J. theor. Biol. (1991) **151,** 323–331

Periodicity in DNA Coding Sequences: Implications in Gene Evolution

ANASTASIOS A. TSONIS†, JAMES B. ELSNER‡ AND
PANAGIOTIS A. TSONIS§

† *Department of Geosciences, University of Wisconsin-Milwaukee, Milwaukee, WI* 53201, ‡ *Department of Meteorology, Florida State University, Tallahassee, FL* 32201 *and* § *Department of Biology, The University of Dayton, Dayton, OH* 45469, *U.S.A.*

(*Received on* 21 *March* 1990, *Accepted on* 25 *March* 1991)

In this paper we have employed Fourier analysis of DNA coding and non-coding sequences in an attempt to identify possible patterns in gene sequences. It was found that while intronic sequences show a rather random pattern, coding sequences show periodicities and in particular a periodicity of 3. We were able to reconstruct such patterns by assuming a gene having one codon occuring in about 40% of the sequence. This could indicate that the predominant presence of codons all starting from the same base could confer the observed periodicities. Indeed, it was found that proteins do obey this rule. Implications of this finding in gene evolution are discussed.

Introduction

Since the early 1970s, scientists have attempted to derive algorithms to study information in the DNA sequences. Such algorithms were meant to discriminate coding from non-coding regions, finding translation initiation sites or explore the function of genes. Specifically Reichert *et al.* (1973); Wong *et al.*, (1974) first devised algorithms to study information by aligning two sequences. Such algorithms were limited at the time due to the lack of sequencing methods and availability of long sequences. Subsequently Stormo *et al.* (1982, 1986) used algorithms to study the translational initiation sites from all other sites in *Escherichia coli* by inspecting over 78 000 nucleotides of mRNA. Their methods could give evaluation of a sequence's activity depending on the function or alignment to other sequences. Other algorithms to discriminate between coding and non-coding regions based on preference of the codon usage have been reported by Fickett (1982) and by a series of papers by Staden & McLachlan (1982) Staden (1984*a*, *b*). More recently Nazarea *et al.* (1985) have used Fourier analysis on tRNA and rRNA sequences. Their analysis showed periodicity by those sequences, that as it was proposed reflect a closed triadic set of tandem repeat lengths in a class of ancestral macromolecules. The authors suggested that these "homologies" of common ancestry was conserved during evolution and was indicated by a periodicity of spacing.

The evolution of the genetic information and the generation of genes is one of the most challenging problems for evolutionary and molecular biologists. The

0022-5193/91/150323+09 $03.00/0

mechanisms by which nature produced the genetic information and subsequent generation of genes are not well-understood. Modern genetic information carrying DNA sequences are thought to have evolved from ancestral primordial blocks. The leading theories suggest that the first building blocks were single-stranded (RNA) oligonucleotides produced by random polymerization of nucleotides. Such theories have in fact been substantiated by the discovery of RNA introns, that can act as enzymes (Cech, 1985). Therefore, the early primordial blocks could be replicated or fused to produce repeating units without the use of proteins. Ohno (1988*a*, *b*) has proposed that these short repeating units formed the basis for the generation of longer sequences by becoming progressively longer and less homologous to each other. In fact he has shown that in the case of two related modern sequences, the alpha A-crystallin and the small heat shock protein (32% homology) an original heptameric repeating unit can be identified. So far, limited studies at the theoretical or statistical level have given clues about possible rules obeyed during the generation of longer sequences from basic primordial blocks. In order to determine whether or not coding, or genomic sequences in general, follow any principle we have undertaken a detailed analysis by applying mathematical concepts.

Selection of Sequences and Analysis

Our analysis involves genomic and mRNA sequences. The rationale for this is to clarify any possible difference between coding and non-coding sequences. Thus, we have selected long sequences [over 6000 base pairs (bp)], the human c-myc proto-oncogene gene (Colby *et al.*, 1983) (8082 bp), the human b-nerve growth factor gene (Ullrich *et al.*, 1983) (11 594 bp), the rat cartilage proteoglycan core protein mRNA (Doege *et al.*, 1987) (6554 bp), the human apolipoprotein mRNA (Chen *et al.*, 1986) (14 070 bp), and the human dystrophin mRNA (Hoffman *et al.*, 1987) (13 957 bp). The first two sequences contain in more than 90% intronic sequences while the other three are only coding sequences. Then we generated an imaginary gene which we call the random gene. This gene was generated as follows: an integer number in the interval [1, 4] is selected at random if the number is one then the first base of the sequence is assumed to be A. If the number is two then the first base is assumed to be G, if the number is three then the first base is assumed to be C and if the number is four the first base is assumed to be T. We then repeat this procedure to determine which will be the second base and we continue until we have generated a random sequence of As, Gs, Cs, and Ts of length 8000. The reason for considering the random gene will become clear below.

For mathematical purposes any gene sequence may be thought of as a sequence of integers in the interval [1, 4]. Representing non-coding or coding sequences in this way allows us to consider it as an "observable" and to apply certain statistical mathematical techniques.

A very powerful tool for treating observables is spectral density function. The spectral density is very useful in isolating periodicities. The presence of a cycle and its frequency are shown by the spectral density function. For a purely random

process the spectral density oscillates randomly about a constant value indicating that no particular frequency explains any more of the variance of the sequence than any other frequency. For periodic or quasi-periodic sequences only peaks at certain frequencies (which are all related) exist. At any other frequency the value of the spectral density function is zero.

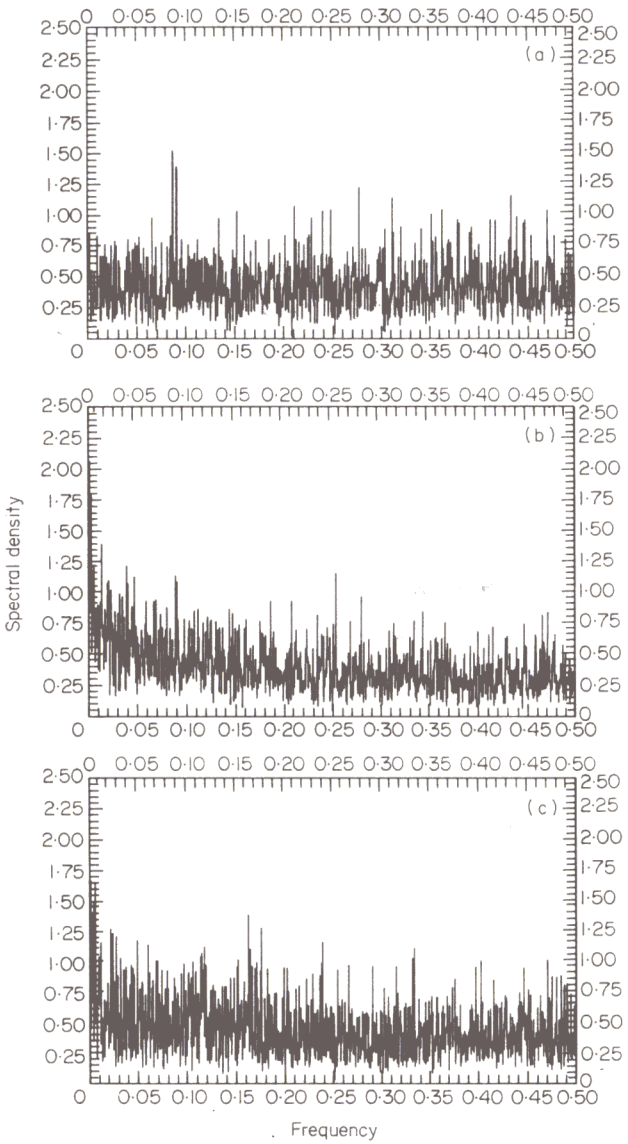

FIG. 1. Spectral density for genomic sequences (a), random sequences, (b) c-myc gene, (c), β-nerve growth factor gene.

Figures 1 and 2 demonstrate the above by showing the spectral density function for the random gene, the myc gene, the nerve growth factor gene (Fig. 1), the dystrophin mRNA, the core protein mRNA and the apoliprotein mRNA (Fig. 2). In the case of the random gene as expected the spectral density fluctuates randomly about a value near 0·45 [Fig. 1(a)]. Comparing Fig. 1(b) and (c) we observe

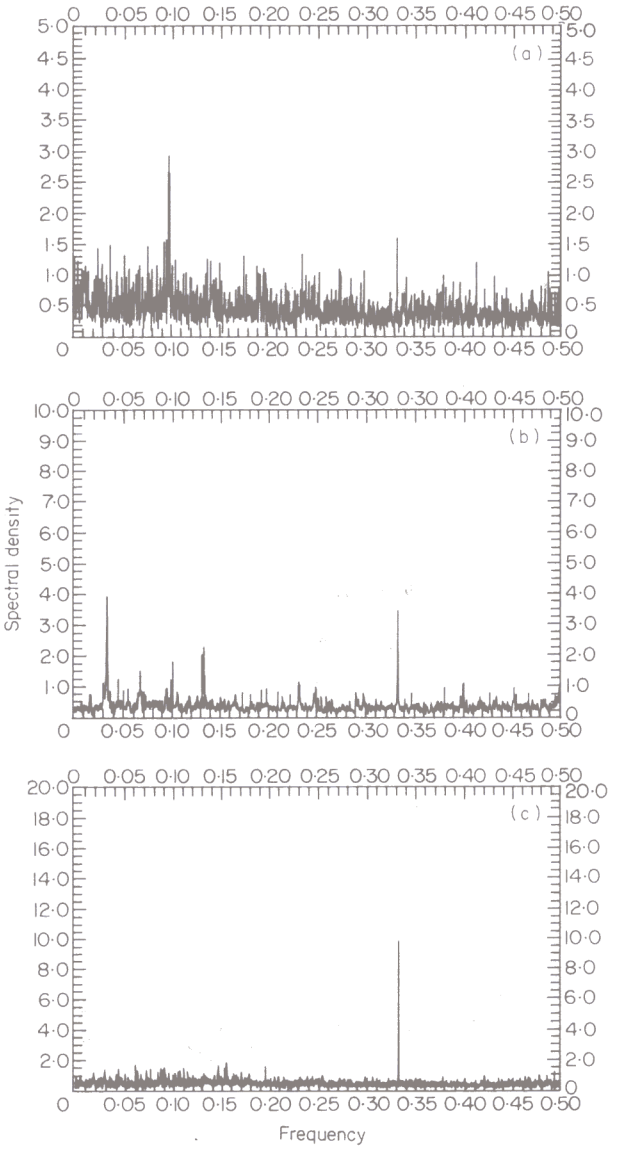

FIG. 2. Spectral density for coding mRNA sequences. (a) Dystrophin mRNA, (b) proteoglycan core protein mRNA, (c) apoliprotein mRNA

qualitative similarities between the spectral density of genomic sequences (containing mostly non-coding sequences) and a random sequence. Some small differences exist close to the origin of the graphs. This indicates that the sequence depicted by myc and nerve growth factor genes (non-coding sequences) may not be completely random but it has some very short memory (at best a low-order autoregressive process). In other words, some underlying process that at a certain instance determines the value of the next few steps may be present. Such processes would not result in pattern regularities or hidden structures in the sequence and therefore no periodicity can be observed. Figure 2(a), (b) and (c) (for the three mRNA sequences) appear to be more exciting. The spectral density functions present several pronounced peaks on a continuous background similar to the continuous background found in random sequences (note, however, that the scale is not the same in all cases). These results indicate that certain periodicities are involved in the coding sequences. More specifically a periodicity of three is most evident in all mRNA sequences examined. In dystrophin mRNA a periodicity of ten is also observed (0·1 value). Based on the above observations from the Fourier spectra we now propose a procedure to reconstruct from "primitive sequence" the dynamics involved in the spectra.

Elaboration on the Periodicity Found in the Spectra

In order to completely understand the properties of the sequences that are imprinted on the Fourier spectra let us consider the periodic sequence A--A--A--A-- . . . where the blanks can be filled randomly by A, C, G, or T. This sequence shows a periodicity of three because of the repetition of the base A. The spectral density of such sequence is significantly non-zero only at one frequency (0·33) which corresponds to the perfect periodicity of base A ($1/0{\cdot}333 = 3{\cdot}0$). In other words for this sequence a continuous background does not exist [Fig. 3(a)]. This definitely is not the case with the coding sequences where the spectral density because: (a) has a much lower value at frequency 0·333 and (b) shows significant activity at all frequencies. Let us return to our sequence A--A--A--A-- . . . and now we destroy the perfect repetition of base A by randomly replacing it with G, C, or T. Figure 3(b) shows the spectral density of the sequences shown in Fig. 3(a) but with 57% of the As randomly replaced by G, C, or T. Obviously now the spectral frequency value at frequency 0·33 has been reduced to a value very similar to that observed in real coding sequences [see e.g. Fig. 2(c)], while at the same time we observe activity at all frequencies due to the random break-up of the perfect repetition of base A. In fact we have reconstructed spectra that exhibit similar properties with the spectra of real coding sequences.

Therefore, the results reported in Fig. 2 indicate that certain periodicities are detectable in coding sequences. In addition, as it is demonstrated in Fig. 3 this periodicity is not perfect (as expected for real coding sequences), but broken in many places and provide a mechanism in explaining the global structure of coding sequences. Having observed the spectra and reconstructing the dynamics involved in the coding sequences we can elaborate on a possible evolutionary scenario.

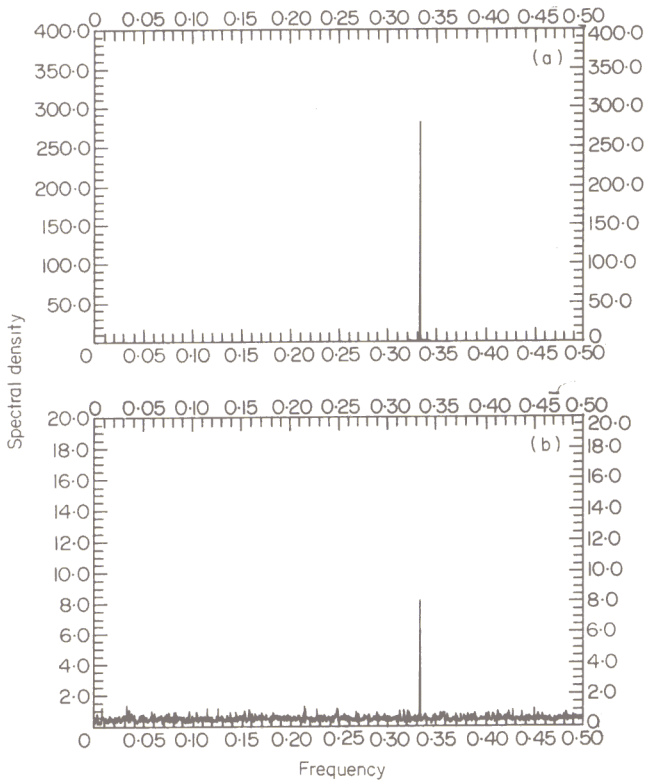

FIG. 3. Spectral density of periodic sequences. (a) A--A--A--A-- . . . (b) A--A--A--A--, but 57% of As have been randomly replaced by G, C, or T.

Periodicity of Three and Implications in Gene Evolution and Protein Structure

The search for periodicities in gene sequences has in the past produced some interesting ideas on the origin of the genetic code and the principles underlying its construction. Shepherd has analyzed coding sequences for *RNY* periodicity, *R* being a purine, *Y* a pyrimidine and *N* any nucleotide (Shepherd, 1982). Codons of *RNY* form seem to be predominant suggesting that it could be the most primitive codon. It is interesting that most of the amino acids coded by an *RNY* codon are most likely to have been generated by prebiotic synthesis and have also been frequently found in meteorites (for discussion see Watson *et al.*, 1987). Recently, Cocho & Rius have observed three-periodicities, with a different approach which takes into consideration an algorithm related to the relative strengths of G:C and A:T bonds (Cocho & Rius, 1989). Fickett (1982) has also reported that oligonucleotides in sequences tend to be repeated with a periodicity of 3. Ohno's search for primordial blocks, however, has involved the base trimer CTG included in the primordial repeating units (Ohno, 1988*a*, *b*). CTG is the preponderant leucine codon from *E. coli* to man. Based on the recurrence of this base trimers Ohno proposed

that TG/CT excess-TA/CG deficiency might be a universal rule of the coding sequence construction (Ohno 1988*a*, *b*). Recently Yomo & Ohno (1989) extended this observation to noncoding regions as well. These rules are the result of the presence or absence of particular codes from the coding sequences. For example, the TG/CT excess is observed because the CTG codon for Leu (the most prominent residue in proteins) is well-used. The TA/CG deficiency could result form the low usage of Arg and the termination codes TAA and TAG. These rules, however, do not interfere with the periodicity of 3, since C or G could form the basis for it (see below). The amino acid preference and the unequal usage of codons has been the subject of extensive theoretical studies. Staden & McLachlan (1982), and McLachlan *et al.* (1984) have studied the effects of codon preference and its use to identify protein coding regions in long DNA sequences. They described a method that based on the codon preference (occuring only in coding sequences) they were able to distinguish between intons and exons. In other words, they arrive to the conclusion drawn by Fickett (1982) that the unequal frequency of codon usage could give certain structural features including periodicity to distinguish between potential coding sequences. Furthermore, Rowe & Trainor (1983) using the procedure by Gatlin (1972) have analyzed the existence of strong codon biases in viral DNA. Their study employing correlation plots of dinucleotides, such as TT, CC, AA, GG shows that non-coding regions do not show any correlations while correlation in coding regions have a period of 3 as a result of codon bias. Weber (1987) has also reported that in the codons of the human malaria parasite *Plasmodium falciparum* T or A in the third position were strongly preferred. Such a codon usage could also confer a periodicity in the coding sequences of the parasite.

Our results show clearly that coding sequences do possess a periodicity of 3. Other periodicities could be observed occasionally probably due to possible length of a repeating block, or due to specific structural motifs that might exist in the particular protein. However, when genomic sequences were examined, where non-coding sequences are present in more than 90% of the total sequences, the periodicity is lost and the sequences show a rather random distribution. Why intronic sequences follow this pattern? Any intronic sequence is also comprised of the four bases A, G, C, T and if someone takes them every 3 to form a codon a putative protein could be made. This putative protein could also be lengthy (interruption of stop codons would not really interfere with our reasoning). So if these intronic sequences could theoretically create a protein and, therefore, act as coding sequences, why do they not show periodicities? The answer could lie in the structure of proteins themselves. Even though it has not explicitly been studied, it is accepted that each protein contains a few amino acids predominantly. In other words, sometimes two, three or four amino acids compose more than 30% of all amino acids in a protein. Consequently, if the codons for these amino acids start from the same first base the periodic nature of the coding sequences becomes apparent according to our data (Fig. 3). Therefore, we extended our studies to examine whether or not in a sample of 100 proteins this is the case. The proteins were selected randomly from the PIR sequence library (by choosing numbers, while precaution was taken not to choose sequences belonging to the same family) and examined using GENEPRO (Riverside

Scientific Enterprises). Such a sampling shows that most of a protein sequence is made of a few amino acids and that the coding mRNAs will show a periodicity of 3 because the codons for these amino acids have the first base in common. In fact we found that in 80% of the examined proteins four amino acids (all having the same first base in their codons, which could differ from protein to protein) compose 32% of each protein sequences. In the rest of the cases one, two or three amino acids compose almost 40% of the proteins. Leucine (starting letter C) was the most frequent, but amino acids whose codons are starting from G were almost as many as the ones starting from C. In the particular examples used in Fig. 2 the periodicity of 3 can be attributed to the base C for dystrophin (32·1% of amino acids have C in the first place of their codons), G for core protein (40% of the amino acids have G in the first place of their codons), A for apolipoprotein (35·7% of the amino acids have A in the second place of their codons). The introns, however, were not evolved to code for proteins and this principle does not apply and that is why they do not show periodicity of 3. This in turn could suggest that the primitive genetic code was a repetition of one codon or that the most primitive proteins were polymers of one amino acid. Subsequent degeneration of the repeated codons and mutations would have caused the generation of different proteins. Alternatively, this also could imply that amino acids themselves could have determined the codon specific for each one of them. In this case each amino acid could interact specifically with RNA oligomers and create a progenitor of the genetic code. Such ideas have been presented by several scientists in the past and recently (Woese, 1982; Hicke *et al.*, 1989; Yarus & Christian, 1989). Different ideas involving specific anticodon-amino acid interactions have also been developed (Shimizu, 1987*a*, *b*; Jungck *et al.*, 1982). Specific interactions of anticodonic dinucleoside monophosphates and their cognate amino acids have been detected by using ultraviolet absorbance photometry, circular dichroism and by fast bombardment-mass spectroscopy. This kind of specificity, we believe, should have arisen later in evolution to facilitate protein synthesis. By accepting these specific interactions, however, we can use the same reasoning for the generation of the primitive genetic code (as a repetition of one or few codons), because substituting the codons with anticodons would not affect the periodicity which in turn suggests predominant use of one or few amino acids for the generation of a primitive genetic code. All in all, viewing the periodicities from another angle it is tempting to speculate that this RNA—amino acid collaboration played a big role in the genesis of the genetic code, and to propose that random base assembly in primitive RNA oligomers alone was not the basis for the construction of modern sequences.

REFERENCES

CECH, T. R. (1985). *Internat. Rev. Cytol.* **93**, 3-22.

CHEN, S. H., YANG, C. Y., CHEN, P. F., SETZER, D., TANIMURA, M., LI, W. H., GOTTO, A. M. JR. & CHAN, L. (1986). *J. biol. Chem.* **261**, 12918-12921.

COCHO, G. & RIUS, J. L. (1989). *Presented at the Conference on Fundamental Problems of Evolutionary Biology.* Moscow.

COLBY, W. W., CHEN, E. Y., SMITH, D. H. & LEVISON, A. D. (1983). *Nature, Lond.* **301**, 722-725.

DOEGE, K., SASAKI, M., HORIGAN, E., HASSELL, J. R. & YAMADA, Y. (1987). *J. biol. Chem.* **362,** 17757-17767.
FICKETT, J. W. (1982). *Nucl. Acids Res.* **10,** 5303-5318.
GATLIN, L. L. (1972). *Information Theory and the Living System.* New York: Columbia University.
HICKE, B. J., CHRISTIAN, E. L. & YARUS, M. (1984). EMBO J. **8,** 3843-3851.
HOFFMAN, E. P., MONAKO, A. P., FEENER, C. C. & KUNKEL, L. M. (1987). *Science* **238,** 347-350.
JUNGCK, J. R., DICK, G. & DICK, G. (1982). *Biosystems* **15,** 259-273.
MCLACHLAN, D. A., STADEN, R. & BOSWELL, D. R. (1984). *Nucl. Acids Res.* **12,** 9567-9575.
NAZAREA, A. D., BLOCH, D. P. & SEMRAU, A. C. (1985). *Proc. natn. Acad. Sci. U.S.A.* **82,** 5337-5341.
OHNO, S. (1988*a*). *Proc. natn. Acad. Sci. U.S.A.* **85,** 4378-4382.
OHNO, S. (1988*b*). *Proc. natn. Acad. Sci. U.S.A.* **85,** 9630-9634.
REICHERT, T. A., COHEN, D. N. & WONG, A. K. C. (1973). *J. theor. Biol.* **42,** 245-261.
ROWE, G. W. & TRAINOR, L. E. H. (1983). *J. theor. Biol.* **101,** 151-170.
SHEPHERD, J. C. W. (1982). *CSH Symp. quant. Biol.* **47,** 1099-1108.
SHIMIZU, M. (1987*a*). *J. Phys. Soc. Jap.* **56,** 43-45.
SHIMIZU, M. (1987*b*). *J. Phys. Soc. Jap.* **56,** 893-896.
STADEN, R. (1984*a*). *Nucl. Acids Res.* **12,** 521-538.
STADEN, R. (1984*b*). *Nucl. Acids Res.* **12,** 551-567.
STADEN, R. & MCLACHLAN, A. D. (1982). *Nucl. Acids Res.* **10,** 141-156.
STORMO, G. D., SCHNEIDER, T. D. & GOLD, L. (1986). *Nucl. Acids Res.* **14,** 6661-6679.
STORMO, G. D., SCHNEIDER, T. D., GOLD, L. & EHRENFEUCHT, A. (1982). *Nucl. Acids Res.* **10,** 2997-3011.
ULLRICH, A., GRAY, A., BERMAN, C. & DULL, T. J. (1983). *Nature, Lond.* **303,** 821-825.
WATSON, J. D., HOPKINS, N. H., ROBERTS, J. W., STEITZ, J. A. & WEINER, A. M. (1987). In: *Molecular Biology of the Gene.* pp. 459-461. Menlo Park, CA: Benjamin/Cummings.
WEBER, J. L. (1987). *Gene* **52,** 103-109.
WOESE, C. R. (1982). In: *Ribosomes: Structure, Function and Genetics* Chambliss G. *et al.*, eds) pp. 357-376 Baltimore, MD: University Park Press.
WONG, A. K. C., REICHERT, T. A., COHEN, D. N. & AYGUN, B. O. (1974). *Comput. Biol. Med.* **4,** 43-57.
YARUS, M. & CHRISTIAN, E. L. (1989). *Nature, Lond.* **342,** 349-350.
YOMO, T. & OHNO, S. (1989). *Proc. natn. Acad. Sci. U.S.A.* **86,** 8452-8456.