Flexibility of the genetic code with respect to DNA structure

Pierre-François Baisnée[1], Pierre Baldi[1,], Søren Brunak[2] and Anders Gorm Pedersen[2]*

[1]*Department of Information and Computer Science, University of California, Irvine, CA 92697-3425, USA and* [2]*Center for Biological Sequence Analysis, The Technical University of Denmark, DK-2800 Lyngby, Denmark*

ABSTRACT

Motivation: The primary function of DNA is to carry genetic information through the genetic code. DNA, however, contains a variety of other signals related, for instance, to reading frame, codon bias, pairwise codon bias, splice sites and transcription regulation, nucleosome positioning and DNA structure. Here we study the relationship between the genetic code and DNA structure and address two questions. First, to which degree does the degeneracy of the genetic code and the acceptable amino acid substitution patterns allow for the superimposition of DNA structural signals to protein coding sequences? Second, is the origin or evolution of the genetic code likely to have been constrained by DNA structure?

Results: We develop an index for code flexibility with respect to DNA structure. Using five different di- or trinucleotide models of sequence-dependent DNA structure, we show that the standard genetic code provides a fair level of flexibility at the level of broad amino acid categories. Thus the code generally allows for the superimposition of any structural signal on any protein-coding sequence, through amino acid substitution. The flexibility observed at the level of single amino acids allows only for the superimposition of punctual and loosely positioned signals to conserved amino acid sequences. The degree of flexibility of the genetic code is low or average with respect to several classes of alternative codes. This result is consistent with the view that DNA structure is not likely to have played a significant role in the origin and evolution of the genetic code.

Contact: pfbaldi@ics.uci.edu; baisnee@ics.uci.edu; brunak@cbs.dtu.dk; gorm@cbs.dtu.dk

INTRODUCTION

The primary function of DNA is to carry genetic information through the genetic code. DNA, however, contains a variety of other signals related, for instance, to reading frame, codon bias, pairwise codon bias, splice sites and transcription regulation, nucleosome positioning and DNA structure. Because multiple signals (Schaap, 1971; Trifonov, 1989) subject to different evolutionary pressures coexist along the DNA molecule, it is important to investigate the constraints that each signal poses on the others. The focus here is on the relationship between the genetic code and DNA structure.

There is mounting evidence that DNA structure, beyond the double-helical pattern, plays a fundamental role in a number of biological processes. This is not surprising given the packing requirements for chromosomal DNA and the interactions of the DNA molecule with its complex cellular environment. DNA structure has been implied, for instance, in DNA–protein interactions, gene regulation, nucleosome positioning, and even genetic instabilities (Baldi *et al.*, 1996; Pazin and Kadonaga, 1997; Tsukiyama and Wu, 1997; Werner and Burley, 1997; Pedersen *et al.*, 1998; Baldi *et al.*, 1999). It has been shown that the exact DNA sequence influences the three-dimensional structure of DNA (Brukner *et al.*, 1990; Bolshoy *et al.*, 1991; Hassan and Calladine, 1996; Olson *et al.*, 1998; Sinden *et al.*, 1998). Because structural and protein coding signals may be superimposed in coding regions, we would like to assess the relationship between the genetic code and DNA structure. In particular, it is reasonable to hypothesize that the genetic code ought to display a good degree of structural flexibility with respect to protein coding. If an amino acid or amino acid class is to be encoded by codons located in DNA regions with different structural properties, the standard genetic code must have a substantial degree of flexibility with respect to DNA structure. We expect a hydrophobic amino acid, for instance, to be realizable with codons covering a wide spectrum of DNA bendability structural properties, from very stiff to very bendable.

*To whom correspondence should be addressed.

Here we assess, quantify, and confirm this hypothesis and its limitations, using a number of computational models of DNA structure, based on different experimental or theoretical approaches, which relate the nucleotide sequence to DNA flexibility and curvature (Ornstein *et al.*, 1978; Satchwell *et al.*, 1986; Goodsell and Dickerson, 1994; Sinden *et al.*, 1998; Brukner *et al.*, 1995; Hassan and Calladine, 1996; Hunter, 1996; Baldi *et al.*, 1998; Baldi and Baisnée, 2000). These models are typically in the form of di-nucleotide or tri-nucleotide scales (see Ponomarenko *et al.*, 1999, for a long but non-exhaustive list). While there is no final consensus regarding these models, it is likely that each one provides a slightly different and partially complementary view of DNA structure. It is worth noting that DNA structure is a complex phenomenon with multiple facets at different length scales. The models we use are meant primarily to address relatively local effects. These models are described in the next section.

Finally, when two codes are superimposed, a second more speculative issue is whether the constraints one code places on the other may have had a strong causal influence on the origin or evolution of the other. In particular, we discuss whether the genetic code may have evolved without being significantly constrained by structural DNA requirements.

MATERIALS AND METHODS

DNA structure

Among the previously mentioned sequence-dependent structural models of double-stranded DNA, we retained the following five di- or tri-nucleotide scales.

Bendability (B). One tri-nucleotide B model has been derived using the cutting frequencies of the DNase I enzyme (Brukner *et al.*, 1995). DNase I is known to preferably bind and cut DNA that is bent, or bendable, towards the major groove (Lahm and Suck, 1991; Suck, 1994). Thus DNase I cutting frequencies on naked DNA can be interpreted as a quantitative measure of major groove compressibility or anisotropic B. Such data have been used to calculate B parameters for the 32 complementary tri-nucleotide pairs.

Position preference (PP). From experimental investigations of the positioning of DNA in nucleosomes it has been found that certain tri-nucleotides have a strong preference for being positioned in phase with the helical repeat. Depending on the exact rotational position, such triplets will have minor grooves facing either towards or away from the nucleosome core (Satchwell *et al.*, 1986). Based on the premise that flexible sequences can occupy any rotational position on nucleosomal DNA, these preference values can be used as measures of DNA flexibility. Hence, in this model, all triplets with close to zero preference are assumed to be flexible, while triplets with preference for facing either in or out are taken to be more rigid. Note that we do *not* use this scale as a measure of how well different triplets form nucleosomal DNA. Instead, the absolute value, or unsigned nucleosome positioning preference, is used here as a measure of DNA flexibility (Pedersen *et al.*, 1998).

Propeller twist angle (PT). The di-nucleotide PT scale of Hassan and Calladine (1996) is based on X-ray crystallography of DNA oligomers. Di-nucleotides with a large PT tend to be more rigid than di-nucleotides with a low PT.

Protein deformability (PD). The PD scale of Olson *et al.* (1998) is a di-nucleotide scale derived from empirical energy functions extracted from the fluctuations and correlations of structural parameters in DNA–protein crystal complexes. In our work, we have found that this scale is highly, but not fully, correlated with the PT scale.

Base stacking energy (BS). The BS di-nucleotide scale of Ornstein *et al.* (1978) is derived from approximate quantum mechanical calculations on crystal structures. Values are measured in kilocalories per mole.

With the exception of BS, all the scales were determined by purely experimental observations of sequence-structure correlations. In previous studies, we found these models useful (Baldi *et al.*, 1996; Pedersen *et al.*, 1998; Baldi *et al.*, 1999; Pedersen *et al.*, 2000; Baldi and Baisnée, 2000), in particular for the detection of putative new structural signatures associated with an increase of B in downstream regions of RNA polymerase II promoters. A similar approach was used in Liao *et al.* (2000) to analyze the structure of insertion sites for P transposable elements in *Drosophila melanogaster* and suggest that the corresponding transposition mechanism recognizes a structural signature rather than a specific sequence motif.

Structural signals in DNA sequences might be of a 'punctual' nature, thus corresponding to specific scale values at a given di- or tri-nucleotide step. They might also extend over several steps, and could thus correspond either to specific profiles of scale values over a given window size or—assuming that the structural properties of successive steps are additive and that signals are likely to correspond to extreme values—to a specific average value over that window. Here, for each codon XYZ in the genetic code, as well as for stop signal triplets, we compute several numerical values, one for each structural model. For tri-nucleotide scales, the value is the one provided by the model itself. For di-nucleotide scales, we use the sum of the values associated with the pair of di-nucleotides XY and YZ as discussed in Baldi and Baisnée (2000). In Baldi *et al.* (1998) and Baldi and Baisnée (2000), we showed that, overall, the structural values

provided by the previous models, as well as their linear sum over short nucleotide sequences, are statistically fairly independent of each other and of A + T content.

Amino acid categories

The flexibility of the code with respect to DNA structure can be analyzed at the level of single amino acids. But it is also possible to group amino acids into broad categories, such as hydropathy or size. Several categorizations have been proposed and none is universally applicable, although many of them concur on the classification of a particular amino acid. Here we tested three categorizations (size, polarity and charge, and hydropathy) using the following more or less standard classifications.

(1) Size classification:
Small = {A, C, D, G, N, P, S, T, V}.
Not Small = {E, F, H, I, K, L, M, Q, R, W, Y}.
(2) Polarity and charge classification:
Hydrophobic = {A, F, I, L, M, P, V, W}.
Hydrophilic, negative charge (acid) = {D, E}.
Hydrophilic, no charge (non-acidic) = {C, G, N, Q, S, T, Y}.
Hydrophilic, positive charge (basic) = {H, K, R}.
(3) Hydrophobicity classification:
Hydrophilic = {D, E, H, K, N, P, Q, R, S, T}.
Hydrophobic = {A, C, F, G, I, L, M, V, W, Y}.

Assessment of the flexibility of genetic codes with respect to DNA structure

For the purpose of comparing the flexibility of the standard code to that of alternative codes, we introduced a simple 'code flexibility' index, defined as the average of the ranges spanned by each coded item (amino acids and stop signal), per scale or averaged over different scales. This average can be calculated over the whole set of coded items, or over a particular subset of amino acids. When considering such subsets, we also used the total range spanned by the corresponding codons as an indicator. In order to compare code flexibility indices over different structural models, one can normalize the average range by the full range of the corresponding structural scale. Lastly, for the purpose of assessing the standard genetic code with respect to a background set of alternative codes, one can use its Z-score, i.e. its distance from the average in the reference set of codes, normalized by the corresponding standard deviation.

Provided with some assumptions on amino acid and codon distributions, one can also calculate the flexibility that a code provides with respect to DNA structure for the shifted triplets in the two non-coding frames. We explored this issue using uniform distributions. The flexibility index for a non-coding frame is then defined as the range of structural values spanned by the set of triplets that occur in the frame when varying the codons of two given consecutive amino acids, averaged over all amino acid pairs. Lastly, in order to investigate to which degree a tolerance of a few steps positionwise might increase the possibility of superimposing a 'punctual' tri-nucleotide structural signal on a coding sequence, we also calculated a flexibility index over the four frames corresponding to two consecutive amino acids. In this case we consider the range of structural values spanned by all the triplets that occur in any of the four frames.

Alternative codes and model for the distribution of structural values

In order to assess the flexibility of the standard genetic code with respect to DNA structure, we used two different approaches. First, we analytically calculated the expected range of structural values for each coded item, using a very simple (uniform) statistical model for the distribution of structural values over triplets. Second, we computed our code flexibility index for three sets of alternative, artificially generated 'genetic codes'.

Uniform model. We considered a model for hypothetical codes in which the 64 triplets corresponding to codons and stop signals would be distributed *uniformly* over the range of each structural scale. Such a model obviously ignores the actual distribution of triplets in each scale, which, rather than uniform, is roughly Gaussian (as shown in Baldi and Baisnée, 2000). Note that we only use this model as an extreme comparison point, because of its simple analytical treatment.

Random codon codes. In order to account for the actual distribution of the 64 triplets (codons and stop signals) over the range of each structural model, we first constructed a set of alternative codes such as those used in Alff-Steinberger (1969). It consists of 100 000 almost *completely* random codes, all retaining the same number of codons per amino acid as the standard code, but in which the 64 triplets were randomly (uniformly) assigned to amino acids and stop signals. Note that this set does not take into consideration any constraint of the genetic code, other than the number of codons per amino acid. For conciseness, *all figures displaying random code distributions are based on this set.*

Permuted codes. In order to retain more constraints or features of the genetic code, such as the lower degree of variability of the first two letters of the codons coding for any given amino acid, we constructed a second set of codes. It includes the 23 additional codes derived from the standard genetic code by permuting the four letters of the DNA alphabet, and mapping these permutations to codons. For instance, the first permutation of the alphabet 'ACGT' is 'ACTG', and the codon assigned to Methionine in the first permuted code therefore is 'AGT', instead of 'ATG' in the standard genetic code.

Random block codes. Lastly, we constructed a set of 100 000 random codes following the method used in Haig and Hurst (1991). In this set—which retains more features of the standard code than the random codon code set— each block of synonymous codons is randomly assigned to any of the 20 amino acids, while the codons coding for stop signals are conserved. For obvious reasons, each code in this set has the same flexibility index as the standard code. However, both the total range of structural values and the average range per amino acid vary when considering *subsets* of amino acids.

Independence of DNA structural values in coding regions and properties of encoded proteins

To investigate whether the physical properties of polypeptide sequences are correlated to the structural properties of the corresponding encoding DNA, we selected two scales measuring amino acid properties: hydrophobicity (Kyte and Doolittle, 1982) and isoelectric point (pI).

We then extracted the 6898 (respectively, 34 465) non-overlapping 150 (30) amino acid long sequences from the complete set of *Escherichia coli* coding sequences (CDS), as well as the corresponding 450 (90) bp long DNA sequences (Blattner *et al.*, 1997). We calculated the average theoretical hydrophilicity (Kyte and Doolittle, 1982) and pI values for the polypeptide chains, as well as the average of DNA structural values over all successive overlapping di- or tri-nucleotides in the DNA sequences. Lastly, for each window we plotted the average values obtained for polypeptide sequences against those obtained for the DNA sequences, and computed the corresponding Pearson linear correlation coefficient.

RESULTS

Flexibility of the standard genetic code with respect to DNA structure at the level of single amino acids

Figure 1 displays the range of structural values spanned by each set of synonymous codons and by stop signals, for each of the five structural models considered. (The numerical values for the 64 possible triplets are easily derived from the tables provided in Baldi and Baisnée, 2000.) The main observations are as follows:

- As one would expect, there is a substantial correlation between the number of synonymous codons for a given amino acid and the range they cover. (Correlation coefficients between range and number of codons are as follows for the five models considered: B, 0.57; PP, 0.57; BS, 0.48; PT, 0.77; PD, 0.59.) With the obvious exception of the two amino acids that are coded for by only one codon, M (Met) and W (Trp), the genetic code thus provides some flexibility with respect to DNA structure: codons coding for a given amino acid span a fraction of the total range of the structural scales.

- However, a visual inspection of Figure 1 suggests that the standard genetic code is far from providing a maximal range to the nine amino acids that are coded for by only two synonymous codons; this is particularly clear in the case of the PD scale.

In order to further assess the flexibility of the standard genetic code with respect to DNA structure, we calculated the structural ranges that one could expect for amino acids and stop signals under the various assumptions and for the artificially generated 'genetic codes' described in the Materials and Methods.

In the case of the 'uniform' model, a simple calculation shows that the expected range for an amino acid coded by N synonymous codons is given by $(N - 1)/(N + 1) * R$, where R represents the maximal range for a given structural scale. We found that the ratio of the observed range to the expected range, averaged over all coded items, is lower than unity for every structural model (data not shown). It averages to 0.67 over the five models considered. This relatively low value is partly due to the fact that the structural values of the 64 triplets, instead of being evenly distributed, have a bell-shaped distribution. The uniform model therefore overestimates the expected structural ranges, and underestimates the ratio of the observed range to the expected range.

The results concerning the sets of alternative codes are shown in Figures 2 and 3, which compare the flexibility of the standard genetic code to that of the random codon and permuted code sets, respectively. The main observations relative to these two figures are as follows:

- With the exception of the PP scale, the genetic code exhibits a low flexibility, when compared to the set of almost *completely* random codes. (The corresponding Z-scores for the scales B, PP, BS, PT, PD are respectively -1.15, 0.58, -3.55, -4.45, -3.08. The ratio of the observed to the expected range averages to 0.81 over the five models.)

- This is particularly striking in the case of the di-nucleotide scales (BS, PT, PD), with a code flexibility index several standard deviation units away from average. The latter result is not surprising if one considers the way we calculate the structural value of a triplet for a di-nucleotide scale (as the sum of the two overlapping di-nucleotides it contains), and the structure of the code, i.e. the fact that for every amino acid coded by 2, 3 or 4 codons, the first two codon positions are always occupied by the same di-nucleotide, while for the three amino acids coded by 6 codons, only two di-nucleotides are assigned to these same positions. This obviously limits the structural range of every amino acid and that of the stop signals in the case of di-nucleotide scales.
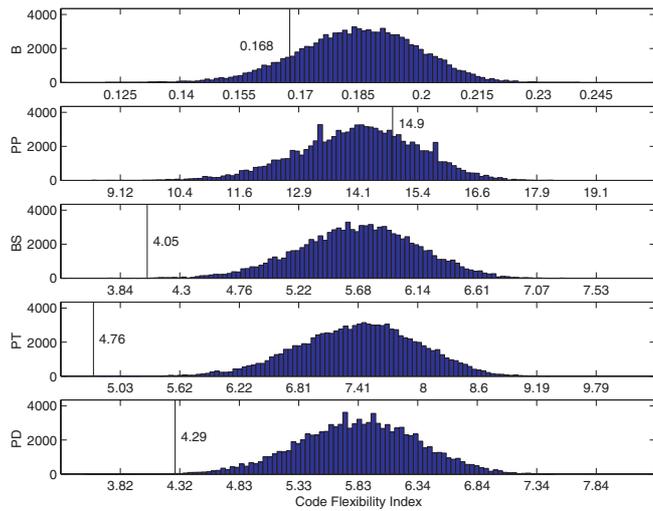
Fig. 1. Scale values of amino acid codons and stop signals. Amino acids and stop signals are identified by their one letter code on the *x*-axis of the bottom plot. The number of synonymous codons for a given amino acid is shown on the *x*-axis of the topmost plot. Each plot corresponds to a structural scale. For each amino acid, a black line represents the range of values spanned by synonymous codons, while markers show the actual values for individual codons. The shape and face of the markers identify groups of four codons sharing the same di-nucleotide in their first two positions. For instance, the white circles identify the four Arginine codons that are of the form CGN, where N can be any of the four letters in the DNA alphabet.

- When limiting the analysis to the set of permuted codes represented in Figure 3, which, among other features, retain the lower variability of the first two codons letters for every amino acid, it appears that the range of variation of the code flexibility index is much smaller and that its average value is much lower than for random codon codes, again excepting the PP scale. The standard genetic code appears to have a code flexibility index close to the permuted codes average for every structural scale.

Lastly, we calculated the average flexibility the code provides, with respect to DNA structure, in the coding and non-coding frames spreading over the triplets of two consecutive amino acids. The calculation method is explained in Section Assessment of the flexibility of the genetic codes with respect to DNA structure, and the results are plotted in Figure 4. The main observations are as follows:

- The permuted codes appear to offer a strikingly reduced flexibility in frames 2 and 3, when compared to random codon codes. This can be attributed to the 'block-structure' of the permuted codes, i.e. the low variability of the first two nucleotides in the triplets coding for any given amino acid. In the case of di-nucleotide scales, the way we calculate the structural value of a triplet contributes to the low flexibility value

Fig. 2. Flexibility of the standard genetic code with respect to DNA structure compared to that of 100 000 random codon codes. Each plot corresponds to a structural scale. In each plot, the *x*-axis represents code flexibility, defined as the average range spanned by a coded item (amino acid or stop signal). The histogram represents the distribution of code flexibility values for 100 000 random codon codes. All plots are centered on the estimated mean, and extend over five standard deviations on each side of the mean. A vertical line represents the code flexibility value of the standard genetic code. See text for a description of the random codes.

Fig. 3. Flexibility of the standard genetic code with respect to DNA structure, compared to that of the 23 alternative codes obtained through alphabet permutation. In each plot, the *x*-axis represents code flexibility, defined as the average range spanned by a coded item (amino acid or stop signal). A filled circle represents the standard genetic code, while plain circles represent alternative codes. The vertical line represents the average flexibility, around which plots are centered. The *x*-axis extends 2.5 standard deviations away from average in each direction. See text for a description of the permuted codes.

for such codes, as previously noted. It also explains why the highest average flexibility is observed in the second frame. In the case of tri-nucleotide scales, this result suggests that there is a form of continuity between scale values and the letter content of the corresponding triplets.

- When considering the global flexibility provided over all four frames (see labels 'F1–4' in Figure 4), one can note that it approaches the full range of structural values in the five models considered. More precisely, it corresponds to 63–83% of that range, depending on the scale. In other words, should a punctual structural signal (corresponding to a single tri-nucleotide) need to be superimposed to a coding sequence with some tolerance positionwise, most conceivable codes would be likely to provide sufficient flexibility to position this signal within at most two frames from any given target point, on average.

Flexibility of the standard genetic code with respect to DNA structure at the level of amino acid categories

Because amino acids belonging to the same broad category to a certain extent can be substituted one for another
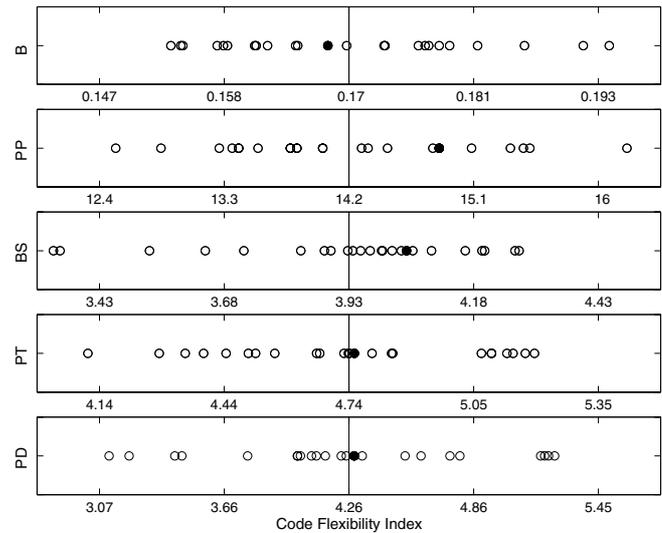
in proteins, it is at the level of such broad categories, rather than single amino acids, that one would expect the genetic code to provide a wide range of structural values, under the assumption that structural constraints exist in coding regions. We investigate here whether it is the case for the three aforementioned amino acid classifications (see Materials and Methods), and we compare the standard code to our sets of random and permuted codes under the perspective of the hydropathy classification.

The range of structural values spanned by each amino acid category is plotted in Figure 5. The main observations are as follows:

- Broad categories of amino acids in the three aforementioned classification systems span almost the whole range of structural values, for all five models considered. Amino-acids belonging to the same broad category, which to some extent can be regarded as functionally equivalent as far as protein structure is concerned, can therefore be coded for with structurally widely differing codons.

- One notable exception is the class corresponding to polar and negatively charged amino acids, which only contains the amino acids D (Asp) and E (Glu), and the corresponding four codons.
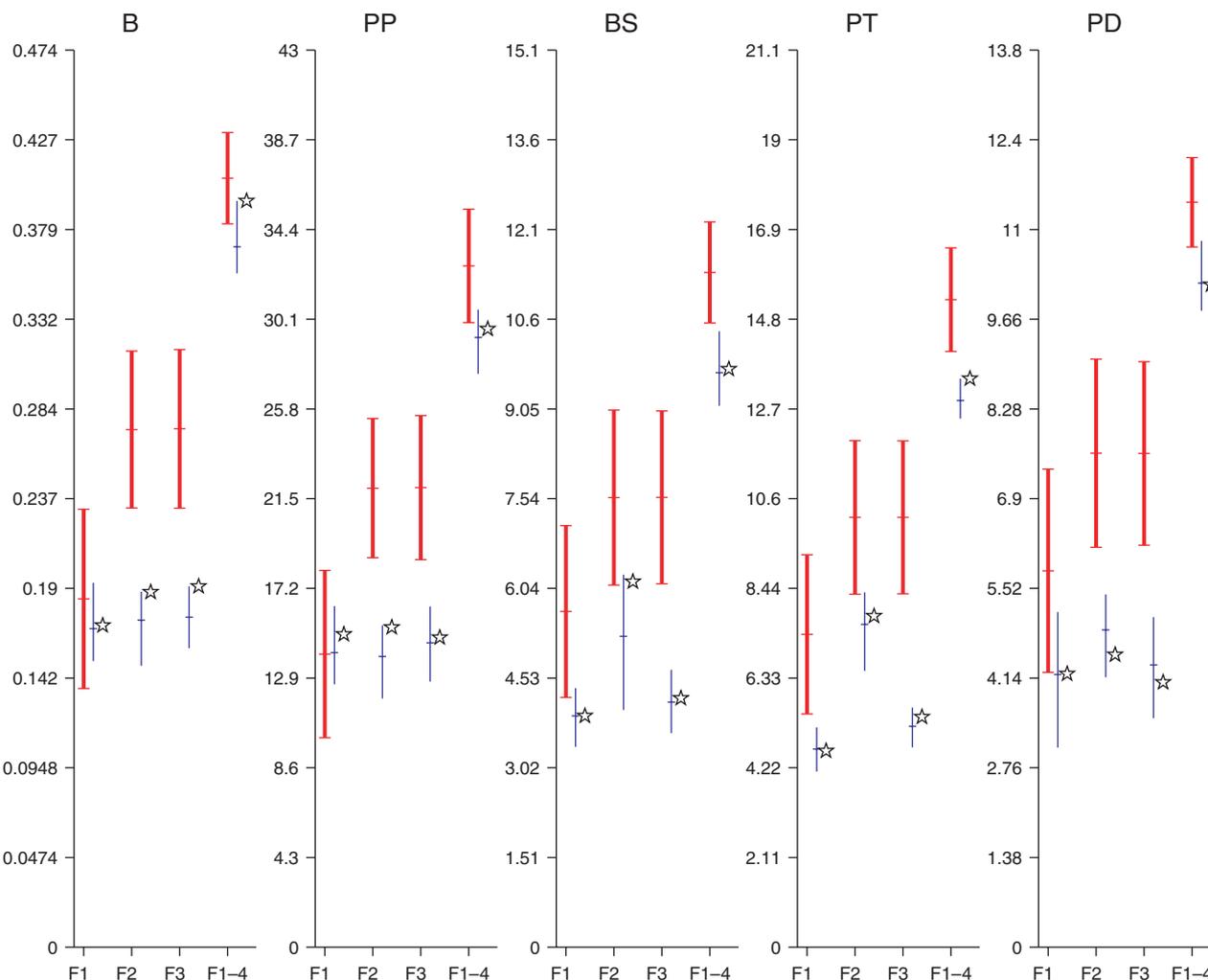
Fig. 4. Flexibility of the standard genetic code with respect to DNA structure over coding and non-coding frames for two successive amino acids, compared to that of random codon and permuted codes. Each plot corresponds to a structural scale. In each plot, the y-axis represents code flexibility, measured as the range of structural values spanned by the set of triplets that occur at a given position when varying the codons of two given consecutive amino acids, averaged over all amino acid pairs. Along the x-axis, F1 represents the coding frame for the first amino acid (corresponding to the results shown in Figures 2 and 3, excepting that the triplets coding for a stop signal were here excluded from the calculation). F2 and F3 represent the two non-coding frames. F1-4 represents the global flexibility calculated over all four possible frames, i.e. considering the range of values occurring in any of the four triplet steps. For each frame, the first (thick) line schematically represents the values obtained for random codon codes and ranges from −3 to +3 standard deviation units around the average; the second (thin) line represents the full range of values observed for permuted codes, the average being signaled by a tick mark; the star represents the standard genetic code. The y-axis spans the full range of structural values for each of the five models. Its tick marks represent 10% increments.

The range of structural values spanned by each amino acid hydropathy category in the standard genetic code is compared in Figure 6 to that observed for random codon codes. The main observations are as follows:

- Both hydrophilic and hydrophobic categories span the maximum possible range in the standard code for three out of the five structural models considered; for the two remaining scales, each category spans a range which is very close to the maximum range.

- However, Figure 6 shows that most random codon codes also span a maximal range or a very high one. This is also true of random block codes.

When considering the more limited set of permuted codes, we observed that the range spanned by each
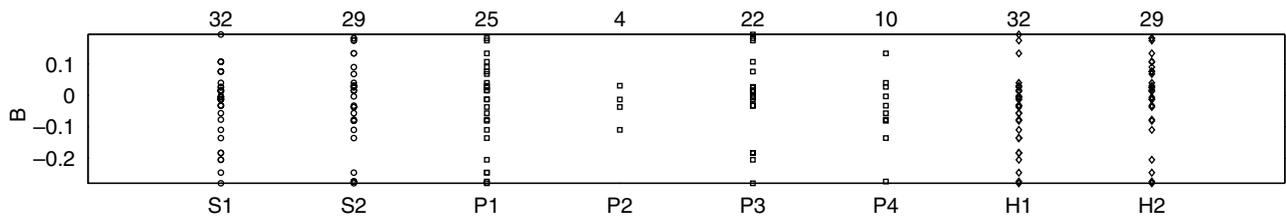
Fig. 5. B values of broad amino acid categories according to three classification systems. The three classification systems are displayed using different marker shapes. From left to right: size (S1 = small; S2 = not small); polarity and charge (P1 = hydrophobic; P2 = hydrophilic, negatively charged (acidic); P3 = hydrophilic, no charge (non-acidic); P4 = hydrophilic, positive charge (basic); hydropathy (H1 = hydrophilic; H2 = hydrophobic). The number of codons in each category is shown on the topmost plot. For each category, markers show the actual values for individual codons. Very similar plots were obtained for the four other structural scales.
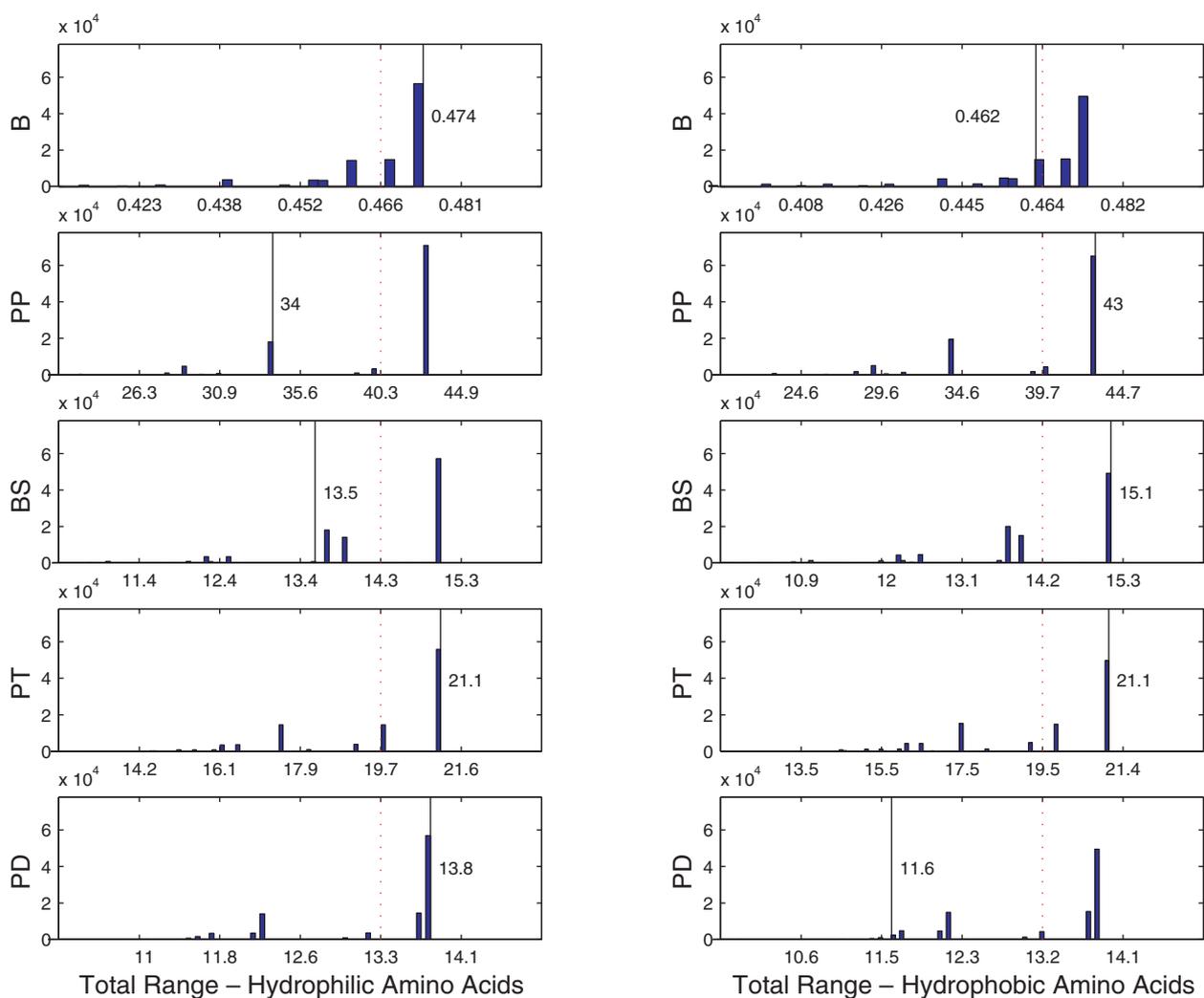
Fig. 6. Total range of structural values spanned by codons coding for hydrophilic and hydrophobic amino acids. Standard code compared with a set of 100 000 random codon codes. Each plot corresponds to a structural scale. In each plot, the *x*-axis represents the total range spanned by a hydropathy category. The histogram represents the distribution of range values for 100 000 random codon codes. A dotted vertical line represents the estimated mean range for random codes. A vertical plain line represents the range of the standard genetic code. The *x*-axis extends from −4 to +2 standard deviation units around mean. See text for a description of the random codes.
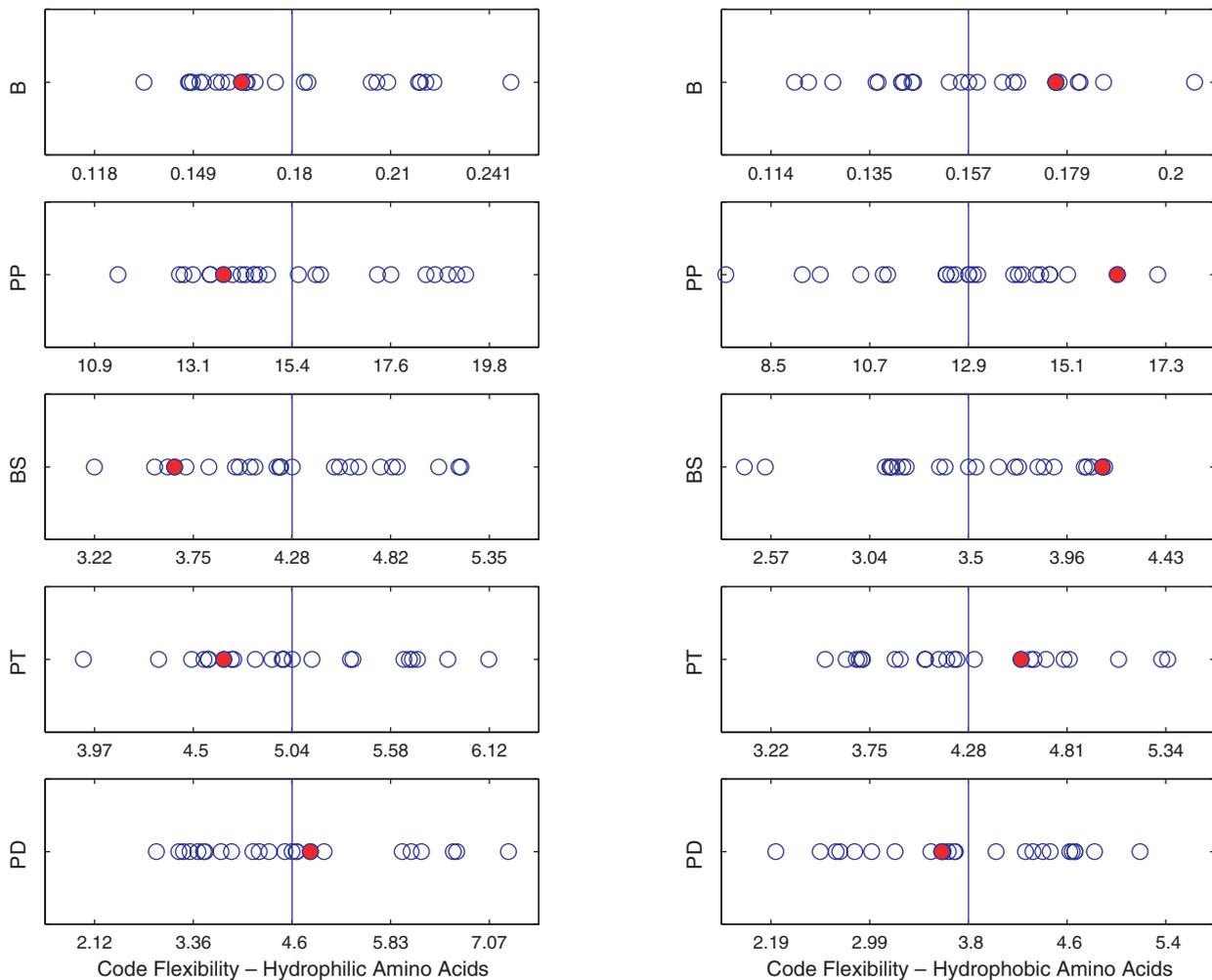
Fig. 7. Average structural range per amino acid for hydrophilic and hydrophobic categories. Each plot corresponds to a structural scale. In each plot, the *x*-axis represents code flexibility, defined as the average range spanned by the coded items in each category. A filled circle represents the standard genetic code, while plain circles represent alternative codes. The vertical line represents the average code flexibility, around which plots are centered. The *x*-axis extends 2.5 standard deviations away from average in each direction. See text for a description of the permuted codes.

hydropathy category is very high for every permutation, and that at least one permutation achieves the maximal possible range for each scale and each category (data not shown). The range covered by the standard genetic code fluctuates from scale to scale and category to category within a narrow absolute interval around the average.

Lastly, we computed our code flexibility index separately for each hydropathy category. Figure 7 compares the code flexibility achieved for these categories by the standard and the permuted codes. Interestingly, and excepting the PD scale for which the standard code is very close to average, hydrophobic amino acids span on average a range significantly higher than expected for the permuted codes, whereas the converse is true for hydrophilic amino acids.

Overall, the hypothesis that the genetic code is rather flexible with respect to DNA structure is confirmed at the level of broad amino acid categories: codons coding for amino acids belonging to the same class tend to be widely spread over the spectrum of each structural model. In other words, should a strong structural constraint exist at a given coding position, there would be an amino acid satisfying that constraint in every broad class. However, it was observed that the two hydropathy categories also span an optimal or very high range for most random and all permuted codes. Rather than the result of a hypothetical evolutionary force, flexibility at the level of broad categories can therefore be considered as the mere result of chance and the high number of amino acids they contain.

Independence of DNA structural values and protein properties in *E.coli*

The code flexibility we observed at the level of broad amino acid categories means that there is *in principle* no correlation between the structural properties of coding DNA and the physical properties of the corresponding coded amino acids and proteins. In other words, genomic regions with similar structural properties need not encode similar proteins, and proteins of a given type might be coded for by structurally widely differing stretches of DNA. However, it is not obvious that the lack of correlation between DNA structure and amino acid category observed at the code level translates into a lack of correlation in actual biological sequences. Notably, the fact that some amino acids and some codons are more frequent than others could introduce a correlation. In order to investigate this issue, we analyzed all coding regions in the entire *E.coli* genome (see Materials and Methods). This analysis showed that the lack of correlation, which reflects code flexibility, is also observed in real biological sequences. There is practically no correlation between codon usage and codon structural value (Pearson linear correlation coefficients are low for all models, ranging from 0.07 to 0.23). Furthermore, there is no correlation between protein structural characteristics (pI and hydrophobicity), on one hand, and the five DNA structural characteristics, on the other hand (Pearson correlation coefficients vary between −0.13 and 0.21).

DISCUSSION

Using five di-nucleotide or tri-nucleotide models of DNA structure, we have shown that the genetic code is rather flexible with respect to DNA structure at the level of broad amino acid categories, although its flexibility at the single amino acid level is only mild, and appears to be low or average with respect to several classes of artificial codes.

Evolution of the genetic code

Given the low average range spanned by single amino acids compared to that covered in random codon codes, and given the wide range spanned by broad amino acid categories for most random and permuted codes, it is tempting to speculate that DNA structural flexibility has not played a major role in the origin and evolution of the genetic code. Other constraints besides structure are likely to have played a greater role and actually *reduced* the degree of flexibility of the genetic code with respect to DNA structure at the level of single amino acids, notably by favoring the assignment of codons sharing the same first two nucleotides to a single amino acid. A number of scenarios can accommodate such a speculation.

First, one can hypothesize that DNA structure, as captured by the five models examined here, did not correspond to any *early* selective pressure, and did not impose strong constraints over coding regions while the genetic code was evolving. This would obviously have been the case if the code had evolved in a primitive RNA or pre-DNA world (as suggested, for instance, in Freeland *et al.*, 1999). It could also have been the case in an early DNA world. Selective pressures linked to DNA structure, if any, might indeed have emerged at a late stage of evolution, once the code had already developed, and in relation with functions arising in complex organisms, such as gene regulation and tight packing of large DNA pieces within membrane compartments.

Second, even if early selective pressures existed in relation to DNA structure in coding regions, the flexibility offered at the level of broad categories of early amino acids might have already been high enough to cope with such pressures.

In the absence of strong structural constraints, other forces could have freely driven the evolution of the code, corresponding for instance to the key factors generally put forward by the main theories on the origin of the genetic code (see Di Giulio, 1997, for a review). The resulting degeneracy pattern could then have been targeted at the adaptation to (or modulation of) constraints (or functions) more closely related to the translation machinery or to protein function than DNA structure. The stereochemical theory thus proposes that current codon assignments reflect primordial RNA–amino acid affinities (see for instance Yarus, 1998). The co-evolution theory (Wong, 1975) suggests that codons, originally assigned to prebiotic precursor amino acids, were progressively assigned to new amino acids derived from the precursors as biosynthetic pathways evolved. In line with the physico-chemical or ambiguity reduction theories, it has often been conjectured that the genetic code has been optimized with respect to error minimization (see for instance Freeland *et al.*, 2000), or that it provides through mutation events both a wide dynamic range (high 'changeability') and a smooth path (high 'mutational robustness') in the protein property space (Aita *et al.*, 2000). Alternatively, it has been suggested that the code has evolved so as to minimize the non-linearity to decoding (Tolstrup *et al.*, 1994). Lastly, it has also been proposed that the correlation between the physico-chemical properties of an amino acid and that of its codons—such as the well known correlation between the hydrophobicity of the amino acid and that of the first two bases of the anticodon from $3'$ to $5'$, or the correlation between the free energy change of anticodon–codon association and the volume of the corresponding amino-acid—might optimize the polypeptide polymerization process (Lehmann, 2000).

The previous discussion of a hypothesized relation between DNA structural parameters and the evolution of the genetic code cannot be straightforwardly extended to

RNA. Indeed, double-stranded RNA is generally found in the so-called A-conformation, which is fatter and less flexible than the usual B-form of DNA (Tinoco *et al.*, 1987; Hagerman, 1997). Furthermore, the apparent lack of correlation between the sequence and conformation of double-helical RNA (Holbrook *et al.*, 1981; Hagerman, 1997), along with the suggested single-strandedness of the primordial RNA-genomes (Lazcano *et al.*, 1988; Szathmary and Smith, 1993), makes it unlikely that double-stranded aspects of RNA-structure had a significant impact on the evolution of the genetic code.

DNA sequences

Even if DNA structural constraints may have had a minor impact on the evolution of the genetic *code*, their pressure is still likely to have influenced the evolution of DNA *sequences*, together with other forces. In particular, for coding regions, the flexibility that the code provides at the level of broad amino acid classes with respect to DNA structure is thus interesting from an evolutionary standpoint. Indeed, such flexibility allows for the superimposition of protein encoding signals and DNA structural signals (possibly related to regulatory phenomena) extending over any number of nucleotides, with very few restrictions. Moreover, provided a tolerance of a few frames position-wise, any punctual signal limited to a single tri-nucleotide step could in many instances be superimposed to a coding sequence while conserving not only the amino acid categories, but the amino acids themselves.

It is however likely that evolution has taken advantage of the code degeneracy and shaped codon usage so as to fine-tune several biological functions, in addition to those related to DNA structure. Synonymous codons, for instance, provide flexibility with respect to $G + C$ content (notably through the choice of the third codon letter for every amino acid provided with two codons or more), which can be used to accommodate global or local genomic constraints. There is also some evidence that the palette of synonymous codons might actually be used to modulate translation in general and gene expressivity or protein folding in particular: highly expressed genes contain almost exclusively 'fast' codons in some organisms (Lafay and Sharp, 1999; Gouy and Gautier, 1982; Grantham *et al.*, 1981), while complex folding protein domains seem to be coded by 'slow' codons (Thanaraj and Argos, 1996). Similarly, it has been found that the third (silent) nucleotide in those codons which encode different amino acids frequently located at the edge of protein secondary structures, is substantially conserved and therefore represents a signal at the DNA or RNA levels which correlates with the structure of the protein (Brunak and Engelbrecht, 1996). Lastly, Hartl *et al.* (1994) have observed a codon bias for conserved amino acids with conserved codons in enteric bacteria, and suggest that this bias might be linked to secondary struc-ture constraints in the corresponding mRNA. Whereas we have here shown that structural signals can be superimposed to protein coding sequences, it remains to be determined to which degree they can coexist with such other signals or adjustments, also mediated by synonymous codon or equivalent amino acid usage in coding regions. Given the high flexibility the code provides at the level of broad amino acid categories with respect to DNA structure, a substantial degree of superposition can be expected.

Future work is however needed to verify the latter assertion. In order to further explore the biological significance of DNA structure as captured by di- or tri-nucleotide scales, it might also be interesting to analyze pairwise codon bias and investigate whether it relates to specific patterns of structural values. Another possibility is to look for statistical evidence, in conserved sequences occurring in widely different genomic and structural contexts, that synonymous codons and equivalent amino acids are used to modulate structural properties of DNA or to conserve structural signals. In particular, it has been observed that, for repeated and conserved amino acids in conserved sequences, codon preference is dependent on the order in which amino acids appear (Tyson and Dhindsa, 1995). Such patterns might be linked to DNA structure constraints, and might be worth investigating under this perspective.

REFERENCES

Aita,T., Urata,S. and Husimi,Y. (2000) From amino acid landscape to protein landscape: analysis of genetic codes in terms of fitness landscape. *J. Mol. Evol.*, **50**, 313–323.

Alff-Steinberger,C. (1969) The genetic code and error transmission. *Proc. Natl Acad. Sci. USA*, **64**, 584–591.

Baldi,P. and Baisnée,P.-F. (2000) Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics*, **16**, 865–889.

Baldi,P., Brunak,S., Chauvin,Y. and Krogh,A. (1996) Naturally occurring nuclesome positioning signals in human exons and introns. *J. Mol. Biol.*, **263**, 503–510.

Baldi,P., Brunak,S., Chauvin,Y. and Pedersen,A.G. (1999) Structural basis for triplet repeat disorders: a computational analysis. *Bioinformatics*, **15**, 918–929.

Baldi,P., Chauvin,Y., Pedersen,A.G. and Brunak,S. (1998) Computational applications of DNA structural scales. In *Proceedings of the 1998 Conference on Intelligent Systems for Molecular Biology (ISMB98)*. The AAI Press, Menlo Park, CA, pp. 35–42.

Blattner,F., Plunkett,G., Bloch,C., Perna,N., Burland,V., Riley,M., Collado-Vides,J., Glasner,J., Rode,C., Mayhew,G., Gregor,J., Davis,N., Kirkpatrick,H., Goeden,M., Rose,D., Mau,B. and

Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.

Bolshoy,A., McNamara,P., Harrington,R.E. and Trifonov,E.N. (1991) Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc. Natl Acad. Sci. USA*, **88**, 2312–2316.

Brukner,I., Jurukovski,V. and Savic,A. (1990) Sequence-dependent structural variations of DNA revealed by DNase I. *Nucleic Acids Res.*, **18**, 891–894.

Brukner,I., Sánchez,R., Suck,D. and Pongor,S. (1995) Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J.*, **14**, 1812–1818.

Brunak,S. and Engelbrecht,J. (1996) Protein structure and the sequential structure of mRNA: α-helix and β-sheet signals at the nucleotide level. *Proteins*, **25**, 237–252.

Di Giulio,M. (1997) On the origin of the genetic code. *J. Theor. Biol.*, **187**, 573–581.

Freeland,S., Knight,R. and Landweber,L. (1999) Do proteins predate DNA? *Science*, **286**, 690–692.

Freeland,S.J., Knight,R.D., Landweber,L.F. and Hurst,L.D. (2000) Early fixation of an optimal genetic code. *Mol. Biol. Evol.*, **17**, 511–518.

Goodsell,D.S. and Dickerson,R.E. (1994) Bending and curvature calculations in B-DNA. *Nucleic Acids Res.*, **22**, 5497–5503.

Gouy,M. and Gautier,C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.*, **10**, 7055–7074.

Grantham,R., Gautier,C., Gouy,M., Jacobzone,M. and Mercier,R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic Acids Res.*, **9**, r43–r74.

Hagerman,P. (1997) Flexibility of RNA. *Annu. Rev. Biophys. Biomol. Struct.*, **26**, 139–156.

Haig,D. and Hurst,L. (1991) A quantitative measure of error minimization in the genetic code. *J. Mol. Evol.*, **33**, 412–417.

Hartl,D.L., Moriyama,E.N. and Sawyer,S.A. (1994) Selection intensity for codon bias. *Genetics*, **138**, 227–234.

Hassan,M.A.E. and Calladine,C.R. (1996) Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J. Mol. Biol.*, **259**, 95–103.

Holbrook,S., Sussman,J. and Kim,S. (1981) Absence of correlation between base-pair sequence and RNA conformation. *Science*, **212**, 1275–1277.

Hunter,C.A. (1996) Sequence-dependent DNA structure. *Bioessays*, **18**, 157–162.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Lafay,B. and Sharp,P.M. (1999) Synonymous codon usage variation among *Giardia lamblia* genes and isolates. *Mol. Biol. Evol.*, **16**, 1484–1495.

Lahm,A. and Suck,D. (1991) DNase I-induced DNA conformation: 2 Å structure of a DNase I-octamer complex. *J. Mol. Biol.*, **222**, 645–667.

Lazcano,A., Guerrero,R., Margulis,L. and Oro,J. (1988) The evolutionary transition from RNA to DNA in early cells. *J. Mol. Evol.*, **27**, 283–290.

Lehmann,J. (2000) Physico-chemical constraints connected with the coding properties of the genetic system. *J. Theor. Biol.*, **202**, 129–144.

Liao,G., Rehm,E.J. and Rubin,G.M. (2000) Insertion site preferences of the P transposable element in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA*, **97**, 3347–3351.

Olson,W.K., Gorin,A.A., Lu,X., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11 163–11 168.

Ornstein,R.L., Rein,R., Breen,D.L. and MacElroy,R.D. (1978) An optimised potential function for the calculation of nucleic acid interaction energies. I. Base stacking. *Biopolymers*, **17**, 2341–2360.

Pazin,M.J. and Kadonaga,J.T. (1997) SWI2/SNF2 and related proteins: ATP-driven motors that disrupt protein-DNA interactions? *Cell*, **88**, 737–740.

Pedersen,A.G., Baldi,P., Brunak,S. and Chauvin,Y. (1998) DNA structure in human RNA polymerase II promoters. *J. Mol. Biol.*, **281**, 663–673.

Pedersen,A.G., Jensen,L.J., Brunak,S., Staerfeldt,H.H. and Ussery,D.W. (2000) A DNA structural atlas for *Escherichia coli*. *J. Mol. Biol.*, **299**, 907–930.

Ponomarenko,M.P., Ponomarenko,J.V., Frolov,A.S., Podkolodny,N.L., Savinkova,L.K., Kolchanov,N.A. and Overton,G.C. (1999) Identification of sequence-dependent DNA sites interacting with proteins. *Bioinformatics*, **15**, 687–703.

Satchwell,S.C., Drew,H.R. and Travers,A.A. (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.*, **191**, 659–675.

Schaap,T. (1971) Dual information in DNA and the evolution of the genetic code. *J. Theor. Biol.*, **32**, 293–298.

Sinden,R.R., Pearson,C.E., Potaman,V.N. and Ussery,D.W. (1998) DNA: structure and function. *Adv. Gen. Biol.*, **5A**, 1–141.

Suck,D. (1994) DNA recognition by Dnase I. *J. Mol. Recognit.*, **7**, 65–70.

Szathmary,E. and Smith,J. (1993) The evolution of chromosomes. II. molecular mechanisms. *J. Theor. Biol.*, **164**, 447–454.

Thanaraj,T.A. and Argos,P. (1996) Ribosome-mediated translational pause and protein domain organization. *Protein Sci.*, **5**, 1594–1612.

Tinoco,I., Davis,P., Hardin,C., Puglisi,J., Walker,G. and Wyatt,J. (1987) RNA structure from A to Z. In *Cold Spring Harbor Symposia on Quantitative Biology,* Vol. LII, pp. 135–146.

Tolstrup,N., Toftgard,J., Engelbrecht,J. and Brunak,S. (1994) Neural network model of the genetic code is strongly correlated to the GES scale of amino acid transfer free energies. *J. Mol. Biol.*, **243**, 816–820.

Trifonov,E.N. (1989) The multiple codes of nucleotide sequences. *Bull. Math. Biol.*, **51**, 417–432.

Tsukiyama,T. and Wu,C. (1997) Chromatin remodeling and transcription. *Curr. Opin. Genet. Dev.*, **7**, 182–191.

Tyson,H. and Dhindsa,R. (1995) Codon usage in plant peroxidase genes. *DNA Seq.*, **5**, 339–351.

Werner,M.H. and Burley,S.K. (1997) Architectural transcription factors: proteins that remodel DNA. *Cell*, **88**, 733–736.

Wong,J.T. (1975) A co-evolution theory of the genetic code. *Proc. Natl Acad. Sci. USA*, **72**, 1909–1912.

Yarus,M. (1998) Amino acids as RNA ligands: a direct-RNA-template theory for the code's origin. *J. Mol. Evol.*, **47**, 109–117.