possibility is to postulate that certain kinds of pairing geometry have even or odd grading, the motivation being the $gl(1/1)$ dynamical symmetry analogue of supersymmetric quantum mechanics, whereby an even Watson-Crick (ground) state would lie below an odd (excited) wobble state in the case of $U(34) \cdot R(III)$ or $G(34) \cdot Y(III)$ pairs. The following, speculative grading of pairs

$$\begin{aligned} \{G \cdot C, C \cdot G, A \cdot U, U \cdot A, U \cdot C, I \cdot U\} & \text{ even} \\ \{G \cdot U, U \cdot G, U \cdot U, I \cdot C, U \cdot C, I \cdot A\} & \text{ odd} \end{aligned}$$

within the VMC and EC is compatible with the supermultiplet structure described previously in §1.2.

Finally we wish to emphasise the following point: the differences in codon-anticodon *binding energies* are several orders of magnitude less than differences in codon *reading rates*. The connection with the “spectroscopic” theme of earlier sections does not lie directly within the bound states sketched in the potentials of Figure 1.11(c)-(d), which are indicative of only the first stage of a multi-step kinetic pathway. Rather the “spectrum”, if there is one, is in terms of *total* reading reaction rates, analogously to the “potentiation” concept of Takai (2006).

1.5 Conclusions

Regularities inherent in the genetic code “alphabets” allow discussion of codon-amino acid relationships to be abstracted from a biochemical setting to a mathematical/logical one. In this review we have attempted to present insights into the form (and possible evolution) of the genetic code, borrowing group theory concepts from spectroscopy. In §1.2 an argument for an evolutionary role of continuous and/or graded symmetries was made, in comparison with different models in the (biological) literature. Section 1.3 provided some numerical support for this view, via fits of physico-chemical properties of amino acids to those of codons. Finally in §1.4 we proposed a possible role for quantum processes in codon reading. Our hope is that an eventual, dynamical code model will bear out the preliminary steps taken here in this direction.

Acknowledgements:

This research was partly funded by Australian Research Council grant DP0344996. JDB wishes to thank K. Takai for helpful discussions. The assistance of Elizabeth Chelkowska in formatting the 3D plots of genetic code data is gratefully acknowledged.

Short autobiography

Jim Bashford is presently an ice sheet data analyst at the Australian Government Antarctic Division. He graduated with a PhD in theoretical physics from the University of Adelaide in 2003. Recent research interests have included modelling of codon-amino acid degeneracy, oligomer thermodynamics, nonlinear models of DNA dynamics and phylogenetic entanglement.

Peter Jarvis is at the School of Mathematics and Physics, University of Tasmania. His main interests are in algebraic structures in mathematical physics and their applications, especially combinatorial Hopf algebras in integrable systems and quantum field theory. In applications of group theory to physical problems, aside from the work on supersymmetry in the genetic code, recent papers have included applications of classical invariant theory to problems of quantum physics (entanglement measures for mixed state systems), and also to phylogenetic reconstruction (entanglement measures,

including distance measures, for taxonomic pattern frequencies).

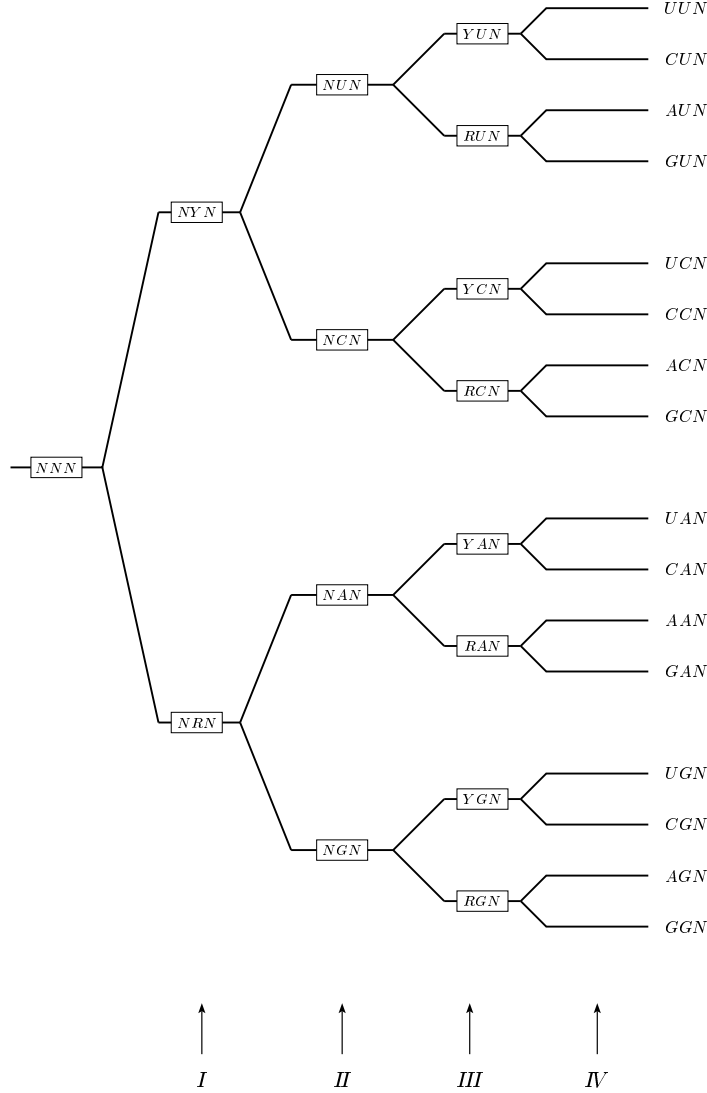


Fig. 1.4 Code evolution steps according to Jiménez-Montaña 1999 [Jiménez-Montaña (1999)] transcribed into a group branching scheme in the $sl(6/1)$ chain. Dynamical symmetry breaking stages: I : $sl(5/1) \times gl^{(2)}(1)_{m-d}$. II : $sl(4/1)^{(3,1)} \times gl^{(2)}(1)_{m-d} \times gl^{(2)}(1)_{m+d}$. III : $sl(3/1) \times gl^{(1)}(1)_{m-d} \times gl^{(2)}(1)_{m-d} \times gl^{(2)}(1)_{m+d}$. IV : $sl(2/1)^{(3)} \times gl^{(1)}(1)_{m-d} \times gl^{(1)}(1)_{m+d} \times gl^{(2)}(1)_{m-d} \times gl^{(2)}(1)_{m+d}$.

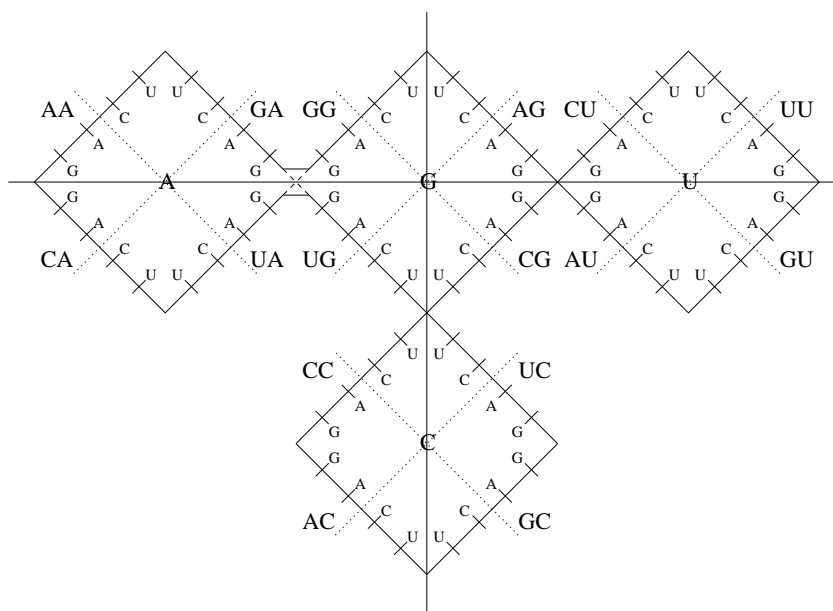


Fig. 1.5 Diamond presentation of codon labelling $NNN = (d_1, m_1, d_2, m_2, d_3, m_3)$

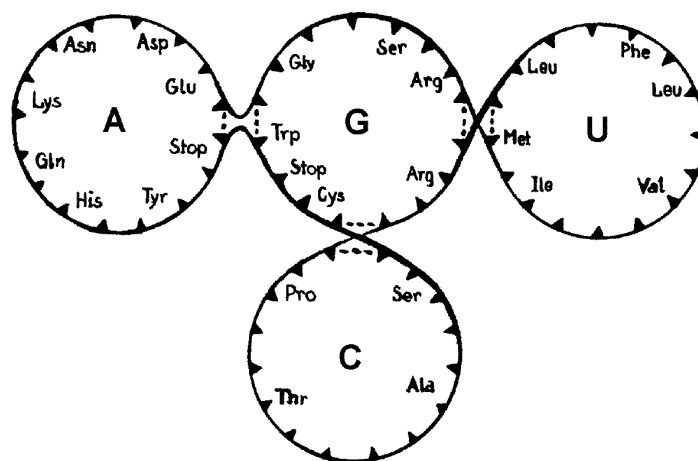


Fig. 1.6 Genetic code 'mutation rings' according to Siemion. The 'mutation number' $0 \leq k \leq 63$, labels the looping closed path around the main second base letter rings, with excursions into and out of the G ring starting with GAU (*Asp*).

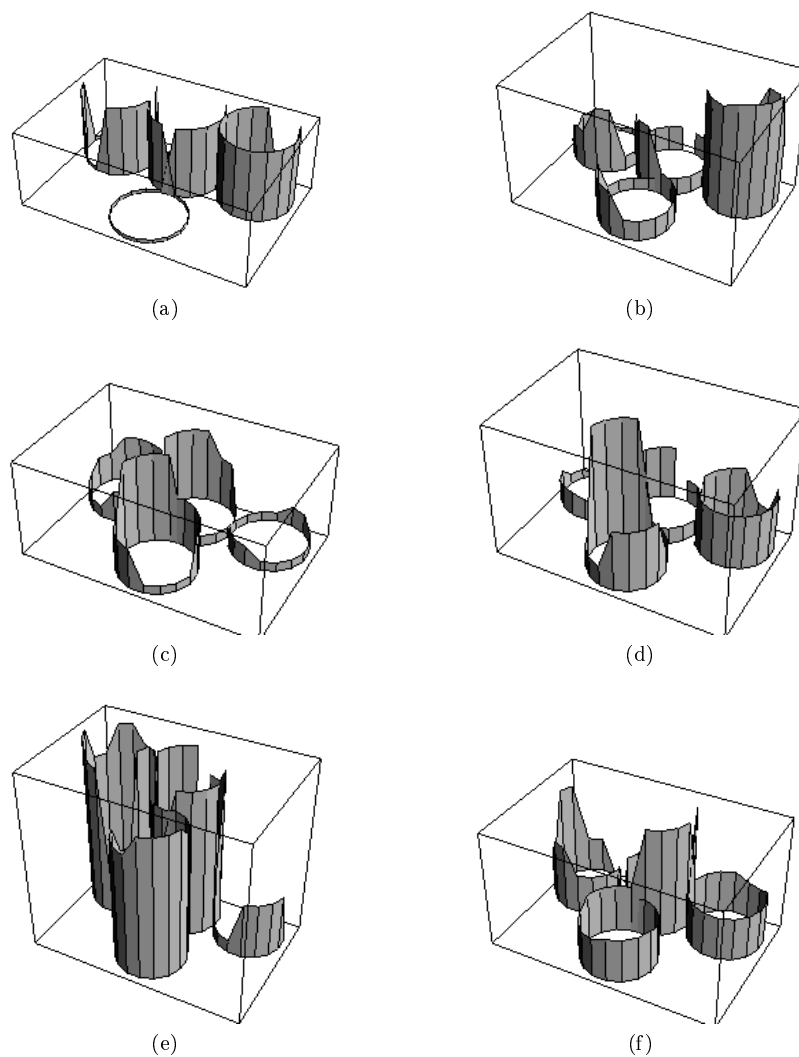


Fig. 1.7 Histograms of physicochemical parameters superimposed upon Siemion's rings: (a) *aaRS* synthetase class I=0, II=1; Chou-Fasman parameters relating (b) to beta sheets and (c) coils [Jiang *et al.* (1998)]; (d) *pKb* (a measure of codon polarity); [Sober (1970)] (e) hydrophobicity [Bull and Breese (1974)] and (f) isoelectronic potential [Sober (1970)].

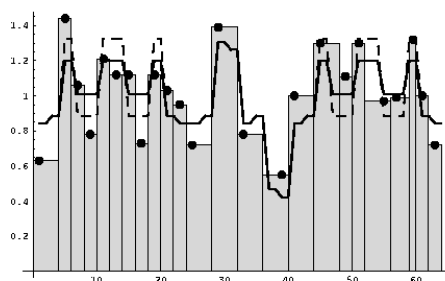


Fig. 1.8 P^α vs k . Histogram: data; solid & dashed curves: polynomial fits (four parameters); dots: preferred codon positions.

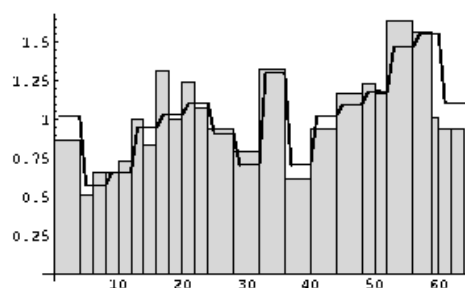


Fig. 1.9 P^β vs k . Histogram: data; solid curve: polynomial least squares fit (five parameters).

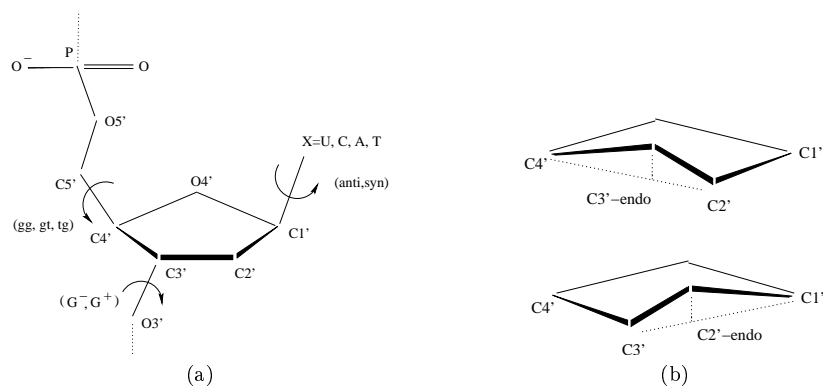


Fig. 1.10 (a) RNA backbone showing rotamer degrees of freedom. (b) Preferred ribose buckling conformations in A-form RNA.

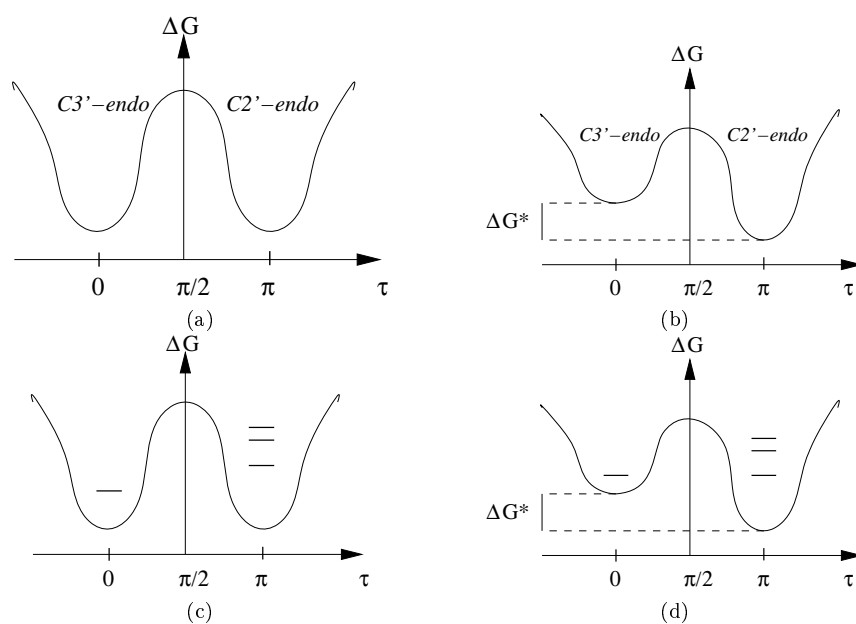


Fig. 1.11 (a) Sketch of Uridine ribose pseudorotation potential, showing equally stable *C2'-* and *C3'-endo* conformers. (b) Effect of xo^5 modification to Uridine on potential. (c) Same as (a) but with hypothetical (conformer) bound states imposed. (d) Same as (b) showing hypothetical bound states.

Bibliography

- Agris, P. (1991). Wobble position modified nucleosides evolved to select transfer RNA codon recognition: a modified wobble hypothesis, *Biochimie* **73**, pp. 1345–1349.
- Agris, P. (2004). Decoding the genome: a modified view, *Nucleic Acids Research* **32**, pp. 223–238.
- Altona, C. and Sundralingam, M. (1972). Conformational analysis of the sugar ring in nucleotides and nucleosides: a new description using the concept of pseudorotation, *Journal of the American Chemical Society* **94**, pp. 8205–8212.
- Andachi, Y., Yamao, F., Muto, A. and Osawa, S. (1989). Codon recognition patterns as deduced from sequences of the complete set of transfer RNA species in *mycoplasma capricolum*: resemblance to mitochondria, *Journal of Molecular Biology* **209**, pp. 37–54.
- Bashford, J. and Jarvis, P. (2000). The genetic code as a periodic table: algebraic aspects, *BioSystems* **57**, pp. 147–161.
- Bashford, J., Tsohantjis, I. and Jarvis, P. (1997). Codon and nucleotide assignments in a supersymmetric model of the genetic code, *Physics Letters A* **233**, pp. 481–488.
- Bashford, J., Tsohantjis, I. and Jarvis, P. (1998). A supersymmetric model for the origin of the genetic code, *Proceedings of the National Academy of Sciences, USA* **95**, pp. 987–992.
- Baumann, U. and Oro, J. (1993). Three stages in the evolution of the genetic code, *BioSystems* **29**, pp. 133–141.
- Beuning, P. and Musier Forsyth, K. (1999). Transfer RNA recognition by aminoacyl-tRNA synthetases, *Biopolymers* **52**, pp. 1–28.
- Biro, J., Benyó, B., Szlávecz, Á., Fördös, G., Micsik, T. and Benyó, Z. (2003). A common periodic table for codons and amino acids, *Biochemical and Biophysical Research Communications* **306**, pp. 408–415.
- Bock, A., Forchhammer, K., Heider, J., Leinfelder, W., Sawers, G., Veprek, B. and Zinoni, F. (1991). Selenocysteine: the 21st amino acid, *Molecular Microbiology* **5**, pp. 515–520.
- Bull, H. and Breese, K. (1974). Surface tension of amino acid solutions: a hy-

- drophobicity scale of the amino acid residues, *Archives of Biochemistry and Biophysics* **161**, pp. 665–670.
- Cavalcanti, A., Leite, E., Neto, B. and Ferreira, R. (2004). On the classes of aminoacyl-tRNA synthetases, aminoacids and the genetic code, *Origins of Life and Evolution of the Biosphere* **34**, pp. 407–420.
- Chechetkin, V. (2006). Genetic code from tRNA point of view, *Journal of Theoretical Biology* **242**, pp. 922–934.
- Crick, F. (1966). Codon-anticodon pairing: the wobble hypothesis, *Journal of Molecular Biology* **19**, pp. 548–555.
- Crick, F. (1968). The origin of the genetic code, *Journal of Molecular Biology* **38**, pp. 367–379.
- Di Giulio, M. (2003). The early phases of the genetic code origin: conjectures on the evolution of coded catalysis, *Origins of Life and Evolution of the Biosphere* **33**, pp. 479–489.
- Elsen, J. v. d., Wu, H. and Bagby, S. (2005). Evolution of the genetic triplet code via two types of doublet codons, *Journal of Molecular Evolution* **61**, pp. 54–64.
- Eriani, G., Delarue, M., Poch, O., Gangloff, J. and Moras, D. (1990). Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs, *Nature* **347**, pp. 203–206.
- Fabrega, C., Farrow, M., Mukhopadhyay, B., de Crecy-Lagard, V., Ortiz, A. and Schimmel, P. (2001). An aminoacyl tRNA synthetase whose sequence fits into neither of the two known classes, *Nature* **411**, pp. 110–114.
- Forger, M., Hornos, Y. and Hornos, J. (1997). Global aspects in the algebraic approach to the genetic code, *Physical Review E - Statistical Physics* **56**, 6, pp. 7078–7082.
- Forger, M. and Sachse, S. (2000a). Lie superalgebras and the multiplet structure of the genetic code.I. Codon representations, *Journal of Mathematical Physics* **41**, 8, pp. 5407–5422.
- Forger, M. and Sachse, S. (2000b). Lie superalgebras and the multiplet structure of the genetic code.II. Branching schemes, *Journal of Mathematical Physics* **41**, 8, pp. 5423–5444.
- Frappat, L., Sciarrino, A. and Sorba, P. (2001). Crystalizing the genetic code, *Journal of Biological Physics* **27**, 1, pp. 1–34.
- Frappat, L., Sorba, P. and Sciarrino, A. (1998). A crystal base for the genetic code, *Physics Letters A* **250**, 1-3, pp. 214–221.
- Freeland, S., Wu, T. and Keulmann, N. (2003). The case for an error minimising standard genetic code, *Origins of Life and Evolution of the Biosphere* **33**, pp. 457–477.
- Gamow, G. (1954). Possible relation between deoxyribonucleic acid and protein structures, *Nature* **173**, p. 318, 13 February 1954.
- Grover, L. (1997). Quantum mechanics helps in searching for a needle in a haystack, *Physical Review Letters* **79**, 2, p. 325.
- Hornos, J. and Hornos, Y. (1993). Algebraic model for the evolution of the genetic code, *Phys. Rev. Lett.* **71**, 26, pp. 4401–4404.
- Hornos, J. E. M., Hornos, Y. M. M. and Forger, M. (1999). Symmetry and sym-

- metry breaking: an algebraic approach to the genetic code, *International Journal of Modern Physics B* **13**, 23, pp. 2795–2885.
- Inagaki, Y., Kojima, A., Bessho, Y., Hori, H., Oham, T. and Osawa, S. (1995). Translation of synonymous codons in family boxes by *mycoplasma capricolum* tRNAs with unmodified Uridine or Adenosine at the first anticodon position, *Journal of Molecular Biology* **251**, pp. 486–492.
- Jiang, B., Gou, T., Peng, L.-W. and Sun, Z.-R. (1998). Folding type-specific secondary structure propensities of amino acids, derived from α -helical, β -sheet, α/β , and $\alpha+\beta$ proteins of known structures, *Biopolymers* **45**, pp. 35–49.
- Jiménez-Montaña, M. A. (1999). Protein evolution drives the evolution of the genetic code and vice versa, *Biosystems* **54**, pp. 47–64.
- Jiménez-Montaña, M. A., de la Mora Basáñez, C. R. and Pöschel, T. (1996). The hypercube structure of the genetic code explains conservative and non-conservative amino acid substitutions *in vivo* and *in vitro*, *BioSystems* **39**, pp. 117–125.
- Jiménez Sanchez, A. (1995). On the origin and evolution of the genetic code, *Journal of Molecular Evolution* **41**, pp. 712–716.
- Kano, A., Andachi, Y., Ohama, T. and Osawa, S. (1991). Novel anticodon composition of transfer RNAs in *micrococcus luteus*, a bacterium with a high GC genomic content: correlation with codon usage, *Journal of Molecular Biology* **221**, pp. 387–401.
- Kawashima, S. and Kanehisa, M. (2000). AAindex: Amino acid index database. *Nucleic Acids Research* **28**, p. 374, <http://www.genome.ad.jp/dbget/>.
- Kent, R. D., Schlesinger, M. and Wybourne, B. G. (1998). On algebraic approaches to the genetic code, *Canadian J Phys* **76**, pp. 445–52.
- Knapp, M. J. and Klinman, J. P. (2002). Environmentally coupled hydrogen tunneling: linking catalysis to dynamics, *European Journal of Biochemistry* **269**, pp. 3113–3121.
- Kurland, C., Hughes, D. and Ehrenberg, M. (1996). *Limitations of translational accuracy. In Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology. Volume I (FC Neidhart et al. eds.)* (American Society for Microbiology Press).
- Lim, V. and Curran, J. (2001). Analysis of codon :anticodon interactions within the ribosome provides new insights into codon reading and the genetic code structure, *RNA* **7**, pp. 942–957.
- Mac Dónaill, D. (2003). Why nature chose A,C,G and U/T: an error-coding perspective of nucleotide alphabet composition, *Origins of Life and Evolution of the Biosphere* **33**, pp. 433–455.
- Mac Dónaill, D. and Manktelow, M. (2004). Molecular informatics: quantifying information patterns in the genetic code, *Molecular Simulation* **30**, 5, pp. 267–272.
- Négadi, T. (2003). Rumer's transformation, in biology, as the negation, in classical logic, *International Journal of Quantum Chemistry* **94**, pp. 65–74.
- Nieselt Struwe, K. and Wills, P. (1997). The emergence of genetic coding in physical systems, *Journal of Theoretical Biology* **187**, pp. 1–14.

- Ninio, J. (2006). Multiple stages in codon-anticodon: double-trigger mechanisms and geometric constraints, *Biochimie* **88**, pp. 963–992.
- Ogle, J., Carter, A., and Ramakrishnan, V. (2003). Insights into the decoding mechanism from recent ribosome structures, *Trends in Biochemical Sciences* **28**, pp. 259–266.
- Osawa, S., Jukes, T. H., Watanabe, K. and Muto, A. (1992). Recent evidence for evolution of the genetic code, *Microbiological Reviews* **56**, pp. 229–264.
- Patel, A. (2001a). Quantum algorithms and the genetic code, *Pramana - Journal of Physics* **56**, 2-3, pp. 367–381.
- Patel, A. (2001b). Testing quantum dynamics in genetic information processing, *Journal of Genetics* **80**, 1, pp. 39–43.
- Patel, A. (2001c). Why genetic information processing could have a quantum basis, *Journal of Biosciences* **26**, 2, pp. 145–151.
- Patel, A. (2005). The triplet genetic code had a doublet predecessor, *Journal of Theoretical Biology* **233**, 4, pp. 527–532.
- Poole, A., Jeffares DC and Penny, D. (1996). The path from the RNA world, *Journal of Molecular Evolution* **46**, pp. 1–17.
- Ronneberg, T., Landweber, L. and Freeland, S. (2000). Testing a biosynthetic theory of the genetic code: fact or artifact? *Proceedings of the National Academy of Sciences, USA* **97**, pp. 13690–13695.
- Rumer, Y. (1966). Systematization of the codons of genetic code, *Translated from Doklady Akademii Nauk SSSR* **167**, pp. 1393–1394.
- Saenger, W. (1984). *Principles of Nucleic Acid Structure* (Springer-Verlag, New York).
- Seligmann, H. and Amzallag, G. (2002). Chemical interactions between amino acid and rna: multiplicity of the levels of specificity explains origin of the genetic code, *Naturwissenschaften* **89**, pp. 542–551.
- Shabalina, S., Ogurtsov, A. and Spiridonov, N. (2006). A periodic table of mRNA secondary structure created by the genetic code, *Nucleic Acids Research* **34**, pp. 2428–2437.
- Shapira, D., Shimon, Y. and Biham, O. (2005). Algebraic analysis of quantum search with pure and mixed states, *Physical Review A* **71**, 4, pp. 1–8.
- Siemion, I. (1994). The regularity of changes of Chou-Fasman parameters within the genetic code, *BioSystems* **32**, pp. 25–35.
- Sober, H. E. (1970). *C. R. C. Handbook of Biochemistry : Selected data for molecular biology* (Chemical Rubber Co, Cleveland), 2nd edition.
- Srinivasan, G., James, C. M. and Krzycki, J. A. (2002). Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA, *Science* **296**, pp. 1459–1462.
- Swanson, R. (1984). A unifying concept for the amino acid code, *Bull Math Biol* **46**, p. 187.
- Szathmáry, E. (1999). The origin of the genetic code: amino acids as cofactors in an RNA world, *Trends in Genetics* **15**, pp. 223–229.
- Takai, K. (2006). Classification of the possible pairs between the first anticodon and the third codon positions based on a simple model assuming two geometries with which the pair effectively potentiates the decoding complex,

- Journal of Theoretical Biology* **242**, pp. 564–580.
- Takai, K. and Yokoyama, S. (2003). Roles of 5'-substituents of tRNA wobble Uridines in the recognition of purine-ending codons, *Nucleic Acids Research* **31**, pp. 6383–6391.
- Weber, A. and Lacey, J. (1978). Genetic code correlations: amino acids and their anticodon nucleotides, *Journal of Molecular Evolution* **11**, pp. 199–210.
- Weberndorfer, G., Hofacker, I. L. and Stadler, P. F. (2003). On the evolution of primitive genetic code, *Origins of Life and Evolution of the Biosphere* **33**, pp. 491–514.
- Woese, C., Dugre, D., Saxinger, W. and Dugre, S. (1966). The molecular basis for the genetic code, *Proceedings of the National Academy of Sciences, USA* **55**, pp. 966–974.
- Wong, J. (1976). The evolution of a universal genetic code, *Proceedings of the National Academy of Sciences, USA* **73**, 7, pp. 2336–2340.
- Yang, C. (2005). On the structural regularity in nucleobases and amino acids and relationship to the origin and evolution of the genetic code, *Origins of Life and Evolution of Biosphere* **35**, 3, pp. 275–295.
- Yokoyama, S. and Nishimura, S. (1995). Modified nucleosides and codon recognition, in D. Söll and U. RajBhandary (eds.), *tRNA: Structure, Biosynthesis and Function* (American Society for Microbiology, Washington DC), pp. 207–223.
- Yokoyama, S., Watanabe, T., Murao, K., Ishikura, H., Yamaizumi, Z., Nishimura, S. and Miyazawa, T. (1985). Molecular mechanisms of codon recognition by tRNA species with modified Uridine in the first position of the anticodon, *Proceedings of the National Academy of Sciences, USA* **82**, pp. 4905–4909.

Index

- algorithm
 - Grover's, 9, 25
- anticodon, 3, 25, 28
- base
 - complementary, 4
 - nucleotide, 3
- base, classification, 6
- base, conformational states, 25
- box, 6, 15, 24
 - family, 4, 18, 28
 - mixed, 4, 19, 28
- branching rule, 12, 15, 22
- genetic code
 - doublet predecessor, 5
 - hypotheses for evolution of, 7
 - mitochondrial, 28
 - mitochondrial, 5
 - universal (eukaryotic), 5
 - variations, 19
- operator
 - Casimir, 10
 - Hamiltonian, 20, 21, 28
- protein
 - assembly, 8
 - structure, 24
- proteins, 5, 11
 - amino-acyl tRNA synthetase (aaRS), 4, 8, 18, 25, 34
- elongation factor (EF-Tu), 3, 25
- superposition, 9, 26
- superposition, 25
- supersymmetric quantum mechanics, 20
- supersymmetry
 - algebras, 10
- symmetry breaking, 10, 20, 29
 - partial, 19
- symmetrybreaking, partial, 20
- Uridine, 18, 27
 - modified, 27
- weight label, 13, 15, 22
- wobble pairing, 4, 5, 18, 27