“Word” Preference in the Genomic Text and Genome Evolution: Different Modes of n-tuplet Usage in Coding and Noncoding Sequences

J Mol Evol (2005) 61:23–35
DOI: 10.1007/s00239-004-0209-2

JOURNAL OF MOLECULAR EVOLUTION

"Word" Preference in the Genomic Text and Genome Evolution: Different Modes of *n*-tuplet Usage in Coding and Noncoding Sequences

Christoforos Nikolaou, Yannis Almirantis

Institute of Biology, National Research Center for Physical Sciences "Demokritos," 15310, Athens, Greece

Received: 7 July 2004 / Accepted: 2 February 2005 [*Reviewing Editor:* Dr. Brian Morton]

Abstract. Extensive work on *n*-tuplet occurrence in genomic sequences has revealed the correlation of their usage with sequence origin. Parallel to that, there exist different restrictions in the nucleotide composition of coding and noncoding sequences that may result in distinct modes of usage of *n*-tuplets. The relatively simple approaches described herein focus on such differences. They are based on simple summation measures of *n*-tuplet frequencies, computed after filtering the background nucleotide composition. Among the main targets of this work is to draw some conclusions on the qualitative differences in the composition of genomic sequences depending on their functionality. Moreover, an evolutionary model is formulated, including simple forms of ubiquitous events of genome dynamics: genomic fusions, genome shuffling due to transpositions, replication slippage, and point mutations. This model is shown to be able to reproduce all the statistical features of genomic sequences discussed herein.

Key words: "Word" frequencies — *n*-tuplet occurrence — Coding potential — Genome evolution

Introduction

The nucleotide composition of genomic sequences has been shaped in evolutionary time through various procedures, reflecting the constitution of protobiotic macromolecules, the occurrence of ancestral mixing events, and the necessity for physicochemical stability as well as the interplay with protein molecules and parts of cellular "machinery." The combination of the factors mentioned above with selection has resulted in a great compositional variety among the known organisms, even at the level of mononucleotide frequencies. The very distinct G + C genomic contents of bacterial genomes (see, e.g., Li 1997) and the isochore structure of several vertebrate taxa (Bernardi 1989, 1993) stand out as the most obvious examples. Early studies of the nucleotide composition of genomic sequences have pointed out the existence of selective pressures, and other types of biases, relating local sequence constitutional preferences with higher-level chromatin structure (Holmquist 1989; Zuckerkandl 1992).

Furthermore, soon after extensive sequencing of genomic sequences had started, observations of special characteristics regarding the occurrence of dinucleotides began to accumulate (see, e.g., Burge et al. 1992; Karlin and Burge 1995; Nussinov 1981). Nearest neighbor patterns were shown to exist in all known organisms, thus affecting the di-, tri-, and, less strongly, higher *n*-nucleotide content of their genomes. These observations soon led to the formation of the concept of "genomic signature" (Karlin et al. 1994; Karlin and Mrazek 1997),

Correspondence to: Yannis Almirantis; *email:* yalmir@bio.demokritos.gr

through which a genomic identity could be ascribed to each organism based on a vector of the relative abundances of the 16 dinucleotides. The genomic signature has been used to derive evolutionary relationships, which have been proved to be in accordance with conventional taxonomy. Species dependence was made apparent at the level of codon preference as soon as the formulation of efficient codon usage tables had taken place (Karlin and Mrazek 1997; Nakamura et al. 1996). Recently, Qi et al. (2004) have extended such applications using amino acid residues instead of nucleotides.

Observed patterns in species *n*-tuplet usage have been directly visualized through approaches like the "Chaos Game Representation" (Jeffrey 1990; Goldman 1993) and "genomic portraits" (Hao 2000a, b), and in most cases, the *n*-tuplet occurrences are not correlated with the functional role of the examined sequences. There are, however, certain aspects of the genomic text that differ between coding and noncoding sequences and may, to some extent, affect their compositional profile. The coding procedure imposes several restrictions on sequences, which are reflected upon codon and amino acid preferences and this is very likely to lead to avoidance of specific oligonucleotides. On the other hand, noncoding sequences have been shown to exhibit distinct compositional tendencies resulting in the clustering of same (Dechering et al. 1998) and similar (purine or pyrimidine) nucleotides at several length scales (Peng et al. 1992; Almirantis and Provata 1997).

This work mainly focuses on an effort to capture and quantify possible different tendencies and patterns in the *n*-tuplet composition between sequences of different functionality. The approaches presented herein are not attempting the development of a tool able to determine coding–noncoding partition of a given set of genomic sequences. However, approaches along the same line of research have shown that such partitions are possible, which may lead to relatively satisfactory degrees of specificity between coding and noncoding sequences (Nikolaou and Almirantis 2003, 2004; Almirantis and Nikolaou 2005). Contributions in this field are usually based on more elaborate techniques, often requiring combination of algorithms carefully selected and trained. The approaches presented and discussed in the following may be seen as part of an attempt to reveal attributes of sequences that would primarily aim at shedding light on genomic structure and evolutionary history.

Addressing tendencies and constraints in the oligonucleotide usage, is made easier when subtracting the background composition, in order to reveal the relative preferences or avoidances to their entireness. The approaches described in the following do so with the use of the simplest form of odds ratios (Blaisdell 1986; Brendel et al. 1986; Stuckle et al. 1990, 1992), which ascribe a suitable statistical weight in the value of each observed frequency of occurrence through a simple division over expected values based on mononucleotide composition. In this way oligonucleotides with occurrence that exceeds or falls short of the one expected statistically may be detected. Oligonucleotides with extreme over- or under-representation in genomic sequences are primary candidates to be essential to the cell. Recent studies have been able to detect subsequences of regulatory function through the examination of their content in such over- or underrepresented oligonucleotide strings (Frith et al. 2004).

When measures of *n*-tuplet frequencies are summed up, a suppression of species-specific features in the *n*-tuplet distribution is found to occur, thus revealing other characteristics of the sequence, which can be correlated with its functional role. In a different approach, odds ratio-based measures of *n*-tuplet frequencies have been introduced in two-dimensional representations of sequences, thus modifying the original concept of "genomic portraits." In this way, existent patterns of over- or underrepresentations that are specific to coding or noncoding sequences may be revealed in a pictorial representation. As will be seen in the sequel, a quantification of such representations is also possible.

Furthermore, statistical attributes in the use and extent of preferred and avoided *n*-tuplets may be revealed through observation of the rank distributions of *n*-tuplet occurrence in a so-called Zipf plot. (Mantegna et al. 1994; Schmitt and Herzel 1997).

The methodology developed thus far is used in order to access the product of a proposed evolutionary model, consisting of biologically plausible events at the molecular level. It is pointed out that a simple genomic dynamics may account for the reproduction of all the essential compositional features, proper to real genomes discussed herein.

Sequences and Data Handling

Two sequence collections, representative of different functionalities, have been formed. Eukaryotic CDS and intronic sequence with mean length ~4000 nucleotides (~4×10^3 nt) were collected. These original collections were completely cleaned for redundancy, thus constituting a reliable reference set. Finally, each of the two was "trimmed" to a size of 500 sequences. A surrogate sequence collection was also formed using the following convention: For each intronic sequence a surrogate sequence was constructed with the same length and nucleotide constitution using a standard random number generator (Knuth 1981). A surrogate sequence collection based on nucleotide composition of CDSs was also formed. Its distributions are not depicted, as they were found to coincide with the ones obtained for the intronic surrogate collection.

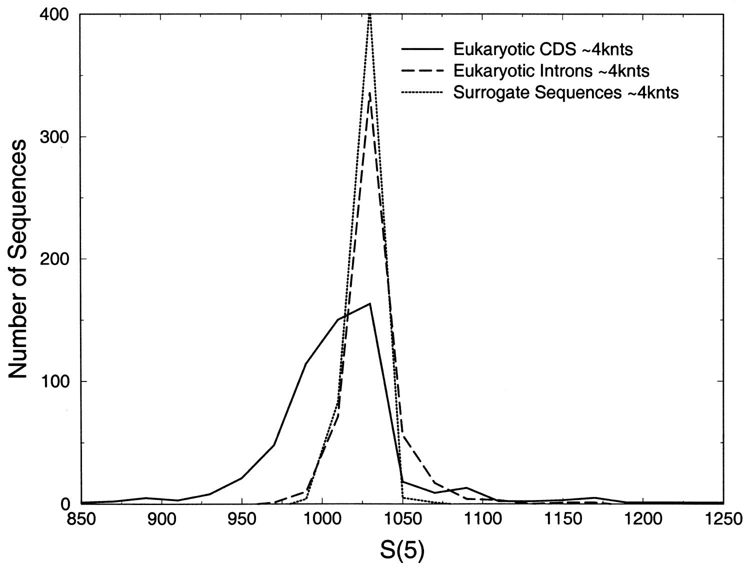

Fig. 1. $S^{(5)}$ value distribution curves for coding and intronic sequences from higher eukaryotes with mean length ~4000 nt. A random sequence collection of the same mean length is also drawn for comparison.

Results and Discussion

The Sum of Odds Ratio n*-tuplet Frequencies Distinguishes Weakly Between Coding and Noncoding DNA*

For the counting of simple frequencies of occurrence of n-tuplets a virtual reading head of length n was "sliding" through the sequence under examination, moving one nucleotide in each step. The formation of the frequency of occurrence measures and their incorporation in a summing-up quantity were carried out as follows.

(A) In a sequence with length L nucleotides the number of occurrences of each n-tuplet was counted, for n ranging from 1 to 7.

(B) In the next step, the simple frequency of occurrence of each one of the above n-tuplets was calculated by simple division over $L - n$ for each value of n. This provided seven matrices containing simple frequencies of occurrence: $f_i{}^{(1)}$, $f_{ij}{}^{(2)}$, $f_{ijk}{}^{(3)}$, etc., each of the indices $i, j, k\ldots$ denoting one of the four nucleotides (A, G, C, or T).

(C) In order to eliminate biases due to the background nucleotide composition, each frequency of occurrence $f_{ijk\ldots}{}^{(n)}$was divided over the product of the corresponding mononucleotide frequencies

$$R_{ijk\ldots q}{}^{(n)} = f_{ijk\ldots q}{}^{(n)}/f_i{}^{(1)}f_j{}^{(1)}f_k{}^{(1)}\ldots f_q{}^{(1)}$$

(D) Subsequently, the simple sum of all odds ratio frequencies was computed for each n:

$$S^{(n)} = \sum R_{ijk\ldots q}{}^{(n)} \quad (i,j,k,\ldots n = A,G,C,T)$$

Low values of n ($n = 2,3$) are found to give poor results in distinguishing coding and noncoding sequences, as they mostly provide information which is organism (origin)-correlated as has been shown in previous works (Gentles and Karlin 2001; Karlin and Mrazek 1997). Oligonucleotides with length >4 are more likely to bear less dependence on the origin of the sequence than on its functionality, as they remain, on one hand, dependent on the short scale grammar of the genetic code, but, on the other, less affected by special patterns of nearest neighbors or mutational tendencies. However, the observation of patterns in n-tuplet usage becomes incomplete for $n \geq 6$ or more, due to finite size effects, thus $n = 5$ is the optimal choice.

A graphical representation of the $S^{(5)}$-value distributions of the completely nonredundant reference set (eukaryotic CDS and Introns with mean length ~4000 nt) is depicted in Fig. 1. For both genomic sequence collections the distribution's maximum is located at the value expected under randomness (1024), coinciding with the clear peak of the very narrow-shaped surrogate sequences' distribution, also included in Fig. 1. The non-coding sequences' distribution is overlapping to a great extent with the random one, apart from a "tail" toward higher values. On the other hand, the coding sequence collection clearly deviates from randomness, with its most extended skew being toward the low $S^{(5)}$-value region. The CDS' distribution shape is obviously due to specific constraints imposed by the coding procedure as well as other forms of encoded information (Trifonov 1989). This leads to an overall significant deviation from randomness. The reason underlying the low $S^{(n)}$ value tendency of coding sequences is probably the avoidance of specific n-tuplets (words) imposed by the coding character of the sequences through the genetic code grammar. These avoidances are reflected upon underrepresented words, which shift the distribution to lower values. The codon usage skews in each organism have two major components. The one is the skew, met universally at various extents, toward the RNY (R, purine in 5′; N, any base in the middle; Y, pyrimidine in 3′prime) codon pattern (Crick et al. 1976; Eigen and Schuster 1977), while the other consists of organism dependent tendencies. Furthermore, *n*-tuplets containing stop codons are likely to be underrepresented (in some cases, even in the noncoding strand, see Yomo and Urabe [1994]). Notice that repetitive RNY motifs (...RNYRNYRNY...), when followed consistently, make *n*-tuplets including homo-purine/pyrimidine strings scarce in both strands. The above results may be seen as a generalization, in a form of an asymmetric distribution, of previous findings on the existence of concrete highly avoided "words" in coding sequences (Brendel et al. 1986; Hao 2000a, b).

A small but not insignificant percentage (~9%) of coding sequences presented high $S^{(5)}$-values do not abide by the aforementioned general tendency for underrepresentations. A careful investigation of the sequences responsible for this distribution's tail extending to the right revealed that they consist of genes coding for proteins either bearing tandem aminoacid repeats or being extremely rich in specific amino acids, often juxtaposed in homopolymeric residues. This is bound to lead to overrepresentation of the triplets encoding for such repeated patterns and consequently to overrepresentation of the corresponding 5-tuplets.

It is reasonable to suggest that the intronic distribution's long "tail" pointing toward high values is due to the aforementioned relative abundance of same or similar nucleotides in nocoding sequences (see also in the sequel).

Genomic "Portraits" Using Odds Ratio Frequencies

Having carried out a first approach to the distribution of *n*-tuplets in sequences of different functionality, we attempted to directly visualize the odds ratio *n*-tuplet usage pattern, modifying the concept of genomic "portraits" as introduced by Hao (2000a, b). In a genomic portrait, originally a whole sequence is depicted on a matrix of 4^n squares, each representing the frequency of occurrence of a specific *n*-tuplet of order *n*. The four corners of the square image are ascribed to the four nucleotides following the pattern $\begin{bmatrix} c & g \\ a & t \end{bmatrix}$. Then the frequency of each of the 4^n *n*-tuplets is assigned to the corresponding subsquare inside the "portrait" following the $\begin{bmatrix} c & g \\ a & t \end{bmatrix}$ pattern in a self-similar way as formulated by Jeffrey (1990). In this way the four corner subsquares represent the occurrences of the four homonucleotides (AA..., TT..., GG..., CC...), while the diagonal ones correspond to homopurine and homopyrimidine *n*-tuplets.

In their original form, genomic portraits incorporated simple frequencies of occurrence. We chose to use odds ratios in an attempt to "filter" species-specific skews of nucleotide constitution, thus trying to reveal patterns that are related to the sequences' functionality. In the following we present the results obtained for 5-tuplets. The coloring of the subsquares is related to the value which $R_{ijklm}{}^{(5)}$ takes for each *n*-tuplet, gradually changing from red to violet for increasing $R_{ijklm}{}^{(5)}$ (see legend to Fig. 2 for details). The discussion of Fig. 2c, which corresponds to a model generated sequence, is postponed to a later section.

Although in their original form, genomic "portraits" were used for either complete genomes or at least quite long sequences (of the order of 10^4 nucleotides) and each portrait represented one examined sequence, in Figs. 2a and b, we have chosen to present data from all 500 sequences of our CDS and intronic collections in a cumulative way. The value represented in each subsquare is the mean value obtained from the total of 500 sequences of each collection.

Noticeable differences in the color patterns between the two images were observed. In the CDS collection there is no clear over- or underrepresentation of a specific kind of *n*-tuplets. In the intronic collection, on the other hand, one can easily observe the overrepresentation of highly clustered *n*-tuplets. This means mostly *n*-tuplets consisting of strings of similar nucleotides (homopurine and homopyrimidine *n*-tuplets), as represented by the "cold" (blue and violet) colored corners and diagonals of the image (subsquares near the four corners represent *n*-tuplets rich in the corresponding nucleotide). The residual overrepresentation of homopurine and homopyrimidine *n*-tuplets visible in the coding portrait seems to have functional implications as indicated by several studies (Bucher and Yagil 1991; Raghavan et al. 2000). Comparison of the upper and lower sides (GC and AT rich tracts, respectively) shows them to be quite different in introns while remaining practically equivalent in CDSs. This would be expected, on the basis that, A + T-rich tracts have been observed to occur more frequently in introns, possibly due to replicational slippage that leads to their extension (Deschering et al. 1998).

Furthermore, a persistent feature of eukaryotes, the underrepresentation of the dinucleotide CpG, appears in the image of the intronic collection, in a self-similar manner. Subdomains of the image that correspond to oligonucleotides containing CpG all present the particular "warm" (red, orange) color pattern. Additionally, in both figures there is a moderate underrepresentation of TpA-containing pentanucleotides.

The above observations agree with results from previous contributions in Chaos Game Representa-

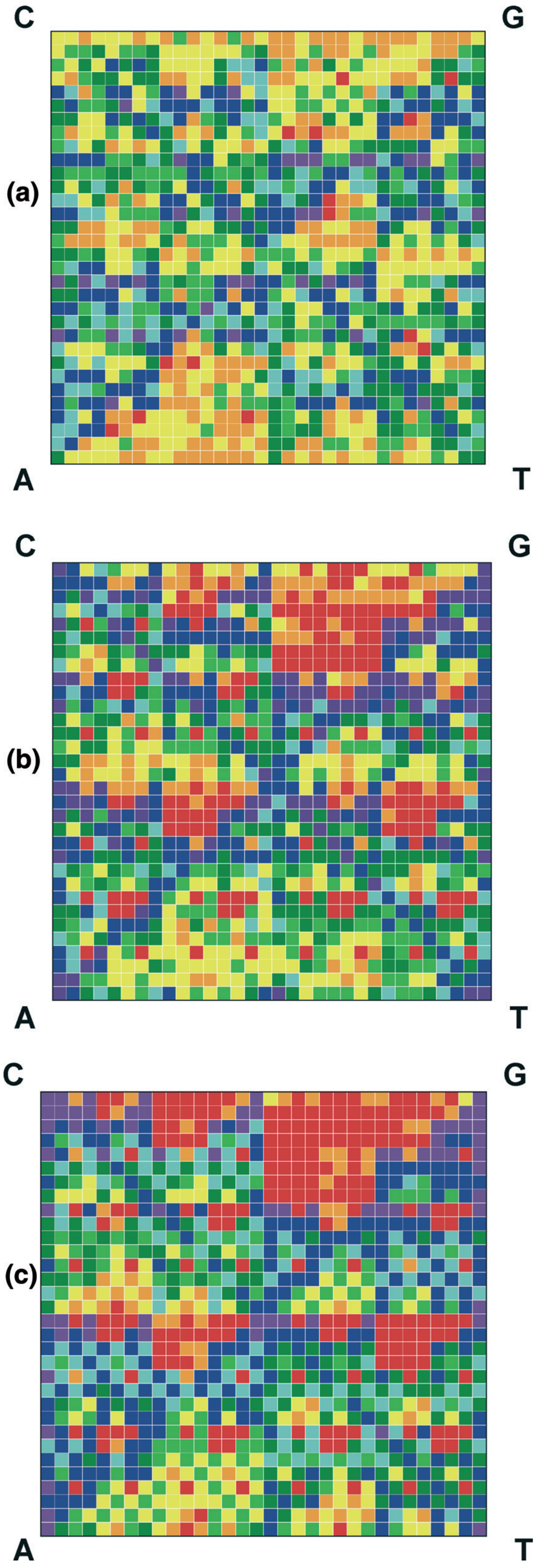

Fig. 2. "Genomic portraits" of (**a**) 500 coding sequences from higher eukaryotes with mean length ~4000 nt, (**b**) 500 intronic sequences of the same origin and mean length, and (**c**) model-generated sequence. Warm colors (red, orange, yellow, light green) correspond to underrepresented 5-tuplets, while cold ones (green, cyan, blue, violet) indicate over represented 5-tuplets with increasing values of odds ratios: 0 ≤ red < 0.4 ≤ orange < 0.6 ≤ yellow < 0.8 ≤ light green < 1.0 ≤ dark green < 1.2 ≤ cyan < 1.3 ≤ blue < 1.7 ≤ violet.

tions and genomic portraits. Nevertheless, even recent studies on this matter (Deschavanne et al. 1999), miss the aspect of the strong dependence of the resulting image on the functionality of the examined sequence, which is apparent here. Eubacterial genomes are expected to produce more coding-like images because of the high density of coding sequences. Chromosomes from higher eukaryotes, on the contrary, will be depicted in images resembling the noncoding structure, with expressed diagonal and corner patterns (see Figs. 3a and 4 in Deschavanne et al. 1999).

An essential difference in our approach lies in the use of the simple "filter" of odds ratios on the background nucleotide composition, which promotes the emergence of patterns that are correlated with the sequence functionality and not its special origin. Notice that figures corresponding to 2a and b, but formed using simple frequencies, do not present the aforementioned visual pattern of CpG underrepresentation and similar nucleotide clustering to the same degree and are therefore omitted herein.

On the other hand, the use of odds ratios based on Markov models of higher order in the genomic portraits leads to the disappearance of the observed patterns. This makes the use of odds ratios over theoretical values corresponding to the mononucleotide constitution the best choice in the present approach. That may be explained by the fact that homo-purinic/pyrimidinic clustering is not specific to a particular length (e.g., $n = 5$) but it appears in all cluster sizes. The odds ratios, which are based on higher-order Markov models, would thus smooth out the X-shaped pattern. The same holds for the "fractal" pattern proper to dinucleotide scarceness. A similar observation was made in a previous work (Nikolaou and Almirantis 2003, Table 2, columns B and C), where quantification of coding/noncoding distinction was optimal when based on a zeroth-order Markov process.

Quantifying the Differences in the Genomic Portrait Patterns Between Coding and Noncoding Sequences

As discussed, one of the main oligonucleotide usage patterns observed in Figs. 2a and b is the tendency for clustered n-tuplets in noncoding DNA. Our aim was to go further by using this property for a clearer distinction between sequences of coding and noncoding DNA. This was achieved using highly clustered n-tuplets as indices of an assumed noncoding character. A new quantity. $S_c^{(n)}$, is defined in an appropriate way, which is described as follows.

A). For each sequence, all 1024 odds ratios for 5-tuplets were calculated and then two subsets

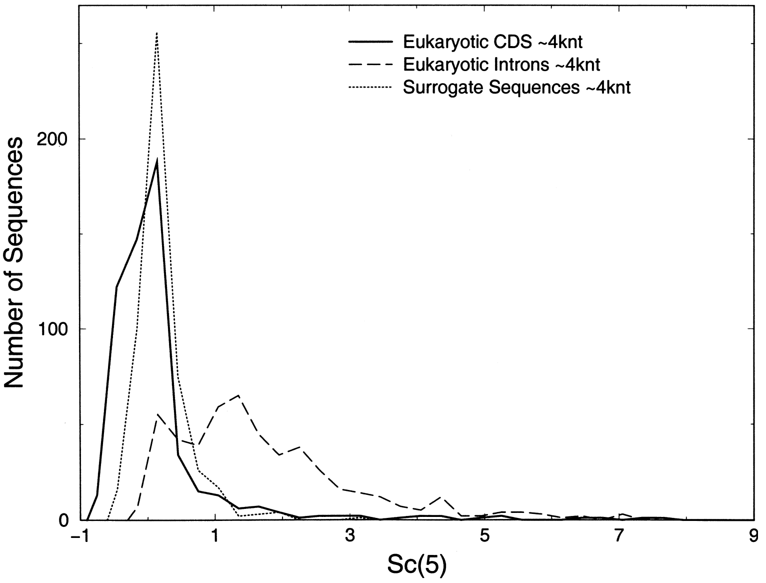

Fig. 3. $Sc^{(5)}$ value distribution curves for coding and intronic sequences from higher eukaryotes with mean length ~4000 nt. A random sequence collection of the same mean length is also drawn for comparison.

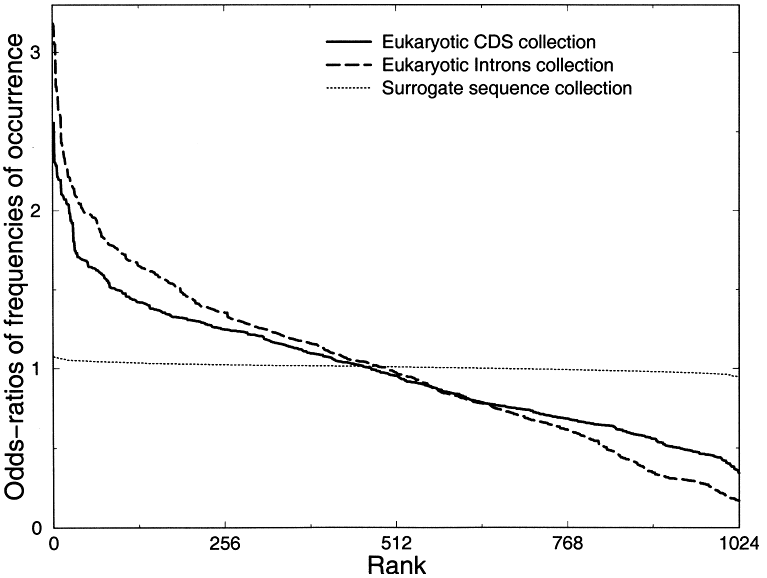

Fig. 4. Rank distribution diagram (using odds ratios) for 5-tuplets. Five hundred-sequence collections of CDSs and introns from higher eukaryotes with mean length ~4000 nt and random sequence collection are included.

considered. One includes the four completely homologous strings (AAAAA, GGGGG, CCCCC, TTTTT). The mean of these values was considered the high-clustering component $C_h^{(5)}$. The mean value of the nondiagonal 960 odds ratio frequencies represented the low-clustering component $C_l^{(5)}$. The 60 remaining diagonal n-tuplet values constituting homopurine and homopyrimidine clusters were not used.

(B). Next, the result of the subtraction $S_c^{(5)} = C_h^{(5)} - C_l^{(5)}$ was computed. The observed overabundance of same-nucleotide strings may be principally attributed to replication slippage favoring the extension of short repetitive DNA, particularly in the noncoding.

A graphic representation of $S_c^{(5)}$-value distributions for the two completely non-redundant reference collections alongside the surrogate sequence collection is depicted in Fig. 3. One can notice that, regarding the particular statistical feature of the abundance of homonucleotide clusters, it is the coding sequences that now appear to be near-random, in contrast to noncoding genomic segments, which tend to deviate from random behavior. Such a “transition” of random behavior under a different prospect (compare with $S^{(n)}$ in Fig. 1) is indicative of the complexity in the organization of genomic sequences in general. Both quantities $S_c^{(5)}$ and $S^{(n)}$ serve to illustrate qualitatively statistical features of genomic sequences.

Rank Distributions of n-tuplet Measures of Frequency

Several analogies have been drawn between natural languages and the genomic text. Noncoding genomic sequences have been shown to follow to some extent a kind of Zipf's law, meaning to tend to produce a linear slope, when plotting the values of n-tuplet frequencies against their rank in a double logarithmic plot. Mantegna et al. (1994) have pointed out the existence of differences in the rank distributions between coding and noncoding sequences, using simple frequencies of occurrence, while Schmitt and Herzel (1997) assessed the entropy of genomic sequences starting from a similar observation. Figure 4 represents rank distributions of 5-tuplet odds ratio frequencies obtained from the eukaryotic CDS and intronic sequence collections. For each collection, all sequences are concatenated and the rank distribution is formed according to the total counts of odds ratio 5-tuplet frequencies. The distribution curves are plotted on a linear scale, alongside the curve corresponding to the surrogate (random) sequence collection used above.

Overall, the deviation from random behavior is greater in the case of the noncoding sequences. The noncoding sequences exhibit more extreme over-and underrepresentations, since the corresponding curve's absolute distance from the random sequences' distribution is always greater than that of coding sequences. Notice that the region above unity is corresponding to overrepresentation, while the one below it is indicative of underrepresented n-tuplets. Thus, noncoding sequences are more abundant in both over- and under represented n-tuplets.

Attempting quantification of the distance between the two rank distributions, we apply the Spearman rank-order correlation coefficient (r_s) (see Press et al. 1986 and references therein). Rank-order correlation is a nonparametric statistics particularly suitable in cases where the type of distribution is not known. r_s is calculated via the formula: $r_s = 1 - \left[6D/\left(N^3 - N\right)\right]$, where D is the sum squared difference of ranks, defined as: $D = \sum_{i=1}^{N} (R_i - S_i)^2$and $N = 1024$. r_s equals 0.726 for the rank distributions in Fig. 4, with absolute correlation corresponding to $r_s = 1$ and absence of correlation corresponding to $r_\mathrm{s} = 0$. For reasons of comparison, we calculated interspecies values of r_s. The largest contigs from human chromosome 1 (H.sap) mouse chromosome 5 (M.mus) and chicken chromosome 1 (G.gal) were taken into consideration. We obtained $r_{s\,\mathrm{H.sap-M.mus}} = 0.970$, $r_{s\,\mathrm{H.sap-G.gal}} = 0.956$, and $r_{s\,\mathrm{G.gal-M.mus}} = 0.958$.

In Fig. 4, the over represented n-tuplets lie in the part of the curve located above the random one, while the part below it is due to underepresentation. Measurement of the corresponding areas may assess the impact of each of these regions. For the coding sequence, the part with the greatest area is the one corresponding to the underrepresented n-tuplets. The opposite holds for the noncoding sequences. This may serve as an explanation for the shift of coding sequences to lower $S^{(5)}$ values observed in Fig. 1, since the contribution of low, underrepresented values is greater than the corresponding, derived from high, overrepresented ones.

The above observations show that there are two distinct modes of usage of n-tuplets corresponding to coding and noncoding sequences. Avoidance of specific words is a dominant characteristic of sequences coding for proteins, as we already discussed in relation to RNY pattern, stop codon scarceness, and features specific to the encoded protein. The abundance of same- or similar-nucleotide clusters in noncoding has been widely reported (see cited references) and is also demonstrated in the genomic portrait image presented in Fig. 2b. This particular abundance in a specific type of n-tuplets is probably the main reason for the overrepresentation pattern of the noncoding sequences. Regarding the underrepresented n-tuplets in non-coding sequences, there is a clear avoidance of n-tuplets containing the two most underrepresented dinucleotides in the human genome, namely, CpG and TpA (Karlin and Ladunga 1994; Karlin et al. 1994. Karlin and Mrazek 1997; Gentles and Karlin 2001), which is, again, in agreement with the very expressed underrepresentation of CpG and TpA containing n-tuplets shown in Fig. 2b. Although these are general genomic features, at least CpG scarceness remains functionality correlated, as it appears to be twice as pronounced in the noncoding, most probably because of the lack of coding constraints. In parallel with the above results comes the finding of long-rangeness in the distributions of n-tuplets rich in these two dinucleotides in long human genomic contigs (Katsaloulis et al. 2002). As those authors comment, these features may be related to the abundance of CpG and TpA in regulatory consensus sequences and, consequently, in a scarceness of them in the rest of the genome.

An Evolutionary Scenario Which May Account for the Observed Statistical Properties of Genomic Sequences

In previous work we have put forward an evolutionary scenario having as the principal goal to reproduce quantitatively the high clustering of similar nucleotides at several length scales, occurring principally in the noncoding, using a suitable quantity (the modified standard devation; MSD[m] [Almirantis 1999; Nikolaou and Almirantis 2002]) as the measure of this genomic property. That scenario included the incorporation into an initial large random sequence of other random sequences (in general, with different nucleotide constitutions), supposed to model intra- or interchromosomal or even interorganismic transfer of genomic material. After every such genomic-fusion event, extensive shuffling occurred through transpositions at random (Almirantis 1999). Moreover, it has been pointed out (Provata 1999) that fusion and mixing events of segments characterized by short-range size distributions finally engender sequences with long-range properties, a feature shared by the noncoding segment size distributions of higher eukaryotic genomes (Almirantis and Provata 1999; Provata and Almirantis 2000).

Here we propose an extension of that scenario, which, in addition to the above events, includes (i) extension of same-nucleotide tracts, mimicking replication slippage known to favor the extension of short repetitive DNA, particularly in the noncoding; (ii) point mutations, especially transitions; and (iii) suppression of the dinucleotides TpA and CpG, in order to reproduce their natural underrepresentation. These types of events occur mostly simultaneously in real time during the evolutionary genome dynamics. Here, for reasons of simplicity, we apply them consequently to the initial random sequences and then we study the properties of the resulting "artificial genome."

Notice that in the following, sequence and symbol operations modeling the real genomic events are kept as simple as possible and the same holds for the parameters used (genomic constitutions, probabilities of a homonucleotide tract doubling its length, etc.). Our main purpose is not to reproduce the exact features of any particular genome or to follow closely the (very complicated and in many respects still unclear) detailed dynamics of real genomes in evolutionary time. We just aim to point out that combination of these types of simple operations suffices to make several main quantitative aspects of genomic statistics appear.

The artificial sequence was modeled in the following way.

(A) **Genomic fusion.** Two subsequences, with length 10^6 nt each, were constructed using a random number generator for the juxtaposition of nucleotides: one AT-rich, with A = 40%. T = 40%. G = 10%. C = 10% (designated **A1**) and one GC-rich, with A = 10%. T = 10%. G = 40%. C = 40% (**A2**). Next, the two subsequences were concatenated, simulating an event of genomic fusion.

(B) **Genome Shuffling.** Extensive shuffling of the chimeric sequence, through a procedure which simulated genomic transposition, followed. A shuffling event included the cutting of a segment with length chosen at random in a range between 500 and 1000 nucleotides and then randomly reinserting it in a different position. In total 150,000 such events took place, for a sequence of length 2×10^6 nt.

(C) **Replication slippage.** In the next step, homonucleotide strings were extended in a way simulating strand slippage during replication. Strings of identical nucleotides were doubled in length in a percentage depending on their initial length. Starting from 4-plets, they were expanded to 8-plets in a percentage of 8%, 5-plets were expanded to 10-plets in a percentage of 10%, and so on, up to 10-plets, which were expanded to 20-plets in 20%. This would be an upper boundary, as no homonucleotide strings longer than 10 bases were found in the initial sequence. See the appendix for a discussion about the choice of these parameter values.

(D) **Point Mutations.** The next step included the occurrence of point mutations in the resulting sequence. The determination of the ratio transitions/transversions (ti/tv) has been the subject of several works (see, e.g., Yang and Yoder 1999). We have tried several ti/tv values ranging between 50 and 2, but the results remain qualitatively the same. So only transitions were finally implemented in the evolutionary scenario presented here, for the sake of simplicity and maintenance of the minimal prerequisites. All possible transitions were carried out at random points, with the same probability equal to 0.25 each. The choice of this probability value is justified in the Appendix.

(E and F) **CpG and TpA depletion.** In the final step, directed mutagenesis was incorporated in our model. It is well known that the two most underabundant dinucleotides in eukaryotic genomes are TpA and CpG, the latter being the scarcer of the two. The simulation of mutational processes leading to the depletion of CpG and TpA was carried out in the following way. Initially CpG dinucleotides were mutated to TpA with a probability of 0.95. In the next step, TpA was mutated to all possible dinucleotides with differential probabilities. In particular, TpA was changed to TpT and ApA, being among the most abundant dinucleotides in eukaryotic genomes, with a probability of 0.1 each. The set of the remaining four dinucleotides needing alteration in only one position to result in TpA (namely, TpC, TpG, CpA, and GpA) was obtained with a probability of 0.08 each. Finally, the remaining eight dinucleotides apart from CpG (ApG, ApC, ApT, GpC, GpT, GpG, CpT, and CpC) were incremented with a probability of 0.02 each. The above choices of mutation probabilities were fine-tuned in order to produce odds ratio values of TpA and CpG met in the human genome (see Gentles and Karlin 2001).

The formation of two subsequences with clear abundance in a pair of complementary nucleotides each was carried out to simulate exchange in genomic material between genomes, or genomic regions, which radically differ in their nucleotide constitution. Extreme abundances in A + T or G + C are still observed in bacteria and other organisms (see, e.g., for *Plasmodium falciparum*, Dechering et al. [1998]), while "isochores," regions of distinct G + C content with sharp boundaries are common feature of the genomes of vertebrates. Moreover, the sequences were conformed, with the Chargaff second parity rule (Chargaff 1951; Lin and Chargaff 1967). As we discuss in more detail elsewhere (Almirantis 1999; Almirantis and Provata 2001), there is evidence of both interspecies and intragenomic (from distant places having different constitutions) occurrences of large-scale mixing, which is mimicked in our model by the "genomic fusion" event. Moreover, several causes of maintenance of constitutional divergence may exist in genomes placed under the generic term "ecology of the noncoding DNA" by Holmquist (1989). These, as yet unresolved, mechanisms generate and conserve large-scale genomic inhomogeneities, with the most prominent the isochore structure present in several taxa (see, e.g., Bernardi 1989, 1993), shorter-scale purine/pyrimidine clustering (Peng et al. 1992), and several other similar genomic features.

Transpositions and replication–insertion events are also a major component of genomic dynamics. Here, without loss of generality, we include only simple transpositions. Variations of the presented algorithm, including copying and random insertion, give qualitatively the same results.

Many authors have investigated the implications for the genome structure of the supposed ubiquitous replication slippage events, in a very extended range of organisms (see, e.g., Hancock 1993; Tautz et al. 1986). Slippage, while able for both contraction and expansion of repeated sequences, has a net result of expansion for moderate length repeats. In some cases this is found to be due to reduced proofreading efficiency in expanding events compared to deletion ones (Gragg et al. 2002). Thus we have modeled the complicated realistic slippage only with length doubling steps, and we have restricted it only to tracts of homonucleotides. Notice that the longer a cluster of tandem repeats (or a homonucleotide tract), the more probable the slippage event (Gragg et al. 2002). The choice to allow duplications for $n \geq 4$ is based on the observation (Dechering et al. 1998) that homopolymer AT-tracts occur with frequencies equal to the expected for $n < 8$–10, while they are overrepresented, in the noncoding, for higher values. It is extensively discussed in the literature that slippage combined with point mutations plays a key role in the stabilization of large amounts of "simple" DNA in genomes and as a source of new genomic material (see, e.g. Harr et al. [2000] and, for an earlier discussion, Tautz et al. [1986]).

The ability of the model to reproduce the statistical features of natural sequences was monitored through rank distribution diagrams and genomic portraits.

In Fig. 5 the rank distribution diagrams of the sequences generated after each step (A–F) of the proposed evolutionary model are presented alongside the corresponding curve obtained from eukaryotic intronic sequences. The inclusion of a random sequence would produce an almost-perfect straight line at 1 (see Fig. 4).

Curve **A1** in Fig. 5 corresponds to one of the two concatenated random sequences before shuffling. While nucleotides are not equiprobable, oligonucleotides are still present at frequencies that are in conformity with randomness. Thus, the calculated odds-ratios for pentanucleotides fall around the expected value of 1. Notice that the corresponding rank distribution obtained from simple frequencies of occurrence for a nonequiprobable random sequence would have produced a step-like curve, spanning a very wide range of values from extreme overrepresentation to extreme underrepresentation, as illustrated for artificial sequences constructed through markovian processes (Mantegna et al. 1994).

Sequence **B** in Fig. 5, corresponding to the shuffled, concatenated random sequences exhibits a clear step-like pattern, which is an indication of its distance from stationarity, meaning that frequencies of individual pentanucleotides do not conform to their expected values, based on sequence constitution.

The following operations of the model make the resulting artificial sequences to gradually approach the curve of the concatenated genomic (intronic) sequences. The artificial sequence produced by the proposed evolutionary model appears to share the characteristics of natural sequences to a great extent. Its range of values is very close to those of the presented intronic collection, and, also, to those of other noncoding sequences not presented in this figure, while its slope lacks the discontinuities of the shuffled sequence, apart from a slight drop in the region of underepresentations. However, this particular drop is also a characteristic of the distribution of the intronic sequence collection and is due to the marked underrepresentation of CpG dinucleotides.

The proximity of the rank profiles of the model-generated and the genomic sequences has its origin in the same- and similar-nucleotide clustering produced by the simulation events of mixing of initially heterogeneous parts and the subsequent extensive shuffling. Clustering is further enforced by the simulation of slippage-like amplification of same-nucleotide tracts in combination with point mutations.

The genomic portraits as described in the previous section are used as a suitable way to test the degree of resemblance between model-generated and genomic

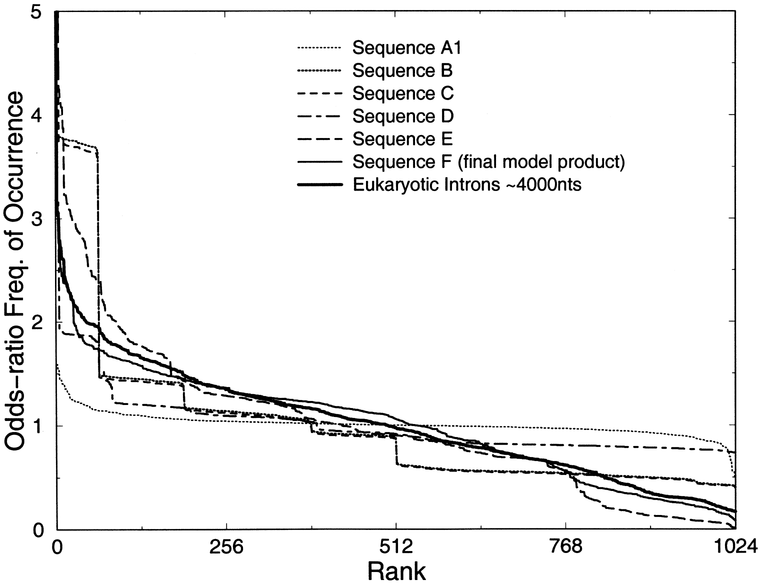

Fig. 5. Rank distribution diagram (using odds ratios) for 5-plets. Sequences **A1** to **F** represent products of the successive steps of the proposed model. The concatenation of eukaryotic intronic sequences is also included. The model-generated sequences gradually approach the genomic one.

sequences, with respect to a detailed analysis of *n*-tuplet occurrences. Odds ratios of 5-plet frequencies calculated for the artificial sequence are depicted in the genomic portrait in Fig. 2c, with the same coloring convention that was used for the CDS and intronic sequence collections in Figs. 2a and b. One can see a marked similarity of the artificial sequence to the intronic ones. The self-similar pattern of CpG underrepresentation is obvious in both the noncoding and the artificial sequence. Moreover, the characteristic X-shaped overrepresentation pattern of the diagonals is also apparent. This particular pattern, attributed to the overabundance of purine and pyrimidine nucleotide clusters, has arisen in the artificial sequence as a result of the combined procedures of replication slippage, expanding same nucleotide tracts, and point transitions, interrupting such tracts. The incidence of point mutations into repeated elements, expanded through slippage, alongside their scattering through transposition throughout the genome, has been proposed as procedures stabilizing genomic evolution (Harr et al. 2000; Kruglyak et al. 2000; Lovett 2004).

In Fig. 6, we present, in a quantitative way, the gradual convergence of the model-generated sequences (**A1**–**F**) to the genomic ones, through Spearman rank ($1 - r_s$) and δ-distance as defined by Karlin et al. (1994) and Karlin and Mrazek (1997) (regarding 5-plets herein). The Spearman rank correlation coefficient quantifies the distances between rank distributions visualized by Fig. 5, while δ-distance serves as a measure of the similarity between two genomic portraits (Figs. 2b and c). The product of these two quantities is additionally used as an overall measure of sequence convergence between the model-generated and the genomic ones. Notice that in Fig. 6, the initial and final values are of greater significance, as the separation of the individual events mimicking molecular dynamics is conventional (not occurring in true evolution). This separation was adopted in order to monitor the impact of each step of the whole procedure.

Concluding Remarks

It is well known that *n*-tuplet usage is strongly correlated with the sequence's origin. Nevertheless, compositional constraints differ between coding and noncoding sequences and are, therefore, expected to be expressed through different patterns of *n*-tuplet usage distributions. The approaches described above are able to capture such differences, which remain species-independent.

Patterns of *n*-tuplet usage in coding sequences are mainly affected by the protein-encoding procedure, being of great diversity and reflecting the wealth of protein structures and functions, exhibited in any genome. On the other hand, noncoding *n*-tuplet usage patterns reflect, more directly, the overall genomic dynamics, with the same- or similar-nucleotide clustering at several length scales being the most recognizable feature. Under this prospect, coding and noncoding sequences are shown to behave in a random or a nonrandom manner alternatively, depending on the compositional aspect examined.

The evolutionary scenario formulated herein aims at demonstrating that, in principle, all the above modalities in the *n*-tuplet occurrence patterns may be viewed as expressions of a simple genomic dynamics including the most typical forms of events at the molecular level. Attempts have been carried out by several authors (some references have been

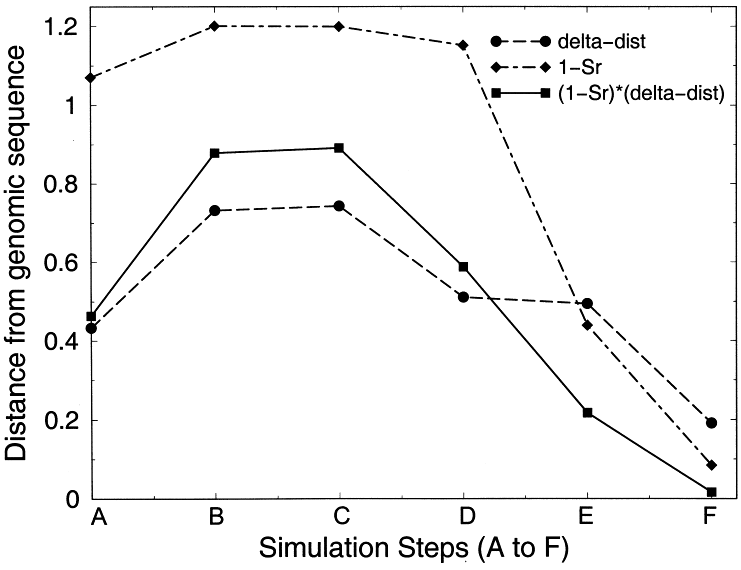

Fig. 6. Quantitative monitoring of the simulation process, using 1–Spearman rank correlation coefficient and δ-distance, as well as their product.

included above), aiming at the elucidation of specific types of genomic dynamics: replication slippage, recombination dynamics, transposition dynamics, origin of isochore structure, etc. Most of these processes of genome evolution are not tackled exhaustively in the presented framework, in order to avoid a detailed construction with an excess of *ad hoc* hypotheses. Here we have focused on pointing out that information derived from the study of n-tuplet usage may serve for a better understanding of both the common properties of coding and non-coding sequences, attributed through evolution, and their differential characteristics that have been shaped through the divergence imposed by differences in function.

Appendix: Estimation of Optimal Parameter Values for the Proposed Model

Here we have focused in some detail on quantitative aspects of the proposed evolutionary model. First, let us notice that the fused sequences are both products of a random symbol generator. Each is characterized by A = T and G = C, while (without loss of generality) for the result of their mixing, A = T = G = C holds. The order of magnitude for the number of shuffling events is related to the results of our previous works, where such a hierarchy of events in an (initially) random sequence is shown to generate nucleotide clustering similar to the one of genomic sequences (Almirantis, 1999; Nikolaou and Almirantis, 2002).

As already discussed, slippage events are reported to occur more often in longer homonucleotide clusters. For the probability of a slippage event on a homonucleotide cluster of length n, we adopt the form $P_n = l(n/10)$, where l is considered a free model parameter to be determined. The search for the optimal value for l is carried out through examination of the difference between the $S_c^{(5)}$ values of model-generated (for several values of l) and a typical genomic sequence. As genomic sequence is taken the concatenation of the 500 introns of the collection used throughout this work. $S_c^{(5)}$ was chosen for this evaluation, as it reflects the relative difference in the occurrence of homonucleotides and other nonclustered pentanucleotides. From Fig. 7 it is obvious that $l = 0.2$ is an optimum (zero distance from natural sequence).

In the next step of the model in its simplified form, point mutations occur, only in the form of transitions. The determination of the optimal value for the transition rate is illustrated in Fig. 8. The quantity measuring the difference between the sequence - product of the model for various rates and the genomic sequence is $d = \bar{D} - \bar{C}$, where $\bar{D}$ is the mean value of the frequencies of occurrence values of the 60 nonhomopolymeric, homopurinic, and homopyrimidinic and $\bar{C}$ is the mean value of the frequencies of occurrence of the four homopolymeric pentanucleotides (AAAAA, TTTTT, GGGGG, CCCCC). This measure is suitable for our purposes, as it quantifies the degree of formation of the X-shaped pattern in the genomic portraits. As depicted in Fig. 8, the optimal value (zero distance from natural sequence) of the transition rate is close to 0.25.

As mentioned earlier, the parameter values for the CpG and TpA depletion have been selected under the constraint of reaching the relative dinucleotide frequencies of occurrence in the human genome.

An exhaustive examination of the model's robustness has not been included herein. However, parameter values chosen in a wide, nonetheless

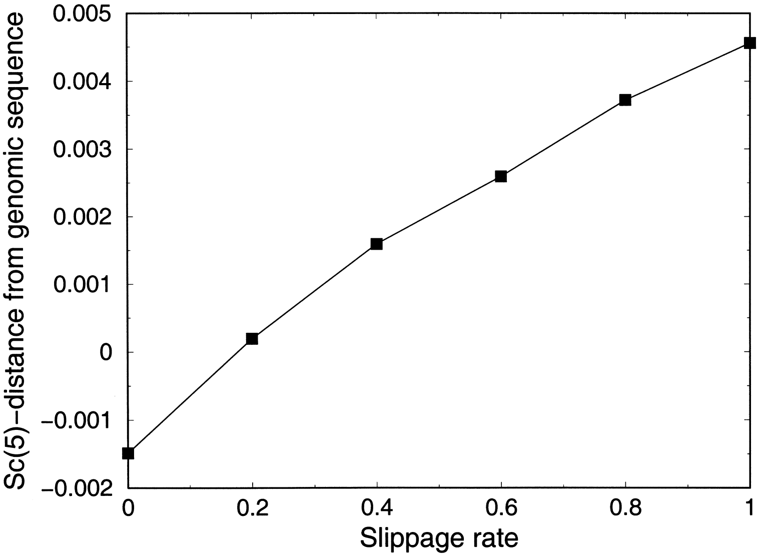

Fig. 7. Optimization of the slippage rate free parameter *l*, through distance of the $S_c(5)$ value between genomic and model-produced sequences.

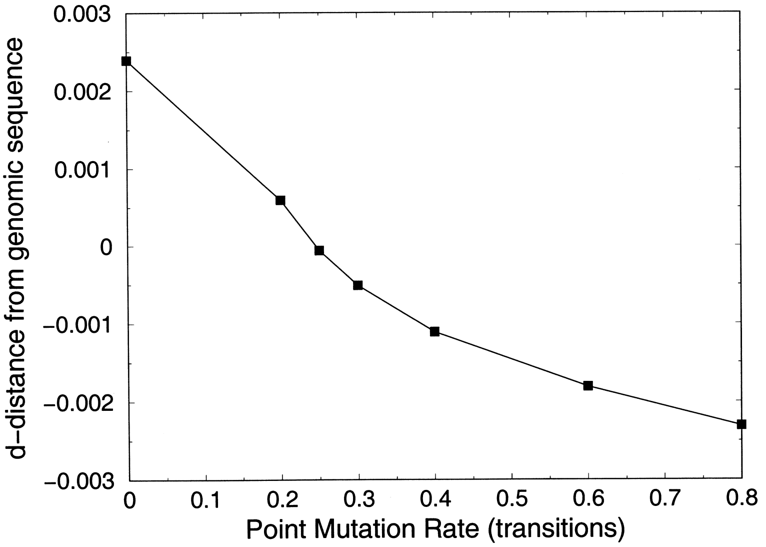

Fig. 8. Optimization of the mutation rate (only transitions considered), through distance of the *d*-distance value between genomic and model-produced sequences.

biologically plausible, range lead to model-generated sequences whose qualitative properties remain close to those observed in genomic sequences.

References

Almirantis Y (1999) A standard deviation based quantification differentiates coding from noncoding DNA sequences and gives insight to their evolutionary history. J Theor Biol 196: 297–308

Almirantis Y, Nicolaou C (2005) Multi-criterial coding sequence prediction. Combination of GeneMark with two novel, coding-character specific quantities. Comput Biol Med 35:627–643

Almirantis Y, Provata A (1997) The "clustered structure" of the purines/pyrimidines distribution in DNA distinguishes systematically between coding and noncoding sequences. Bull Math Biol 59:975–992

Almirantis Y, Provata A (1999) Long- and short-range correlations in genome organisation. J Stat Phys 97:233–239

Almirantis Y, Provata A (2001) An evolutionary model about the origin of non-randomness, long-range order and fractality in the genome. Bioessays 23:647–656

Bernardi G (1989) The isochore organization of the human genome. Annu Rev Genet 23:637–661

Bernardi G (1993) The isochore organization of the human genome and its evolutionary history—A review. Gene 135:57–66

Blaisdell BE (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. Proc Natl Acad Sci USA 83:5155–5159

Brendel V, Beckmann JS, Trifonov EN (1986) Linguistics of nucleotide sequences: morphology and comparison of vocabularies. J Biomol Struct Dyn 4:11–21

Bucher P, Yagil G (1991) Occurrence of oligopurine. oligopyrimidine tracts in eukaryotic and prokaryotic genes. DNA Seq 1:157–172

Burge C, Campbell AM, Karlin S (1992) Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci USA 89:1358–1362

Chargaff E (1951) Structure and function of nucleic acids and mechanism of their enzymic degradation. Experientia 6:201–209

Crick FH, Brenner S, Klug A, Pieczenik G (1976) A speculation on the origin of protein synthesis. Orig Life 7:389–397

Dechering KJ, Cuelenaere K, Konings RN, Leunissen JA (1998) Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. Nucleic Acids Res 26:4056–4062

Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol Biol Evol 16:1391–1399

Eigen M, Schuster P (1977) The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. Naturwissenschaften 60:541–565

Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z (2004) Detection of functional DNA motifs via statistical over-representation. Nucleic Acids 32:1372–1381

Gentles AJ, Karlin S (2001) Genome-scale compositional comparisons in peukaryotes. Gen Res 11:540–546

Goldman N (1993) Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. Nucleic Acids Res. 21:2487–2491

Gragg H, Harfe BD, Jinks-Robertson S (2002) Base composition of mononucleotide runs affects DNA polymerase slippage and removal of frameshift intermediates by mismatch repair in *Saccharomyces cerevisiae*. Mol Cell Biol 24:8756–8762

Hao BL (2000) Fractals from genomes. Modern Phys Lett B 14:871–875

Hao BL (2000) Fractals from genomes—Exact solutions of a biology-inspired problem. Physica A 282:225–246

Hancock JM (1993) Evolution of sequence repetition and gene duplications in the TATA-binding protein TBP (TFIID). Nucleic Acids Res 21:2823–2830

Harr B, Zangerl B, Schlotterer C (2000) Removal of microsatellite interruptions by DNA replication slippage: phylogenetic evidence from Drosophila. Mol Biol Evol 7:1001–1009

Holmquist GP (1989) Evolution of chromosome bands: molecular ecology of noncoding DNA. J Mol Evol 28:469–486

Jeffrey HJ (1990) Chaos game representation of gene structure. Nucleic Acids Res 18:2163–2170

Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11:283–290

Karlin S, Ladunga I (1994) Comparisons of eukaryotic genomic sequences. Proc Natl Acad Sci USA 91:12832–12836

Karlin S, Mrazek J (1997) Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci USA 94:10227–10232

Karlin S, Ladunga I, Blaisdell BE (1994) Heterogeneity of genomes: measures and values. Proc Natl Acad Sci USA 91:12837–12841

Katsaloulis P, Theoharis T, Provata A (2002) Statistical distribution of oligonucleotide combinations: applications in human chromosomes 21 and 22. Physica A 316:380–396

Knuth DE (1981) The art of computer programming. Addison–Wesley, Chicago

Kruglyak S, Durrett R, Schug MD, Aquadro CF (2000) Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations. Mol Biol Evol 8:1210–1219

Li WH (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Lin HJ, Chargaff E (1967) On the denaturation of deoxyribonucleic acid. II. Effects of concentration. Biochim Biophys Acta 145:398–409

Lovett ST (2004) Encoded errors: mutations and rearrangements mediated by misalignment at repetitive DNA sequences. Mol Microbiol 5:1243–1253

Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng CK, Simons M, Stanley HE (1994) Linguistic features of noncoding DNA sequences. Phys Rev Lett 73: 3169–3172

Nakamura Y, Wada K, Wada Y, Doi H, Kanaya S, Gojobori T, Ikemura T (1996) Codon usage tabulated from the international DNA sequence databases. Nucleic Acids Res 24:214–215

Nikolaou C, Almirantis Y (2002) A study of the middle-scale nucleotide clustering in DNA sequences of various origin and functionality by means of a method based on a modified standard deviation. J Theor Biol 217:479–942

Nicolaou C, Almirantis Y (2003) Mutually symmetric and complementary triplets: differences in their use distinguish systematically between coding and non-coding genomic sequences. J Theor Biol 223:477–487

Nicolaou C, Almirantis Y (2004) Measuring the coding potential of genomic sequences through a combination of triplet occurrence patterns and RNY preference. J Mol Evol 59:309–316

Nussinov R (1981) Eukaryotic dinucleotide preference rules and their implications for degenerate codon usage. J Mol Biol 149:125–131

Peng C-K, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, Stanley HE (1992) Long range correlations in nucleotide sequences. Nature 356:168–170

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1986) Numerical recipies—The art of scientific computing. Cambridge University Press, Cambridge

Provata A (1999) Random aggregation models for the formation and evolution of coding and non-coding DNA. Physica A 264:570–580

Provata A, Almirantis Y (2000) Cantor fractal properties of DNA sequences. Fractals 8:15–27

Qi J, Wang B, Hao B-L (2004) Whole proteome prokaryote phylogeny without sequence alignment: A k-string composition approach. J Mol Evol 58:1–11

Raghavan S, Hariharan R, Brahmachari SK (2000) Polypurine polypyrimidine sequences in complete bacterial genomes: preference for polypurines in protein-coding regions. Gene 242:275–283

Schmitt AO, Herzel H (1997) Estimating the entropy of DNA sequences. J Theor Biol 188:369–377

Stuckle EE, Emmrich C, Grob U, Nielsen PJ (1990) Statistical analysis of nucleotide sequences. Nucleic Acids Res 18:6641–6647

Stuckle EE, Nielsen PJ, Grob U (1992) Probability of occurrence of specific oligomers. J Theor Biol 159:299–306

Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322:652–656

Trifonov EN (1989) The multiple codes of nucleotide sequences. Bull Math Biol 51:417–432

Yang Z, Yoder AD (1999) Estimation of the transition/transversion rate bias and species sampling. J Mol Evol 48:274–283

Yomo T, Urabe I (1994) A frame-specific symmetry of complementary strands of DNA suggests the existence of genes on the antisense strand. J Mol Evol 38:113–120

Zuckerkandl E (1992) Revisiting junk DNA. J Mol Evol 34:259–271