JOURNAL OF MOLECULAR EVOLUTION

© Springer-Verlag New York Inc. 1999

Physicochemical Optimization in the Genetic Code Origin as the Number of Codified Amino Acids Increases

Massimo Di Giulio,[1] Mario Medugno[2]

[1] International Institute of Genetics and Biophysics, CNR, Via G. Marconi 10, 80125 Naples, Napoli, Italy
[2] Centro di Ricerca per il Calcolo Parallelo e Supercalcolo, CNR, Via Diocleziano 328, 80124 Naples, Napoli, Italy

Received: 25 October 1998 / Accepted: 19 February 1999

Abstract. We have assumed that the coevolution theory of genetic code origin (Wong JT, Proc Natl Acad Sci USA 72:1909–1912, 1975) is essentially correct. This theory makes it possible to identify at least 10 evolutionary stages through which genetic code organization might have passed prior to reaching its current form. The calculation of the minimization level of all these evolutionary stages leads to the following conclusions. (1) The minimization percentages increased linearly with the number of amino acids codified in the codes of the various evolutionary stages when only the sense changes are considered in the analysis. This seems to favor the physicochemical theory of genetic code origin even if, as discussed in the paper, this observation is also compatible with the coevolution theory. (2) For the first seven evolutionary stages of the genetic code, this trend is less clear and indeed is inverted when we consider the global optimisation of the codes due to both sense changes and synonymous changes. This inverse correlation between minimization percentages and the number of amino acids codified in the codes of the intermediate stages seems to favor neither the physicochemical nor the stereochemical theories of genetic code origin, as it is in the early and intermediate stages of code development that these theories would expect minimization to have played a crucial role, and this does not seem to be the case. However, these results are in agreement with the coevolution theory, which attributes a role to the physicochemical properties of amino acids that, while important, is nevertheless subordinate to the mechanism which concedes codons from the precursor amino acids to the product amino acids as the primary factor determining the evolutionary structuring of the genetic code. The results are therefore discussed in the context of the various theories proposed to explain genetic code origin.

Key words: Genetic code theories — Intermediate evolutionary stages — Error minimization — Coevolution — Polarity and molecular volume of amino acids

Working Plan

The coevolution theory of genetic code origin (Wong 1975) seems to be the best hypothesis we have to explain the origin of genetic code organization (Di Giulio 1997a, b, 1999). This theory suggests that early on in the genetic code, only precursor amino acids were codified (Wong 1975). As product amino acids evolved from these through biosynthetic pathways, part or all of the codon domain of precursor amino acids was conceded to product amino acids (Wong 1975). This theory therefore makes specific predictions on the relative evolutionary times at which amino acids first appeared and, consequently, on their entry into the organization of the genetic code. Hence this theory furnishes a unique opportunity to follow the evolution of genetic code structuring.

Moreover, it is known that the physicochemical properties of amino acids played an important role in organizing the genetic code (for references, see Szathmary 1993; Di Giulio 1997a). Therefore, if we also consider

Correspondence to: M. Di Giulio; *e-mail:* digiulio@iigb.na.cnr.it

certain distance functions, an expression of the physicochemical properties of amino acids, as the number of amino acids codified in the evolving genetic code varies (in accordance with the coevolution theory), we will be able to investigate the evolutionary route that followed the structuring of the genetic code. Moreover, we might be able to understand better the interactions between the two main forces held to be responsible for genetic code organization: the biosynthetic relationships between amino acids and their physicochemical properties (for references, see Szathmary 1993; Di Giulio 1997a). This working plan seems to provide a unique opportunity furnished by the coevolution theory, as it makes it possible to follow the evolution of a biological organization as the number of its constituents varies, which is very often impossible in biology because of the lack of information on the intermediate evolutionary stages.

It should, therefore, be worthwhile following the evolutionary structuring of the genetic code, i.e., finding out how the physicochemical variables of amino acids behave as the number of amino acids codified in the evolving genetic code varies. This might reveal interesting aspects of genetic code origin, and for these reasons, we decided to conduct such an analysis.

Materials and Methods

Readers who are also interested in the strictly technical aspects encountered in this paper are referred to the relevant literature (Wong 1980; Di Giulio 1989; Di Giulio et al. 1994; Di Giulio and Medugno 1998). However, to make the present paper "self-standing," some brief information is given below.

The distance functions (Δ) used are, in one case, given simply by the absolute value of the difference between the polarity values of amino acids weighted with values that can be extracted from the genetic code structure (Di Giulio 1989; Di Giulio et al. 1994; Di Giulio and Medugno 1998). (These weights are given simply by the number of times the codons of a certain amino acid transform by single-base substitution into those of another amino acid according to the genetic code structure). In the second case, the function (Δ) combines the absolute values of the differences between the polarity values of amino acids, normalized with the relative standard deviation, with those of the molecular volumes of amino acids (again, normalized with the relative standard deviation) and the resulting distances are always weighted with values that can be extracted from the genetic code structure (Di Giulio et al. 1994; Di Giulio and Medugno 1998).

For a general introduction to the problem of minimizing the physicochemical distances between amino acids in genetic code origin, see Wong (1980) and Di Giulio (1989). In particular, the values of Δ_{mean}, Δ_{code}, and Δ_{low} make it possible to calculate the level reached by the minimization of certain distances in the code being studied. These values (Wong 1980; Di Giulio 1989; Di Giulio et al. 1994; Di Giulio and Medugno 1998) are the value that the distance function (Δ) assumes: (1) in the mean random code (Δ_{mean}), (2) in the code being studied (Δ_{code}), and (3) in the particular amino acid configuration of the code being studied that minimizes the distance function value (Δ_{low}). These three values make it possible to calculate the minimization percentage $[=(\Delta_{mean} - \Delta_{code})/(\Delta_{mean} = \Delta_{low}) \times 100]$ representing the optimization level of that particular code (Wong 1980; Di Giulio 1989; Di Giulio et al. 1994; Di Giulio and Medugno 1998).

All calculations were automated through the use of ad hoc programs. In particular, (1) to calculate Δ_{mean}, we wrote a program, named MEA, which calculates the mean of the desired distances; (2) to calculate Δ_{code}, we wrote a program, named EC, which evaluates the distance function value for a specific amino acid configuration for a certain code; and (3) to calculate Δ_{low}, we wrote a program named EXAS (see the Appendix) capable of generating all the possible $n!$ permutations of an array of n elements and, therefore, of conducting an exhaustive search for the distance function minimum when the number of permutations was not very high, while we used the simulated annealing technique (Di Giulio et al. 1994) when this number was too high, i.e., for codes codifying for 13 or more amino acids; and (4) to estimate the mean number of synonymous changes present in the random code of a certain code having a specific plurality of codons, we wrote a program named ESE (see the Appendix). Five million random codes were generated to estimate this mean number.

All these programs run on PC and are available upon request from the authors.

Results

By following the coevolution theory (Wong 1975) it is a fairly simple task to define the probable intermediate stages through which genetic code organization passed prior to reaching its current form. In particular, we used Wong's Fig. 1 (1975, p 1910) and the considerations referred therein, the comments made in another paper by Wong (1988), and the comments and Fig. 1 for the number of biosynthetic steps reported by Taylor and Coates (1989). If we follow the indications given in these three papers (Wong 1975, 1988; Taylor and Coates 1989), we can reasonably define 10 intermediate stages before the code reaches its current form (Fig. 1). A few brief comments are needed to clarify certain choices made in Fig. 1. In Fig. 1a Ser occupies almost the whole first row of the code. This was necessary (1) because Ser is held to be a very ancient amino acid (Wong 1988) and is the precursor amino acid of Cys and Trp (Wong 1975); (2) because it is unlikely that Phe and Tyr were used early on in the code (Wong 1988); (3) because there seems to be a correlation between amino acids in biosynthetic relationships and the rows of the genetic code (Taylor and Coates 1989); and finally, (4) because Phe and Tyr, through phospho*enol*pyruvate and phosphoglycerate, could be in biosynthetic relationship with Ser (Taylor and Coates 1989). The other choice that needs to be explained regards Fig. 1a. The pyruvate family includes the amino acids Ala, Val, and Leu (Wong 1988; Taylor and Coates 1989). Of these, Ala and Val are considered to be extremely ancient amino acids (Wong 1988) and therefore might both be included in Fig. 1a. Indeed, these are both included in Fig. 1b. Nevertheless, we have chosen to include only Ala in Fig. 1a because this amino acid plays a central role in the metabolism of amino acids (Wong 1975) and because, in this first stage of genetic code evolution (Fig. 1a), we thought it best not to subdivide the domain of codons codifying for precursor amino acids as imposed by the coevolution theory. However, we have performed the calculations with Val in-

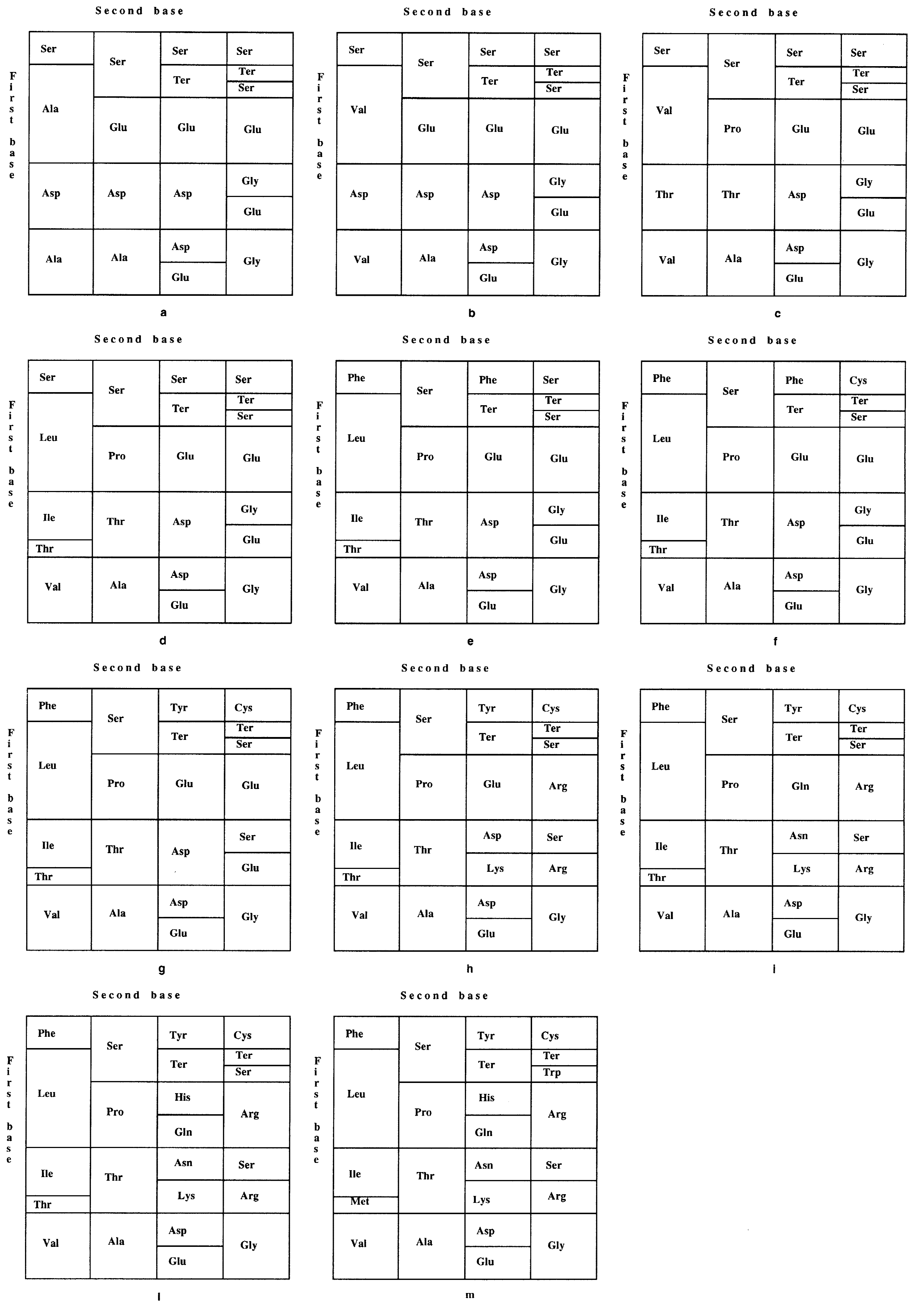

Fig. 1. This shows the probable evolutionary stages through which genetic code organization passed, as identified on the basis of the coevolution theory of genetic code origin (Wong 1975). See text for further information.

Table 1. Summary of all the important variables concerning the function that uses amino acid polarity distances[a]

Code in Figs. 1 and 2	Number of amino acids in the code	Without synonymous changes				With synonymous changes				Synonymous changes only (minimization percentage)
		Δ_{mean}	Δ_{code}	Δ_{low}	Minimization percentage	Δ_{mean}	Δ_{code}	Δ_{low}	Minimization percentage	
a	5	3.40	3.51	2.90	−22.0	2.71	1.73	1.22	65.8	87.8
b	6	3.59	3.87	2.72	−32.2	2.94	1.97	1.29	58.8	91.0
c	8	3.06	2.98	2.11	8.4	2.67	1.70	1.10	61.8	53.4
d	10	3.13	3.00	2.21	14.1	2.79	1.85	1.25	61.0	46.9
e	11	3.05	2.93	2.16	13.5	2.77	1.90	1.21	55.8	42.3
f	12	2.97	2.97	1.92	0.0	2.72	1.96	1.20	50.0	50.0
g	13	2.84	2.90	1.61	−4.9	2.60	1.96	0.95	38.8	43.7
h	15	2.99	2.39	1.82	51.3	2.81	1.69	1.18	68.7	17.4
i	17	2.97	2.00	1.74	78.9	2.81	1.44	1.21	85.6	6.7
l	18	2.90	1.96	1.71	79.0	2.75	1.45	1.22	85.0	6.0
m	20	2.82	1.94	1.61	72.7	2.69	1.44	1.06	76.7	4.0

[a] Each row refers to the values of the variables referring to one of the codes in Figs. 1 and 2, as indicated in column 1. The rest of the table is self-explanatory. See the text for comments and further information.

stead of Ala in Fig. 1a, obtaining equivalent results (data not shown).

We feel that Fig. 1 does not seem to require further comments, as it essentially follows Wong's Fig. 1 (1975, p. 1910) and his considerations. Nevertheless, for clarity's sake, it is worth following the evolution of a group of amino acids in biosynthetic relationships: the Asp family.

As can be seen Asp is initially codified by 14 codons (Fig. 1a) (Wong 1975). At the eight-amino acid stage, Asp concedes eight codons to Thr (Wong 1975) (Fig. 1c) both because Thr is considered to be an amino acid that entered the code relatively early on (Wong 1988) and because Thr is separated from Asp by just a few biosynthetic steps (Taylor and Coates 1989). At the 10-amino acid stage, Thr concedes three codons to Ile (Fig. 1d) because Ile is considered to be an amino acid that developed relatively early on (Wong 1988). This situation for the Asp family remains stable until the 15-amino acid stage, when Lys, a late amino acid (Wong 1988), developed from Asp (Fig. 1h), whereas in the subsequent stage Asn developed from Asp (Fig. 1i). The choice to let Lys enter the code before Asn seems paradoxical because Asn is separated from Asp by just one biosynthetic step, while Lys is separated from Asp by several biosynthetic steps (Taylor and Coates 1989) and both these amino acids are considered to be late developers (Wong 1988), whereas this choice seems to be justified with the coevolution theory (Wong 1975) because, if Lys is an amino acid produced by Asp, for Asp to concede codons to Lys, it must have maintained the codons AAU and AAC until all the amino acids produced by Asp developed, and only at the end did it concede these codons to Asn (Wong 1975). This seems to justify the entry of Lys into the evolving genetic code before Asn (Figs. 1h and i). Finally, the development of Met from Thr (Wong 1975) is hypothesized as the last stage (Fig. 1m) both because Met is considered to be a late amino acid (Wong 1988) and because it is codified by a single codon. Similar considerations are made for the other families of precursor amino acids and seem to justify the succession of evolutionary stages of the genetic code reported in Fig. 1.

We then conducted an analysis to calculate the minimization percentages. In particular, for each of the configurations reported in Fig. 1 we calculated the corresponding Δ_{code} value (see Materials and Methods) both for the polarity distances of amino acids (Table 1) and for the polarity and molecular volume distances (Table 2), following the distance functions already introduced (Di Giulio 1989; Di Giulio et al. 1994; Di Giulio and Medugno 1998). For each of the configurations (Fig. 1) we then calculated the Δ_{low} value (see Materials and Methods), i.e., the amino acid configuration minimizing the objective function value both for polarity distances (Fig. 2) and for those combining polarity and molecular volume (Fig. 3). These configurations (Figs. 2 and 3) were identified using two programs (see Materials and Methods). The first program, used for codes of up to 12 amino acids (Figs. 1a–f), conducts an exhaustive search on all possible permutations and identifies the one with the minimum value in the objective function (see the Appendix); the second, used for codes of from 13 to 20 amino acids (Figs. 1g–m), identifies the configuration with a minimum value in the objective function by means of the simulated annealing technique (Di Giulio et al. 1994). Figure 2 reports the amino acid configurations that minimize the function value for polarity distances, while Fig. 3 reports those that minimize the function value for combined polarity and molecular volume distances.

For each set of amino acids (Fig. 1) we went on to calculate the corresponding Δ_{mean} values (see Materials and Methods) and, finally, the minimization percentages (Tables 1 and 2; without synonymous changes).

Analogously to this analysis, which does not take into account the synonymous changes in the codes being

Second base / First base

Glu Gly 7.9	Glu 12.5	Glu Ter	Glu Ter Glu
	Ser 7.5	Ser	Ser
Ala 7.0	Ala	Ala	Asp Ser
Gly	Gly	Ala Ser	Asp 13.0

a

Second base / First base

Val Ala 7.0	Val 5.6	Val Ter	Val Ter Val
	Gly 7.9	Gly	Gly
Ser 7.5	Ser	Ser	Glu Gly
Ala	Asp 13.0	Ser Gly	Glu 12.5

b

Second base / First base

Gly Ser 7.5	Gly 7.9	Gly Ter	Gly Ter Gly
	Glu 12.5	Pro 6.6	Pro
Ala	Ala 7.0	Val 5.6	Thr Pro
Ser	Asp 13.0	Val Pro	Thr 6.6

c

Second base / First base

Ser Gly 7.9	Ser 7.5	Ser Ter	Ser Ter Ser
	Ile 4.9	Pro 6.6	Pro
Asp 13.0 Val	Val 5.6	Thr 6.6	Ala Pro
Glu12.5	Leu 4.9	Thr Pro	Ala 7.0

d

Second base / First base

Ser 7.5 Gly 7.9	Val 5.6	Ser Ter	Val Ter Val
	Ile 4.9	Pro 6.6	Pro
Asp 13.0 Phe	Phe 5.0	Ala 7.0	Thr Pro
Glu 12.5	Leu 4.9	Ala Pro	Thr 6.6

e

Second base / First base

Glu 12.5 Ser 7.5	Ile 4.9	Glu Ter	Asp 13.0 Ter Ile
	Cys 4.8	Pro 6.6	Pro
Val 5.6 Phe	Phe 5.0	Thr 6.6	Ala Pro
Gly 7.9	Leu 4.9	Thr Pro	Ala 7.0

f

Second base / First base

Cys 4.8 Leu 4.9	Ala 7.0	Asp 13.0 Ter	Glu 12.5 Ter Ala
	Val 5.6	Pro 6.6	Pro
Phe 5.0 Tyr	Tyr 5.4	Thr 6.6	Ala Pro
Ile 4.9	Ser 7.5	Thr Pro	Gly 7.9

g

Second base / First base

Cys 4.8 Leu 4.9	Pro 6.6	Asp 13.0 Ter	Glu 12.5 Ter Pro
	Val 5.6	Gly 7.9	Ala 7.0
Phe 5.0 Tyr	Tyr 5.4	Arg 9.1 Lys 10.1	Pro Ala
Ile 4.9	Thr 6.6	Arg 9.1 Gly 7.9	Ser 7.5

h

Second base / First base

Cys 4.8 Leu 4.9	Thr 6.6	Arg 9.1 Ter	Gly 7.9 Ter Thr
	Val 5.6	Gln 8.6	Ala 7.0
Phe 5.0 Tyr	Tyr 5.4	Asn 10.0 Lys 10.1	Thr Ala
Ile 4.9	Pro 6.6	Glu 12.5 Asp 13.0	Ser 7.5

i

Second base / First base

Cys 4.8 Leu 4.9	Thr 6.6	Arg 9.1 Ter	Gly 7.9 Ter Thr
	Pro 6.6	Asn 10.0 Lys 10.1	Ala 7.0
Phe 5.0 Tyr	Tyr 5.4	Gln 8.6 His 8.4	Thr Ala
Ile 4.9	Val 5.6	Glu 12.5 Asp 13.0	Ser 7.5

l

Second base / First base

Asn 10.0 Gln 8.6	Ala 7.0	Trp 5.2 Ter	Lys 10.1 Ter Asp 13.0
	Pro 6.6	Phe 5.0 Cys 4.8	Ser 7.5
Arg 9.1 Glu 12.5	Val 5.6	Met 5.3 Tyr 5.4	Ala Ser
His 8.4	Thr 6.6	Leu 4.9 Ile 4.9	Gly 7.9

m

Fig. 2. This shows the amino acid code configurations that minimize the objective function value for polarity distances (Δ_{low}). These configurations are related to the corresponding intermediate code (same letter) identified by the coevolution theory (Fig. 1). The numbers indicate the amino acid polarity values (Woese et al. 1966). See text for further information.

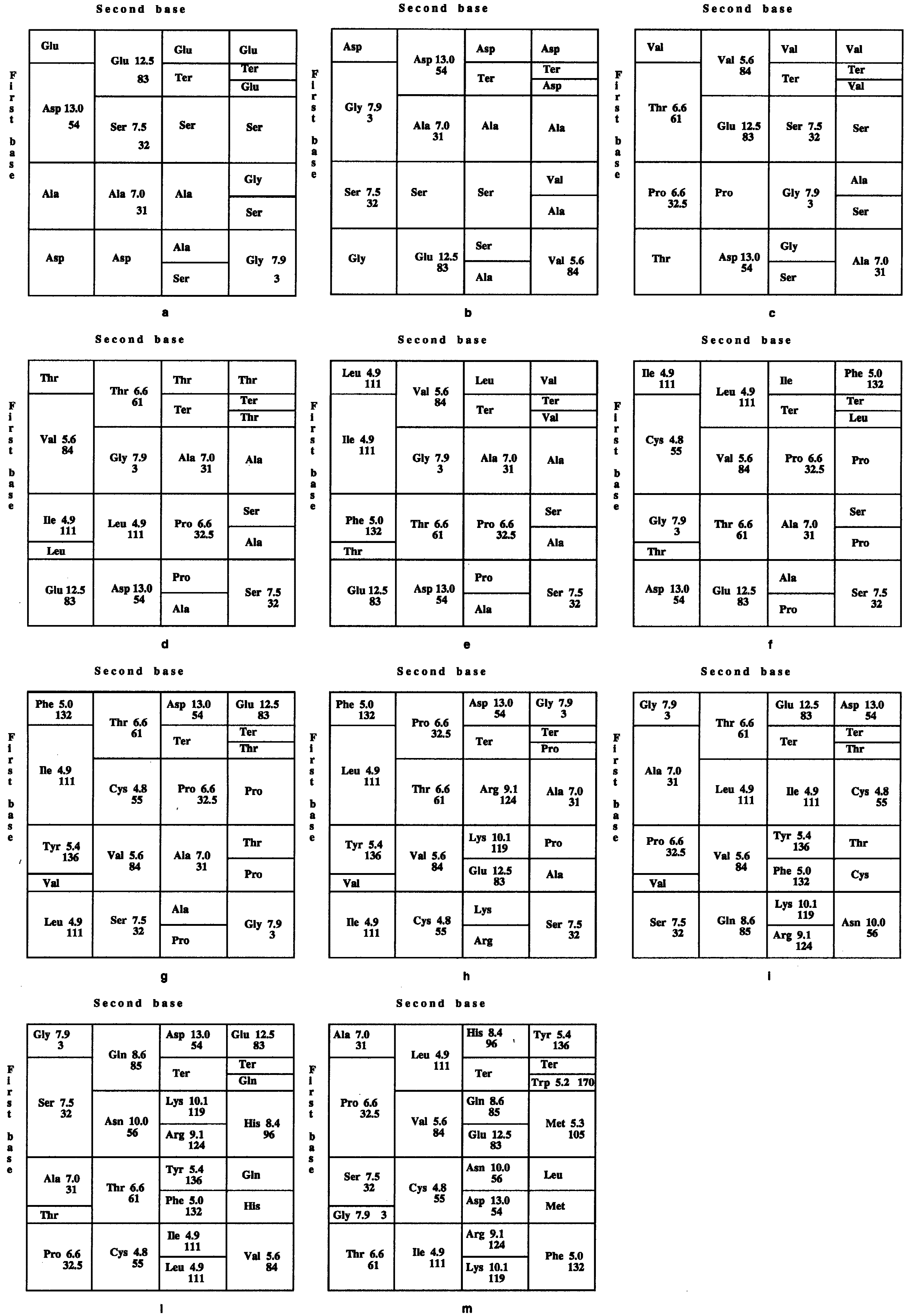
Second base
First base
a
Glu
Glu 12.5 83
Glu
Glu
Ter
Ter
Glu
Asp 13.0 54
Ser 7.5 32
Ser
Ser
Ala
Ala 7.0 31
Ala
Gly
Ser
Asp
Asp
Ala
Ser
Gly 7.9 3
b
Asp
Asp 13.0 54
Asp
Asp
Ter
Ter
Asp
Gly 7.9 3
Ala 7.0 31
Ala
Ala
Ser 7.5 32
Ser
Ser
Val
Ala
Gly
Glu 12.5 83
Ser
Ala
Val 5.6 84
c
Val
Val 5.6 84
Val
Val
Ter
Ter
Val
Thr 6.6 61
Glu 12.5 83
Ser 7.5 32
Ser
Pro 6.6 32.5
Pro
Gly 7.9 3
Ala
Ser
Thr
Asp 13.0 54
Gly
Ser
Ala 7.0 31
d
Thr
Thr 6.6 61
Thr
Thr
Ter
Ter
Thr
Val 5.6 84
Gly 7.9 3
Ala 7.0 31
Ala
Ile 4.9 111
Leu
Leu 4.9 111
Pro 6.6 32.5
Ser
Ala
Glu 12.5 83
Asp 13.0 54
Pro
Ala
Ser 7.5 32
e
Leu 4.9 111
Val 5.6 84
Leu
Val
Ter
Ter
Val
Ile 4.9 111
Gly 7.9 3
Ala 7.0 31
Ala
Phe 5.0 132
Thr
Thr 6.6 61
Pro 6.6 32.5
Ser
Ala
Glu 12.5 83
Asp 13.0 54
Pro
Ala
Ser 7.5 32
f
Ile 4.9 111
Leu 4.9 111
Ile
Phe 5.0 132
Ter
Ter
Leu
Cys 4.8 55
Val 5.6 84
Pro 6.6 32.5
Pro
Gly 7.9 3
Thr
Thr 6.6 61
Ala 7.0 31
Ser
Pro
Asp 13.0 54
Glu 12.5 83
Ala
Pro
Ser 7.5 32
g
Phe 5.0 132
Thr 6.6 61
Asp 13.0 54
Glu 12.5 83
Ter
Ter
Thr
Ile 4.9 111
Cys 4.8 55
Pro 6.6 32.5
Pro
Tyr 5.4 136
Val
Val 5.6 84
Ala 7.0 31
Thr
Pro
Leu 4.9 111
Ser 7.5 32
Ala
Pro
Gly 7.9 3
h
Phe 5.0 132
Pro 6.6 32.5
Asp 13.0 54
Gly 7.9 3
Ter
Ter
Pro
Leu 4.9 111
Thr 6.6 61
Arg 9.1 124
Ala 7.0 31
Tyr 5.4 136
Val
Val 5.6 84
Lys 10.1 119
Glu 12.5 83
Pro
Ala
Ile 4.9 111
Cys 4.8 55
Lys
Arg
Ser 7.5 32
i
Gly 7.9 3
Thr 6.6 61
Glu 12.5 83
Asp 13.0 54
Ter
Ter
Thr
Ala 7.0 31
Leu 4.9 111
Ile 4.9 111
Cys 4.8 55
Pro 6.6 32.5
Val
Val 5.6 84
Tyr 5.4 136
Phe 5.0 132
Thr
Cys
Ser 7.5 32
Gln 8.6 85
Lys 10.1 119
Arg 9.1 124
Asn 10.0 56
l
Gly 7.9 3
Gln 8.6 85
Asp 13.0 54
Glu 12.5 83
Ter
Ter
Gln
Ser 7.5 32
Asn 10.0 56
Lys 10.1 119
Arg 9.1 124
His 8.4 96
Ala 7.0 31
Thr
Thr 6.6 61
Tyr 5.4 136
Phe 5.0 132
Gln
His
Pro 6.6 32.5
Cys 4.8 55
Ile 4.9 111
Leu 4.9 111
Val 5.6 84
m
Ala 7.0 31
Leu 4.9 111
His 8.4 96
Tyr 5.4 136
Ter
Ter
Trp 5.2 170
Pro 6.6 32.5
Val 5.6 84
Gln 8.6 85
Glu 12.5 83
Met 5.3 105
Ser 7.5 32
Gly 7.9 3
Cys 4.8 55
Asn 10.0 56
Asp 13.0 54
Leu
Met
Thr 6.6 61
Ile 4.9 111
Arg 9.1 124
Lys 10.1 119
Phe 5.0 132

Fig. 3. This shows the amino acid code configurations that minimize the objective function value for distances combining polarity and molecular volume of amino acids (Δ_{low}). These configurations are related to the corresponding intermediate code (same letter) identified by the coevolution theory (Fig. 1). The numbers indicate the polarity values above (Woese et al. 1966) and the molecular volume values below (Grantham 1974). See text for further information.

Table 2. Summary of all the important variables concerning the function that uses the distances combining amino acid polarity and molecular volume values

		Without synonymous changes				With synonymous changes				
Code in Figs. 1 and 3	Number of amino acids in the code	Δ_{mean}	Δ_{code}	Δ_{low}	Minimization percentage	Δ_{mean}	Δ_{code}	Δ_{low}	Minimization percentage	Synonymous changes only (minimization percentage)
a	5	3.20	3.24	2.76	−9.1	2.55	1.60	1.17	68.8	77.9
b	6	3.14	3.22	2.63	−15.7	2.57	1.64	1.22	68.9	84.6
c	8	2.76	2.83	2.22	−13.0	2.40	1.61	1.10	60.8	73.8
d	10	2.76	2.64	2.08	17.6	2.46	1.63	1.19	65.4	47.8
e	11	2.72	2.49	2.04	33.8	2.47	1.61	1.14	64.7	30.9
f	12	2.64	2.51	1.90	17.6	2.41	1.66	1.15	59.5	41.9
g	13	2.60	2.31	1.76	34.5	2.38	1.56	1.06	62.1	27.6
h	15	2.77	2.35	1.93	50.0	2.60	1.66	1.28	71.2	21.2
i	17	2.85	2.29	1.90	58.9	2.69	1.66	1.34	76.3	17.4
l	18	2.86	2.31	1.90	57.3	2.71	1.71	1.38	75.2	17.9
m	20	2.76	2.15	1.74	59.8	2.63	1.60	1.29	76.9	17.1

[a] Each row refers to the values of the variables referring to one of the codes in Figs. 1 and 3, as is indicated in column 1. The rest of the table is self-explanatory. See the text for comments and further information.

studied, we conducted another analysis which does (Di Giulio 1989). In particular, for each of the configurations reported in Fig. 1, we built up the corresponding configuration which maximizes the number of synonymous changes. In other words, the synonymous codons were allocated through a simple manual procedure in such a way as to maximize the number of synonymous changes. Moreover, in choosing these configurations, we kept the termination codons constant in terms of both their number (three) and their arrangement in the actual genetic code.

For each of these configurations (not shown) we calculated the Δ_{low} values which, combined with the maximum number of synonymous changes (Di Giulio 1989; Di Giulio et al. 1994) calculated for a specific code, furnished the Δ_{low} values (Tables 1 and 2; with synonymous changes). The Δ_{code} values (Tables 1 and 2; without synonymous changes) were transformed into the corresponding values by combining the synonymous changes and thus generating new Δ_{code} values (Tables 1 and 2; with synonymous changes). Finally, the Δ_{mean} values (Tables 1 and 2; without synonymous changes) were combined with the number of synonymous changes estimated from a sample of 5 million random codes using the ESE program (see the Appendix) to generate the new Δ_{mean} values (Tables 1 and 2; with synonymous changes). We then went on to calculate the minimization percentages (Tables 1 and 2; with synonymous changes).

Finally, we calculated another type of minimization percentage obtained from the simple difference between the minimization percentage with synonymous changes and the one without synonymous changes (Tables 1 and 2; synonymous changes only).

Discussion

In attempting to predict the behavior of the minimization percentage as the number of amino acids codified in the evolving genetic code increases, we encountered certain difficulties. While one train of thought leads us to predict a linear increase in the minimization percentage as the number of amino acids codified in the evolving genetic code increases, another leads us to conclude that there might be an inverse correlation between these two variables since the minimization percentage in the early and intermediate stages of genetic code evolution nevertheless turn out to be bolstered by a large number of synonymous changes. These data seem to us to show evidence of both these behaviors.

If, as seems to be the case, optimization of the physicochemical properties of amino acids played an important role in structuring the genetic code (for references see Szathmary 1993; Di Giulio 1997a; for more recent articles see Ardell 1998; Freeland and Hurst 1998), we would expect there to be an increase in the optimization level as the number of amino acids codified in the evolving genetic code increases. This is observed both by correlating the minimization percentages (without synonymous changes) referred to the polarity distances with the number of amino acids codified in the corresponding code [Table 1; rows a–m, $r = +0.910$, $n = 11$, $F = 43.5$, df $= (9,1)$, $P < 10^{-4}$] and by making the same correlation but this time referred to the polarity and molecular volume distances [Table 2; rows a–m, $r = +0.951$, $n = 11$, $F = 85.5$, df $= (9,1)$, $P < 10^{-4}$].

These strong correlations allow us to say that, in general, there has been a linear increase in the minimization percentage as the number of amino acids codified in the evolving code increases. Clearly there must have been a constant and strong selective pressure to reduce the deleterious effects of translation errors and/or mutation. This seems to supply evidence in favor of the physicochemical hypothesis of the genetic code (Sonneborn 1965; Woese 1965; Woese et al. 1966; Fitch and Upper 1987). However, these observations are also compatible with the coevolution hypothesis (Wong 1975), i.e., as the precursor amino acids conceded part of their codon domain to the product amino acids, the latter were attributed with codons in such a way that similar amino acids were assigned to similar codons and increasing the minimization percentage in the corresponding code as a consequence of selective pressure to reduce the deleterious effects of genetic message translation errors (Di Giulio 1997a, 1998). Moreover, these correlations might also be the effect of a physicochemical interaction between amino acids and anticodons (Weber and Lacey 1978; Jungck 1978; Lacey and Mullins 1983; Lacey et al. 1992; Di Giulio 1996) taking place on hairpin RNA structures (Di Giulio 1998) which are the ancestors of tRNA (Di Giulio 1992) and house anticodons in their stem (Di Giulio 1998). Therefore, in accordance with the latter interpretation, the apparent selective pressure to reduce the deleterious effects of translation errors might be the result of amino acid–anticodon interactions and might not be easily distinguishable from these interactions. Even from the latter viewpoint, the coevolution hypothesis is still compatible with the physicochemical hypothesis (Di Giulio 1998), although it is the evolutionary development of the code as indicated in Fig. 1 which attributes a secondary role to the latter hypothesis because this development (Fig 1) is based substantially on the coevolution hypothesis. Finally, it seems to us that the evolutionary development of the code (Fig. 1) does not lend itself to the stereochemical hypothesis of genetic code origin (Woese 1967; Balasubramanian et al. 1980; Simizu 1982), at least as regards the one that Yarus (1998) calls the strong or exuberant form of this theory, and thus we feel that the above-reported correlations are not an expression of this theory. The stereochemical hypothesis (Woese 1967; Balasubramanian et al. 1980; Shimizu 1982) suggests that the origin of the genetic code lies in the strong interactions that took place between codons or anticodons and amino acids which somehow (poorly specified in this theory) promoted peptide synthesis early on. Now it is the evolutionary development of the code itself (Fig. 1) which goes against the stereochemical hypothesis. Figure 1a, for instance, shows that Asp is codified by 14 codons: Should we conclude, if the stereochemical hypothesis is true, that Asp is in a stereochemical relationship at this evolutionary stage (Fig. 1a) with 14 codons (or corresponding anticodons)? This seems to be unthinkable, whereas the above-reported correlations might be interpreted through what Yarus (1998) calls the weak stereochemical hypothesis, as the latter does not seem to be dissimilar from the one that is more generally known as part of the physicochemical hypothesis (Weber and Lacey 1978; Jungck 1978; Lacey and Mullins 1983; Lacey et al. 1992), which, as we have already seen, is able to explain these correlations (Di Giulio 1998). Therefore, although the above-reported correlations might be interpreted through a weak interpretation of the stereochemical hypothesis, it is the very scenario of code evolution (Fig. 1) that does not lend itself to a strong stereochemical interpretation, and therefore these correlations are not, in our opinion, an expression of this theory.

In the strong correlations reported above we can observe certain behaviors. If we consider only the first seven evolutionary stages (Figs. 1a–g), we do not find any significance for polarity distances in the correlation between minimization percentages (without synonymous changes) and the number of amino acids codified in the corresponding code [Table 1; rows a–g, $r = +0.626$, $n = 7$, $F = 3.2$, df = (5,1), $P = 0.13$], whereas the same correlation is significant for polarity and molecular volume distances combined [Table 2, rows a–g, $r = +0.893$, $n = 7$, $F = 19.6$, df = (5,1), $P = 0.0069$]. [It is therefore possible that in the early and intermediate stages of code evolution (Figs. 1a–g), other trends were in progress (see below)]. Although this behavior is unexpected considering that the final minimization level of polarity distances only is higher than that of polarity and molecular volume distances combined (Di Giulio *et al.*, 1994) (Tables 1 and 2; without synonymous changes, row m), it becomes comprehensible if we consider that the molecular volume or more generally the “size” of amino acids might have been an important physicochemical variable of amino acids in early and intermediate evolutionary stages (Figs. 1a–g) because the ‘‘size’’ seems to reflect the β-sheets of proteins (Di Giulio 1996).

As regards the correlations which include synonymous changes between minimization percentages and the number of codons codified in the corresponding codes (Tables 1 and 2; with synonymous changes), these do not seem to be significant or are only marginally so. Indeed, for polarity distances we obtain a nonsignificant correlation coefficient between these two variables [Table 1; rows a–m, $r = +0.522$, $n = 11$, $F = 3.4$, df = (9,1), $P = 0.099$], while for polarity and molecular volume distances combined, we obtain only a marginally significant correlation coefficient [Table 2; rows a–m, $r = +0.607$, $n = 11$, $F = 5.3$, df = (9,1), $P = 0.048$]. This behavior seems to be justified, at least as far as polarity distances are concerned, by an inverse correlation between these two variables for the first seven evolutionary stages of the code (Figs. 1a–g), which indeed show a negative and significant correlation coefficient [Table 1; rows a–g, $r = -0.819$, $n = 7$, $F = 10.2$, df = (5,1), $P = 0.024$], whereas the same correlation for polarity and molecular volume distances combined turns out to be only marginally significant [Table 2; rows a–g, $r = -0.732$, $n = 7$, $F = 5.8$, df = (5,1), $P = 0.061$]. [We feel that the latter two correlation coefficients cannot show a complete significance because the corresponding coefficients, which do not include synonymous changes, have the opposite sign (see the two previous coefficients referred to seven pairs of data): i.e., there are forces in play that act in opposite directions.]

Our interpretation of these observations is that the global level of optimization, due to the minimization percentages including synonymous changes, decreased in the early and intermediate stages of code evolution (Tables 1 and 2; with synonymous changes, rows a–g), and this behavior must seemingly be attributed to the high minimization percentage due to the high number of synonymous changes present in the codes of these stages (Figs. 1a–g; Tables 1 and 2; synonymous changes only, rows a–g) which thus enabled this decrease in the minimization percentages. In other words, these data seem to be interpretable through a decrease in the minimization percentages up to 13 amino acids (Tables 1 and 2; with synonymous changes, rows a–g), corresponding, again, to a high buffering of errors due to the presence of a high number of synonymous changes (Figs. 1a–g; Tables 1 and 2; synonymous changes only, rows a–g). After this, as the number of these changes continued to decrease (Tables 1 and 2; synonymous changes only, rows h–m), an increase in the minimization percentage became necessary and did indeed take place (Tables 1 and 2; with and without synonymous changes, rows h–m). Therefore, the high number of synonymous changes present in these codes (Figs. 1a–g) was *per se* able to ensure a buffering of the translation errors because this allowed a decrease in the minimization percentages.

These decreases in the minimization percentages in the early and intermediate stages of code evolution (Tables 1 and 2; with synonymous changes, row a–g) favor neither the physicochemical (Sonneborn 1965; Woese 1965; Woese et al. 1966; Fitch and Upper 1987) nor the stereochemical (Woese 1967; Balasubramanian et al. 1980; Shimizu 1982) hypotheses, because it is in these very stages that these hypotheses attribute a crucial role to distance minimization, whereas the exact opposite seems to take place. In other words, if, as expected by the physicochemical hypothesis, minimization was the main adaptive theme promoting the origin of the genetic code, then in any phase of its evolution the code should display increasing or constant minimization percentages, whereas here we seem to see a decrease, i.e., the opposite of what this hypothesis would expect. Furthermore, the real increase in minimization percentages seems to have taken place only in the final four stages of code evolution (Tables 1 and 2; with and without synonymous changes, rows h–m), and therefore, the physicochemical hypothesis is unable to explain code evolution in the early and intermediate stages. Moreover, these behaviors are what we would expect if the physicochemical properties of amino acids played a role that, while important, was subordinate to that played by the mechanism of codon concession from precursor amino acids to product amino acids, as predicted by the coevolution hypothesis (Wong 1975, 1980; Di Giulio 1997a), to structure the genetic code.

In conclusion, we feel that the correlations reported above are the expression of the coevolution theory of the origin of the genetic code (Wong 1975), because although they also involve the physicochemical properties of amino acids, in the final analysis these correlations depend on the evolutionary stages of the code (Fig. 1), which are the manifestation of the predictions of this theory. At the same time, these correlations seem to shed more light on the interconnections between the physicochemical and the coevolution hypotheses of genetic code origin.

Appendix

The DPERM Algorithm

An algorithm is devised to generate once and only once all the possible n! permutations p of an array with n elements. We can consider the array element $p_1, \ldots, p_n$ as the digits of a number in base $n + 1$, where the leftmost elements are the most significant; then the largest permutation has $p_1 > p_2 \ldots > p_n$. Given a permutation p_i, the algorithm generates the permutation $p_{i+1} < p_i$ and the decreasing permutation sequence $p_1 = (n, n - 1, \ldots, 1), \ldots, p_{n!} = (1, 2, \ldots, n)$ can be generated. The steps involved in this algorithm are as follows.

- Swap $p(n - 1)$, $p(n)$ if $p(n - 1) > p(n)$ else, in the largest position $x < n - 1 : p(x) \neq 1$
- Swap $p(x)$ and $p(m) = \max_{x < y \leq n} p(y) : p(y) < p(x)$.
- Sort $p(y)$, $\forall\ x < y \leq n$.

The ESE Algorithm

Let there be n amino acids $m_1, \ldots, m_n$ and 61 codons codifying these amino acids, an array $c = c_1, \ldots, c_{61}$ can then describe the *cells* of the genetic code. Moreover, we give a matrix B where the elements $b_{j,c} \neq 0$ describe the *communicating cells* with cell c.

We can remap array c as a 16 × 4 matrix or as a 4 × 4 block matrix with block size 4 × 1. With respect to cell c, we define communicating cells as

- the cells in the same 4 × 1 block to which cell c belongs,
- the cells in the same row of cell c, and
- the cells in the 4 × 1 blocks of the same column of cell c and the same block relative position.

Given the number of synonymous codons codifying each amino acid, we can make the set of codons CO; then we randomly allocate the codons over the cells of the genetic code table by calling the Allocell routine described below. After a complete allocation of codons to the cells, the number of synonymous changes G is the count of the times an amino acid occurs in its communicating cells. If we call com(c_i) the set of cells communicating with c_i and $m(c_i)$, the amino acid allocated in cell c_i, G is defined by

$$G = \sum_{i=1}^{61} \# \text{ of occurrences of } m(c_i) \text{ in } \mathrm{com}(c_i)$$

We estimated the mean values of G over 5 million randomly generated codes.

The Allocell Algorithm

An array L describes the set of free cells in which no amino acid has been allocated and is initialized with the values $L(i) = i$. A variable PL can be used to describe the number of actual parts in the set. If a random number $r \in [1,PL]$ is generated, the cell randomly chosen from the allocation of an amino acid is $L(r)$. The random allocation of the codon CO_k to the cell $c_{L(r)}$ is obtained by this type of assignment.

The deletion of the element $L(r)$ consists of shifting leftward the elements $L(i)$ with $r < i < PL$ and decrementing PL.

References

Ardell DH (1998) On error minimization in a sequential origin of the standard genetic code. J Mol Evol 47:1–13

Balasubramanian R, Seetharamulu P, Raghunathan G (1980) A conformational rationale for the origin of the mechanism of nucleic acid-directed protein synthesis of ‘living’ organisms. Origins Life 10:15–30

Di Giulio M (1989) The extension reached by the minimization of polarity distances during the evolution of the genetic code. J Mol Evol 29:288–293

Di Giulio M (1992) On the origin of the transfer RNA molecule. J Theor Biol 159:199–214

Di Giulio M (1996) The β-sheets of proteins, the biosynthetic relationships between amino acids, and the origin of the genetic code. Origins Life Evol Biosph 26:589–609

Di Giulio M (1997a) On the origin of the genetic code. J Theor Biol 187:573–581

Di Giulio M (1997b) The origin of the genetic code. Trends Biochem Sci 22:49

Di Giulio M (1998) Reflections on the origin of the genetic code: A hypothesis. J Theor Biol 191:191–196

Di Giulio M (1999) The coevolution theory of the origin of the genetic code. J Mol Evol 48:253–255

Di Giulio M, Medugno M (1998) The historical factor: The biosynthetic relationships between amino acids and their physicochemical properties in the origin of the genetic code. J Mol Evol 46:615–621

Di Giulio M, Capobianco MR, Medugno M (1994) On the optimization of the physicochemcial distances between amino acids in the evolution of the genetic code. J Theor Biol 186:43–51

Fitch W, Upper K (1987) The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. Cold Spring Harbor Symp Quant Biol 52:759–767

Freeland SJ, Hurst LD (1998) The genetic code is one in a million. J Mol Evol 47:238–248

Grantham R (1974) Amino acid different formula to help explain protein evolution. Science 185:862–864

Jungck JR (1978) The genetic code as a periodic table. J Mol Evol 11:211–224

Lacey JC Jr, Mullins DW Jr (1983) Experimental studies related to the origin of the genetic code and the process of protein synthesis—a review. Origins Life 13:3–42

Lacey JC Jr, Wickramasinghe NSMD, Cook GW (1992) Experimental studies on the origin of the genetic code and the process of protein synthesis: A review updata. Origins Life Evol Biosph 22:243–275

Shimizu M (1982) Molecular basis for the genetic code. J Mol Evol 18:297–303

Sonneborn TM (1965) Degeneracy of the genetic code: Extent, nature, and genetic implications. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. New York, Academic Press, pp 377–397

Szathmary E (1993) Coding coenzyme handles: A hypothesis for the origin of the genetic code. Proc Natl Acad Sci USA 90:9916–9920

Taylor FJR, Coates D (1989) The code within the codons. BioSystems 22:177–187

Weber AL, Lacey JC Jr (1978) Genetic code correlations: amino acids and their anticodon nucleotides. J Mol Evol 11:199–210

Woese CR (1965) On the origin of the genetic code. Proc Natl Acad Sci USA 54:1546–1552

Woese CR (1967) The genetic code. Harper & Row, New York

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

Wong JT (1975) A co-evolution theory of the genetic code. Proc Natl Acad Sci USA 72:1909–1912

Wong JT (1980) Role of minimization of chemical distances between amino acids in the evolution of the genetic code. Proc Natl Acad Sci USA 77:1083–1086

Wong JT (1988) Evolution of the genetic code. Microbiol Sci 5:174–182

Yarus M (1998) Amino acids as RNA ligands: A direct-RNA-template theory for the code’s origin. J Mol Evol 47:109–117