Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 4378–4382, June 1988
Evolution

Codon preference is but an illusion created by the construction principle of coding sequences

(periodic-to-chaotic transition/recurring base oligomers)

SUSUMU OHNO

Beckman Research Institute of the City of Hope, Duarte, CA 91010

Contributed by Susumu Ohno, February 22, 1988

ABSTRACT **Modern coding sequences are in the periodic-to-chaotic transition. In the case of two related sequences for lens αA-crystallin and small heat shock protein, the original repeating units were heptameric in length. Accordingly, base trimers that were parts of heptameric units recurred far more frequently than those that were not included. In the crystallin coding sequence, CTG trimer recurred 21 times, and TCT and TCC trimers recurred 17 times each. By contrast, CTA and TCG, although related to the above, recurred only 4 and 3 times, respectively. It is a small wonder that 10 of the 16 leucine residues were encoded by CTG, whereas none was encoded by CTA, and that 17 of the 23 serine residues were encoded either by TCT or by TCC, whereas only 1 was encoded by TCG. In the small heat shock protein coding sequence, however, AGC became parts of the two prominent heptameric recurring units. Not surprisingly, 10 of the 22 serine residues were now encoded by AGC. In conclusion, the so-called codon preference is a mere reflection of the construction principle of coding sequences and has very little to do with selection *per se*.**

In translating the informational content of base sequences to the amino acid sequence of proteins, the choice of base triplets as codons inevitably created the redundancy that is most pronounced in the case of leucine, serine, and arginine, each being encodable by six codons. Yet, these synonymous codons are seldom utilized to the same extent. Of the six codons for leucine, the correlation between the most frequently utilized codon and the greatest abundance of tRNA species designed to recognize that codon has been most convincingly demonstrated by Grantham's group (1, 2). In *Escherichia coli*, the preferred leucine codon was CTG (1), whereas it was TTG in bakers' yeast (2). Since such biased usage of synonymous codons has been observed in all organisms studied, the idea of codon preference has gained acceptance. Inasmuch as the codon preference implies selection among randomly sustained synonymous base substitutions, Kimura (3), in finalizing his neutral theory of evolutionary compatible mutations, devoted a great deal of thought to this problem. Yet, there is an inherent weakness in this widely accepted notion of codon preference. The organism can easily adjust relative abundance of synonymous tRNA species by increasing or decreasing the number of gene copies, for example. This immediately suggests an alternative and perhaps more plausible explanation. Extremely biased appearances of synonymous codons might have become *fait accompli* early in evolution of each coding sequence. The presently existing order of abundance among synonymous tRNA species then represents merely the subsequently adjusted compromise to these *fait accompli*.

Keeping the above noted alternative in mind, I shall now examine two related coding sequences rich in leucine, serine, and arginine codons, one for αA-crystallin of vertebrate eyes and the other for more ubiquitous small heat shock protein.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Preponderances of Leucine Codon CTG and Serine Codons TCT and TCC in the Hamster αA-Crystallin Coding Sequence Are but Two Sides of the Same Coin

Lens αA-crystallins of vertebrate eyes are 173 residues long, mainly β-sheet-forming proteins that have been conserved in evolution rather stringently. In this paper, I shall deal with hamster αA-crystallin and its coding sequence (4). Included in its 173 residues are 16 leucine, 23 serine, and 13 arginine. As shown in Fig. 1, the so-called codon preference was very evident with regard to leucine as well as serine, whereas not so pronounced among arginine codons. CTG encoded 10 of the 16 leucine residues, whereas TCT and TCC codons accounted for 17 of the 23 serine residues. At this point, the hamster αA-crystallin coding sequence appeared to have represented a classical case for codon preference. Yet, when the recurrence rate of each relevant base trimer was computed, this notion broke down completely. Of the 12 kinds of base trimers that can serve as leucine and serine codons, 3 recurred most frequently: CTG 21 times and TCT and TCC 17 times each. Only half of these recurring base trimers were utilized as leucine and serine codons, whereas the remainder constituted parts of two neighboring codons. In sharp contrast, this 519-base-long coding sequence contained only one TTA base trimer. It is a small wonder that TTA was never utilized as leucine codon. As shown in the middle of Fig. 1, repeating units were longer than trimeric in length. Of the 21 CTG trimers, 8 recurred as TCTG, 7 recurred as CCTG, and 4 recurred as GCTG tetramers. Interestingly, 5 of the 7 CCTGs were translated in the same reading frame to yield leucine, whereas 7 of the 8 TCTG were translated in another reading frame to encode serine, as shown in the left and the center of Fig. 1, middle. The TCT portion of these TCTG tetramers accounted for 7/9th of serine codon TCTs. Observing the top of Fig. 1, one anomaly can be noted. In spite of the fact that CTT trimers recurred 13 times, this trimer was utilized only once as leucine codon. Of 13 CTTs, 8 recurred as CTTC tetramer, and 6 of these were translated in the same reading frame to yield phenylalanine and none yielded leucine, as shown at the right of Fig. 1, middle. Indeed, of 14 phenylalanine residues included in hamster αA-crystallin, only 4 were encoded by TTT, and the rest were encoded by TTC. These three recurring base tetramers shown in the middle of Fig. 1 were actually parts of the recurring base heptamers. Three of these recurring heptamers and their derivatives are shown in three vertical columns at the bottom of Fig. 1. At the top of the left column are two identical copies of the CCTGTCT heptamer encoding a pair of Leu-Ser dipeptides. Immediately 5′ to the second copy is its single base deleted copy CCTG–CT. The new heptamer CCTGCTC is now translated in a different reading frame to encode the 130th to 132nd Ser-Cys-Ser. This derived heptamer

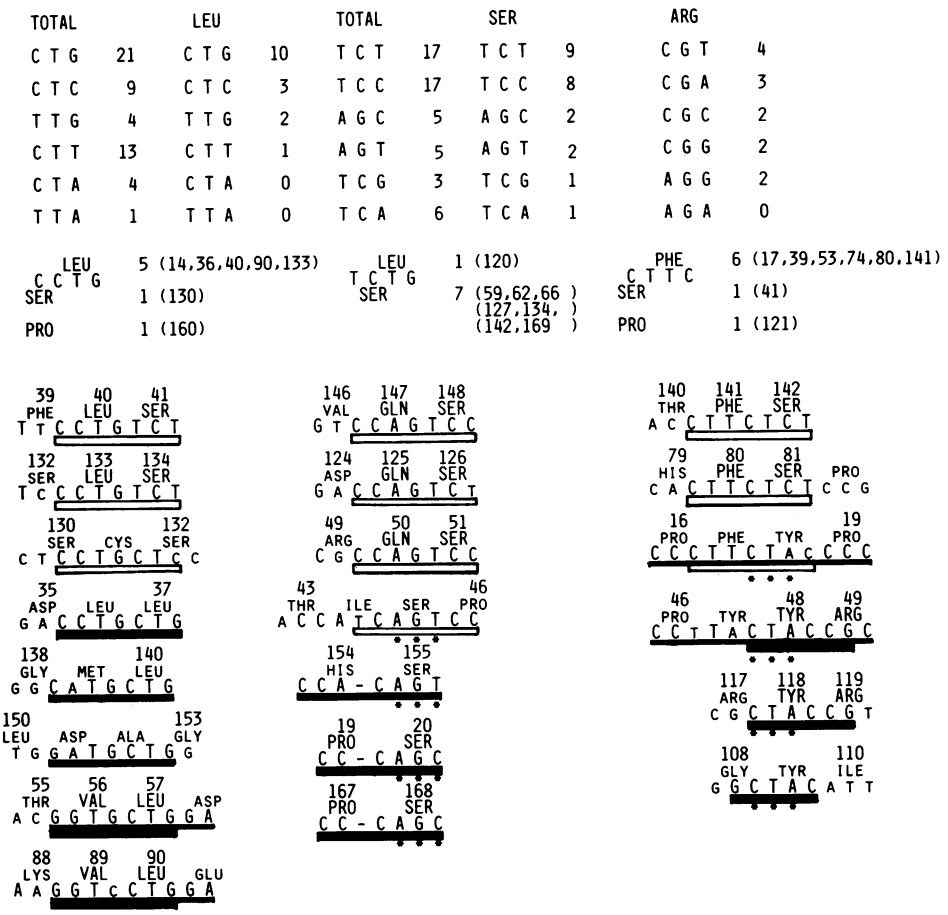

FIG. 1. Codon preferences in hamster αA-crystallin. At the top, six codons each for leucine, serine, and arginine are aligned vertically in the order of frequent usage within the 173-codon-long hamster lens αA-crystallin coding sequence (4). Numbers at the right of each column show incidences as codons. In the case of leucine and serine codons, total incidences of recurrence as base triplets are also shown at the left of each column. In the middle, reading frame choices by three recurring base tetramers are shown. For example, CCTG recurred seven times, five times encoding leucine and once each encoding serine and proline. Numbers within parentheses indicate positions of that residue in the amino acid sequence of hamster αA-crystallin. At the bottom, the three recurring base heptamers and their respective derivatives are aligned in three columns. Recurring heptamers themselves are underlined by thick open bars and each base is shown in large uppercase letters. A single and two base-substituted copies of each are underlined by progressively thinner open bars, and substituted bases are shown in small uppercase letters. Derived heptamers, hexamers, and pentamers that became secondary repeating units are underlined by solid bars. In the center column, AGT and AGC utilized as serine codons are indicated by asterisks and so are the CTA base trimers in the right column. Amino acids of recurring dipeptides are shown in large uppercase letters.

CCTGCTC, in turn, had its own copies. CCTGCTG and its single-base-substituted copy CATGCTG encoded Leu-Leu (positions 36–37) and Met-Leu (positions 139–140), whereas a single base substitution of CATGCTG translated in a different reading frame encoded ASP-ALA (positions 151–152). This heptamer by a further single base substitution also encoded a pair of Val-Leu dipeptides. It should be noted that derivatives of one recurring heptamer shown at the bottom left column of Fig. 1 accounted for 5 of the 7 recurring CCTG as well as all 4 recurring GCTG. Shown in the bottom center column of Fig. 1 is another recurring heptamer, CCAGTCC. Its 2 invariant and 2 single-base-substituted copies encoded a triplet of Gln-Ser dipeptides in one reading frame and the 44th to 46th Ile-Ser-Pro in another. In the vicinity of the last copy for Ile-Ser-Pro emerged a new repeating unit, CCATCAGT. Its single-base-deleted copy CCACAGT encoded His-Ser (positions 154–155). A pair of Pro-Ser dipeptides was encoded by its single-base-deleted derivative CCCAGC. In this manner, derivatives of another recurring heptamer, CCAGTCC, shown in the bottom center column accounted for all 5 AGT base trimers, 2 of which served as serine codons as well as both AGC serine codons.

Part of the reason for the relative abundance of phenylalanine (14 residues) and the relative paucity of tyrosine (6 residues) in hamster αA-crystallin (4) is found in the bottom right column of Fig. 1. Two invariant copies of the third recurring heptamer CTTCTCT encoded a pair of Phe-Ser dipeptides, whereas its two-base-substituted copy yielded Phe-Tyr (positions 17–18). From the vicinity of this two-base-substituted copy emerged the new hexameric repeating unit CTACCG that encoded a pair of Tyr-Arg dipeptides. Contained within the bottom right column are all 4 CTA base trimers, none of which is utilized as a leucine codon. It should also be noted that TAC base trimers contained within this column encoded 5 of the 6 tyrosine residues. The conclusion can thus be drawn that the abundance of CTG, TCT, TCC, as well as TTC base triplets within the hamster αA-crystallin coding sequence is attributable to the fact that these were parts of heptameric repeating units that spawned many copies. The preponderance of CTG among leucine codons, of TCT and TCC among serine codons, as well as of TTC among phenylalanine codons is but a reflection of the above. Conversely, the paucity of AGC and CTA base trimers can readily be attributed to the fact that they were parts of secondary repeating units derived from degenerate copies of the recurring heptamers. These three recurring heptamers shown at the bottom of Fig. 1 partially overlap with each other, thus suggesting their common ancestry.

The Construction Principle of the Hamster αA-Crystallin Coding Sequence

I have proposed that the first set of coding sequences to emerge in the prebiotic world was composed of repeats of base oligomers, for they alone possessed the inherent property of self elongation, and that numbers of bases in oligomeric recurring units were not multiples of 3 (ref. 5). What was the fate of these short exact repeating units? In 1972 Southern deduced that these repeating units were fated to become progressively longer and less exact (6). The above statement by Southern antedated by 5 years the well-known Feigenbaum conjecture on the general property of the periodic-to-chaotic transition applicable to numerous and diverse physical systems (7). It now appears however, that decay of the original periodicity is not always due to doubling and tripling but also by the golden mean (8).

In the case of the hamster αA-crystallin coding sequence (4), the primordial repeating unit appeared to have been heptameric in length, already seen in Fig. 1. Thus, its construction principle was very similar to the porcine muscarinic acetylcholine receptor coding sequence of the rhodopsin family previously analyzed in great detail (9). Accordingly, the description of the construction principle here would be selective rather than exhaustive. In addition to the three identified in the bottom of Fig 1, three more heptameric repeating units are shown in Fig. 2. Overall, the base tetramer CTTC was included in two different recurring heptamers, whereas TCT trimer was part of the three and CTG trimer was part of the two. When the ultimate ancestor of this coding sequence was heptameric repeats eons ago, three consecutive copies of the original primordial heptamer should have given the heptapeptidic periodicity to the ancestral polypeptide. Provided that the first recurring heptamer CCTGTCT represented in the bottom left column of Fig. 1 was the primordial heptamer, the heptad encoded by its three copies should have been Pro-Val-Ser-Cys-Leu-Leu-Ser. It should be noted that the codon for the first serine was TCC, whereas that for the second serine was TCT. The T/CCTGTCT/-CCTGTCT portion of the three consecutive copies survives to this day in the hamster αA-crystallin coding sequence (4). A three-base substituted version of the above, T/CCTGCTC/-CCTGTCT, still encoded the 130th to 134th Ser-Cys-Ser-Leu-Ser as seen in the second and third rows of the bottom left column of Fig. 1. More often, however, tandemly recurring primordial heptamers were seen as parts of the next class of longer repeating units that were related to heptamers by the golden mean (8). Inasmuch as $7 \times (1 + \sqrt{5})/2 = 11.326$, this next class of repeating units, invariably recurring in tandem and still translated in different reading frames, was either 11 or 12 bases long. By observing the third and fourth rows of the bottom center column of Fig. 1, it should be noted that the 44th to 46th Ile-Ser-Pro was encoded by TCAGTCC, which was a single-base-substituted copy of the CCAGTCC unit that in a different reading frame encoded Gln-Ser (positions 50–51). Actual repeating units, however, were 11 bases long, ACCATCAGTCC encoding the 43rd to 46th Thr-Ile-Ser-Pro differing by two base substitutions from ACCGCCAGTCC encoding the 48th to 51st Tyr-Arg-Glen-Ser. Thus, these two copies were separated from each other only by 5 bases.

The very similar situation is seen at the top of Fig. 2 with regard to another recurring heptamer, GTCTGCC. This heptamer was a part of the 12-base-long repeating unit that recurred in tandem, two copies differing only by two base substitutions. They were again translated in different reading frames, and only 10 bases intervened between these two copies. The next class of longer repeating units was either 14

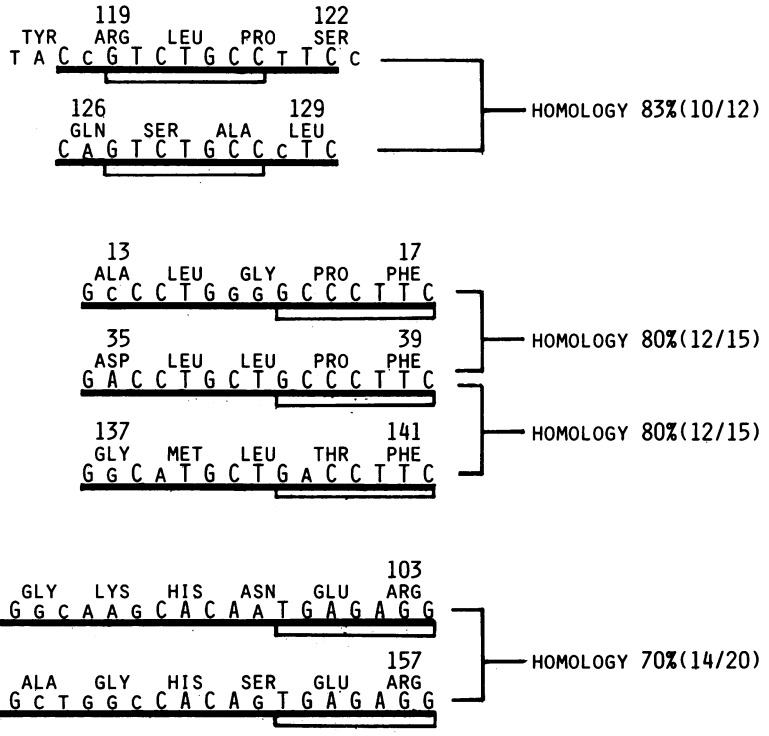

FIG. 2. Representative examples of three classes of longer repeating units contained within the coding sequence for hamster αA-crystallin (4). Each longer repeating unit is underlined by solid bars, and, in the case of pairs, identical bases are shown in large uppercase letters, whereas mismatched bases are shown in small uppercase letters. In the case of one 15-base-long triplet shown in the middle, all of the bases of the middle copy are shown in large uppercase letters, whereas mismatched bases in the other two copies are shown in small uppercase letters. Each pair in these instances shares the identical copy of the three different recurring heptamers underlined by thick open bars. Only in the case of one triplet shown in the middle, a heptamer, in the third copy is a single-base-substituted copy of the one-unit heptamer GCCCTTC. The degree of base sequence identity between members of each pair is shown at the right.

or 15 bases long; thus, this class was of a double periodicity. One triplet of the 15-base-repeating unit is shown in the middle of Fig. 2. Two copies share the identical heptamer GCCCTTC, whereas the heptamer contained in the third copy is a single-base-substituted copy. Copies of this class of repeating units were almost invariably translated in the same reading frame; thus, their presence was felt in the amino acid sequence as oligopeptidic repeats. The longest class of repeating units that still maintained ≥70% base sequence identity between the copies was 18–21 bases long. The periodicity decay following the golden mean should have yielded 18- to 19-base-long units—i.e., $11 \times (1 + \sqrt{5})/2 = 17.79$ and $12 \times (1 + \sqrt{5})/2 = 19.42$. However, the decay by tripling of the original periodicity should also have yielded 21-base-long-units. At the bottom of Fig. 2, one such 20-base-long pair that shared yet another recurring heptamer, TGAGAGG, is shown. This heptamer not only encoded a pair of Glu-Arg dipeptides as shown in Fig. 2 but also its 2-base-substituted copy was responsible for Asp-Lys (positions 69–70) and for a pair of Val-Lys dipeptides at positions 71–78 and 87–88 among others. In addition to six recurring heptamers identified in Fig. 1 and Fig. 2, there were two more. GGAGGAC encoded a pair of Glu-Asp dipeptides at positions 83–84 and 91–92, whereas its single-base-substituted copies yielded Glu-Gly (positions 29–30), Gln-Asp (positions 104–105), Val-Asp (positions 124–125), and Ala-Asp (positions 134–135) dipeptides. The last recurring heptamer was CACCATC encoding Thr-Ile dipeptides at positions 4–5 and 43–44. This heptamer also spawned a sizable number of copies.

Although only one example each of the three classes of longer repeating units is shown in Fig. 2, it should be realized here that the total of 109 involved bases already accounted for 21% of the total of hamster αA-crystallin coding sequence (4). The fact is that within a given coding sequence, every segment is represented elsewhere as its copy or copies.

The Preponderance of Serine Codon AGC and of Arginine Codons CGC and CGG in the Human Small Heat Shock Protein Coding Sequence

Homology between lens αA-crystallins of vertebrate eyes and more ancient and ubiquitous small heat shock proteins was first noted by Ingolia and Craig (10), who studied four small heat shock proteins of *Drosophila melanogaster*. However, the homology with hamster αA-crystallin was considerably higher with human small heat shock protein (11). After the introduction of 26 deletions and 53 insertions to maximize the homology, the 173-residue-long hamster αA-crystallin (4) and the 199-residue-long human small heat shock protein (11) shared identical amino acid residues at 72 positions (32% homology). Contained within the 199-residue-long human heat shock protein were 16 leucine, 22 serine, and 18 arginine residues. Their codon preferences are summarized at the top of Fig. 3. It should be noted that although CTG remained to be the preponderant leucine codon, there occurred marked shifts from the hamster αA-crystallin coding sequence in codon preferences for serine and arginine. The previously prominent TCT was not used even once as serine codon; instead, 10 of the 22 serine residues were encoded by AGC.

Paralleling this marked shift among serine codons, TCT as base trimers recurred only 3 times, whereas AGC recurred all together 18 times. At the same time, CGC and CGG together accounted for 16 of the 18 arginine codons. The reason for this sudden prominence of serine codon AGC as well as of arginine codon CGC is shown in the bottom column of Fig. 3. The 9 recurring base heptamers identified within the human small heat shock protein coding sequence were clearly related to the 8 recurring heptamers already noted within the hamster αA-crystallin coding sequence (Figs. 1 and 2), thus

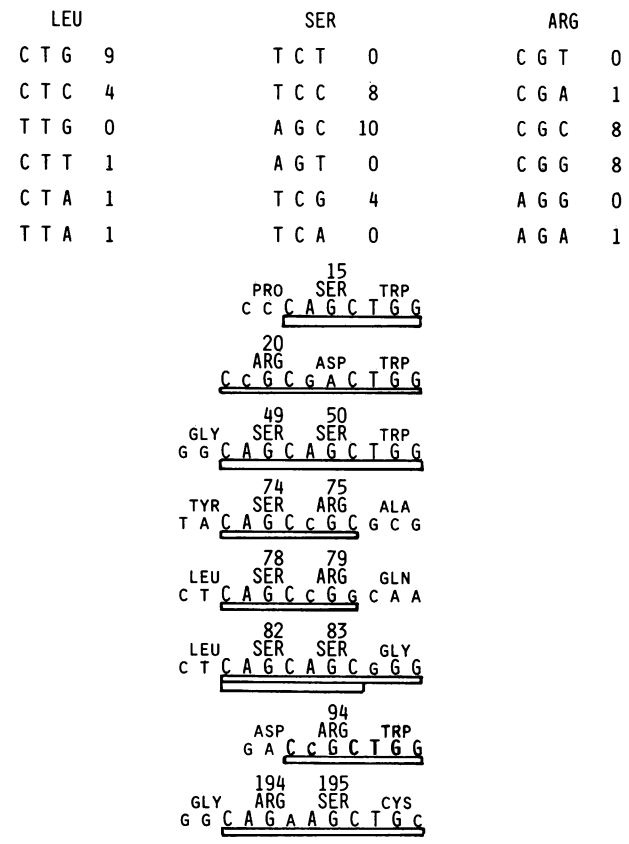

FIG. 3. Codon preference in human small heat shock proteins (199 residues). Various codons for leucine, serine, and arginine found within the 597-base-long human small heat shock protein coding sequence (11) are tabulated at the top. Comparison with Fig. 1 immediately reveals differences in prominent codons with regard to serine and arginine but not with regard to leucine. The preponderance of AGC among serine codons and of CGC among arginine codons was due to the presence of the two partially overlapping recurring base heptamers, CAGCAGC and CAGCTGG. At the bottom, eight copies of these two heptamers and/or of the composite decamer CAGCAGCTGG were listed. Eight of the 10 AGC serine codons are contained within these eight copies. These repeating units also accounted for 3 of the 8 CGC arginine codons. Copies of these two primordial heptamers translated in different reading frames are not shown.

confirming the propinquity of descents between the two. None of these 9 recurring heptamers of the human heat shock protein coding sequence, however, contained the TCT trimer within, GTCTGCC of the crystallin (Fig. 2, top), for example, becoming GGCTGCC. Furthermore, the heptamer CACCATC of the crystallin spawned two recurring heptamers in the small heat shock protein coding sequence, GATCACC encoding a pair of Ile-Thr dipeptides at positions 120–121 and 179–180 and CAGCAGC. It was this latter heptamer together with CAGCTGG that were entirely responsible for the frequent recurrence of the AGC trimer—therefore, its prevalence as serine codon. These two recurring heptamers overlapped with each other once, the resulting decamer encoding the 49th to 51st Ser-Ser-Trp (Fig. 3, bottom column, third row). In addition, its CAGCAGC portion encoded another Ser-Ser-dipeptide at positions 82–83, whereas its CAGCTGG portion was recapitulated to encode a Ser-Trp dipeptide (positions 82–83). It should be pointed out that eight copies of CAGCAGC and CAGCTGG shown in the bottom column of Fig. 3 already contained 8 of the 10 existing AGC serine codons. When the AGC portion of the above heptamers underwent a single-base substitution, it became the prominent arginine codon CGC; three examples are seen in Fig. 3, whereas a two-base substitution yielded CGG for

the 79th arginine. The more frequent source of the arginine codon CGG, however, was the CTG portion of three recurring heptamers that by a single-base substitution became CGG. Neither CAGCAGC nor CAGCTGG was identified in the hamster αA-crystallin coding sequence (4); nevertheless, a single-base-substituted copy of the former CAGCACC yielded Gln-His (positions 7–8) of the crystallin.

1. Grantham, R., Gautier, C., Gouy, M., Jacob, M. & Mercier, R. (1981) *Nucleic Acids Res.* **9,** 543–574.
2. Grantham, R., Gautier, C. & Gouy, M. (1980) *Nucleic Acids Res.* **8,** 1893–1912.
3. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, England).
4. van den Heuvel, R., Hendriks, W., Quaz, W. & Bloemendal, H. (1985) *J. Mol. Biol.* **185,** 273–284.
5. Ohno, S. (1987) *J. Mol. Evol.* **25,** 325–329.
6. Southern, E. (1972) in *Modern Aspects of Cytogenetics: Constitutive Heterochromatin in Man,* Symposia Medica Hoechst, ed. Pfeiffer, R. A. (Schattauer, Stuttgart, F.R.G.), Vol. 6, 19–28.
7. Feigenbaum, M. J. (1979) *J. Stat. Phys.* **21,** 669–706.
8. Gwin, E. G. & Westervelt, R. M. (1987) *Phys. Rev. Lett.* **59,** 157–160.
9. Ohno, S. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 6486–6490.
10. Ingolia, T. D. & Craig, E. A. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 2360–2364.
11. Hickey, E., Brandon, S. W., Potter, R., Stein, G. & Weber, L. A. (1986) *Nucleic Acids Res.* **14,** 4127–4133.