Genomic Signature: Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences

Patrick J. Deschavanne, Alain Giron,† Joseph Vilain,† Guillaume Fagot,† and Bernard Fertil†*

*Laboratoire de Mutagénèse, Institut Jacques Monod, Paris, France; and †Unité INSERM 494, Paris, France

We explored DNA structures of genomes by means of a new tool derived from the ''chaotic dynamical systems'' theory (the so-called chaos game representation [CGR]), which allows the depiction of frequencies of oligonucleotides in the form of images. Using CGR, we observe that subsequences of a genome exhibit the main characteristics of the whole genome, attesting to the validity of the genomic signature concept. Base concentrations, stretches (runs of complementary bases or purines/pyrimidines), and patches (over- or underexpressed words of various lengths) are the main factors explaining the variability observed among sequences. The distance between images may be considered a measure of phylogenetic proximity. Eukaryotes and prokaryotes can be identified merely on the basis of their DNA structures.

Introduction

The recent availability of long genomic sequences opens a new field of research devoted to the analysis of their structure (Beutler et al. 1989; Woese, Kandler, and Wheelis 1990; Charlesworth 1994; Sharp and Matassi 1994; Doolittle 1997; Maley and Marshall 1998). Genomic sequences have already been analyzed with regard to similarities and differences in the relative abundances of oligonucleotides (called ''words'' hereafter) up to four bases long (Phillips, Arnold, and Ivarie 1987; Beutler et al. 1989; Deschavanne and Radman 1991; Bhagwat and McClelland 1992; Burge, Campbell, and Carlin 1992; Karlin, Burge, and Campbell 1992; Blaisdell et al. 1993; Karlin and Burge 1995; Gelfand and Koonin 1997; Karlin, Mràzek, and Campbell 1997). Significant differences in terms of di- and tetranucleotide frequencies were found among genomes of different phyla, allowing the derivation of partial-ordering relationships and leading to the genomic signature concept (Karlin and Burge 1995; Karlin, Mràzek, and Campbell 1997). Meanwhile, singular short-word frequencies have been reported for various species and shown to be species-specific. The counterselection of the CpG dinucleotide in vertebrates and the overrepresentation of the octonucleotide HIP1 in cyanobacteria are two typical examples of this phenomenon (Josse, Kaiser, and Kornberg 1961; Robinson et al. 1995). It was shown recently that the dinucleotide relative abundance values vary less along a genome than among species and that closely related organisms display more similar dinucleotide composition than do distant organisms (Karlin, Mràzek, and Campbell 1997). As an explanation, it has been proposed that DNA structure results in part from an autoorganization process in which DNA replication, recombination, and repair, together with local physicochemical constraints, play predominant roles. In addition, environmental factors contribute by means of natural selection to the current structure of each DNA sequence. Both kinds of factors act on the very basis of genome structure, the usage and ordering of the four bases, and together they build up the genome signature.

If one is interested in words that are 8 nt long, about 65,000 (4^8) different words need to be screened for each genome. The chaos game representation (CGR) of DNA sequence offers, in the form of fractal images, a very handy approach for dealing with such a large amount of data (Jeffrey 1990). We developed a version of the method that allows quantification of observed patterns and fast treatment of very long sequences. Genomic comparisons involving parts of the genome or the whole genome, the detection of special genome features, and the construction of molecular phylogenies are the main issues addressed in this paper.

The Making and Handling of CGR Images

The original CGR method is an algorithm which produces pictures revealing patterns in DNA sequences (Jeffrey 1990, 1992; Burma et al. 1992; Dutta and Das 1992; Hill, Schisler, and Singh 1992; Goldman 1993; Oliver et al. 1993). Basically, the whole set of frequencies of the words found in a given genomic sequence can be displayed in the form of a single image in which each pixel is associated with a specific word (fig. 1). Frequencies of words found in a sequence are displayed in a square image, with the location of a given word being chosen according to a recursive procedure. Thus, the image is divided into four quadrants in which sequences ending with the appropriate base are collected. This gives the base composition of the sequence (fig. 1, upper left panel). The quadrants were chosen in such a way that the lower (A+T) and upper (G+C) halves indicate the base composition and the diagonals indicate the purine/pyrimidine composition. Each quadrant is subsequently divided into four subquadrants, each containing sequences ending with a given dinucleotide (fig. 1), such that sequences differing only in the first letter are in adjacent subquadrants. The sequence is read base by base so that all available words are considered. Word frequencies are displayed by the intensity of each pixel, and a pseudo-three-dimensional representation of the

Key words: structure of genome, genomic signature, characterization of species, frequencies of oligonucleotides, image.

Address for correspondence and reprints: Patrick J. Deschavanne, Unité INSERM 494, 91 Bvd de l'Hôpital 75013, Paris, France. E-mail: deschavanne@imed.jussieu.fr.

Mol. Biol. Evol. 16(10):1391–1399. 1999

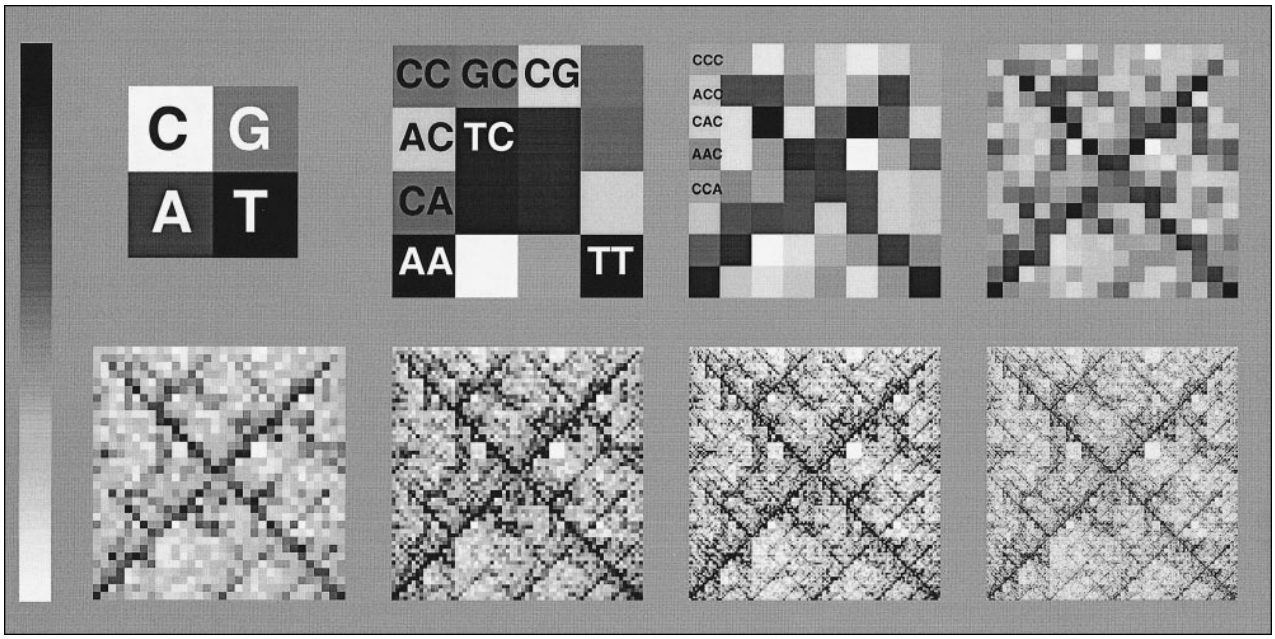


FIG. 1.—The fractal nature of chaos game representation (CGR) images. The frequencies of words up to eight letters long used by the archaebacteria *Archeoglobus fulgidus* are represented from left to right and from top to bottom. For single-letter words, frequencies of letters from only one strand are represented. The gray scale is fitted to the frequency values in order to use its full range of variation for each CGR image.

image highlights unusual frequencies (fig. 2). The gray scale indicates the relative frequency per image of each word: the darker the pixel, the greater the frequency. A colored—and more vivid—version of all pictures can be seen on our web site: http://www.imed.jussieu.fr/u494/equipe_2/mbe.html. Since our method is strand-dependent, the original sequence and its complementary version are concatenated before analysis.

Qualitative and quantitative expressions of the order, regularity, structure, and complexity of DNA sequences are obtained from the CGR image, which simultaneously displays both local and global patterns of the sequence. A Euclidian metric is used to evaluate similarities between images. Each image can be associated with a point in a 4^n-dimensional space (n and 4^n are the size of words and the number of words, respectively), the coordinates of which are the frequencies of words (each word corresponds to a dimension). Comparison between two images can subsequently be achieved by means of the distance between their corresponding points, which characterizes the difference of word frequencies between sequences such that a small distance between images indicates that words are similarly used in the corresponding sequences. Principal component analysis (PCA) was chosen to help display images (and therefore sequences) in plots in which distances among images best match the actual values. PCA provides results under the form of privileged axes (called the principal components), which allows construction of meaningful plots. Examination of these plots usually gives a deeper understanding of the mechanisms underlying the structure of the data. A thorough discussion of PCA can be found in, e.g., Kendall (1975).

The genomic sequences required for our studies were gathered from several data banks, including TIGR, Sanger Institute, and GenBank. Complete genomes, complete chromosomes, and sequences at least 10 kb long were analyzed. Sequences shorter than 100 kb from the same species were combined to allow for meaningful comparison with long sequences from other species. The CGR images were processed on Apple Macintosh computers using a set of programs made with C++, Hypercard, and Matlab. Typical runtime was about 2 s for the processing of the two strands of a 1-Mb sequence, whatever the resolution of CGR image.

Mastering CGR Images

The dictionary of all words 1–8 nt long for the 2.2-Mb genome of *Archeoglobus fulgidus* is shown in figure 1. The resolution of the CGR image is related to the lengths of words: for example, frequencies of the 16 dinucleotides (two-letter words) are displayed on a 4×4 pixel image, while a 256×256 pixel image is required for eight-letter words. It should be pointed out that the length of the sequence determines the operative value of the corresponding image. For example, for seven-letter words, a two-stranded sequence of 100 kb will lead to 12 occurrences of each word on average. If length of sequences is such that the mean density per pixel is small (lower than four), overexpressed words can still be observed, whereas counterselected words can no longer be detected.

The most striking feature of these images is the structure that progressively appears and sharpens as a function of word length. The main structure is quite well established for words longer than 5 nt. It essentially re-

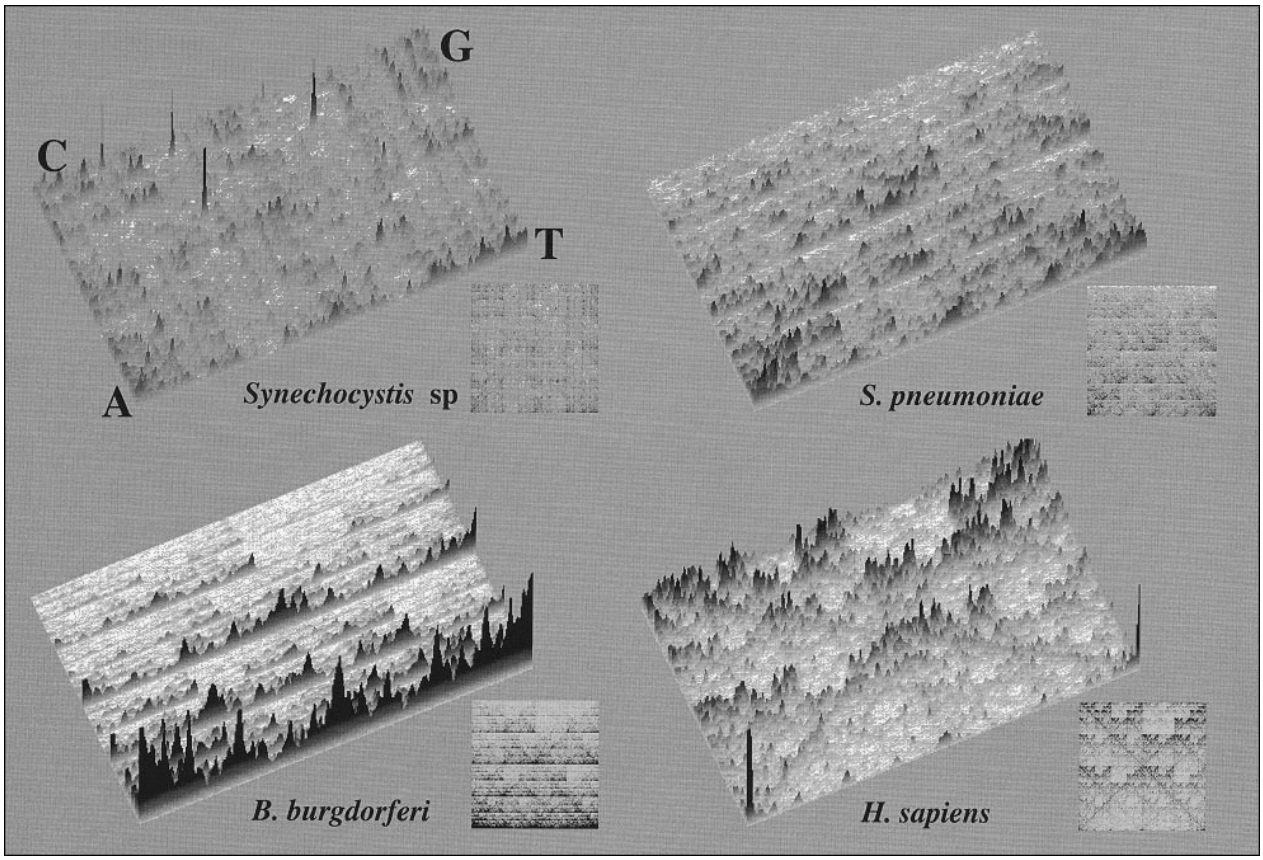


FIG. 2.—Three-dimensional display of seven-letter frequencies of four sequences (100 kb each). The corresponding two-dimensional chaos game representation images are given in the insert.

sults from an abundance of words composed of purine and pyrimidine stretches, clearly visible as diagonally oriented lines, and a deficit in some tetranucleotide words, appearing as white squares. The white square corresponding to the word CTAG in the *A. fulgidus* genome illustrates such a deficit (fig. 1). Moreover, words conserve frequency properties of imbedded shorter words, leading to the fractal nature of the image.

The frequencies of words found in a given species may undergo strong variations (table 1). Some words may never be encountered in a given genome (about 300 eight-letter words are represented only once or not at all in the genome of *Clostridium acetobutylicum*), while others are very common (the highest frequencies can reach 60 times the median; table 1). Given the strong distortion in the high frequency domain—a small number of words are heavily represented—an adequate display of the images requires a little overhead processing. A low-pass cutoff and a log transformation applied successively to the word frequency were found to shape the data properly, so that those patterns are well displayed. Extremely high frequencies cannot subsequently be observed as such from these two-dimensional CGR plots. In contrast, a pseudo-three-dimensional representation of these images displays the full range of word frequencies found in a given sequence and points out outliers (fig. 2). The *Synechocystis* sp. image is a good example of the power of this analysis. The three-dimensional representation reveals peaks of frequencies far above the mean. In fact, four peaks are observed in this image,

Table 1
Distribution of Eight-Letter Words in Terms of Quantiles for Seven Species

SPECIES	SIZE (Mb)	QUANTILES									
		Mean	Minimum	0.50%	2.50%	10%	Median	90%	97.50%	99.50%	Maximum
Escherichia coli	4.65	142	0	3	13	37	115	282	415	596	1,503
Bacillus subtilis	4.23	129	2	12	19	35	98	253	422	qr 670	1,764
Mycobacterium tuberculosis	4.46	136	0	1	4	12	73	323	656	1,155	3,573
Clostridium acetobutylicum	4.03	123	0	1	2	8	55	300	654	1,202	3,198
Archeoglobus fulgidus	2.20	67	0	2	6	15	51	136	223	329	798
Mus musculus	1.31	40	0	0	1	4	34	83	127	201	2,805
Homo sapiens	1.02	31	0	0	0	2	23	70	116	201	2,595

Note.—Meaning of values is as follows: for example, 0.50% of the words (i.e., 327 words = 65,536 × 0.5%) appear ≤3 times in the genome of *E. coli*.

illustrating the high frequency of four seven-letter words. It has been noted that this particular species exhibits a very strong selection for the palindromic decanucleotide GGCGATCGCC, which shows up as four seven-letter words (Robinson et al. 1995). The two- and three-dimensional representations are complementary, and both are needed to establish a genomic signature.

Genomic Signature: Variability Along a Genome and Among Genomes

One may ask: to what extent do CGR images vary along the genome? In fact, images obtained from parts of a genome present the same structure as that of the whole genome (fig. 3*a*). The pictures of the short sequences are blurred compared with that of the long sequence but remain similar whatever the part of the genome studied. Local modifications observed among CGR images from different parts of a genome obey the rules of the general design of the whole genome. The variation of CGR images along a genome and/or between genomes can be evaluated in terms of distance between them. In general, distance between images along a genome was smaller than distance among genomes (fig. 3*b*). The mean intragenomic distance is about 700, compared with 1,900 for the mean intergenomic distance (values obtained from 13 species; fig. 3*b*). These facts strongly support the concept of genomic signature and qualify the CGR representation as a powerful tool to unveil it. Analysis of parts of a genome allows one to generate a satisfactory genomic signature. Sequences of several dozen kilobases in length display a good part of the features of the whole genome. Sequences of 100 kb, for example, are adequate to investigate seven-letter words. This fact gives rise to the possibility of comparing nonhomologous genomic sequences when only parts of the genomes are available. In particular, it is interesting to see that the origins of small sequences of DNA can be ascertained merely on the basis of the distribution of words composing them. The diversity of CGR images is considerable. In figure 4 (see also figs. 1–3), we present a selection of typical images for 20 species belonging to the major phyla. The variety of patterns confirms the potential of the CGR approach to get a genomic signature. The three vertebrate images exhibit CpG depletion (visible as scoops of different sizes corresponding to CG-embedding sequences). The two human sequences illustrate the variation in base composition of vertebrate sequences due to the presence of isochores (fig. 4). These sequences, although possessing the same main pattern, have reversed dark zones in the pictures: the first image comes from a presumed AT-rich isochore, and the second comes from a GC-rich isochore. It can be seen in this particular example that variation in base composition does not jeopardize the overall pattern found in human DNA.

In addition to CG scoops, main features of CGR images of DNA sequences include diagonals (stretches of purines and pyrimidines: *Arabinopsis thaliana, Sulfobolus solfataricus, Thermotoga maritima, Methanococcus jannaschii*) or their absence (*Mycobacterium leprae*), horizontal lines with decreasing density from bottom to top or the reverse (AT- or GC-rich genome with long stretches of the main bases: *Schizosaccharomyces pombe, Streptococcus pneumoniae*), empty patches and word-rich regions of different shapes (triangles [*Deinococcus radiodurans*], diagonals [*Agrobacterium tumefasciens*], amorphous [but specific] patches [*Drosophila melanogaster, Emericella nidulans*]). These features can combine into more complex patterns (*Leishmania major, Borrelia burgdorferi*). The similarities observed between some eubacteria, archaebacteria, and eukaryotes displaying the CG depletion are clearly detectable and raise questions about the underlying mechanisms responsible for this depletion (Karlin and Burge 1995). Similarly, CTAG is highly counterselected in some eubacteria and archaebacteria, while the rest of the image can be very different from one species to another, highlighting the complexity of genome structure and organization (*Escherichia coli* and *A. fulgidus*; fig. 3*a*).

Mapping the World of Genomic Signatures

We can use distances between images to build a representation of DNA sequence similarities, just as the mere knowledge of distances between towns allows one to construct a fairly good map of the corresponding area. In this study, we chose 36 images of sequences with the variety of patterns as the unique criterion. An ad hoc implementation of PCA was used to analyze the matrix of distances between all images, and the space generated by the three first principal components (or axes) was chosen to display images with respect to their relative distances. The ranks of components are directly related to their ability to express variability among images. The first component is grounded to the base composition of species; it is highly correlated with A+T concentration (fig. 5*a*), in agreement with its role in species differentiation (Muto and Osawa 1987; Forsdyke 1995, 1996). Our results indicate that this parameter is the most important one for explaining variations between images of genomes. In contrast, components 2 and 3 are not correlated with base composition of species (fig. 5*a*). The second component opposes sequences with purine and pyrimidine runs to sequences with A+T or G+C runs (diagonal lines vs. horizontal lines). The third component qualifies the homogeneity of images: the sizes of the clear (depleted) areas in images are inversely correlated with the lengths of uncommon words (fig. 5*b*). For example, the CG depletion in vertebrates is expressed by a large clear square in the associate images. Therefore, we may consider base composition, stretches, and patches basic characteristics of variability of genome signature.

It is interesting to note that PCA analysis allows one to observe biologically meaningful grouping of species: the three major domains of life—eubacteria, archaebacteria, and eukaryotes—are identifiable (Woese, Kandler, and Wheelis 1990). This is especially visible on the plane spanned by components 2 and 3 (fig. 5*b*). It must be pointed out that the three- domain scenario was recently questioned by Gupta (1998*a*, 1998*b*), who claimed that the life's third domain (Archaea) may be an ''endangered paradigm.'' In fact, our results support

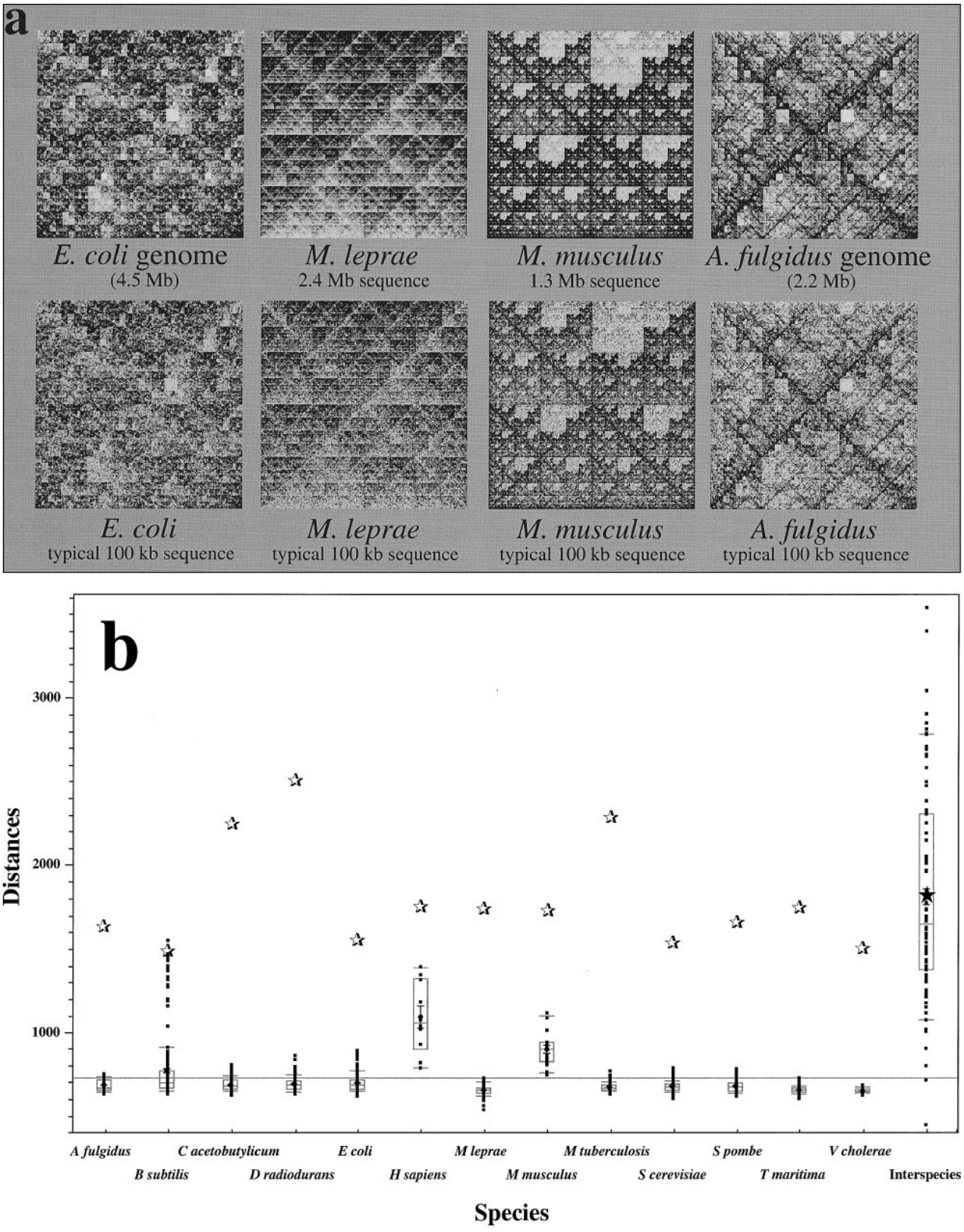


FIG. 3.—Chaos game representation images (seven-letter words) within and among genomes. *a,* Top line: whole sequences. Bottom line: samples corresponding to a randomly selected 100-kb subsequence. *b,* Distribution of Euclidean distances between independent 100-kb subsequences within each species. The distribution of distances between 13 species (using a randomly selected 100-kb subsequence per species) is given on the right of the figure (interspecies distance). The mean distance of a 100-kb subsequence of each species to other species is indicated by a star. Means, standard deviations, and quantile boxes (10%,25%,75%,90%) of distance distribution are shown.

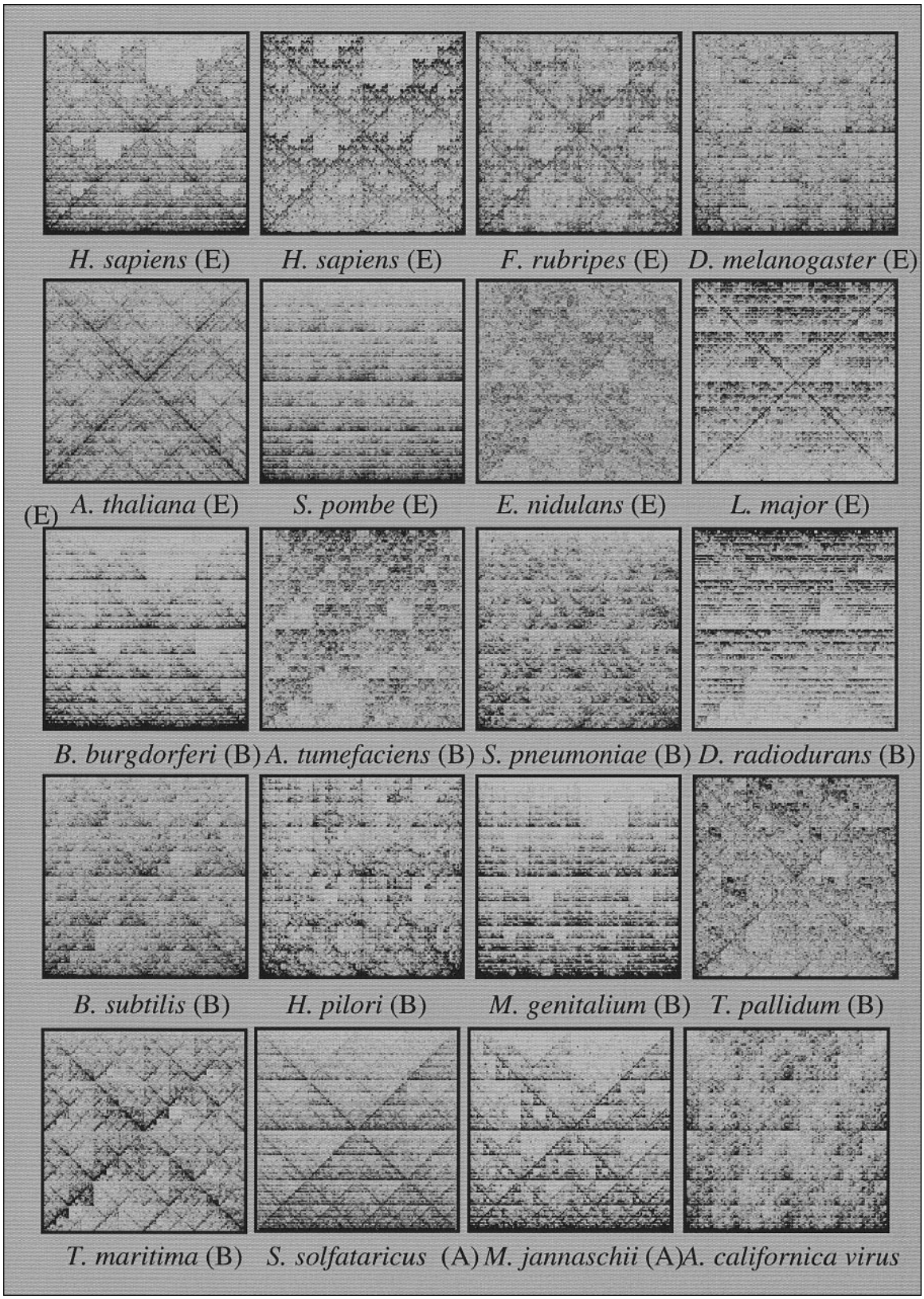
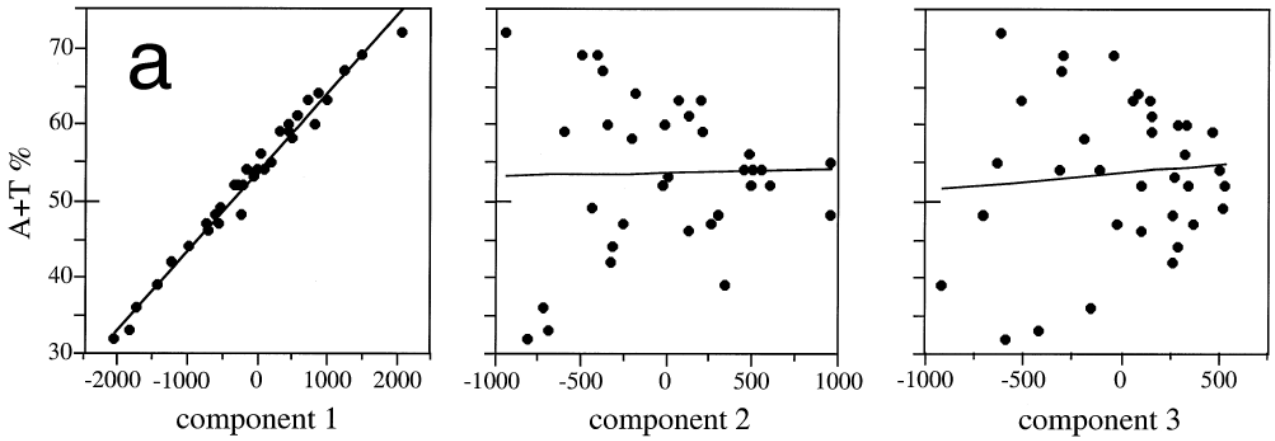


FIG. 4.—Diversity of chaos game representation images (seven-letter words). Images of 100-kb subsequences are displayed for 20 species. Associated letter codes for domain of life are as follows: E, eukaryote; B, eubacteria; A, archaebacteria.



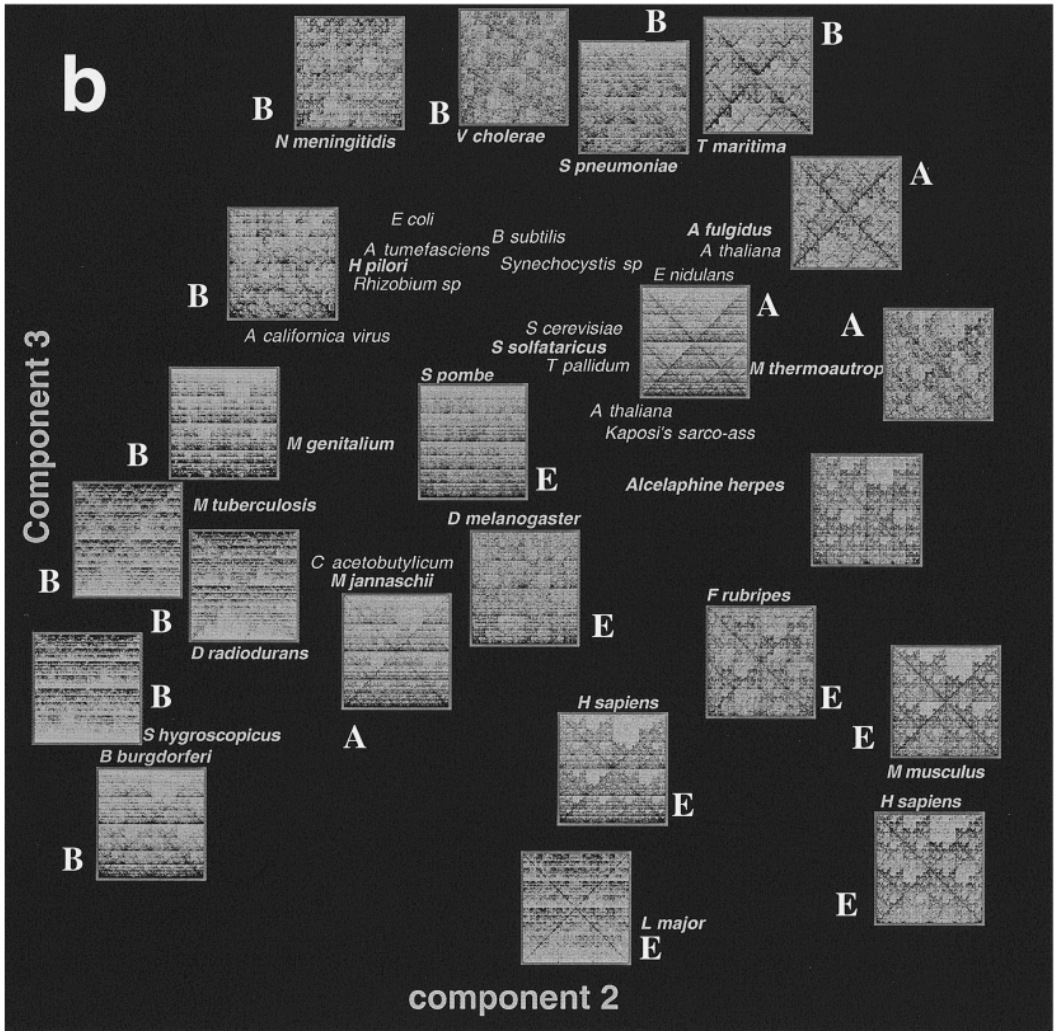


FIG. 5.—Representation of DNA sequences similarities (100 kb, seven-letter words). *a*, Relationships between base compositions of species and their coordinates in the principal component analysis (PCA) representation. The probability that there is no correlation (under the null hypothesis) is <0.00001, 0.914, and 0.615 for components 1, 2, and 3, respectively. *b*, Thirty-six sequences are displayed in the plane spanned by components 2 and 3 of PCA analysis of their Euclidean distances. Only a few images are given, in order to avoid disturbing overlaps. Bold names refer to illustrated sequences. Coding for domain of life is the same as in figure 4.

the ''Gupta'' scenario as well. In our hands, archaebacteria lie between prokaryotes and eukaryotes, so it is just a matter of boundaries to validate either of these two scenarios. Another interesting grouping concerns vertebrates, which are found to be neighbors, as confirmed by a study based on 12 sequences for which the mean distance was 1,200, compared with 1,100 for the intrasequence distance for man and mouse (fig. 3*b*) and

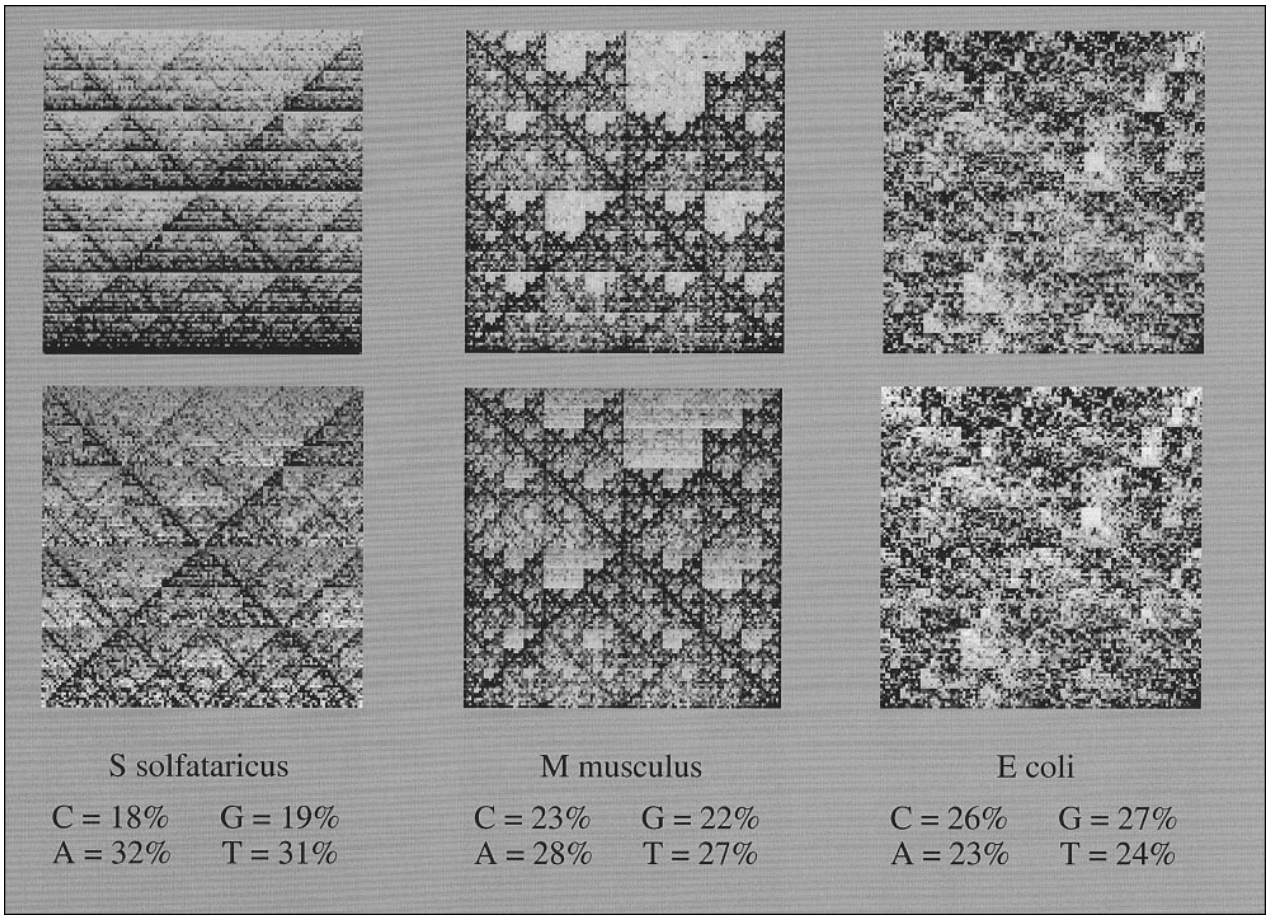


FIG. 6.—Contribution of base composition to chaos game representation (CGR) images. Top row: original 100-kb CGR images. Bottom row: CGR images after subtraction of a mock image resulting only from bias in base composition. The single-strand base composition is given below each image.

2,200 for the mean distance between the 36 sequences used for figure 5. It is worth noting that interspecies distance may sometimes be of the same order of magnitude as intrasequence distance: the distance between *S. pombe* and *Saccharomyces cerevisiae* is about 900. These last two results allow generalization of the genomic signature to the whole set of words up to seven letters long (Karlin and Ladunga 1994; Karlin and Burge 1995).

One may wonder to what extent CGR images and, consequently, distances between images depend on nucleotide concentration. Assuming that succession of nucleotides along a sequence follows the so-called random model (a zero-order Markov chain, namely that the probability of finding a nucleotide does not depend on the preceding nucleotides), the probability of observing a given word is the mere product of the probabilities of its constituent letters. The mock images which can be generated are very typical; they display series of horizontal lines expressing the fractal gradient of frequency which directly results from a bias in base composition. By means of a simple subtraction, it is therefore possible to reveal the contribution of words longer than one letter to the generation of CGR images (fig. 6). It can be observed that, with the exception of horizontal lines, most of the details in CGR images are actually preserved. ''Residual'' patterns are still characteristic and constitute, de facto, the basis of the classification which is offered by components 2 and 3 of PCA. As the lack of correlation between base composition and components 2 and 3 of PCA indicates, C+G concentration does not play any role in the plot presented in figure 5*b*. *Streptomyces hygroscopicus* and *B. burgdorferi* have very different C+G concentrations (0.68 and 0.28, respectively) and are found to be close neighbors in this plot. *Thermotoga maritima* and *Mus musculus*, which have similar C+G concentrations (0.45), have been assigned quite different locations. Although nucleotide concentration is an important factor of species characterization, it appears to be less important when dealing with species clustering. Using a much larger set of species, we are currently working on this interesting question, which deserves specific attention.

Conclusions

The genomic signature as expressed in terms of short nucleotide usage extends and generalizes the genomic signature concept originally proposed by Karlin and co-workers (Karlin and Burge 1995; Karlin, Mràzek, and Campbell 1997). It takes advantage of whole-genome

data and reveals genomewide trends. The structure of DNA is specific to each species and undergoes only slight variations along the whole genome. Diversity among species is considerable and is primarily a consequence of base concentration, stretches of bases, and patches of words with unusual frequencies. The establishment of the dictionary of the words used by a species, together with their frequencies of occurrence, allows one to point out the basic words of the genome. This will help to derive detailed correlations between their use and species characteristics. As a matter of fact, the discrimination which can be achieved between eukaryotes and prokaryotes using the genomic signature gives a genomic basis to the phylogenetic root of classification of species. Moreover, distances between sequences of phylogenetically close species are of the same order as the distances between subsequences of a genome of these species. It was possible to observe that the distribution of word frequency was very broad. Unusual frequencies were detected at first glance. The genomic signature appears as a powerful tool for investigating the mechanisms of DNA maintenance from which the DNA structure results.

Acknowledgments

A lot of credit must be given to Dr. H. Joel Jeffrey for being the first to apply the CGR approach to the analysis of DNA sequences. The authors also wish to thank Dr. Richard D'Ari for helpful discussions and comments.

LITERATURE CITED

BEUTLER, E., T. GELBART, J. HAN, J. A. KOZIOL, and B. BEUTLER. 1989. Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. Proc. Natl. Acad. Sci. USA **86**:192–196.

BHAGWAT, A. S., and M. MCCLELLAND. 1992. DNA mismatch correction by very short patch repair may have altered the abundance of oligonucleotides in the E. coli genome. Nucleic Acids Res. **20**:1663–1668.

BLAISDELL, B. E., K. E. RUDD, A. MATIN, and S. KARLIN. 1993. Significant dispersed recurrent DNA sequences in the Escherichia coli genome. J. Mol. Biol. **229**:833–848.

BURGE, C., A. M. CAMPBELL, and S. KARLIN. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc. Natl. Acad. Sci. USA **89**:1358–1362.

BURMA, P. K., A. RAJ, J. K. DEB, and S. K. BRAHMACHARI. 1992. Genome analysis: a new approach for visualization of sequence organization in genomes. J. Biosci. **17**:395–411.

CHARLESWORTH, B. 1994. Patterns in the genome. Curr. Biol. **4**:182–184.

DESCHAVANNE, P., and M. RADMAN. 1991. Counterselection of GATC sequences in enterobacteriophages by the components of the methyl-directed mismatch repair system J. Mol. Evol. **33**:125–132.

DOOLITTLE, R. F. 1997. Microbial genomes opened up. Nature **392**:339–342.

DUTTA, C., and J. DAS. 1992. Mathematical characterization of chaos game representation. New algorithms for nucleotide sequence analysis. J. Mol. Biol. **228**:715–719.

FORSDYKE, D. R. 1995. Relative roles of primary sequence and (G+C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. J. Mol. Evol. **41**:573–581.

———. 1996. Different biological species ''broadcast'' their DNAs at different (G+C) ''wavelengths.'' J. Theor. Biol. **178**:405–417.

GELFAND, M. S., and E. V. KOONIN. 1997. Avoidance of palindromic words in bacterial and archeal genomes: a close connection with restriction enzymes. Nucleic Acids Res. **25**: 2430–2439.

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

GUPTA, R. S. 1998*a*. Life's third domain (Archaea): an established fact or an endangered paradigm? Theor. Popul. Biol. **54**:91–104.

———. 1998*b*. What are archaebacteria: life's third domain or monoderm prokaryotes related to gram-positive bacteria? A new proposal for the classification of prokaryotic organisms. Mol. Microbiol. **29**:695–707.

HILL, K. A., N. J. SCHISLER, and S. M. SINGH. 1992. Chaos game representation of coding regions of human globin genes and alcohol dehydrogenase genes of phylogenetically divergent species. J. Mol. Evol. **35**:261–269.

JEFFREY, H. J. 1990. Chaos game representation of gene structure. Nucleic Acids Res. **18**:2163–2170.

———. 1992. Chaos game visualization of sequences. Comput. Graphics **16**:25–33.

JOSSE, J., A. D. KAISER, and A. KORNBERG. 1961. Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxiribonucleic acid. J. Biol. Chem. **236**:874–875.

KARLIN, S., and C. BURGE. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. **11**: 283–290.

KARLIN, S., C. BURGE, and A. M. CAMPBELL. 1992. Statistical analyses of counts and distributions of restriction sites in DNA sequences. Nucleic Acids Res. **20**:1363–1370.

KARLIN, S., and I. LADUNGA. 1994. Comparisons of eukaryotic genomic sequences. Proc. Natl. Acad. Sci. USA **91**:12832–12836.

KARLIN, S., J. MRÀZEK, and A. M. CAMPBELL. 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. **179**:3899–3913.

KENDALL, M. 1975. Multivariate analysis. Griffin, London.

MALEY, L. E., and C. R. MARSHALL. 1998. The coming of age of molecular systematics. Science **279**:505–506.

MUTO, A., and S. OSAWA. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. USA **84**:166–169.

OLIVER, J. L., P. BERNAOLA-GALVÀN, G. GUERRERO, and R. ROMÀN-ROLDÀN. 1993. Entropic profiles of DNA sequences through chaos-game-derived images. J. Theor. Biol. **160**: 457–470.

PHILLIPS, G. J., J. ARNOLD, and R. IVARIE. 1987. Monothrough hexanucleotide composition of the Escherichia coli genome: a Markov chain analysis. Nucleic Acids Res. **15**: 2611–2626.

ROBINSON, N. J., P. J. ROBINSON, A. GUPTA, A. J. BLEASBY, B. A. WHITTON, and A. P. MORBY. 1995. Singular overrepresentation of an octameric palindrome, HIP1, in DNA from many cyanobacteria. Nucleic Acids Res. **23**:729–735.

SHARP, P. M., and G. MATASSI. 1994. Codon usage and genome evolution. Curr. Opin. Genet. Dev. **4**:851–860.

WOESE, C. R., O. KANDLER, and M. L. WHEELIS. 1990. Towards a natural system of organisms: proposal for the domains archea, bacteria and eucarya. Proc. Natl. Acad. Sci. USA **87**:4576–4579.

STANLEY SAWYER, reviewing editor

Accepted June 30, 1999