J Mol Evol (1999) 49:182–192

JOURNAL OF MOLECULAR EVOLUTION

ApA Dinucleotide Periodicity in Prokaryote, Eukaryote, and Organelle Genomes

Masaru Tomita,[1,2] Masahiko Wada,[1,3] Yukihiro Kawashima[1,2]

[1] Laboratory for Bioinformatics, Keio University, 5322 Endo, Fujisawa, Kanagawa 252-8520, Japan
[2] Department of Environmental Information, Keio University, 5322 Endo, Fujisawa, Kanagawa 252-8520, Japan
[3] Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa, Kanagawa 252-8520, Japan

Received: 3 November 1998 / Accepted: 24 March 1999

Abstract. Computer analyses of various genome sequences revealed the existence of certain periodical patterns of adenine–adenine dinucleotides (ApA). For each genome sequence of 13 eubacteria, 3 archaebacteria, 10 eukaryotes, 60 mitochondria, and 9 chloroplasts, we counted frequencies of ApA dinucleotides at each downstream position within 50 bp from every ApA. We found that the complete genomes of all three archaebacteria have clear ApA periodicities of about 10 bps. On the other hand, all of the 13 eubacteria we analyzed were found to have an ApA periodicity of about 11 bp. Similar periodicities exist in the 10 eukaryotes, although higher organisms such as primates tend to have weaker periodic patterns. None of the mitochondria and chroloplasts we analyzed showed an evident periodic pattern.

Key words: Dinucleotide — Sequence periodicity — Complete genome — Mitochondria — Chroloplast — Prokaryote — Eukaryote

Introduction

Recent developments in the sequencing of whole genomes have made it possible to analyze genomewide nucleotide patterns which could never be detected before, such as the GC skew [1, 2]. Here we report novel periodic patterns found to exist widely in bacterial and eukaryote genome sequences. Our computer algorithm, designed to amplify and display even remote periodic patterns, detected adenine–adenine dinucleotide (ApA) periodicity in all of the completed procaryote genomes we analyzed, as well as in many of the eukaryote genomes.

Materials and Methods

Genome Sequences Used. We have analyzed ApA periodicity in each genome of 13 eubacteria, 3 archaebacteria, 16 chromosomes of *S. cerevisiae,* 9 other eukaryotes, 60 mitochondria, and 9 chloroplasts. The complete genome sequences of the following prokaryotes were down-loaded from the databases indicated: *H. influenzae* Rd [3], *M. genitalium* [4], *M. jannaschii* [5], *H. pylori* [6], *A. fulgidus* [7], *T. pallidum* [8], and *B. burgdorferi* [9] (TIGR Microbe Database; http://www.tig.org/tdb/mdb/mdb.html); *Synechocystis* PCC6803 [10] (Cyanobase; http://www.kazusa.or.jp/cyano/cyano.html); *M. pneumoniae* [11] (Mycoplasma Pneumoniae Genome Project; http://mail.zmbh.uni-heidelberg.de/M. pnemoniae/MP_Home.html); *E. coli* K-12 [12] (E. coli Genome Project; http://www.genetics.wisc.edu:80/index.html); and *M. thermoautotrophicum* [13] (The *M. thermoautotrophicum* Genome Database; http://www.biosci.ohio-state.edu/genomes/mthermo/index.html). The complete genome sequence of *S. cerevisiae* [14] was down-loaded from the Saccharomyces Genome Database (http://genome-www.stanford.edu/Saccharomyces/index.html). For other eukaryotes, since complete genomes are not currently available (as of October 1998), all DNA sequences in NCBI-GenBank Flat File Release 94.0 were combined. There were nine eukaryotes (excluding *S. cerevisiae*) whose total nucleotide length is longer than 10 million bp: *H. sapiens, M. musculus, C. elegans, A. thaliana, D. melanogaster, R. norvegicus, F. rubripes, O. sativa,* and *S. pombe.* Complete genome sequences of 60 mitochondria and 9 chloroplasts were also obtained from the GenBank database.

Correspondence to: M. Tomita; *e-mail:* mt@sfc.keio.ac.jp

(a)

```
..AA...A....AA..AA....AA...A....AA.......A......A...
AAA....AA..AA...A....AA..A.....AA.....AAAAA.....A.AA
...A....AA...A.....AA.A.A..AAA...A....AA....A..AAA..
.AA...AAAAAAAA..AA........AA..A.......A..A.....AA....
A.A.AA........A.A...........AA........AA.....A.A....
A.AA..A.....AA...........A..A..A....A.......A.A......
.A..A.....AA..A.A...AA......A.A....A..A.AAA.......AA
...A..AAAA........AA..A...A.AAA..A..
```

(b)

```
AA...A....AA..AA....AA...A....AA.......A......A...AAA.
AA..AA....AA...A....AA.......A......A...AAA....AA..AA.
AA....AA...A....AA.......A......A...AAA....AA..AA...A.
AA...A....AA.......A......A...AAA....AA..AA...A....AA.
AA.......A......A...AAA....AA..AA...A....AA..A.....AA.
AAA....AA..AA...A....AA..A.....AA.....AAAAA.....A.AA..
AA....AA..AA...A....AA..A.....AA.....AAAAA.....A.AA...
AA..AA...A....AA..A.....AA.....AAAAA.....A.AA...A....A
AA...A....AA..A.....AA.....AAAAA.....A.AA...A....AA...
AA..A.....AA.....AAAAA.....A.AA...A....AA...A.....AA.A
AAAAA.....A.AA...A....AA...A.....AA.A.A..AAA...A....AA
AAAA.....A.AA...A....AA...A.....AA.A.A..AAA...A....AA.
AAA.....A.AA...A....AA...A.....AA.A.A..AAA...A....AA..
AA.....A.AA...A....AA...A.....AA.A.A..AAA...A....AA...
AA...A....AA...A.....AA.A.A..AAA...A....AA....A..AAA..
AA...A.....AA.A.A..AAA...A....AA....A..AAA...AA...AAAA
AA.A.A..AAA...A....AA....A..AAA...AA...AAAAAAAA..AA...
AAA...A....AA....A..AAA...AA...AAAAAAAA..AA........AA.
AA...A....AA....A..AAA...AA...AAAAAAAA..AA........AA..
AA....A..AAA...AA...AAAAAAAA..AA........AA..A......A..
AAA...AA...AAAAAAAA..AA........AA..A......A..A.....AA.
AA...AA...AAAAAAAA..AA........AA..A......A..A.....AA..
AA...AAAAAAAA..AA........AA..A......A..A.....AA....A.A
AAAAAAAA..AA........AA..A......A..A.....AA....A.A.AA..
AAAAAAA..AA........AA..A......A..A.....AA....A.A.AA...
AAAAAA..AA........AA..A......A..A.....AA....A.A.AA....
AAAAA..AA........AA..A......A..A.....AA....A.A.AA.....
AAAA..AA........AA..A......A..A.....AA....A.A.AA......
AAA..AA........AA..A......A..A.....AA....A.A.AA.......
AA..AA........AA..A......A..A.....AA....A.A.AA........
AA........AA..A......A..A.....AA....A.A.AA........A.A.
AA..A......A..A.....AA....A.A.AA........A.A...........
AA....A.A.AA........A.A...........AA........AA.....A.A
AA........A.A...........AA........AA.....A.A....A.AA..
AA........AA.....A.A....A.AA..A.....AA...........A..A.
AA.....A.A....A.AA..A.....AA...........A..A..A.....A...
AA..A.....AA...........A..A..A....A.......A.A.......A..
AA...........A..A..A....A.......A.A.......A..A.....AA..
```

(c)

```
         *
         *
         *
         *         *
         *         *
         *         *         *         *
         *         *         *         *         *
         *         *         *         *         *
         *         *         *        **         *
         *         *         **       ***        **
 *       **        **        **  *    ***        **
 * * *   **       ***        **  *    ***       ***
 *****  ***  ***  ***        **  * ** ***       ***
 *****  ***  ***  ***    *  ****** ******   *   ***
 *********** **** ***  ***  ************* *** * ***
 ********************* ********************** * *****
 ****************************************************
-----------------------------------------------------
```

Fig. 1. **a** Example sequence. **b** Subsequent sequence after each ApA. **c** Distribution of ApAs after each ApA.

Algorithm. ApA frequency at each distance from the previous ApA dinucleotide within 50 bp was counted in order to analyze the correlation between frequency and distance. To illustrate our algorithm, consider the example sequence in Fig. 1. Only adenine nucleotides are shown. This sequence has a 10-bp periodicity of ApA dinucleotides, and the following demonstrates how our method can effectively amplify and visualize the periodicity.

We first list the downstream sequences (50 bp) of all the ApA dinucleotides (Fig. 1b), then count the ApA dinucleotides for each position (column) on this list. A clear pattern of ApA dinucleotides with 10-bp periodicity can be observed (Fig. 1c).

These counted numbers are normalized and converted to frequencies as percentages. To offset the potential periodicity of 3 bp caused by codon biases, we take the averages of three consecutive base positions at each position (with its adjacent positions).

This algorithm was applied to analyze the genomes of all the species mentioned above. We also analyzed coding and noncoding regions separately.

Results

The ApA periodic patterns of the 16 prokaryote species are shown in Fig. 2, where solid lines represent the results of coding sequences; dotted lines, the results of noncoding sequences; and dotted–dashed lines, the results of all sequences. For the graphs to be on the same scale, some graphs extend beyond the figure frames.

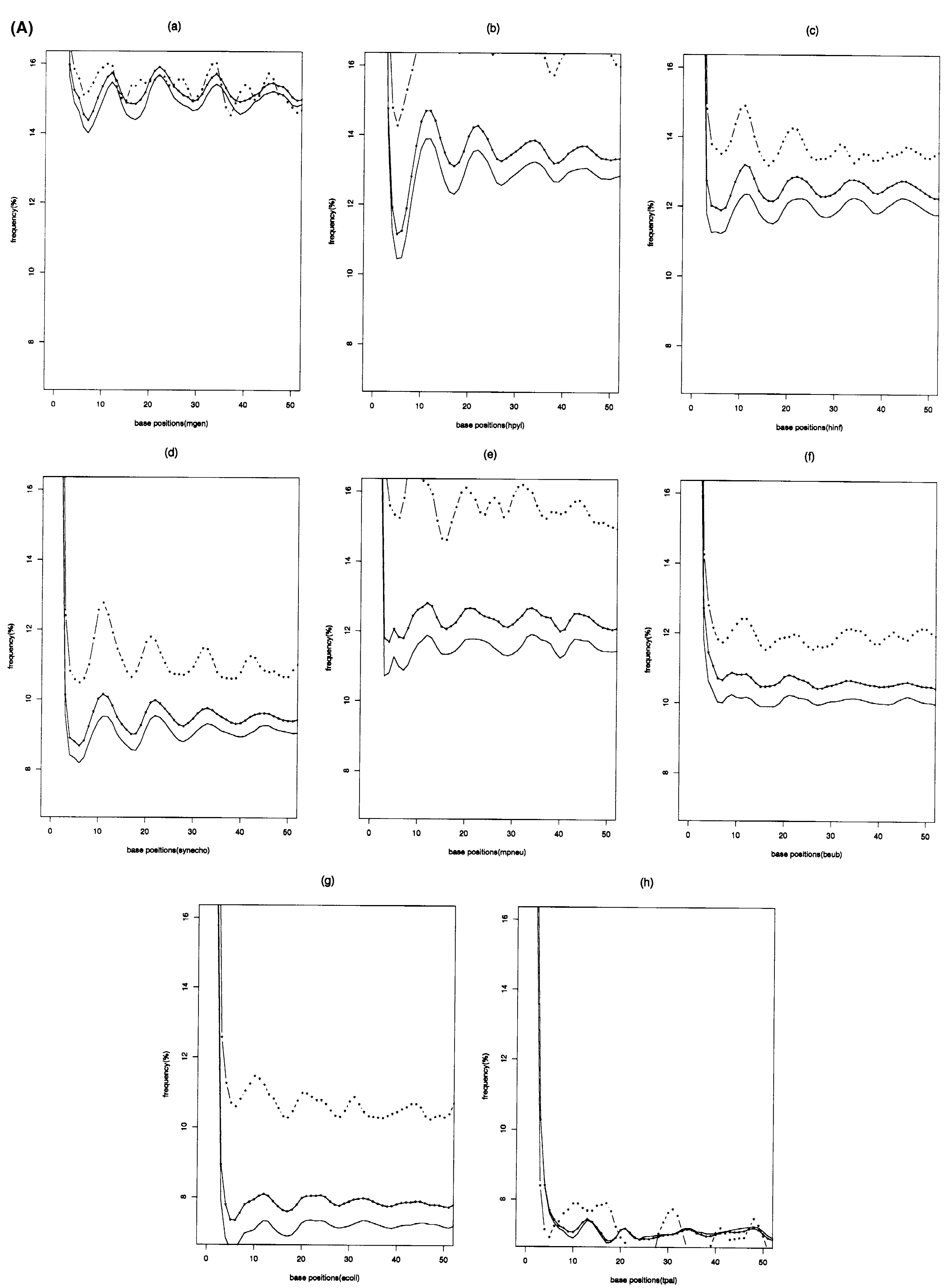

Fig. 2. ApA dinucleotide periodic patterns in prokaryote genomes. Solid lines represent results with only coding sequences; dotted lines, results with only noncoding sequences; and dotted–dashed lines, results with total sequences. All show clear periodic patterns. **a** *Mycoplasma genitalium;* **b** *Helicobacter pylori;* **c** *Haemophilus influenzae* Rd; **d** *Synechocystis* PCC6803; **e** *Mycoplasma pneumoniae;* **f** *Bacillus subtilis;* **g** *Escherichia coli;* **h** *Treponema pallidum;* **i** *Methanococcus jannaschii;* **j** *Archaeoglobus fulgidus;* **k** *Methanobacterium thermoautotrophicum;* **l** *Aquifex aeolicus;* **m** *Borrelia burgdorferi;* **n** *Chlamydia trachomatis;* **o** *Mycobacterium tuberculosis* H37Rv; **p** *Pyrococcus horikoshii.*

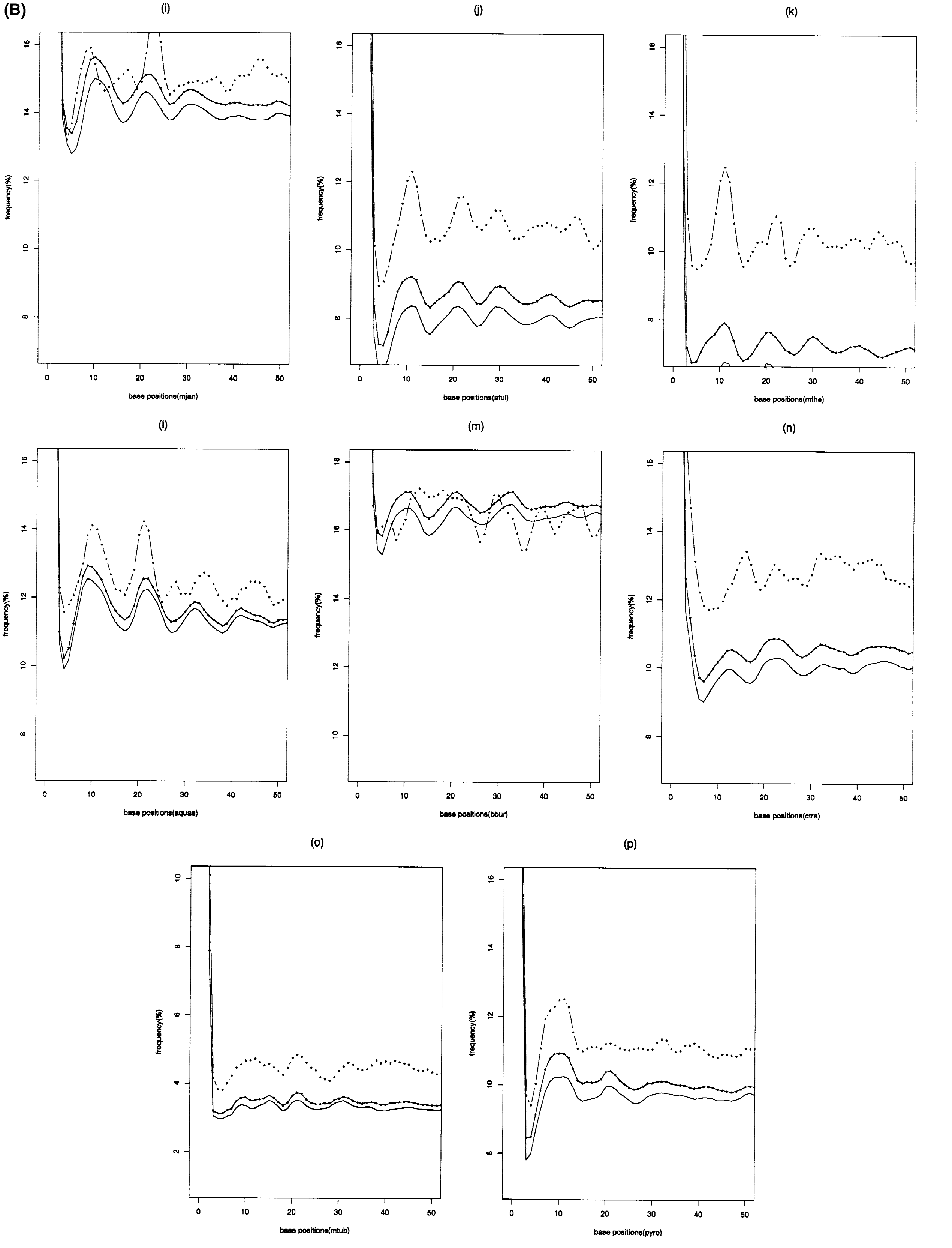

Fig. 2. Continued.

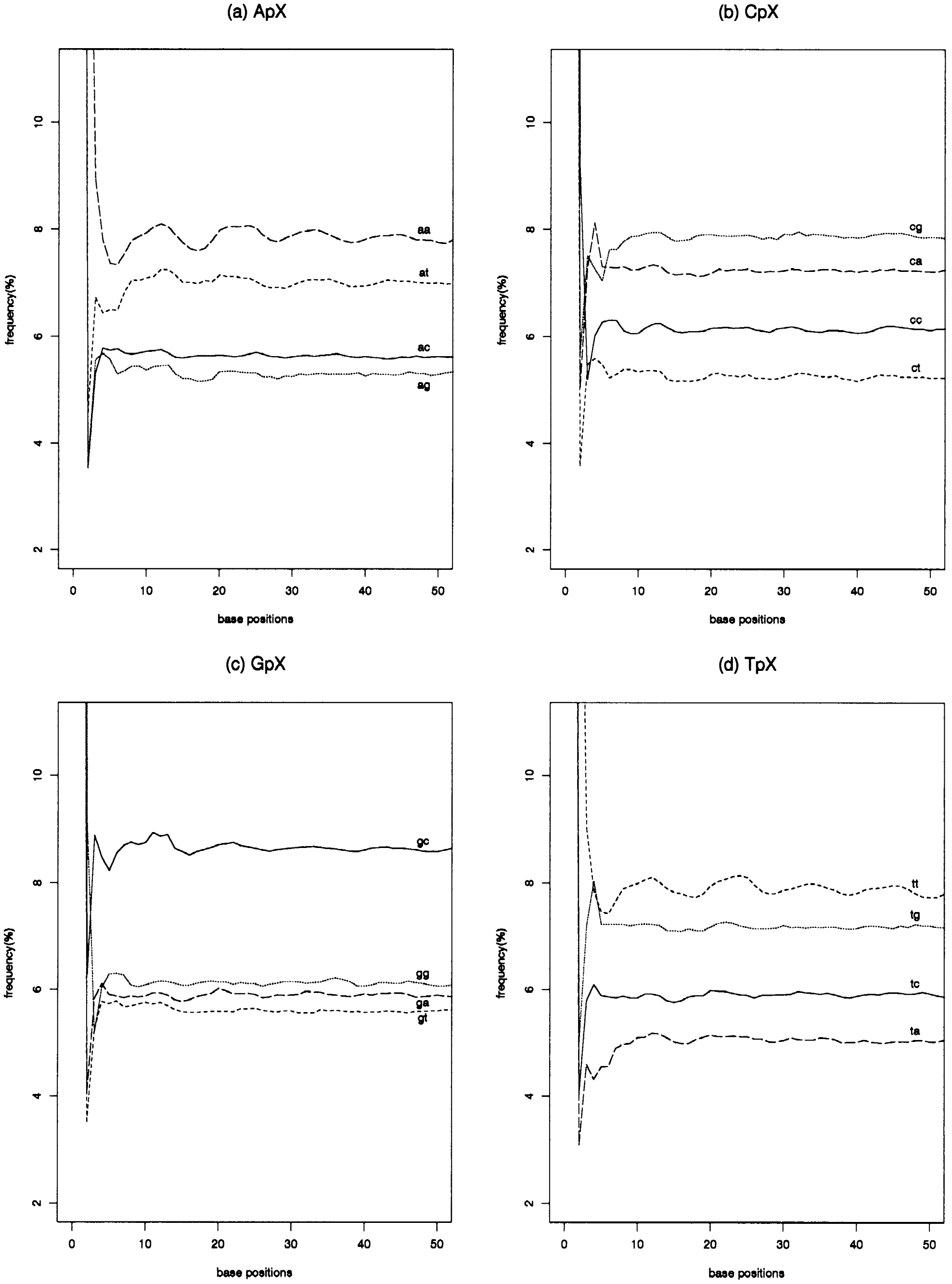

Fig. 3. Dinucleotide periodic patterns in the *E. coli* genome. Only ApA and TpT show clear periodic patterns. **a** ApA, ApC, ApG, and ApT; **b** CpA, CpC, CpG, and CpT; **c** GpA, GpC, GpG, and GpT; **d** TpA, TpC, TpG, and TpT.

M. genitalium, H. pylori, H. influenzae, and *Synecobacter* sp. show a clear periodicity of about 11 bp (Figs. 2a–d), with an amplitude of more than 1%. There is no significant difference in periodic pattern between coding and noncoding regions, indicating that the periodicity is not due to patterns in amino acid sequences.

M. pneumoniae shows a similar pattern, except that, for unknown reasons, a small peak occurs around the fifth base (Fig. 2e). *B. subtilis, E. coli,* and *T. pallidum* also show a periodicity of about 11 bp (Figs. 2f–h), although their amplitudes are significantly lower (less than 1%). The archaebacteria *M. jannaschii, A. fulgidus,* and *M. thermoautotrophicum* have amplitudes higher than 1%, but a shorter periodicity of 10 bp (Figs. 2i–k).

We did the same analysis (with the *E. coli* genome) for all 16 dinucleotides (Fig. 3). Among the 16, ApT, TpA, and GpC appear to have a slight pattern of periodicity. However, none shows a periodic pattern as evident as the patterns of ApA and TpT.

Figure 4 shows the ApA dinucleotide periodic pat-

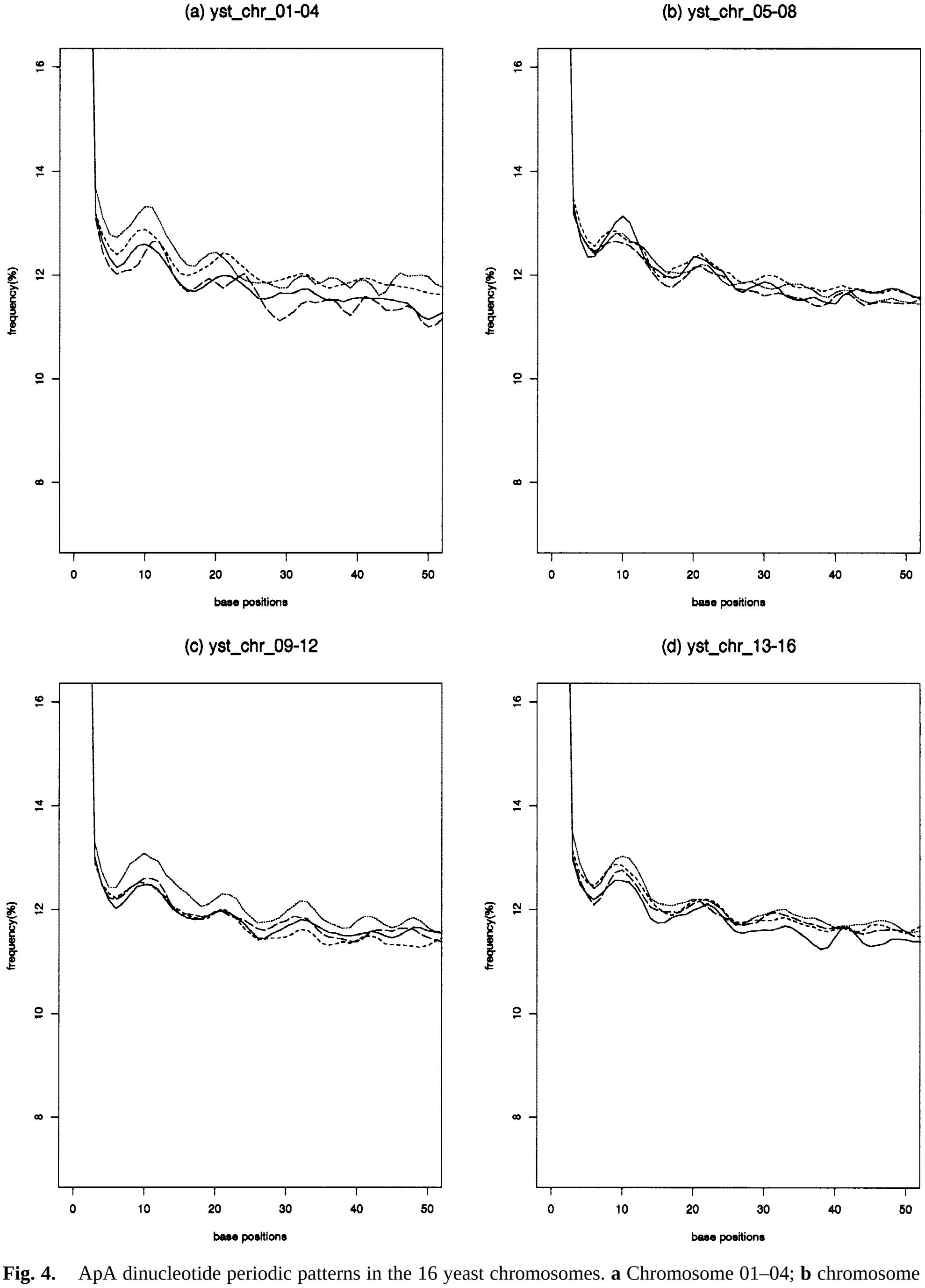

Fig. 4. ApA dinucleotide periodic patterns in the 16 yeast chromosomes. **a** Chromosome 01–04; **b** chromosome 05–08; **c** chromosome 09–12; **d** chromosome 13–16.

terns of the 16 chromosomes of *S. cerevisiae.* It has, in general, a periodicity of about 10.5 bps, although it varies among the 16 chromosomes.

Figure 5 shows the results for nine other eukaryotes. The periodic pattern of *S. pombe* (Fig. 5a) is much weaker than that of *S. cerevisiae.* With respect to plants, *A. thaliana* also shows weak periodicity, while *O. sativa* (Figs. 5b and c) shows no periodic pattern.

There appears to be a tendency among higher animals to show less evident periodic patterns. *C. elegans* (Fig. 5d) has one of the most evident periodic patterns among the animals we examined. *D. melanogaster* (Fig. 5e) has a less evident but still clear periodic pattern. *F. rubripes* (Fig. 5f) is an interesting case in which the coding region shows a very strong periodic pattern, while the noncoding region shows little. The three mammals, *R. norvegicus* (Fig. 5g), *M. musculus* (Fig. 5h), and *H. sapiens* (Fig. 5i), reveal minimal periodic patterns in the coding and noncoding regions.

No clear periodicities of 10–11 bp are observed in the

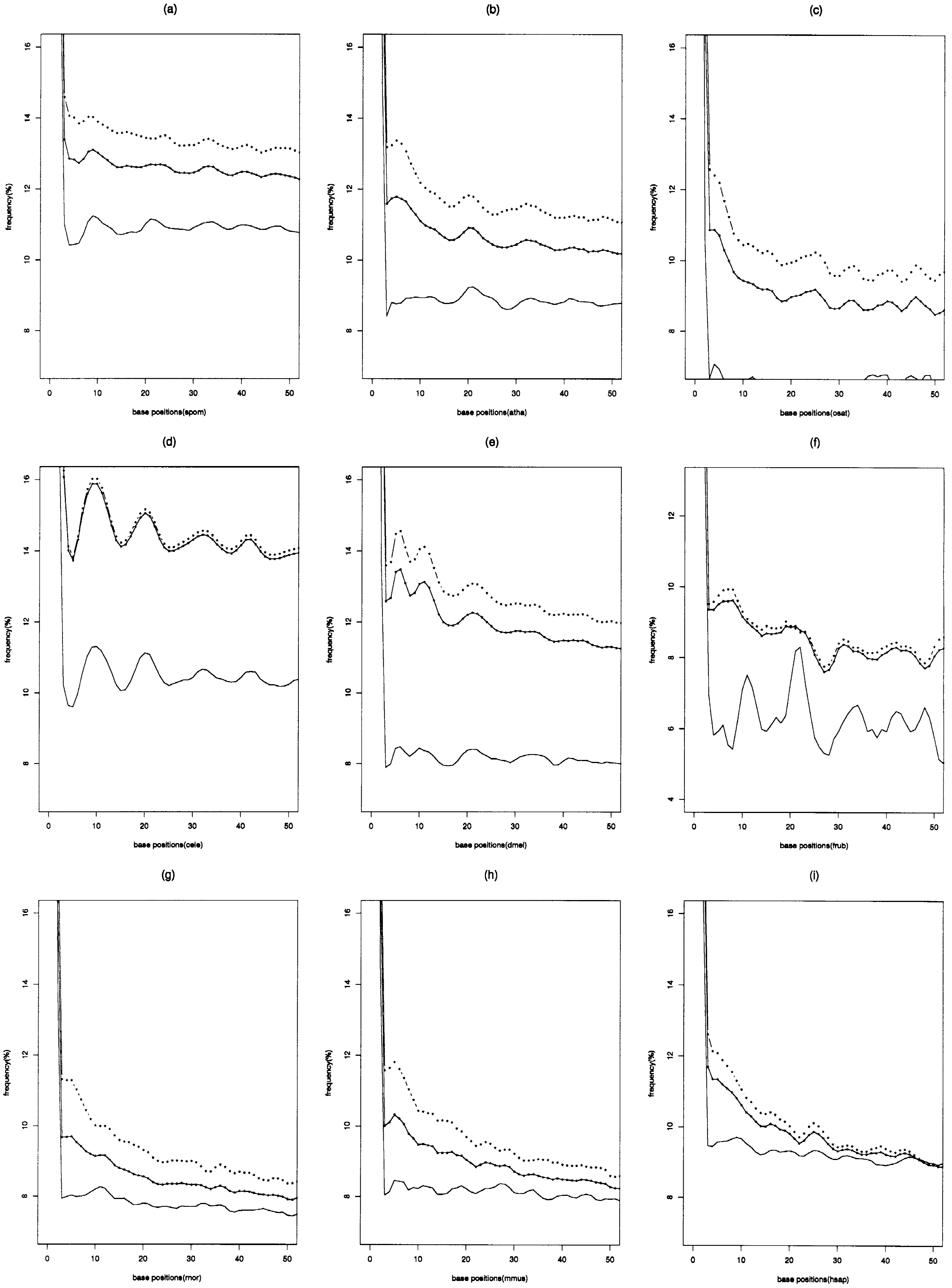

Fig. 5. ApA dinucleotide periodic patterns in eukaryote genomes. Solid lines represent results with only coding sequences; dotted lines, results with only noncoding sequences; and dotted–dashed lines, results with total sequences. Higher animals show weaker patterns. **a** *S. pombe;* **b** *A. thaliana;* **c** *O. sativa;* **d** *C. elegans;* **e** *D. melanogaster;* **f** *F. rubripes;* **g** *R. norvegicus;* **h** *M. musculus;* **i** *H. sapiens.*

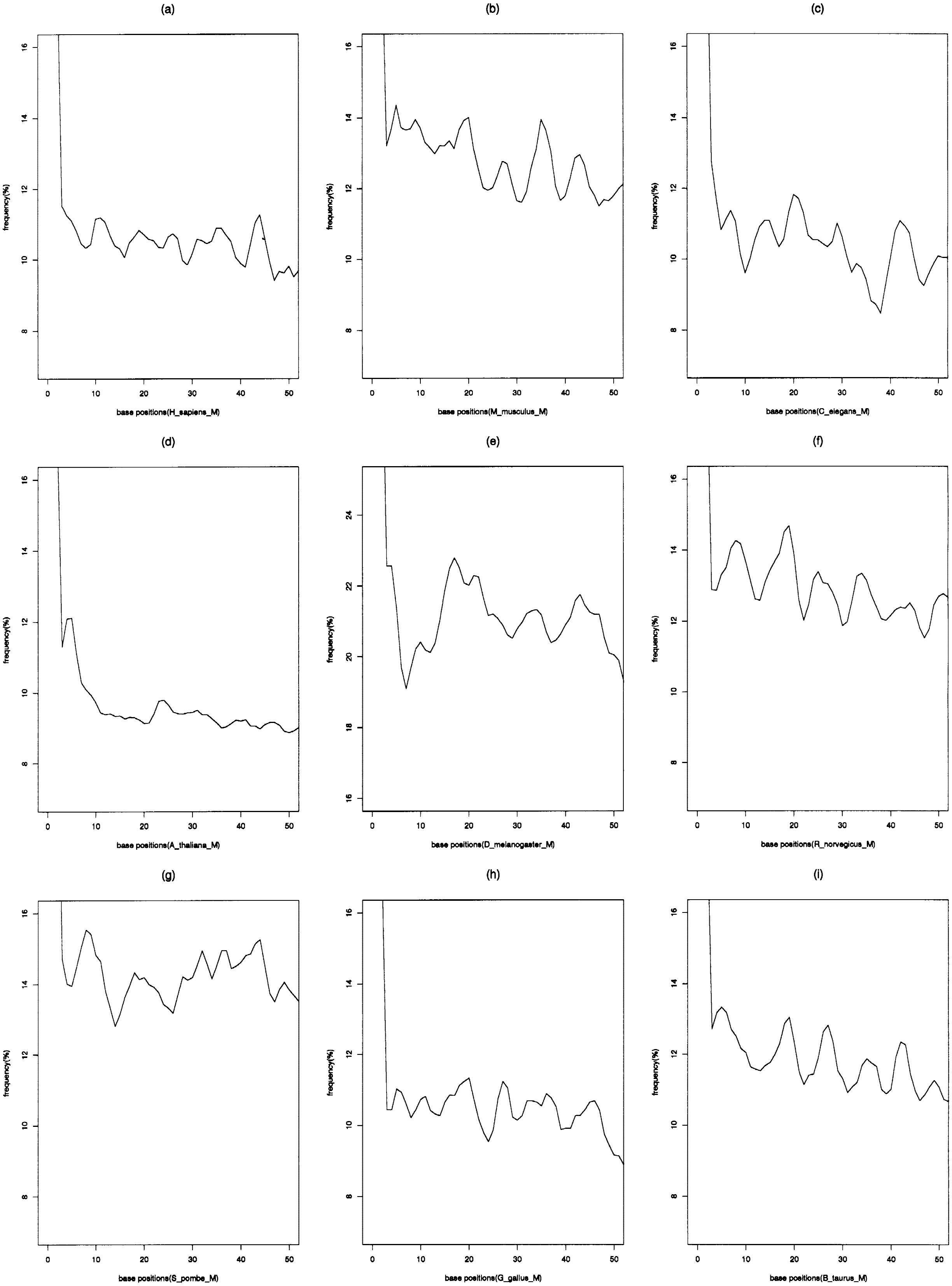

Fig. 6. ApA dinucleotide periodic patterns in mitocondorion genomes. **a** *H. sapiens;* **b** *C. elegans;* **c** *A. thaliana;* **d** *D. melanogaster;* **e** *R. norvegicus;* **f** *S. pombe;* **g** *G. gallus;* **h** *B. taurus;* **i** *P. anserino.*

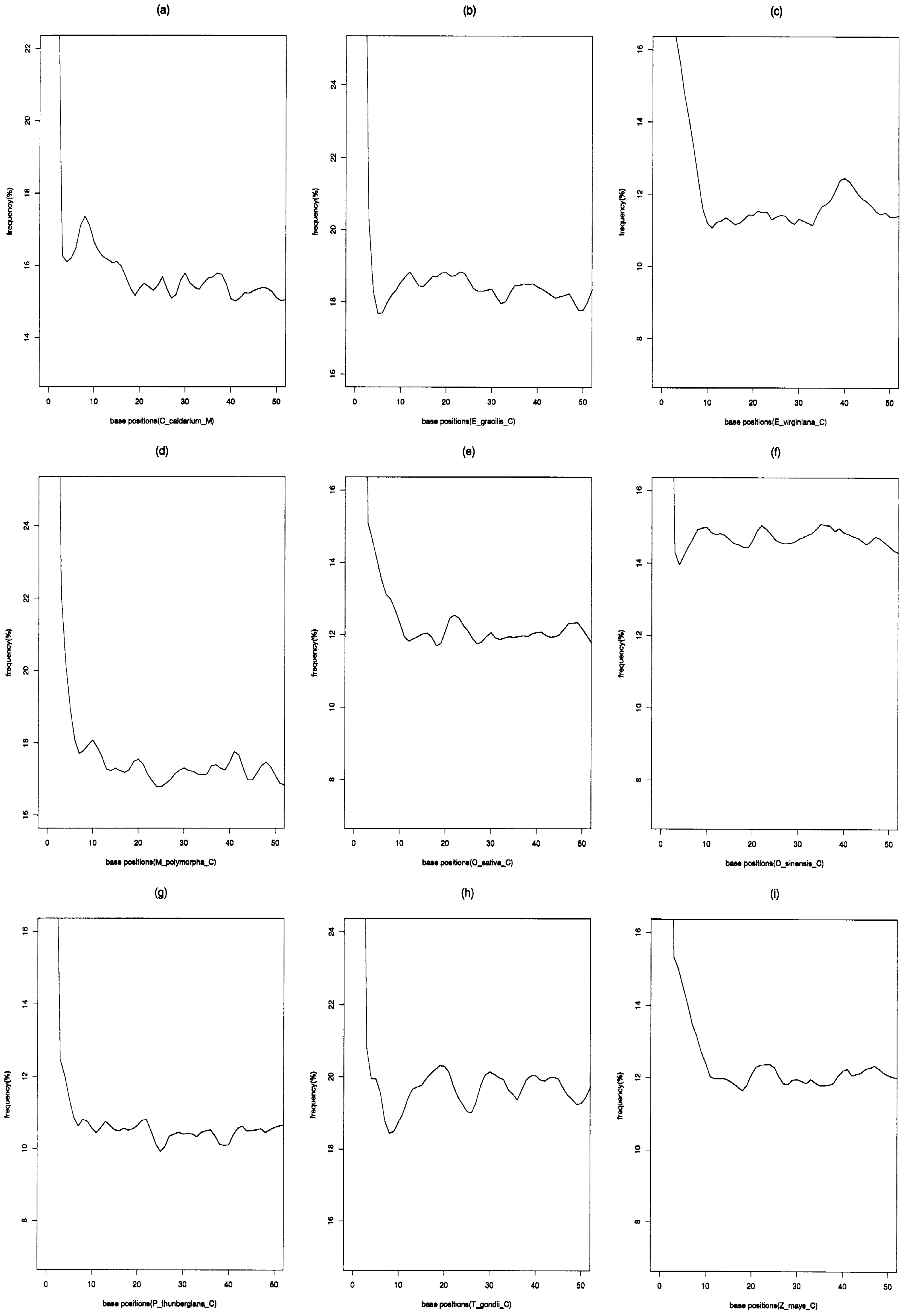

Fig. 7. ApA dinucleotide periodic patterns in chroloplast genomes. **a** *C. caldarium;* **b** *E. gracilis;* **c** *E. virginiana;* **d** *M. polymorpha;* **e** *O. sativa;* **f** *O. sinensis;* **g** *P. thunbergiana;* **h** *T. gondii;* **i** *Z. mays.*

genomes of the 60 mitochondria (Fig. 6; only 9 of 60 are shown) and the 9 chloroplasts (Fig. 7).

Discussion

DNA molecules are in some way anisotropic, due to the variation of base pairs along a DNA double helix. That is, some adjacent base pairs are slightly nonparallel, causing bending of the DNA axis [15]. If several positions of DNA are bent in the same direction at regular intervals, global bending of DNA occurs. In particular, the global bending effect is most prominent when the interval is the same as the helical period, which is about 10.55 bp [16]. It has indeed been shown by in vitro experiments that AA/TT dinucleotides at the periodicity of 10–11 bp cause unidirectional curvature of the DNA [17]. It can therefore be inferred that the 10- to 11-base periodicities we have observed in various genomes are related to DNA bending and thus to the three-dimensional (3D) structure of chromosomes.

Nucleosome sequences in eukaryote genomes were reported to display periodic patterns of ApA dinucleotides, which are believed to be related to DNA bending wrapped around histone octamers in the nucleosome structure [18–22]. It is interesting that our analysis found the periodic patterns of ApA dinucleotides not only in the genomes of eukaryotes, but also in those of eubacteria, whose chromosomes do not involve histone octamers. Recently, Herzel et al. [23, 24] independently performed a similar analysis on bacterial genomes and they also found a nucleotide periodicity of 10–11 bp. Our analysis found that the periodic patterns in many eubacteria (Fig. 2) are more prominent than those in most eukaryotes (Fig. 5). Since eubacterial chromosomes do not possess histone octamers; they may need to rely more on the nucleotide sequence patterns for 3D structure formation of their chromosomes.

Higher eukaryotes such as *H. sapiens* and *M. musculus* show only a slight periodic pattern of the ApA dinucleotide, while lower eukaryotes such as *C. elegance* and *S. cerevisiae* show prominent periodicity, as do prokaryotes. While the reason for this difference is not clear, we can think of two possible explanations. First, gene expression is often regulated by the 3D structure of DNA, and higher eukaryotes involve more complicated gene regulation. This implies that DNA structure dynamically changes more often, lessening the need for a static signal for DNA curvature. Second, genomes of higher eukaryotes are full of transposons, viruses, and other foreign genomic elements, and they may disrupt the periodic patterns of nucleotide sequences.

The difference in periodicity between eubacteria (about 11 bp) and archaebacteria (about 10 bp) is presumably due to the supercoiling of DNA helical structures, as suggested by Herzel et al. [23]. According to Crick's formula for helical DNA trajectories [25], helical periods above 10.55 bp induce negative supercoiling of DNA and periods below 10.55 bp induce positive supercoiling (free DNA has a helical period of 10.55 bp). Thus, the ApA periodicity of about 10 bp in archaebacterial genomes suggests that their DNA has helical periods of about 10 bp and therefore it can be predicted from this periodicity that archaebacterial DNA is positively supercoiled [23]. Although no evidence has been reported for positive supercoiling in the genomic DNA of archaea, there are some strong indications for it. First, it has been found that archaeal plasmids and a virus-like particle from *Sulfologus* are positively supercoiled [26, 27], Second, hyperthermophilic archaea have reverse gyrase, a topoisomerase which introduces positive superhelical turns into DNA [28]. Finally, histonelike proteins found in archaea [29] are reported to bind DNA and result in positive supercoiling [30].

For the genomes of chloroplasts and mitocondria, no clear periodic patterns of 10–11 bp are observed. Some mitochondria, on the other hand, appear to have 7-bp periodicities. However, the significance of these observed patterns is questionable because of their genome sizes, which are one or two orders of magnitude smaller than those of the previously discussed genomes. When we analyzed the first 10 kb of the *E. coli* genome sequence, no clear periodic pattern was observed. Further studies found that 200–300 kb of data is required for the reliable determination of ApA periodcity (data not shown). Since most organella genomes have lengths of less than 200 kb, no conclusion should be drawn from the results of ApA periodicities in these genomes.

Summary

We have conducted comprehensive analyses of ApA dinucleotide periodicities in the genomes of a variety of species. Eubacteria show clear periodicities of about 11 bp, while archaebacteria show shorter periodicities of about 10 bp. *S. cerevisiae* has an average periodicity of 10.5 bp, although periodicities vary among the 16 chromosomes. Higher eukaryotes tend to show less evident periodicities, and mitochondria and chloroplasts show no clear periodicities. We have argued biological significance of these periodicities in relation to the 3D structure of DNA, such as nucleosome and chromosome formation and DNA supercoiling.

References

1. Freeman JM, et al. (1998) Patterns of genome organization in bacteria. Science 279:1827
2. Lobry JR (1996) Origin of replication of Mycoplasma genitalium. Science 272:745–746

3. Fleischmann RD, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512
4. Fraser CM, et al. (1995) The minimal gene complement of Mycoplasma genitalium. Science 270:397–403
5. Blut CJ, et al. (1996) Complete genome sequence of the methanogenic archeon, Methanococcus jannaschii. Science 273:1058–1073
6. Tomb JF, et al. (1997) The complete genome sequence of the gastric pathogen Helicobacter pylori. Nature 388:539–547
7. Klenk HP, et al. (1997) The complete genome sequence of the hyperthermopholic, sulphate-reducing archaeon Archaeoglobus fulgidus. Nature 390:364–370
8. Fraser CM, et al. (1998) Complete genome sequence of Treponema pallidum, the syphilis spirochete. Science 281:375–388
9. Fraser CM, et al. (1997) Genomic sequence of a Lyme disease spirochete, Borrelia burgdorferi. Nature 390:580–586
10. Kaneko T, et al. (1996) Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res 3:109–136
11. Himmelreich R, et al. (1996) Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res 24:4420–4449
12. Blattner FR, et al. (1997) The complete genome sequence of Escherichia coli K-12. Science 277:1453–1474
13. Smith DR, et al. (1997) Complete genome sequence of Methanobacterium thermoautotrophicum delta H: Functional analysis and comparative genomics. J Bacteriol 179:7135–7155
14. Goffeau A, et al. (1997) The yeast genome directory. Nature 387:5–105
15. Trifonov EN, Sussman JL (1980) The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc Natl Acad Sci USA 77: 3816–3820
16. Trifonov EN (1998) 3-, 10.5-, 200- and 400-base periodicities in genome sequences. Physica A 249:511–516
17. Ulanovsky L, et al. (1986) Curved DNA: Design, synthesis, and circularization. Proc Natl Acad Sci USA 83:862–866
18. Trifonov EN (1980) Sequence-dependent deformational anisotropy of chromatin DNA. Nucleic Acids Res 8:4041–4053
19. Drew HR, Travers AA (1985) DNA bending and its relation to nucleosome positioning. J Mol Biol 186:773–790
20. Ioshikhes I, et al. (1992) Prefered positions of AA and TT dinucleotides in aligned nucleosomal DNA sequences. J Biomol Struct Dyn 9:1111–1117
21. Ioshikhes I, et al. (1996) Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. J Mol Biol 262:129–139
22. Bolshoy A, et al. (1997) Enhancement of the nucleosomal pattern in sequences of lower complexity. Nucleic Acids Res 25:3248–3254
23. Herzel H, et al. (1998a) Sequence periodicity in complete genomes of archaea suggests positive supercoiling. J Biomol Struct Dyn 16:341–345
24. Herzel H, et al. (1998b) Interpreting correlations in biosequences. Physica A 249:449–459
25. Crick FHC (1976) Linking numbers and nucleosomes. Proc Natl Acad Sci USA 73:2639–2643
26. Lopez-Garcia P (1997) DNA topology in hyperthermophilic Archaea: Reference states and their variation with growth phase, growth temperature, and temperature stresses. Mol Microbiol 23: 1267–1279
27. Nadal M (1986) Positively supercoiled DNA in a virus-like particle of an archaebacterium. Nature 321:256–258
28. Kikuchi A, Asai K (1984) Reverse gyrase—A topoisomerase which introduces positive superhelical turns into DNA. Nature 309:677–681
29. Pereira SL, et al. (1997) Archaeal nucleosomes. Proc Natl Acad Sci USA 94:12633–12637
30. Musgrave DR, et al. (1991) DNA binding by the archaeal histone HMf results in positive supercoiling. Proc Natl Acad Sci USA 88:10397–10401