J. theor. Biol. (1988) **134,** 1-7

Significance of the Purine-Pyrimidine Motif Present in Most Gene Groups

ATHEL CORNISH-BOWDEN

Centre de Biochimie et de Biologie Moléculaire, Centre National de la Recherche Scientifique, 31 *chemin Joseph-Aiguier, B.P.* 71, 13402 *Marseille Cedex* 9, *France*

(*Received* 11 *January* 1988)

The probability of the sequence YRY(N_i)YRY occurring most frequently with the same i value in seven out of nine gene classes is reassessed and found to be about $1{\cdot}3 \times 10^{-8}$, more than 4000 times greater than the value calculated by Arquès & Michel (1987), but still much too small for chance to be a reasonable explanation for the observation. Even if the sequence YRYNNNNNNYRY were very frequent in the most primitive genes, it would not have survived in a recognizable form to the present day if it were selectively neutral. However, if it is selectively favoured one would expect it to exist regardless of whether it was present in primitive genes.

1. Introduction

Arquès & Michel (1987) recently studied the frequency of occurrence of sequences YRY(N_i)YRY in different classes of gene, for $i = 1$ to 99, and found that seven out of the nine classes considered (2271 eukaryotic protein-coding genes, 1750 kb; 788 prokaryotic protein-coding genes, 768 kb; 121 chloroplast protein-coding genes, 116 kb; 130 mitochondrial protein-coding genes, 117 kb; 50 viral introns, 99 kb; 92 ribosomal RNA genes, 167 kb; 920 tRNA genes, 70 kb; 1182 viral protein-coding genes, 1306 kb; and an unstated number of eukaryotic introns), showed the highest frequency at the same i value, 6. (The exceptions were the viral protein-coding genes, for which the highest frequency was observed at $i = 12$, and the eukaryotic introns, for which no details were given.) They estimated the probability of obtaining such a result by chance to be 3×10^{-12}, and concluded that the sequence YRYNNNNNNYRY can be considered a primitive oligonucleotide, i.e. that it dates from the origin of life, before any molecular evolution occurred.

Unfortunately, the probability calculation of Arquès & Michel (1987) incorporated three separate biases, all in the same direction, and resulted in an estimated probability too small by a factor of about 4000. Nonetheless, the corrected probability is still much too small for the observation to be attributable to chance, and one must still consider whether the existence of the sequence YRYNNNNNNYRY in present-day organisms provides information about primitive sequences.

2. Probability Calculation

(a) *Choice of* $i = 6$

Arquès & Michel (1987) examined the probability that $i = 6$ would give the highest

0022-5193/88/170001+07 $03.00/0

frequency in all classes examined. However, since this value of i was chosen *a posteriori*, i.e. after it was identified as the extreme case and not for any independent reason, this was an improper question to ask. One should instead calculate the probability of observing the same value of i in all classes, without presupposing a particular value. As 99 values of i were allowed, this might imply an underestimate of the probability by a factor of 99, but the calculation is complicated by the facts that (i) there is a near certainty that the value of i giving the maximum frequency will be a multiple of three for protein-coding genes (see Arquès & Michel, 1987), and (ii) only values of i in the range 1 to 29 were considered in the case of the tRNA genes. Thus only nine values of i (3, 6, 9 . . . 27) have an appreciable likelihood of giving the highest frequency in all classes. The true bias is therefore about 9-fold—smaller than 99-fold but far from trivial.

(b) *Treating 7/9 successes as 7/7*

Although it was observed that seven out of nine classes of gene had the highest frequency for $i = 6$, this result was analysed as if it were seven out of seven, i.e. the exceptional classes were omitted from the calculation! For the viral protein-coding genes, this was done in order "to simplify the statistical reasoning", but for the eukaryotic introns no reason was given for the omission. However, obtaining seven wins in nine trials is by no means as improbable as obtaining seven wins in seven trials; if the probability of a single win is small, the probability of obtaining seven wins out of nine approaches 36 times that of obtaining seven wins out of seven. Thus the omission of the exceptional classes leads to a bias approaching 36-fold.

(c) *Choice of YRY(N_i)YRY to be studied*

The choice of YRY(N_i)YRY for study was apparently based on the fact that it was the extreme case out of those examined ("No such rule can be identified with this statistical approach based on the study of the i-motifs which are built with multinucleotides different from YRY"). However, one must not test the extreme case as if it were a typical case: one must **expect** the extreme value in a set of 20, for example, to be "significantly deviant" if tested with a 95% confidence test that ignores the fact that it is the extreme value and is not randomly selected. To estimate the bias resulting from the choice of YRY(N_i)YRY, one needs to know the population of motifs that could have been examined, but this is not clearly specified. I shall assume that it consisted of YYY(N_i)YYY, YYR(N_i)YYR, YYR(N_i)RYY, YRY(N_i)YRY, YRR(N_i)YRR, YRR(N_i)RRY, RYY(N_i)RYY, RYY(N_i)YYR, RYR(N_i)RYR, RRY(N_i)RRY, RRY(N_i)YRR and RRR(N_i)RRR, i.e. that only trinucleotides were considered for the bracketing sequences, that the second trinucleotide was either the same as or the mirror image of the first, and that the only choice was between pyrimidine and purine, so that motifs such as HAC(N_i)HAC (where H = not-G) were not considered either. Thus, if we assume that the most extreme out of 12 possible motifs was tested, the probability calculation was biased by a factor of 12. However, if other more complex motifs were also examined the bias was larger than this (maybe much larger).

(d) *Unbiased calculation of the chance probability*

The points noted above suggest that the value of 3×10^{-12} obtained by Arquès & Michel (1987) was too small by a factor of around $9 \times 36 \times 12$, or about 4000. However, it is not difficult to obtain an unbiased value directly: for the protein-coding genes, if we assume that the probability that the maximum frequency will occur at a value of i that is not a multiple of three is virtually zero, the probability that it will occur with $i = 3$ is 1/33; the probability that it will occur with $i = 6$ is also 1/33; etc. For tRNA genes (for which only values of i up to 29 were considered), the probability of a maximum frequency at any value of i up to 29 is 1/29, but it is 0/29 for $i = 30$ to 99. For the introns and the ribosomal genes the probability is 1/99 for any i from 1 to 99. Thus, for $i = 3$, the probability of obtaining the highest frequency in all nine classes is $33^{-5} \times 29^{-1} \times 99^{-3}$ and one obtains the same result for $i = 6$, 9, 12, 15, 18, 21, 24 or 27. For $i = 30$, 33, 36 . . . 99 the probability is zero, because these values were not possible for the tRNA genes. If all classes except one of the protein-coding genes showed maximum frequency at the same i value, the probability would be 32-fold higher for each, or 5×32-fold higher altogether. If all except the tRNA genes had the same i value, the probability would be 28-fold higher for i less than 29, and $29 \times (24/9)$-fold higher for i greater than 29. If one of the other three classes was the exception the probability would be 3×98-fold higher. These calculations may be extended to deal with the case in which seven out of nine classes have a maximum frequency at the same i value.

Summarizing the results, one finds a probability of $8{\cdot}2 \times 10^{-15}$ for nine classes out of nine, $4{\cdot}6 \times 10^{-12}$ for eight out of nine, and $1{\cdot}09 \times 10^{-9}$ for seven out of nine, or $1{\cdot}10 \times 10^{-9}$ for at least seven out of nine. Multiplying this by 12 to allow for the fact that YRY(N_i)YRY is the extreme case among the 12 kinds of motif tested, one obtains a value of $1{\cdot}3 \times 10^{-8}$ for the likelihood of obtaining by chance a result of the sort reported by Arquès & Michel (1987).

3. Likelihood of Survival of the Sequence Since the Origin of Life

Although the probability calculated above is 4300 times larger than the value estimated by Arquès & Michel (1987), it is still very small, so one must still ask how the motif can have arisen. There are two limiting cases: it may have been an abundant motif in the primitive genes existing before divergence formed the eight classes considered, and despite random neutral mutations since the origin of life, its presence at a higher-than-chance frequency is still detectable; alternatively, its present frequency may reflect a slight selectional advantage. The second possibility cannot easily be tested (except by eliminating alternatives), but the first can be studied by estimating the time needed for random neutral mutations to obliterate all trace of a primitive motif.

It is easy to construct a regular repeating sequence that has a high incidence of motifs with $i = 6$. For example, the sequence $(\mathrm{YRYNNNNNN})_n$, where n is large, gives a very high frequency for $i = 6$. However, the ancestral sequence postulated by Arquès & Michel (1987) cannot be of this type because any such regular repeating sequence leads to high frequencies for i values other than 6. For example, if

NNNNNN in $(YRYNNNNNN)_n$ is defined so that it provides no additional YRY triplets, e.g. if NNNNNN = RRYYRR, then $i = 15, 24, 33 \ldots$ all show as high a frequency as $i = 6$. If each N is defined at random as Y or R the same is true in a statistical sense, but the exact values vary and there is no preference for $i = 6$ over $i = 15, 24, 33$ etc.

An ancestral sequence that leads to a clear preference for $i = 6$ may be constructed by repeatedly replacing dodecanucleotides N_{12} by YRYNNNNNNYRY at randomly chosen loci. However, other values of i still have fairly high frequencies with this type of sequence, and simulations along the lines of those described below lead to fairly rapid loss of a perceptible preference for $i = 6$. To ensure maximum duration of survival for the ancestral motif, one must make the initial frequencies for $i \neq 6$ as low as possible. A sequence that satisfies this requirement is $[YRYR-RYYRRYRY(N_j)]_n$, where $j = 0, 1, 2 \ldots$, increasing by 1 for each repetition of YRYRRYYRRYRY. As YRYRRYYRRYR contains equal numbers of Y and R, the N must also be divided equally between Y and R to give equal numbers in the whole sequence, and if the first 50% of the N are Y and the second 50% are R there are no random occurrences of motifs with $i \neq 6$. To provide 1000 locations at which the possible occurrence of $YRY(N_{99})YRY$ can be examined, the total sequence must have a length of 1104 bases (i.e. $1000 + 99 + 6 - 1$).

A sequence of 1104 bases constructed as described provides 36 occurrences of $i = 6$, 10 occurrences of $i = 87$, and up to 9 occurrences of each other i value. [In all cases, both here and below, only the first 1000 possible loci are counted, so that with $i = 6$, for example, the last 93 (i.e. $99 - 6$) bases in the sequence are ignored. This is the same convention as that defined by Arquès & Michel (1987).] It appears impossible to define an ancestral sequence that gives appreciably more than a 3·5-fold higher frequency of $i = 6$ than of any other i value. By repeatedly introducing mutations R → Y and Y → R at random loci one can assess the period required for the original preference for $i = 6$ to be obliterated.

4. Evolutionary Simulation

An ancestral sequence of 1104 bases (as defined above) was set up in a Digital Equipment Corporation PDP11 computer using a program written in Basic. This was allowed to accumulate substitutions such that each could occur at any locus with a probability of 1/1104. If the site contained R before substitution this was changed to Y, and *vice versa*. Multiple substitutions at the same site were allowed and were not restricted in any way, i.e. the probability of substitution at any locus was maintained at 1/1104, regardless of the previous number of substitutions at that or any other locus. The sequence was allowed to accumulate 1100 substitutions in total, so that the mean number of substitutions at each locus at the end of the simulation was 0·996. The number of motifs for each value of i from 1 to 99 was counted (ignoring the last $(99 - i)$ bases in the sequence, as noted above) after 0, 100, 200, 300 ... 1100 substitutions. The entire simulation was carried out using the same ancestral sequence ten times, each with different random substitutions.

The results are summarized in Fig. 1. The initial high frequency at $i = 6$ always decreased rapidly as substitutions accumulated, and the frequencies at $i \neq 6$ increased. After only 400 substitutions, the preference for $i = 6$ was slight and by 500 substitutions (only 0·45 per locus) it had disappeared. The recovery between 600 and 800 substitutions is most reasonably interpreted as a random walk rather than a consequence of the initial preference for $i = 6$, but even if this recovery was not due to chance it is evident that $i = 6$ was indistinguishable from the general distribution by the end of the simulation, i.e. after 1000 to 1100 substitutions.

This simulation can be considered a sample of ten sequences of 1104, a total of 11 kb. This is less than the number of bases in any of the classes considered by Arquès & Michel (1987), which had a median size of 142 kb (ignoring the eukaryotic introns, for which details were not given). However, increasing the sample size by a factor of about 13 would only improve the statistical uncertainty of the results by a factor of about $13^{0.5}$, or 3·6. It is evident from Fig. 1, that although this might improve the chance of detecting a preference for $i = 6$ after more than 500 substitutions, it would certainly have little effect after 1100 substitutions.

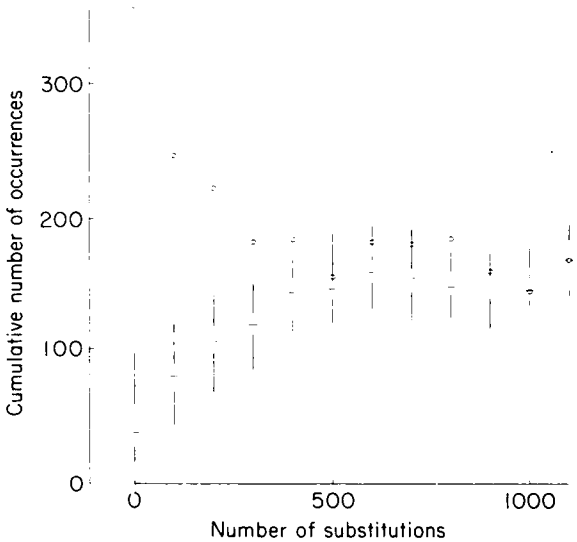

FIG. 1. Elimination of a preferred motif by selectively neutral substitutions. The graph shows the cumulative results of ten simulations of evolution from the ancestral sequence $[YRYR-RYYRRYRY(N_j)]_n$, where $j = 0, 1, 2 \ldots$, increasing by 1 for each repetition of YRYRRYYRRYRY, and n is large enough to give a total length of 1104 bases; the first 50% of the N were defined as Y, and the second 50% as R. In each simulation purines were replaced by pyrimidines and *vice versa* a total of 1100 times at randomly chosen loci, and the frequency of motifs $YRY(N_i)YRY$ were counted for each value of i from 1 to 99, after 0, 100, 200, ... etc. substitutions. The actual counts for $i = 6$ are shown (○), together with distributions for the other 98 values of i. Each distribution is shown as two vertical lines, drawn from the 1st to the 24th (lower quartile) and from the 98th to the 75th (upper quartile) values, and the median is indicated by a horizontal line. This representation is used in accordance with the principles advocated by Tufte (1983), to convey the maximum amount of information with the minimum amount of ink.

One must conclude, therefore, that neutral substitutions at a mean rate of one per locus would be sufficient to eliminate any evidence of a motif of the kind found by Arquès & Michel (1987). Nonetheless, one must still ask how many substitutions ought to have occurred in the time available. Perhaps one per locus is an over- or underestimate. From a variety of evidence (see Kimura, 1983), the neutral substitution rate can be estimated to be around 5×10^{-9} per year, but this includes all base substitutions and must be multiplied by 2/3 to get the transversion rate, i.e. the rate of purine/pyrimidine substitution that is relevant here. This gives a value of 3×10^{-9} per year. According to Dobzhansky *et al.* (1977), quoting Cloud (1974), the diversification of prokaryotes started between $2{\cdot}2 \times 10^9$ and $3{\cdot}3 \times 10^9$ years ago, and we may thus take $2{\cdot}5 \times 10^9$ years as a reasonable estimate of the time available for neutral substitutions to occur. Multiplying this by the transversion rate of 3×10^{-9} per year, one obtains a mean value of seven or eight substitutions per locus. This is much greater than the number used in the simulation, and establishes that all trace of a primitive oligonucleotide should be obliterated if there were no selection.

5. Discussion

Two points emerge from this analysis. First, the fact that YRYNNNNNNYRY is the most frequent of all sequences of the type YRY(N_i)YRY, in seven out of nine classes of gene studied, cannot be ascribed to chance, nor can it reasonably be attributed to the survival of a primitive sequence that is selectively neutral. One cannot discard the possibility that the motif is a selectively favoured primitive sequence, but its existence today provides no information about whether it is primitive or not, because if selectively favoured it will be present today regardless of whether it existed $2{\cdot}5 \times 10^9$ years ago.

It is striking that among the protein-coding genes the preference for $i = 6$ is extremely slight except in the case of those from chloroplasts. For the mitochondrial protein-coding genes, for example, it is barely distinguishable in fig. 1(d) of Arquès & Michel (1987) that the occurrence of $i = 30$ is less than the occurrence of $i = 6$, and $i = 9$ is also very frequent; other multiples of three, for example $i = 48$, are very much less frequent, around 70% of the occurrence of $i = 6$. This type of behaviour is not easy to interpret in relation to the ancestral-gene model. On the other hand, the RNA genes, especially the tRNA genes (Fig. 1(g) of Arquès & Michel, 1987), show a more pronounced preference for $i = 6$, as well as much more suggestion of non-randomness elsewhere in the diagrams. Remarkably, the viral introns (Fig. 1(e) of Arquès & Michel, 1987) also show definite non-randomness in addition to the pronounced preference for $i = 6$; there is a suggestion, weak but clear, of the same sort of regular alternation between high and low values that one sees in the protein-coding genes.

It is not necessary to assume that the same underlying cause explains both the weak preference for $i = 6$ in the protein-coding genes and the stronger preference for $i = 6$ in the others. Given that both ancestral sequence models and models that involve correlations among consecutive bases in more stable sequences are likely

to produce preferences for relatively small i values, one should not be too surprised if two quite different models lead to the same small value.

I am grateful to the referee for correcting an error, and for several useful suggestions that have been incorporated into the discussion.

REFERENCES

ARQUÈS, D. D. & MICHEL, C. J. (1987). *J. theor. Biol.* **128,** 457.

CLOUD, P. (1974). *Amer. Sci.* **62,** 54.

DOBZHANSKY, T., AYALA, F. J., STEBBINS, G. L. & VALENTINE, J. W. (1977). In: *Evolution.* San Francisco: W. H. Freeman.

KIMURA, M. (1983). In: *The Neutral Theory of Molecular Evolution.* Cambridge: Cambridge University Press.

TUFTE, E. R. (1983). In: *The Visual Display of Quantitative Information.* Cheshire, Connecticut: Graphics Press.