J Mol Evol (1991) 32:178–182

Journal of
Molecular Evolution

Codon Swapping as a Possible Evolutionary Mechanism

Eörs Szathmáry

Ecological Research Group, Department of Plant Taxonomy and Ecology, Eötvös University, Kun Béla tér 2, H-1083 Budapest, Hungary

Summary. It is apparent in the genetic code that amino acids of similar chemical nature have similar codons. I show how through successive codon captures (multiple rounds of Osawa–Jukes type reassignments), complete codon swappings in an unfavorable genetic code are evolutionarily feasible. This mechanism could have complemented the ambiguity reduction and the vocabulary extension processes of codon-amino acid assignments. Evolution of wobble rules is implied. Transfer RNA molecules and synthetases may still carry memories of it.

Key words: Genetic code — Codon capture — Codon swapping — Codon shuffling — Ambiguity reduction — Molecular evolution — Origin of life

Introduction

Wong (1980) in his pioneering study on the statistical properties of the universal code showed that minimization of chemical distances between neighboring amino acids (cf. Sonneborn 1965) as measured by the Grantham (1974) index (a combination of composition, polarity, and molecular volume values) could have played only a subsidiary role in evolution. A methodically more careful analysis, based on amino acid polarities, demonstrates, however, that minimization of polarity distances is likely to have reached an evolutionary optimum point where the benefit from further rearrangements would hardly outweigh the costs of the process (Di Giulio 1989).

It was once thought that, contrary to the implicit assumption of Sonneborn (1965), rearrangement of any definite codon configuration to another was impossible because of several deleterious amino acid substitutions occurring simultaneously (Woese 1967; Crick 1968). Data (e.g., Yamao et al. 1985) and speculations have led to the proposal of the Osawa–Jukes mechanism of codon capture (reassignment) without invoking alteration in protein sequences (Osawa and Jukes 1988, 1989). It consists of the following steps: (1) removal of at least one codon from the repertoire of a particular amino acid through directional mutation pressure; (2) confinement of codon–anticodon recognition to the reduced set; (3) reappearance of the missing triplets by reversed mutation pressure; and (4) their recognition by mutant tRNAs with an alternative amino acid assignment. Whether mutual and complete exchanges of codons between amino acids are possible has, however, not been asked, although stop-codon capture has been assigned a hypothetically important role in the evolutionary origin of the genetic code (Lehman and Jukes 1988).

Thus it is interesting to ask if neighboring amino acids could capture each other's codons to the point of complete swapping of assignments. The scenarios I present below show that this is indeed possible with certain restrictions. Codon swappings could have played a subsidiary role in the earliest evolution of the genetic code, although evidence for this may be difficult to obtain.

Codon Swapping Scenarios

I borrow the following component processes of the Osawa–Jukes (1988, 1989) mechanism:

1) Multiple rounds of reversing AT and GC substitution pressures to modify codon usage.
2) The role of nucleotide mutation pressure in altering anticodons.

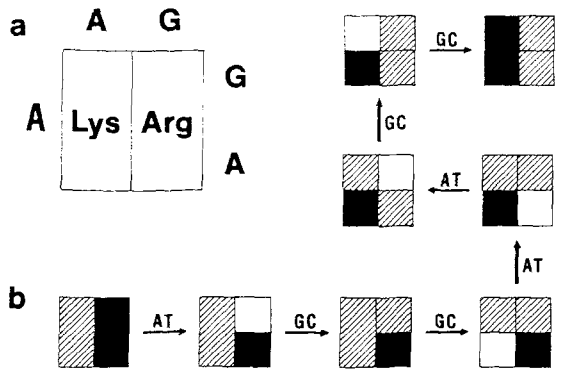

Fig. 1. Complete swapping of codons between two amino acids, lysine (Lys) and arginine (Arg). **a** The initial codon configuration. **b** The process of swapping. Arrows indicate substitution pressure. Striped boxes are occupied by Lys, black ones by Arg, and empty ones are unassigned. Codon AGG is shown to disappear under AT pressure because it is assumed that it is less frequent than codon AAG. In reality, codon disappearance depends on amino acid and codon usage and drift.

3) The appearance of tRNAs with restricted pairing capabilities complementary to the reduced codon repertoire. A few coincident changes independent from substitution pressure are allowed.

I shall use the contemporary wobble rules summarized in Table 1 of Osawa and Jukes (1988), allowing for specific recognition of G, A, U, and of the clusters (GA), (UC), (GAU), (AUC), and (GAUC). The minimal code of Osawa and Jukes (1988, Table 3) will be used as a point of departure, which is very similar to the present-day mammalian mitochondrial code. The examples given in this section are purely didactic (see Discussion for the true evolutionary context).

Figure 1 demonstrates how two neighboring amino acids having two-codon sets with purines in the wobble position of their codons can be swapped in the code through evolution. Amino acids having two-codon sets with pyrimidines in the same position cannot be swapped, however, because we do not know of a specific recognition of C in this case. If only one amino acid of a pair has a pyrimidine in the third position of its codon, complete swapping is still excluded, although it can go almost to completion. Figure 2 shows the swapping of amino acids with neighboring family codon boxes.

Codon Swapping and Component Processes

In discussing one by one the component processes necessary for codon swapping to occur, I first refer to contemporary situations, and then to likely earliest ones.

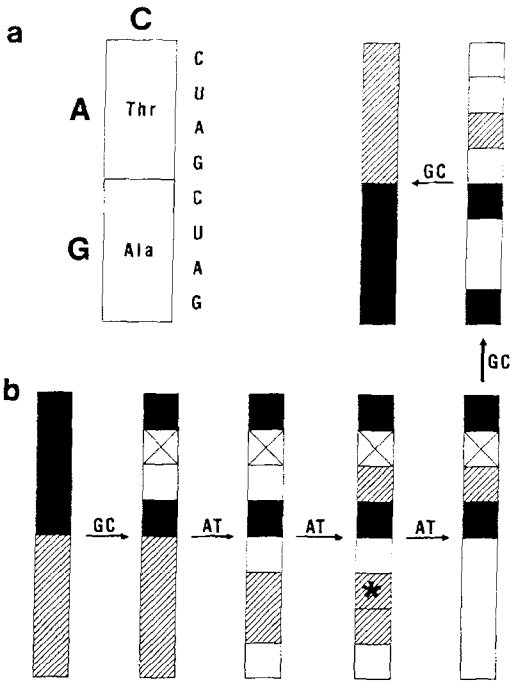

Fig. 2. Codon swapping between the family boxes of threonine (Thr) and alanine (Ala). **a** Initial codon configuration. **b** The process of swapping. Black and striped boxes are occupied by Thr and Ala, respectively. In the fourth step codon GCU (marked with an asterisk) disappears under AT substitution pressure. The following conditions make this process possible: (1) AT pressure reduces the rate of transition to GCC relative to the transversion to GCA; (2) it is more favorable to have GCA than GCU in the fourth stage of the process for Ala, as the former can freely mutate to the captured ACA, whereas ACU cannot be captured (its box is crossed out) because of the ambiguous recognition of ACC. Thus it is favorable to mutate from GCU to GCA, but GCA itself is driven to ACA. Possibility to specifically recognize C would make the process a lot easier.

Wobble Rules

Given the contemporary wobble rules, neighbor swapping for about half of the amino acids in the minimal code (Osawa and Jukes 1988) seems possible. It could be applicable to all amino acids if C were specifically recognizable. Lysyl-C as a specific wobble pair for A was found in *Escherichia coli* in 1988 (Muramatsu et al. 1988a). It is difficult to see why this possibility could not be realized. [The fact that the synthesis of lysyl-C immediately alters the amino acid specificity of the tRNA in the particular case (Muramatsu et al. 1988b) does not imply that this is obligatory. What is needed here is the possibility of specific pairing with A.] Based on plausibility grounds, an unusual wobble pairing (G with A, U, and C) has been invoked in explaining the reassignment of codon AUA from methionine to isoleucine in echinoderms (Osawa et al. 1989).

Strict Watson–Crick type codon–anticodon pairing is not excluded during the emergence of the code. Wobble rules themselves must have undergone evolution. Contemporary wobble rules are thought to

maximize the efficiency of translation (Bulmer 1988). Sixty-one different tRNAs for the 61 codons could function more accurately but less efficiently unless their individual concentrations were equal to the tRNA utilizing wobble rules. This material cost would outweigh the benefits gained from slightly increased accuracy, given the fact that the latter is sufficiently high already. But during the earliest evolutionary period gain in accuracy must have been a very precious thing [the inverse problem of the error catastrophe (Hoffmann 1974; Bedian 1982)].

An example shows why conventional base pairing enhances accuracy. Under the minimal code mentioned before, the UGG anticodon for proline pairs with codons CCA, CCU, CCG, and CCC; i.e., the first anticodon base U pairs by complete wobble. Therefore, $tRNA^{Pro}$ could mispair with all leucine codons CUA, CUU, CUG, and CUC by a single GU mismatch in the middle position. (Symmetrically, a similar problem applies to $tRNA^{Leu}$.) Under Watson–Crick base pairing, however, a $tRNA^{Pro}$ with the anticodon UGG would pair only with the proline codon CCA, and mispairing with any of the leucine codons except for CUA would amount to a double mismatch, which is much more unstable. Similar problems apply to mispairings at the first codon position.

There is evidence that the frequency of mRNA misreading decreases with codon–anticodon mismatch (Bulmer 1988), and the frequency of realized (translationally manifest) mispairing at different sites decreases in the following order: third, first, and second codon base mismatch (Parker 1989). The role of competing abundant noncognate tRNAs in anticodon as well as codon usage has been discussed before (Ninio 1971; Kato 1990), but only for existing codes with wobble.

I suspect that the earliest codes utilized traditional Watson–Crick base pairing with as little wobble as possible. The evolution of the translation apparatus allowed the introduction of wobble later, when overall accuracy was high enough.

Directional Nucleotide Substitution Pressure

My usage of the term substitution pressure is intentional. The former allows for various causes, including mutation pressure arising from the DNA polymerase system (cf. Osawa and Jukes 1988), changes in external temperature (heat selecting for GC; Bernardi and Bernardi 1986), or changes in nucleotide pools (Wolfe et al., cited by Bernardi et al. 1988).

Selection for GC by high temperature deserves some discussion. There are two targets of selection: nucleic acids themselves and the encoded proteins. Increase of GC content in intergenic noncoding segments, in introns, and in third codon positions contribute to the thermal stability of DNA, RNA, and possibly DNA–protein interactions (cf. Wada and Suyama 1986; Bernardi et al. 1988). GC increase in vertebrate coding sequences leads to the increase in frequency of thermally stabilizing amino acids (such as arginine and alanine) and to the decrease in destabilizing ones (like lysine and serine) in proteins (Bernardi and Bernardi 1986). Similar results have come to light on various thermophilic organisms (e.g., Kagawa et al. 1984; Nishiyama et al. 1986; Barstow et al. 1987). Yet there is no unequivocal evidence for a direct link between high GC content and thermophilic organisms. The nonthermophilic *Micrococcus luteus* has 74% GC (Osawa and Jukes 1989). Nevertheless, I suggest that during early evolution selection for nucleic acid stability must have been a primary force of nucleotide substitution pressure. Thus a shift from low to high temperature could have resulted in GC pressure, whereas a reverse shift could have favored AT.

Lower overall fidelity of replication in early systems (Eigen 1971) could have made nucleotide substitution through mutation pressure faster. Early harsh environments could have suffered from wild temperature fluctuations and shifts. By raising the mutation rate, this not only would have meant that optimization of chemical distances of amino acids would have been highly desirable, but would have provided also the means to achieve these alterations, through fueling the codon usage mutations (among other ones) necessary for the rearrangements themselves.

Selection for Better Intermediate Codes

The assumption that the new code is closer to optimum implies that it is selectively advantageous relative to the previous one (Sonneborn 1965). Selection can well aid the process itself. Assume, for example, that in Fig. 1 each original assignment has a score of -1 and each final one is scored $+1$. Then, the given sequence has total scores of -4, -3, -2, -1, $+1$, $+1$, $+3$, $+4$. Thus, all but one step is selectively favored, and one is neutral.

This didactic example must be replaced by explicit calculation of the coevolution of codons and anticodons (cf. Bulmer 1987) in the entire code, but this can be realized by simulations only.

Discussion

Three possible evolutionary mechanisms could have contributed to the early evolution of the genetic code: ambiguity reduction (Woese 1965, 1967; Fitch

1966), for which there is some evidence (Fitch and Upper 1987); pretranslational modification of prebiotically present amino acids leading to new amino acid assignments according to kinship in biosynthesis (Wong 1981); and combinations of codon captures (Lehman and Jukes 1988; Osawa and Jukes 1988, 1989) and swappings (codon shuffling) leading to more optimal codes (cf. Sonneborn 1965). These mechanisms are not entirely mutally exclusive. Reassignment is easier if ambiguity is allowed (Crick 1965). Amino acid codon capture implied by Wong's (1981) mechanism is feasible through transient appearance of stop codons, ready to be captured by entirely novel amino acids in agreement with the mechanism of Lehman and Jukes (1988).

The context in which the emergence of the genetic code took place is likely to have been an RNA world (Orgel 1989). It seems plausible that among the earliest encoded amino acids one finds those that form readily in abiotic experiments (Wong 1981). The codons of these amino acids are likely to have undergone ambiguity reduction, or "the coming into focus of a fuzzy image" (Orgel 1989, p. 471). Whether these amino acids occupied almost the whole codon table or considerably less is not known. Later amino acids arising from inventive biosynthesis (Wong 1981) could have entered the code through the capture of preexisting or transient stop codons (cf. Lehman and Jukes 1988).

Codon swapping could have complemented both phases. Differences might have existed between chemical similarity as expressed in proteins and that felt by assignment catalysts (ancestral to present-day aminoacyl-tRNA synthetases). There is a quantitative difference between chemical similarity of amino acids in proteins (Grantham 1974) and the same in biosynthesis [precursor–product relationships (Wong 1980)]. These differences could have led to a suboptimal initial assignment, and subsequent swapping, of some amino acids in either phase.

Codon swapping requires restricted wobble and excludes recognition of tRNAs by their anticodons by the synthetases. Thus, archetypal and early universal codes of Jukes (1983) with extensive wobble represent relatively evolved states of the genetic code in this respect, as well as in others (cf. Lehman and Jukes 1988). In prearchetypal codes (there must have been a number of them) there were presumably fewer amino acids, but as many tRNAs with matching anticodons as there were sense codons. The fact that the anticodons of tRNAs of only methionine, valine, phenylalanine, glutamine, tyrosine (Yarus 1988), isoleucine (RajBhandary 1988), and arginine (Schulmán and Pelka 1989) are important in recognition by the corresponding synthetases leaves, in principle, freedom to experiment with codon shuffling even today.

Whether there could have been sufficient time for a few codon swappings during the establishment of the genetic code can be considered as follows. In Fig. 1 of Osawa and Jukes (1988), there are rounds of GC, then AT, then GC, then AT pressure along the historical path from the hypothetical minimal code to present-day mitochondrial and chloroplast codes. If we take approximately 3 billion years for this process, then the substantiated reversal of mutation pressure is a rare event. However, it is likely that the habitat of the earliest life forms was harsh and widely fluctuating in temperature as well as in other conditions. If the frequency of, let us say, long-lasting temperature fluctuations was 100–1000 times higher than during the last 3 billion years, then within 0.3 billion years several rounds of AT/GC pressure reversals could have taken place.

These periods could have been utilized only if ancient organisms were able to track these pressures by relatively rapid fixation of appropriate mutations. This is likely to have been possible; as Eigen et al. (1989, p. 678) write: "In the very early phases of evolution, however, error rates and the acceptability of mutations must have been larger than in present organisms, where structures are optimized and error rates are minimized."

As the genetic code is presumably not older than 3.8 ($\pm$0.6) billion years (Eigen et al. 1989), there could have been a time of $\approx$0.3 billion years for the establishment of the code, during which a few codon swappings could have taken place.

Evidence for ancient codon shuffling is obviously difficult to obtain. Data on the discrimination among amino acids by synthetases (Bulmer 1988) and distances among the hypervariable positions in tRNAs of the same species, thought to reflect ancient amino acid-tRNA assignments (Eigen et al. 1989), could give some information on the relative importance of mechanisms thought to have been involved in shaping the genetic code.

Acknowledgments. I thank István Scheuring and Mária Ujhelyi for discussion, and Niles Lehman and an anonymous referee for useful comments. This work was supported by the Hungarian National Research Fund (OTKA).

References

Barstow DA, Murphy JP, Sharman AF, Clarke AR, Holbrook JJ, Atkinson T (1987) Amino acid sequence of the L-lactate dehydrogenase of *Bacillus caldotenax* deduced from the nucleotide sequence of the cloned gene. Eur J Biochem 165:581–586

Bedian V (1982) The possible role of assignment catalysts in the origin of the genetic code. Orig Life 12:181–204

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1–11

Bernardi G, Mouchiroud D, Gautier C, Bernardi G (1988)

Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7–18
Bulmer M (1987) Coevolution of codon usage and transfer RNA abundance. Nature 325:728–730
Bulmer M (1988) Evolutionary aspects of protein synthesis. In: Harvey PH, Partridge L (eds) Oxford surveys in evolutionary biology, vol 5. Oxford University Press, Oxford, pp 1–40
Crick FHC (1965) Codon–anticodon pairing: the wobble hypothesis. Information Exchange Group No 7, Nucleic Acids and the Genetic Code. Scientific Memo 14, June 14, 1965
Crick FHC (1968) The origin of the genetic code. J Mol Biol 38:367–379
Di Giulio M (1989) The extension reached by the minimization of the polarity distances during the evolution of the genetic code. J Mol Evol 29:288–293
Eigen M (1971) Self-organization of matter and the evolution of biological macromolecules. Naturwissenschaften 58:465–523
Eigen M, Lindemann BF, Tietze M, Winkler-Oswatitsch R, Dress A, von Haeseler A (1989) How old is the genetic code? Statistical geometry of tRNA provides an answer. Science 244:673–679
Fitch WM (1966) Evidence suggesting a partial, internal duplication in the ancestral gene for heme-containing globins. J Mol Biol 16:1–12
Fitch WM, 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
Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862–864
Hoffmann GW (1974) On the origin of the genetic code and the stability of the translation apparatus. J Mol Biol 86:349–362
Jukes TH (1983) Evolution of the amino acid code: inferences from mitochondrial codes. J Mol Evol 19:219–225
Kagawa Y, Nojima H, Nukiwa N, Ishizuka M, Nakajima T, Yasuhara T, Tanaka T, Oshima T (1984) High guanine plus cytosine content in the third letter of codons of an extreme thermophile. J Biol Chem 259:2956–2960
Kato M (1990) Codon discrimination due to the presence of abundant non-cognate competitive tRNA. J Theor Biol 142: 35–39
Lehman N, Jukes TH (1988) Genetic code development by stop codon takeover. J Theor Biol 135:203–214
Muramatsu T, Yokoyama S, Hirose N, Matsuda A, Ueda T, Yamaizumi Z, Kuchino Y, Nishimura A, Miyazawa T (1988a) A novel lysine-substituted nucleoside in the first position of the anticodon of minor isoleucine tRNA from *Escherichia coli*. J Biol Chem 263:9261–9267
Muramatsu T, Nishikawa K, Nemoto F, Kuchino Y, Nishimura A, Miyazawa T, Yokoyama S (1988b) Codon and amino acid specificities of a transfer RNA are both converted by a single post-transcriptional modification. Nature 336:179–181
Ninio J (1971) Codon–anticodon recognition: the missing triplet hypothesis. J Mol Biol 6:63–82
Nishimaya M, Matsubara N, Yamamoto K, Iijama S, Uozimi T, Beppu T (1986) Nucleotide sequence of the malate dehydrogenase gene of *Thermus flavus* and its mutation directing an increase in enzyme activity. J Biol Chem 261:14178–14183
Orgel LE (1989) The origin of polynucleotide-directed protein synthesis. J Mol Evol 29:465–474
Osawa S, Jukes TH (1988) Evolution of the genetic code as affected by anticodon content. Trends Genet 4:191–198
Osawa S, Jukes TH (1989) Codon reassignment (codon capture) in evolution. J Mol Evol 28:271–278
Osawa S, Ohama T, Jukes TH, Watanabe K, Yokoyama S (1989) Evolution of the mitochondrial genetic code II. Reassignment of codon AUA from isoleucine to methionine. J Mol Evol 29:373–380
Parker J (1989) Errors and alternatives in reading the universal genetic code. Microbiol Rev 53:273–298
RajBhandary UL (1988) Modified bases and aminoacylation. Nature 336:112–113
Schulman LH, Pelka H (1989) The anticodon contains a major element of the identity of arginine transfer RNAs. Science 246:1595–1597
Sonneborn TM (1965) Degeneracy of the genetic code: extent, nature, and genetic implications. In: Bryson V, Vogel JH (eds) Evolving genes and proteins. Academic Press, New York, pp 377–397
Wada A, Suyama S (1986) Local stability of DNA and RNA secondary structure and its relation to biological function. Prog Biophys Mol Biol 47:113–157
Woese CR (1965) On the evolution of the genetic code. Proc Natl Acad Sci USA 82:1160–1164
Woese CR (1967) The genetic code. Evanston, New York
Wong JTF (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 JTF (1981) Coevolution of the genetic code and amino acid biosynthesis. Trends Biochem Sci 6:33–36
Yamao F, Muto A, Kawauchi Y, Iwami M, Iwagami S, Azumi Y, Osawa S (1985) UGA is read as tryptophan in *Mycoplasma capricolum*. Proc Natl Acad Sci USA 82:2306–2309
Yarus M (1988) tRNA identity: a hair of the dogma that bit us. Cell 55:739–741

Received March 22, 1990/Revised August 7, 1990