Eur. Phys. J. B **3**, 271–276 (1998)

THE EUROPEAN PHYSICAL JOURNAL B

The effect of DNA phase structure on DNA walks

S. Cebrat[1,a] and M.R. Dudek[2,b]

[1] Institute of Microbiology, University of Wrocław, ul. Przybyszewskiego 63/77, 51-148 Wrocław, Poland
[2] Institute of Theoretical Physics, University of Wrocław, pl. Borna 9, 50-204 Wrocław, Poland

Received: 8 January 1998 / Received in final form and Accepted: 28 January 1998

Abstract. We have performed several kinds of DNA walks which often are the first steps for further analysis of DNA structure and long range correlations. The DNA walks analysing frequency of $G + C$ *versus* $A + T$ cannot indicate the coding strand while purine *versus* pyrimidine DNA walks or two-dimensional $(A - T, G - C)$ DNA walks in some instances can indicate the coding strand but cannot resolve the coding frame. The modified two-dimensional $(A - T, G - C)$ DNA walks respecting the three-nucleotide codon structure show very high correlation in nucleotide composition of DNA coding sequences. They can distinguish between coding and non-coding sequences and indicate the strand and the phase in which DNA is coding.

PACS. 02.50.Fz Stochastic analysis – 42.62.Be Biological and medical applications – 02.70.Rw Other computational methods

1 Introduction

The first results of chemical analysis of DNA have shown absolute rules in its nucleotide composition: the number of adenines (A) is equal to the number of thymines (T), the number of guanines (G) is equal to the number of cytosines (C) and the ratio $(A+T)/(G+C)$ is characteristic for the genome of any organism. This is an implication of a structural rule that A on one strand corresponds to T on the opposite strand and G on one strand corresponds to C on the opposite one. Therefore, one strand determines strictly the sequence of the other strand (one strand called "W" for Watson and the other one called "C" for Crick). In a random DNA sequence there should be no statistically significant differences between nucleotide compositions of these two strands. Usually this is also true for the whole natural chromosomes but it may not be true for their fragments. In the fragments, different trends in nucleotide composition of DNA can be observed. Trends or correlations in nucleotide composition of DNA can be shown in terms of different classes of DNA Walks (DW).

DW.G+C

DW.$G + C$ is the simplest kind of a DNA walk which shows the relation between the local content of $(A+T)$ and $(G+C)$ pairs along a DNA molecule. The walker goes "up" if it finds G or C, and it goes "down" if it finds A or T. The ratio $(A + T)/(G + C)$ is a characteristic feature of every genome but locally there are relatively high differences in it. Since there are very simple experimental methods finding sequences rich in $G+C$ (they have higher buoyant density than sequences rich in $A+T$ and can be detected by centrifugation in CsCl gradient) geneticists have looked for their role for a long time. In the first fully sequenced yeast chromosomes the local differences in $(A+T)/(G+C)$ ratio were also found (Oliver *et al.* [1], Dujon *et al.* [2]). In addition, it has been found that in regions richer in $G+C$ density of coding sequences is higher. This is especially well seen in genomes of higher eukaryotes, where in regions very rich in $G+C$ (H3 isochores) coding density is up to 10 fold higher than in regions rich in $A+T$ [3]. Nevertheless, the DNA walks based on $A+T$ *versus* $G+C$ composition cannot give much information on DNA coding properties because:

1. the difference in the global nucleotide composition between coding and non-coding sequences is not very big. For example in the yeast genome the fraction of $G+C$ in known genes is 0.397 and in the intergenic sequences 0.368 (Cebrat *et al.* [4]). Thus, for short sequences there could be no statistically significant difference;
2. such a DNA walk cannot distinguish between coding and non-coding DNA strands. The $(A + T)/(G + C)$ ratio is exactly the same for both DNA strands, while usually only one strand of a DNA sequence is coding.

DW.PP

DW.PP make another class of DNA walks which show the differences in purine *versus* pyrimidine composition of

[a] e-mail: `cebrat@angband.microb.uni.wroc.pl`
[b] e-mail: `mdudek@ift.uni.wroc.pl`

the two DNA strands. The walker goes "up" if it finds G or A, and it goes "down" if it finds C or T. Because of the complementarity rule, the numbers of purines (G and A) and pyrimidines (C and T) in double strands of DNA molecules must be equal, but bias in the two strands with respect to purine/pyrimidine composition is possible. It has been shown many times that coding strands of DNA are richer in purines (Dujon *et al.* [2]) and this feature has been even used for discrimination between coding and non-coding Open Reading Frames (ORF, sequence of triplets which are potentially translatable into protein) (Cebrat *et al.* [5]). Peng *et al.* [6] performed a DNA walk based on the purine *versus* pyrimidine composition of the DNA strand and searched for correlation in it. Usually, DW.PP distinguishes between coding and non-coding DNA strands and indicates the strand which is coding in the analyzed DNA region. The results of DW.PP depend on the bias of two purine/pyrimidine pairs: $A-T$ and $G-C$. It is possible that asymmetries between the DNA strands for these two pairs compensate each other locally even in the same coding sequence [5]. Then, the purine/pyrimidine walk cannot show any correlation. In such a case a DNA walk of the class DW.$GC-AT$ in two-dimensional space ($T-A,\ C-G$) can be used.

DW.GC-AT

DW.$GC-AT$ is the DNA walk in two-dimensional space ($T-A,\ C-G$) performed by Berthelsen *et al.* [7] (see also, *e.g.*, [8]). It can show the bias for each pair.

Both DW.PP and DW.$GC-AT$ can distinguish between the coding and non-coding sequences in some instances, but the conclusions have to be drawn very carefully, since the bias in the purine/pyrimidine composition, (even pair specific) can be introduced by DNA replication process. The process of DNA replication is not equivalent for the two strands. In procaryotic genomes the mode of replication for each strand is determined by the location of start-point of replication (ORI) and terminator of replication (TER). One strand of DNA is synthesized continuously (leading strand) and the other one (lagging) is produced in fragments which are then ligated. The mode of replication is switched between strands at ORI and TER. Since the mode of replication is varying, the different mutational pressures are imposed on each DNA strand. This implicates the purine/pyrimidine bias in non-coding regions, too (Wu and Maeda [9], Francino and Ochman [10] for review).

There is one essential difference between coding and non-coding sequences. While coding sequences may reflect triplet composition of the genetic code there should be no traces of genetic code rules in intergenic sequences (unless they have ancestors among coding sequences). We have shown that respecting the influence of the genetic code structure on a DNA sequence, very strong correlations in coding sequences are observed in DNA walks.

2 Results and discussion

2.1 DNA walks which do not recognize DNA phase structure

Our analysis has been performed on eucaryotic genome of yeast *Saccharomyces cerevisiae* (16 chromosomes, 12 M base pairs) and bacterial genome *Escherichia coli* (4.6 M base pairs). The *S. cerevisiae* genome sequences were downloaded September 23, 1996 from: genome-ftp.stanford.edu. The *Escherichia coli* genome sequence was downloaded from: http://genom4.aist-nara.acjp May 9 1997. After the retrieval, data have not been updated. We have not performed the DW.$G+C$ walks because it is obvious that these walks do not distinguish between coding and non-coding strands. Both strands have exactly the same composition and the results don't depend on the direction of the walk. Nevertheless, it was observed in several genomes that coding regions have higher $(G+C)/(A+T)$ ratio than the whole genome (see Gardiner [3] for review). For the yeast genome the correlation between "coding density" and $(G+C)/(A+T)$ ratio has been shown by Sharp and Lloyd [11]).

During transcription each strand of a DNA sequence can be read only in one direction in natural systems. The RNA strand, which is the matrix for protein synthesis, is complementary to the transcribed DNA strand and resembles the opposite DNA strand. This opposite strand is called the coding strand. Coding strands usually are richer in purines than in pyrimidines. Analysis of long DNA sequences, like whole prokaryotic genomes or eukaryotic chromosomes shows that the role of the strands in respect of the coding/non-coding functions changes many times. In some regions W strand is coding and in the other regions C strand is coding. It is obvious that DW.PP walks (the type of DNA walks defined by Peng *et al.* [6]) performed for each of the two strands have exact mirror symmetry if they start from the same point of a DNA molecule and go in the same direction. The walks show an asymmetry in the purine/pyrimidine composition of DNA strands which is correlated with location of coding sequences. It is necessary to remind that an asymmetry in the purine/pyrimidine composition of DNA strands can be also a result of different mutational pressure during DNA replication. One DNA strand is synthesized without interruption (it is called the leading strand) and the other one is synthesized in fragments (it is called the lagging strand). It could be that this is one of the causes of the asymmetry in the purine/pyrimidine composition of non-coding as well as coding sequences in prokaryotic and eukariotic genomes. The asymmetry can be seen in Figure 1 where DW.PPs performed on strand W of the *E. coli* genome are shown both for coding and non-coding sequences. The coding sequences of the two strands W and C are represented by ORFs longer or equal to 150 codons, *i.e.*, mostly coding sequences, and the non-coding sequences are represented by the regions of W and C outside ORFs longer or equal to 30 codons. In case of the ORFs of strand C the nucleotides corresponding to their antisense (complementary sequence) on strand W

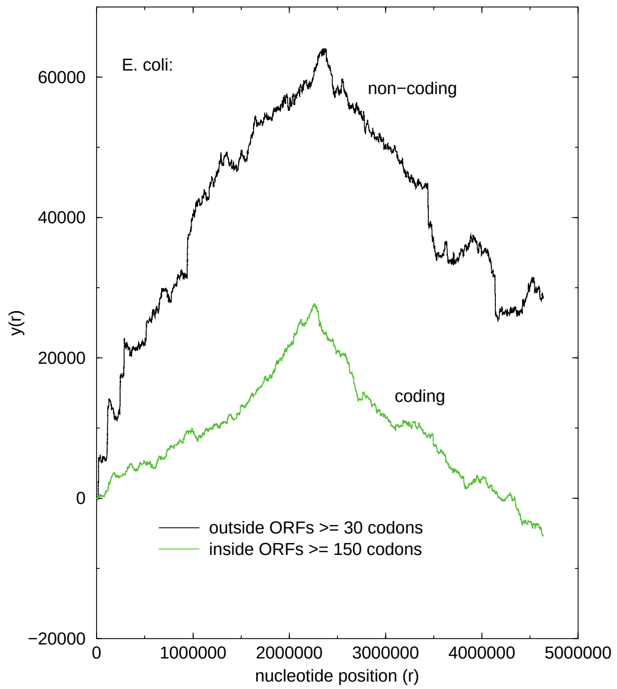

Fig. 1. DW.PPs performed on the strand W of the *E. coli* genome outside the ORFs of the strands W and C longer or equal to 30 codons (upper curve), and inside the ORFs of the strands W and C longer or equal to 150 codons (bottom curve). The walker step along the y (vertical) direction is equal to $\Delta y = \pm\ 16.4306$ and $\pm\ 2.36261$, respectively.

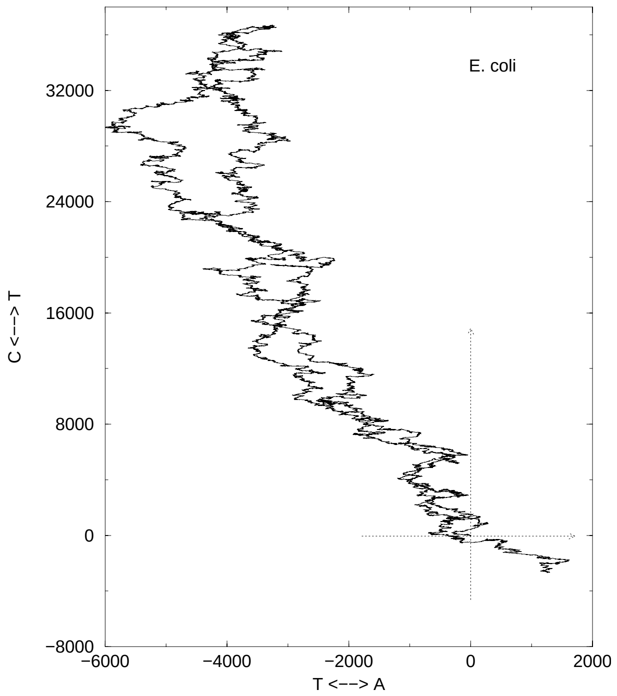

Fig. 2. DW.$GC - AT$ on the entire *E. coli* genome.

contribute to DW.PP. From definition, ORF contains a series of triplets (potentially) coding for amino acids without any termination codons. We have defined the walker step along the y (vertical) direction to be equal to the ratio $\Delta y = \pm\ \frac{\text{total length of the DNA sequence}}{\text{number of nucleotides visited by the walker}}$. The value $|\Delta y|^{-1}$ corresponds to the fraction of nucleotides visited on DNA sequence during DW.PP. If all sites are visited then $\Delta y = \pm\ 1$ as in classical DW.PP (Peng *et al.* [6]). The *E. coli* genome is circular and both DNA walks start from the ORI sequence (the Origin of Replication). It should be noted that the DNA strand is switched from leading to lagging at the origin of replication and at the terminator of replication. The DW.PPs in Figure 1 performed for the coding and non-coding sequences of the *E. coli* genome show trends following the natural division of the strands into leading and lagging fragments. The beginning and the end of the DNA.PP walks at zero and 4.5 M base pairs, respectively, are at the minimum of the plots and correspond to the ORI sequence while the maximum of the plots is reached at the terminator (at 2.3 M base pairs).

More precise information about the nucleotide composition of DNA can be supplied by a walk of the class DW.$GC - AT$ described by Berthelsen *et al.* [7]. In this walk the movement of the walker is associated with a unit shift in the two-dimensional space ($T - A$, $C - G$) depending on the type of nucleotide visited. The shifts are: (0,1) for G, (1,0) for A, $(0, -1)$ for C and $(-1, 0)$ for T. Since in the double strand of DNA the numbers of G and C are equal and the same holds for A and T - any asymmetry between the strands in purine/pyrimidine composition, even if G/C bias compensates A/T bias, is seen. Two-dimensional walks in ($T - A$, $C - G$) space can show strong correlations present in coding sequences [7]. The walks can also distinguish the coding strand in some cases. However, the same type of a DNA walk performed on long DNA sequences usually yields a complex trace of the walker with Brownian-like segments because consecutive coding sequences can compensate each other. The example of DW.$GC - AT$ performed on the entire *E. coli* genome is shown in Figure 2 and a similar walk on the yeast chromosome 4 (1.5 M base pairs) is shown in Figure 3.

2.2 DNA walks respecting DNA phase structure

Since the genetic code is composed of triplets, it is obvious that protein coding information of each strand can be read in three different phases. The term "phase" denotes reading frame: one DNA sequence can be read in six different ways [12–14]. Figure 4 shows the possibilities. It has been assumed that the phase (1) starts from the first nucleotide position in the strand W (triplets in the parenthesis represent the codons), whereas the phases (2) and (3) start from the second and third nucleotide position, respectively. Similarly, the phases (4-6) determine the strand C but they are read in the reverse order with respect to the direction of the strand W. Every sequence starting with codon ATG and ending with one of three stop codons: TAA, TGA or TAG represents ORF. In Figure 4 there are two ORFs in phase (1) and (2). The coding

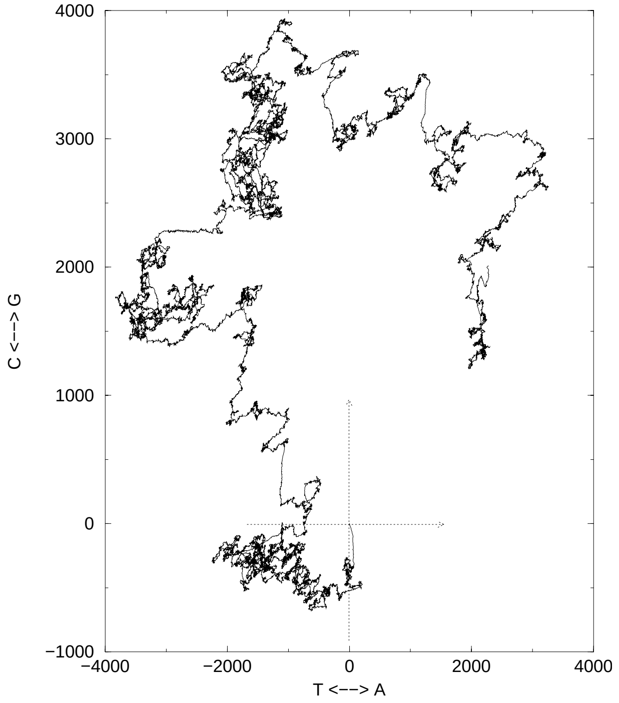

Fig. 3. DW.$GC - AT$ on the entire yeast chromosome 4.

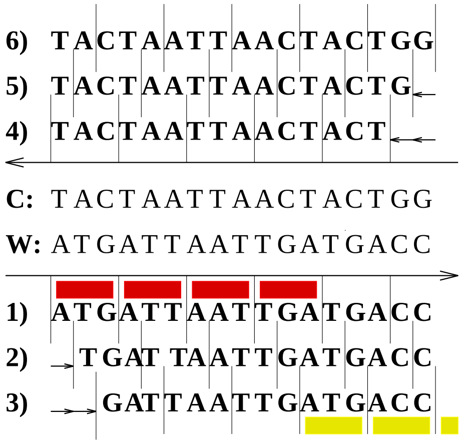

Fig. 4. The phase structure of a DNA molecule: three phases (1-3) of the trinucleotides on the first strand (W) and three phases (4-6) of the trinucleotides on the second strand (C). The blocks mark the codons in the ORFs, *i.e.*, the triplets coding for amino acids. For example the second phase reads $(TGA)(TTA)(ATT)(GAT)(GAC)$.

ORFs have a strong bias-weighting of each nucleotide position in codons [5]. It is important to remember that the nucleotide positions in codons play a different role in coding for amino acids. For example, T in the second position decides about hydrophobic properties of the coded amino acid and A at this position codes for hydrophilic amino acids. Sometimes the asymmetry in the first position is compensated by the asymmetry in the second position. Therefore the best solution is to perform a DNA walk for each of the three positions in codons separately. One can make the analogy with white light and monochromatic lights.

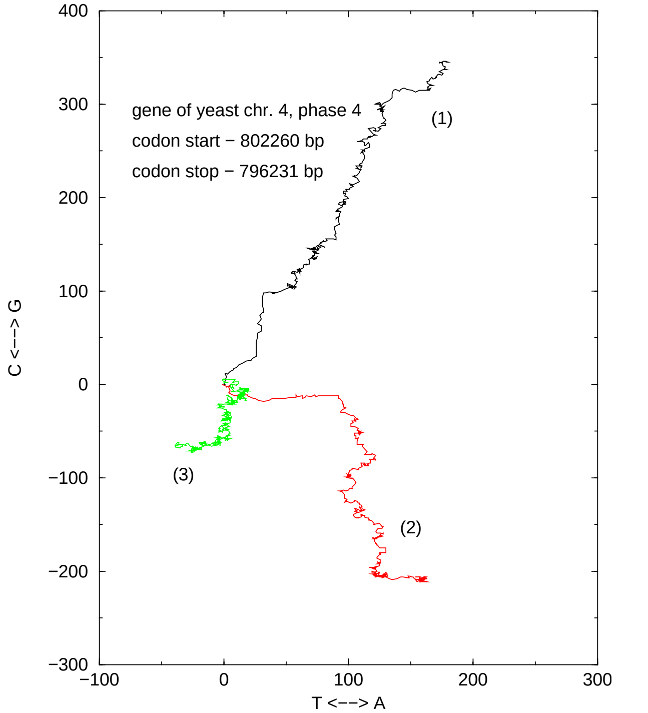

Fig. 5. Spider representing ORF of the yeast chromosome 4 (read in phase 4) with the start at the nucleotide position 796231 and stop at 802260.

The resulting three DNA walks for a typical gene are shown in Figure 5. We have called this DNA walk a "spider" and the individual walks describing the positions in codons are called "spider legs". The trends of the spiders' legs can be used as parameters for discrimination between coding and non-coding sequences (Cebrat *et al.* [4,15]).

Now, if we make spider representation for all long ORFs spliced from both W and C strands of the *E. coli* genome in such a way that they are read in the phases in which they are coding, we will observe a strong anisotropy of every nucleotide position in codons as in Figure 6a (the upper part of Fig. 6). The same spider analysis performed for the same ORFs read in strand W yields a picture as in Figure 6b (the bottom part of Fig. 6). The same result we will obtain for each of the yeast chromosomes. The shrinking of the spider means that the ORFs from one DNA strand compensate the ORFs of the other strand. The compensation is evident in Figure 7 where two spiders: one for all nonoverlapping ORFs longer than 150 codons spliced from strand W and the other one for all nonoverlapping ORFs spliced from strand C (but read in strand W) are shown separately. Now contrary to Figure 6a, one can observe different trends on the spider's legs in the leading and lagging fragments of the *E. coli* genome. Similar case for the entire yeast genome has been shown in Figure 8. In the latter case all 16 yeast chromosomes have been spliced into one big molecule (respecting their phases) which we have called yeast genomosome. In case of the non-coding sequences the effect of shrinking

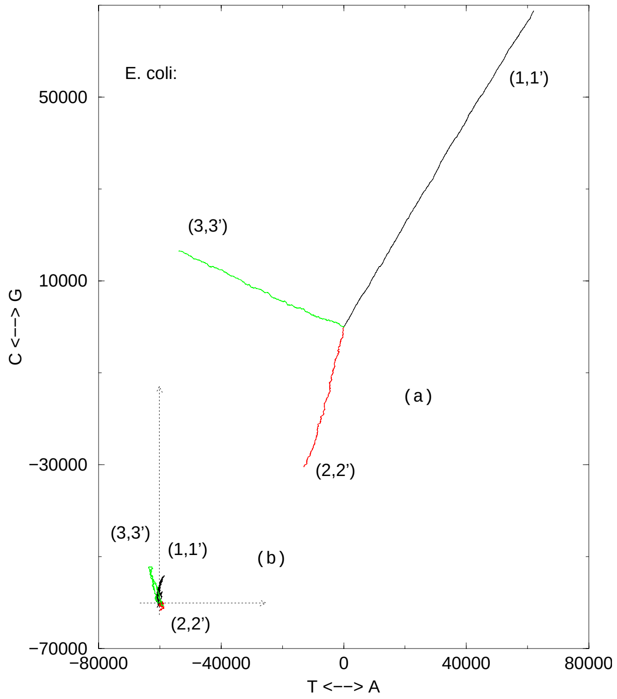

Fig. 6. Spider for *E. coli* representing: (a) nonoverlapping ORFs (longer or equal to 150 codons) spliced from the strand W and C (all ORFs are read in the phases in which they are coding), (b) the same ORFs as in (a) but ORFs of the strand C are represented by their antisense. The spider (b) has been shifted by the vector (−60 000, −60 000) to have clear presentation.

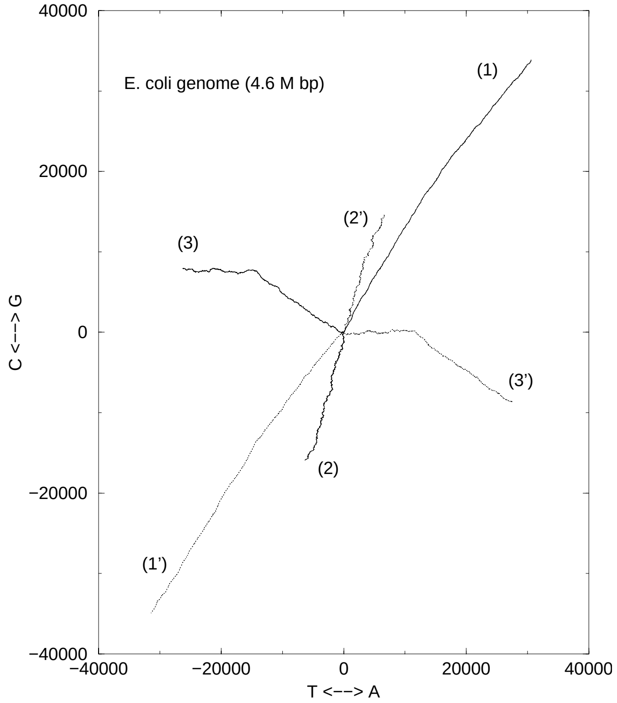

Fig. 7. Spider of *E. coli* representing: (123) nonoverlapping ORFs (longer or equal to 150 codons) spliced from the strand W, (1'2'3') antisense of ORFs ($\geq$ 150 codons) spliced from the strand C.

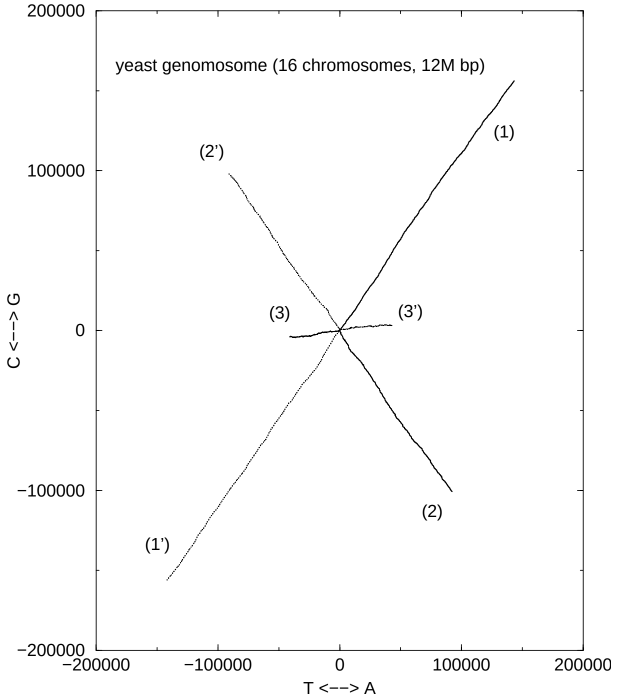

Fig. 8. The same as in Figure 7 but for the yeast genomosome (12 M base pairs).

of the spider is much stronger than for coding sequences. It is obvious that the correlations shown in Figures 6-8 are so clear because they respect the phases in which the coding information has been written.

3 Conclusions

We have shown that statistical analysis of coding properties of the DNA molecule should be done in proper DNA phases. Otherwise coding trends of DNA fragments could compensate each other. DNA walks analyzing purines and pyrimidines depend strongly on whether one looks only at the first position of the codons, only at the second, only at the third, or at all three positions together. The spiders introduced by us – DNA walks respecting positions in codons – make possible to distinguish between coding and non-coding sequences, or coding sequences read in inappropriate phase. We have discussed the possibility in our paper [15].

It is possible that long-range correlations observed in DNA sequence [6,16–18] are connected with strong bias-weighting of genes and intergenic sequences since different mutational pressures can be imposed on:

- leading and lagging DNA strands,
- coding and non-coding sequences,
- different positions in codons.

Once the genes involved in protein synthesis have to carry information on amino acid sequences, and nucleotide composition of genes depends on their position on chromosome, distribution of genes cannot be random.

We would like to thank M. Wolf for many stimulating discussions and D. Stauffer for encouraging us to submit the paper. The work was supported by KBN grants No 2 P30205707, 1016/S/IMI/97, 6 P04A03014.

References

1. Oliver *et al.*, Nature **357**, 38-46 (1992).
2. B. Dujon *et al.*, Nature **369**, 371-378 (1994).
3. K. Gardiner, Trends Genet. **12**, 519-524 (1996).
4. S. Cebrat, M.R. Dudek, P. Mackiewicz, Theory in Bioscience, to be published.
5. S. Cebrat, M.R. Dudek, A. Rogowska, J. Appl. Genetics **38**, 1-9 (1997).
6. C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Nature **356**, 168-170 (1992).
7. Ch.L. Berthelsen, J.A. Glazier, M.H. Skolnick, Phys. Rev. **A45**, 8902-8913 (1992).
8. J.R. Lobry, Biochimie **78**, 323-326 (1996).
9. C-I. Wu, N. Maeda, Nature **327** 169-170 (1987).
10. M.P. Francino, H. Ochman, Trends Genet. **13**, 240-245 (1997).
11. P. Sharp, A. Lloyd, Nucleic Acids Res. **21**, 179-183 (1993).
12. B. Dujon, Trends Genet. **12**, 263-270 (1996).
13. S. Cebrat, M.R. Dudek, Trends Genet. **12**, 12 (1996).
14. B. Dujon, A. Goffeau, Trends Genet. **12** (7), Poster (1996).
15. S. Cebrat, M.R. Dudek, P. Mackiewicz, M. Kowalczuk, and M. Fita, Microb. Comp. Genom. **2**, (1997) 259-268
16. W. Li, Int. J. Bifurc. Chaos **2**, 137-154 (1992).
17. W. Li, K. Kaneko, Europhys. Lett. **17**, 655-660 (1992).
18. R. Voss, Phys. Rev. Lett. **68**, 3805-3808 (1992).