MICROBIAL & COMPARATIVE GENOMICS
Volume 2, Number 4, 1997
Mary Ann Liebert, Inc.

Asymmetry of Coding Versus Noncoding Strand in Coding Sequences of Different Genomes

STANISŁAW CEBRAT,[1] MIROSŁAW R. DUDEK,[2] PAWEŁ MACKIEWICZ,[1] MARIA KOWALCZUK,[1] and MAŁGORZATA FITA[1]

ABSTRACT

We have used the asymmetry between the coding and noncoding strands in different codon positions of coding sequences of DNA as a parameter to evaluate the coding probability for open reading frames (ORFs). The method enables an approximation of the total number of coding ORFs in the set of analyzed sequences as well as an estimation of the coding probability for the ORFs. The asymmetry observed in the nucleotide composition of codons in coding sequences has been used successfully for analysis of the genomes completed at the time of this analysis.

INTRODUCTION

There are many methods to discriminate between coding and noncoding DNA sequences (Fickett, 1996, for review). For nondisrupted genes, one of the better criteria is the length of an open reading frame (ORF). In the yeast genome project (SGD, *Saccharomyces* Genome Database), the lower limit of an ORF length has been set at 100 codons. An additional criterion is the value of the codon adaptation index (CAI) (Sharp and Li, 1987). It has been arbitrarily accepted in SGD that ORFs shorter than 150 triplets with CAI $<$ 0.11 are considered noncoding (Dujon et al., 1994). It has been also accepted in SGD that the longer ORF of a pair of overlapping ORFs is considered coding. Generally, these criteria work well, but some ORFs are shorter than 150 codons with CAI $<$ 0.11 and perform already documented coding functions. Such criteria as CAI and codon bias index (CBI) (Benetzen and Hall, 1982) are based on the observation that codon usage in protein coding sequences does not correspond to codon frequency expected from the nucleotide composition of the genome. Two different forces have been suggested to be responsible for this bias. One is translational selection based on relative concentrations of iso-accepting tRNAs (Ikemura, 1982). The second is mutational pressure that forces a change in overall nucleotide composition of DNA and especially influences the third (silent) positions of codons (Sueoka, 1988; Sharp et al., 1993, for review).

We have assumed that a coding sequence should reflect specific construction of the genetic code, nonrandom (biased) amino acid usage, and physical restrictions of the DNA (RNA) molecule. The most fundamental rule of DNA composition is complementarity of the nucleotide bases, A:T and G:C. In a random sequence, this rule implicates balance in purine/pyrimidine composition of both DNA strands, which is observed in long DNA stretches (as in yeast chromosomes) but not in coding sequences organized in operons (λ phage, for example). Dujon et al. (1994) observed a relative abundance of purine doublets in coding se-

[1]Institute of Microbiology, and [2]Institute of Theoretical Physics, Wrocław University, Wrocław, Poland.

quences of yeast chromosome 2. The same can be concluded from the results presented by Karlin and Burge (1995) for other genomes. This asymmetry in purine/pyrimidine composition of coding vs noncoding DNA strands has been used to discriminate ORFs in the yeast genome (Cebrat et al., 1997a,b). The asymmetry in nucleotide composition of both strands of a protein coding sequence is a sum of the specific asymmetry of each position within codons. Thus, the asymmetry in the first, the second, and the third positions could compensate for each other within a coding sequence, and asymmetry for each position in codons should be analyzed separately. In some aspects, this method resembles the CAI approach, but the results of the analysis are not correlated with the results using CAI; this will be discussed later. Rather, our approach encompasses some rules in amino acid composition of proteins and a highly sophisticated structure for the genetic code.

MATERIALS AND METHODS

Databases and software

The *Saccharomyces cerevisiae* genome sequences were downloaded September 23, 1996, from genome-ftp.stanford.edu. Information on yeast gene function, ORF homology, and their presumed functions was downloaded November 16, 1996, from http://www.mips.biochem.mpg.de. Sequences for all viruses were downloaded May 10, 1997, from ncbi.nlm.nih.gov, sequences for *Escherichia coli* from http://genom4.aist-nara.acjp on May 9, 1997, for *Haemophilus influenzae, Mycoplasma genitalium,* and *Methanococcus jannaschii* from http://www.tigr.org on May 8, 1997, for *Mycoplasma pneumoniae* from http://www.zmbh.uni-heidelberg.de on May 10, 1997, for *Synechocystis* sp.ftp://ftp.kazusa.or.jp on May 8, 1997, and for *Methanobacterium thermoautotrophicum* from ftp://www.genomcorp.com on June 3, 1997. After the retrieval, data were not updated.

In the analyses, we have considered all ORFs found in the completely sequenced genomes longer than 100 codons starting with ATG and ending with one of the three universal stop codons.

The software for all the analyses was written by one of the authors (M.R.D.)

Graphic representation of sequences

To make a graphic representation of a sequence in two-dimensional space, we analyzed the displacement of a DNA walker that checked each position within codons separately. For the DNA walk, we used a modified method of Berthelsen et al. (1992). For each sequence, we performed three DNA walks, independently, for each nucleotide position in codon triplets. The first walker starts from the first nucleotide position of the first codon and then jumps every third nucleotide until the end of the examined sequence (stop codon) has been reached. Similarly, the second and the third walkers start from the second and third nucleotide positions of the first codon, respectively. Every jump of a walker is associated with a unit shift in two-dimensional space 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. Hence, each DNA walk represents a history of nucleotide composition of the first, the second, or the third position of codons along the DNA sequence. The three walks together have been called a *spider* and a single walk has been called a *spider leg*. An example of a spider representing a typical gene in the yeast genome (the multicopy suppressor of sin4, YML109w) is seen in Figure 1a. In Figure 1b, the sequence coding for a hydrophobic protein (vacuolar calcium transporter protein YDL128w) is shown, and in Figure 1c, a spider representing an intergenic sequence of 921 triplets is shown. The spiders depict the nucleotide composition of the three positions in codons, but it is also possible to extract some numerical information from these plots and to characterize whole sets of ORFs by the method.

Distribution of ORFs in a torus projection

For each ORF, we measured (in degrees) the angles of the vector determined by the origin and the end of the spider legs. In fact, the angles equal to arcus tangent [(G-C)/(A-T)] have positive values for the first two quadrants of the plot and negative values for the third and fourth quadrants. This has enabled us to construct a plot where each ORF is represented by a point whose coordinates are (x) the angle represent-

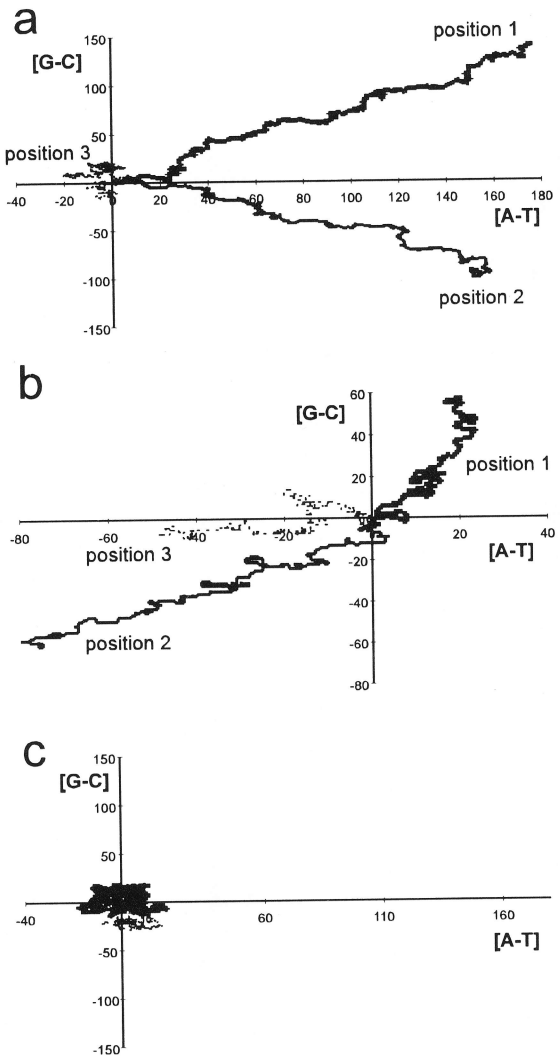

FIG. 1. Two-dimensional representation of DNA walks (spiders) performed for different positions in codons for yeast sequences. **(a)** An example of a spider representing a typical gene in the yeast genome, the sequence coding for a multicopy suppressor of sin4 (YML109w). **(b)** The sequence coding for a hydrophobic protein, vacuolar calcium transport protein (YDL128w). **(c)** A spider representing an intergenic sequence 921 triplets long.

ing the first leg and (y) the angle representing the second leg. It is also possible to use the angle of the third leg as one of the two coordinates or as the third coordinate in three-dimensional space. As both axes of the plot are scaled from -180 degrees to $+180$ degrees, the plot is, in fact, a projection of torus, and its area is finite. The distributions of different sets of ORFs from different genomes are presented in Figures 2 and 3.

Approximation of total number of coding ORFs in a genome

For three genomes, *S. cerevisiae, H. influenzae,* and *E. coli,* we compared the distributions of all ORFs found in the genomes with the distributions of ORFs with already identified functions.

To compare the distributions of coding sequences with the distribution of all ORFs in the same genome, we analyzed for ORFs with known functions the average value and the standard deviations (SD) for angles of the first legs and the average value for angles of the second legs. The average values, $\bar{A}_1$ and $\bar{A}_2$, for the set of genes are considered coordinates of the center of the genes' distribution. Next, we normalized para-

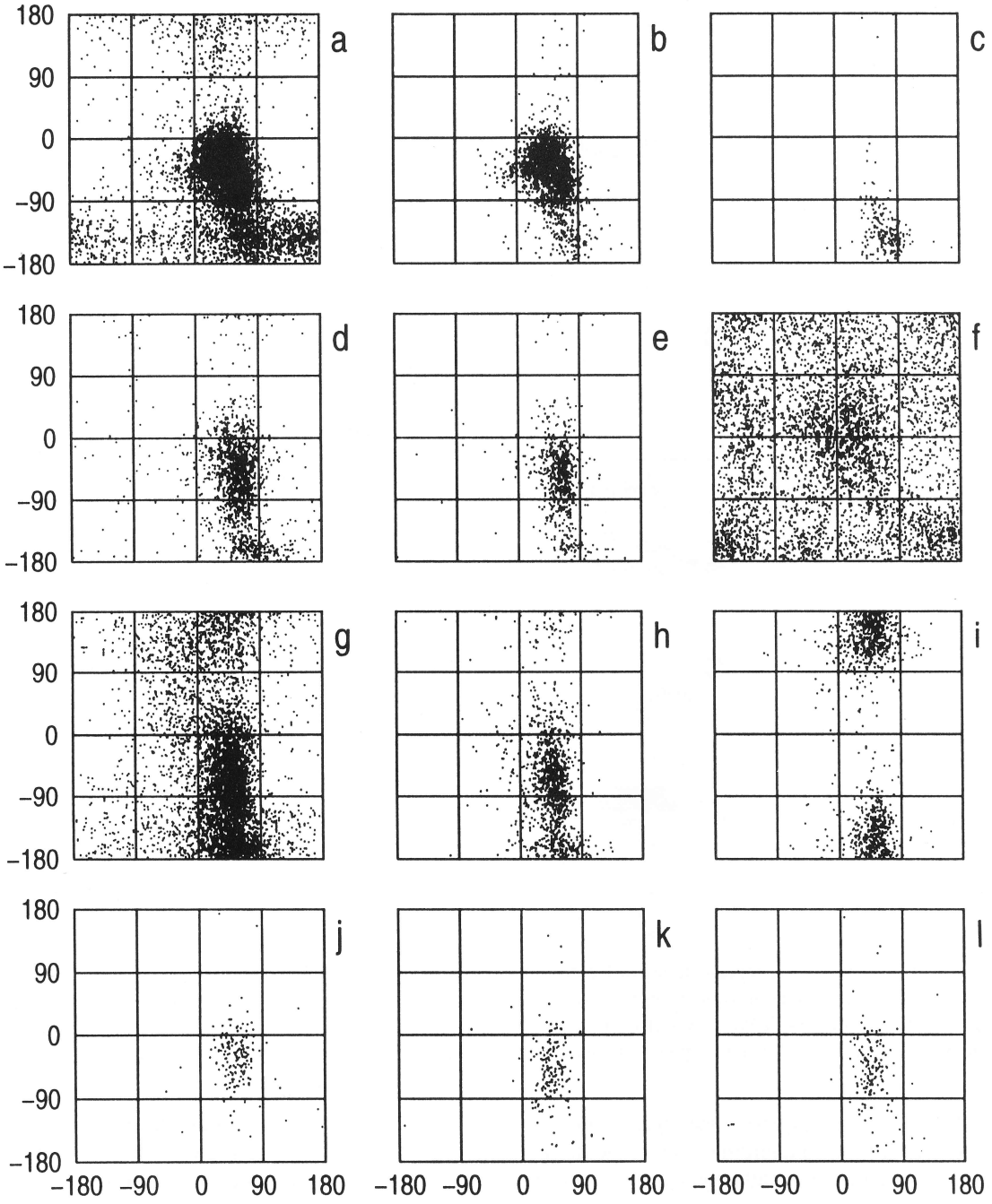

FIG. 2. Distribution of ORFs into the projection of torus where (x) is the angle of the first leg of spider and (y) is the angle of the second leg of spider [an exception is i, where (y) represents the angle of the third leg of the spider]. **(a)** *S. cerevisiae,* all ORFs >100 codons. **(b)** *S. cerevisiae,* ORFs with known function. **(c)** *S. cerevisiae,* ORFs coding for transmembrane proteins. **(d)** *H. influenzae,* all ORFs>100 codons. **(e)** *H. influenzae,* ORFs with known function. **(f)** *S. cerevisiae,* intergenic sequences>100 triplets. **(g)** *E. coli,* all ORFs > 100 codons. **(h)** *E. coli,* ORFs with known function. **(i)** *E. coli,* ORFs with known function (angles 1 and 3). **(j,k,l)** Phage T4 vaccinia, and variola viruses, respectively, all ORFs > 100 codons.

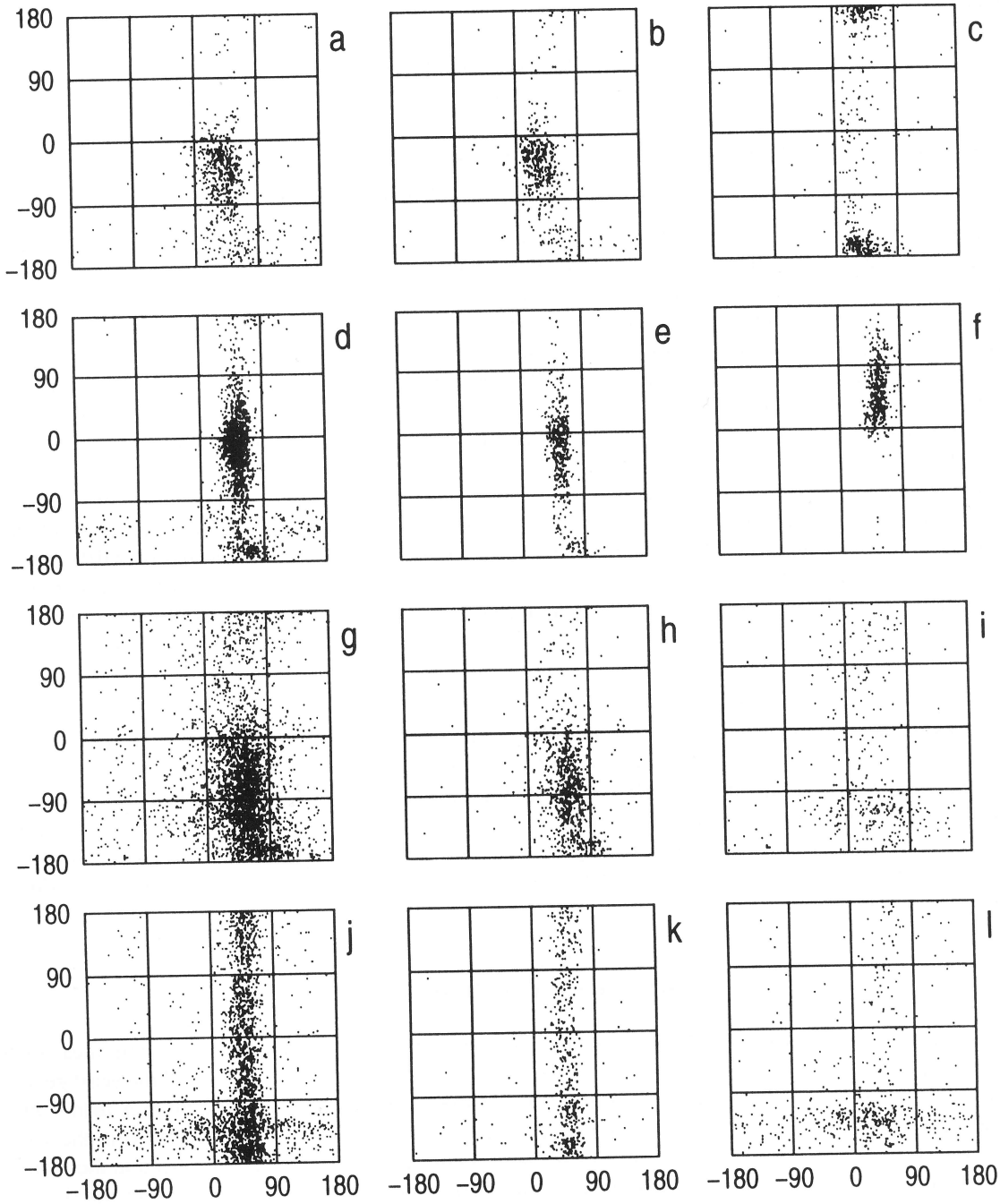

FIG. 3. Distribution of ORFs into the projection of torus where (x) is the angle of the first leg of spider and (y) is the angle of the second leg of spider. **(a)** *M. pneumoniae,* all ORFs > 100 codons. **(b)** *M. genitalium,* all ORFs, angle 1 vs angle 2. **(c)** As in **(b),** but angle 1 vs angle 3. **(d)** *M. jannaschii,* all ORF > 100 codons. **(e)** *M. jannaschii,* ORFs with known functions, angle 1 vs angle 2. **(f)** as in **(e),** but angle 1 vs angle 3. **(g)** *Synechocystis* sp., all ORFs > 100 codons. **(h)** *Synechocystis* sp., ORFs with known function. **(i)** *Synechocystis* sp., shorter ORFs from pairs of overlapping ones. **(j)** *M. thermotrophicum,* all ORFs > 100 codons. **(k)** *M. thermotrophicum,* longer ORFs from pairs of overlapping ones. **(l)** *M. thermotrophicum,* shorter ORFs from pairs of overlapping ones.

meters by giving each ORF values A_1 and A_2, which are equal to the differences (expressed in SD) between the average values and the angle of the first leg and the second leg, respectively, for a given ORF. For each individual ORF, we measured the distance from the center, which is equal to

$$A_i = \sqrt{A_{i1}^2 + A_{i2}^2}$$

To estimate the number of coding ORFs we used the algorithm

$$N_i = ORF_{in} + [(G_{out}/G_{in})(ORF_{in})]$$

where N_i is an assumed maximal number of coding ORFs for a given A_i, ORF_{in} is the number of all ORFs inside the space with a distance to the distribution center $\leq$ Ai, G_{in} is the number of genes (ORFs with known function) with a distance to the distribution center $\leq A_i$, and G_{out} is the number of genes (ORFs with known function) with a distance to center $> A_i$. In this algorithm, for a given A_i, we assumed that all ORFs inside the space determined by A_i are coding and that the fraction of coding ORFs outside the space stays in the same proportion to the number of ORFs inside the space as it is for the distribution of ORFS with the described coding function.

We plotted the N_i values vs distance (A_i). To avoid errors, we cut off 5% of ORFs with maximum values of A_i and 10% with minimal values of A_i. To estimate the approximated number of coding ORFs, we found the extrapolated value of N for $A_i = 0$. In Figure 4 are shown plots for *S. cerevisiae, H. influenzae,* and *E. coli.* For comparison for *S. cerevisiae,* we used as an estimate for coding ORFs the set of all ORFs $\geq$ 100 codons long (7440 ORFs) and the set of 6200 coding ORFs primarily selected by MIPS (Munich Information Centre for Protein Sequences) (http://www.mips.biochem.mpgde). For *E. coli,* we also prepared the distributions of genes and ORFs taking into account the angles of the third legs instead of the second ones. The approximation of total number of coding ORFs done for these distributions is shown in Fig. 4c.

Estimation of coding probability for an ORF

To estimate the coding probability for an ORF, we divided the whole set of all ORFs into classes according to A_i values. For each class, we counted the number of expected coding ORFs (using the method described) and the total number of ORFs found in a given class. The ratio between these values has been assumed as coding probability for ORFs localized in a given class. It is possible to use the coding probability values obtained by this method directly or to plot them against A_i, to make the polynomial approximation, and to describe the probability as a function of A_i. The results are available at our WWW site (angband.microb.uni.wroc.pl).

RESULTS AND DISCUSSION

Because the representation of both parameters describing the ORFs is in degrees ($\pm$180), areas of the plots seen in Figures 2 and 3 are finite (they are the surfaces of the toruses). In the case of the yeast genome, about 6% of this area includes about 75% of all coding ORFs. All these genes have more A than T in the first and second positions of codons, more G than C in the first position, and less G than C in the second position. The ORFs with the lower number of A than T in the second position are localized in the plots below the main cloud of genes. In this latter region are genes coding for transmembrane proteins (information on the set of these genes was kindly supplied by A. Goffeau, Catholique Universitie de Louvain). The shift of these genes below the main cluster is due to the presence of codons coding for hydrophobic amino acids, which are rich in T in the second position, and underrepresentation of codons for hydrophilic amino acids, which are rich in A in the second position. Because transmembrane proteins possess hydrophobic spans, they are relatively rich in T in the second position.

When comparing different classes of yeast ORFs with annotation 1–6 in the MIPS base, it can be seen that classes 2 and 3 are a little more dispersed than class 1 (identified genes have annotation 1 in MIPS), which suggests that there are only a few noncoding ORFs in classes 1, 2, and 3. In the classes above 4 there are many noncoding ORFs. In fact, in class 6, only a few coding ORFs should be expected (data not shown).

The correlation coefficient between the CAI and the distance from the center is close to 0. One would

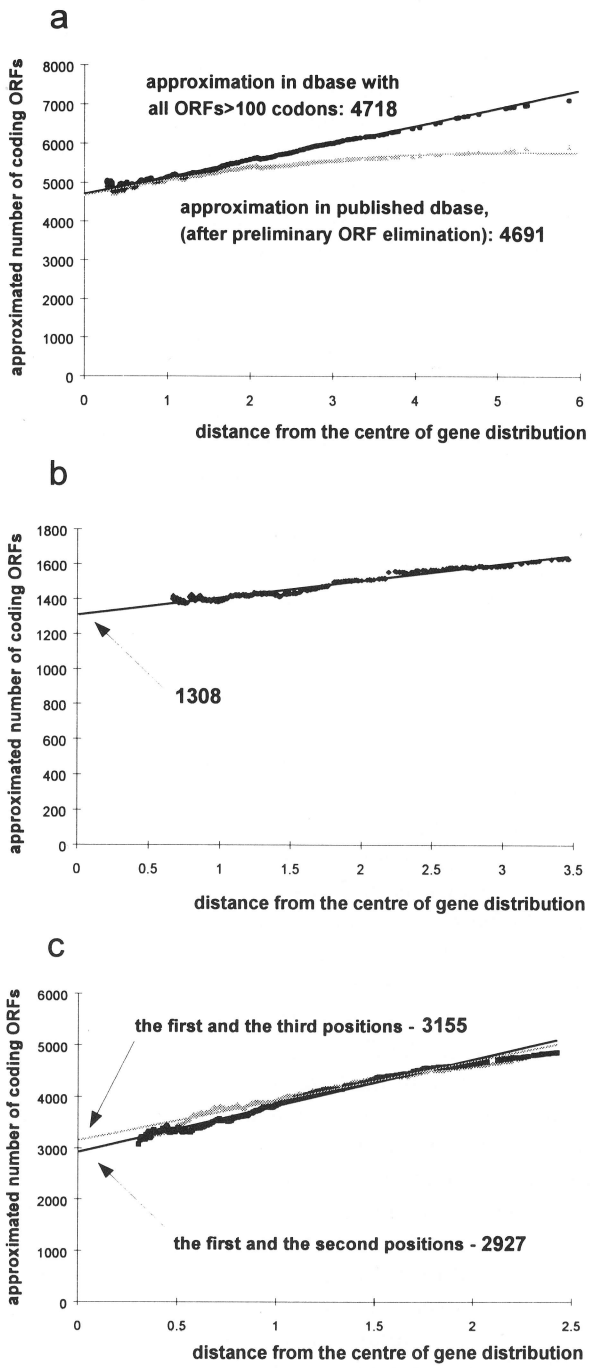

FIG. 4. Approximation of the total number of coding ORFs > 100 codons in genomes of **(a)** *S. cerevisiae*, approximations done for the whole set of ORFs > 100 codons and for ORFs published in the SGD database (after preliminary selection), **(b)** *H. influenzae*, **(c)** *E. coli*, approximations done for distributions prepared for relations between angles of the first and the second legs and relations between angles of the first and the third legs.

expect such a result because the CAI is sensitive to the composition of the third positions of codons, whereas we used parameters measuring the asymmetry of the first two positions. We have observed some correlation (about −0.4) between the distance from the center of the distribution and the length of the ORFs for the yeast. As the distance is reciprocal to the coding probability, a negative correlation should be expected because of two phenomena: (1) noncoding overlapping ORFs or random ORFs are usually shorter, and they are found far from the center of the distribution, and (2) it seems that very long ORFs could be considered as "averaged smaller ORFs." Thus, the SDs for the class of long ORFs should be smaller. To prove the last assumption, we calculated the SDs of the first and the second angles for yeast ORFs longer than 1000 codons and found that they both equal 0.65 of those for the whole set of genes. It is obvious that the correlation between length of ORFs and their coding probability cannot be high because the relation between these parameters is not linear.

The number of coding ORFs in yeast estimated by our method is much lower than that proposed by the SGD program. This number could be underestimated by us if (1) the set of already known genes is not statistically representative for the whole set of coding ORFs in the yeast genome, and it is too homogeneous to be considered a statistically significant sample, or (2) some ORFs with an authentic start translation codon farther downstream from the first ATG have been discarded because the noncoding beginning segment of the sequence would shift the whole ORF farther away from the distribution center. The latter may be true for the shortest ORFs because, as we have shown, the longer ORFs are situated closer to the distribution center. On the other hand, it is less probable for the shorter ORF to have its start translation codon very far downstream from the first ATG. This is the reason that there should be only a small number of ORFs eliminated by this procedure.

The results obtained by this method suggest that the overwhelming number of yeast ORFs with well-established homologies are coding, with only a very few being noncoding. Among ORFs with no known homologies, the noncoding fraction is much larger. As, by definition, these presumptive noncoding ORFs would be counted as orphans, we suggest that the class of orphans is actually much smaller than previously assumed (Dujon, 1996; Casari et al., 1996).

We also estimated the number of coding ORFs > 100 codons in the genomes of *E. coli* and *H. influenzae*. These genomes have different organizations relative to each other. We found that about 85% of the nucleotides of the *H. influenzae* genome is in ORFs of >100 codons. Only about 1% of nucleotides are shared by overlapping ORFs. In the *E. coli* genome, there are many overlapping ORFs (>2000 overlapping ORFs, and 11.5% of all nucleotides are within overlaps). Nonoverlapping ORFs cover about 90% of the *E. coli* genome. Still, when comparing the whole set of ORFs to protein coding ORFs, in *E. coli,* only 48.4% of these ORFs are expected to be coding. In *H. influenzae,* about 77% of all ORFs are expected to be coding. We have also found that in the *E. coli* genome, the composition of the third position in the codons depends strongly on the position of the ORF in the chromosome. Using the first and the third angles as parameters for ORF distribution, followed by approximation of the number of coding ORFs in the *E. coli* genome, we estimate a slightly higher fraction of coding ORFs—52.2% vs 48.4%.

To estimate which position in the codon is the better predictor for a protein coding function, we examined the distributions of the angles for the three spider legs representing codon positions in *S. cerevisiae, E. coli, H. influenzae,* and *M. jannaschii* (Fig. 5). It can be seen that the first position is the best predictor for all the examined genomes. For all examined genomes, the average values of angles for the first positions are between 0 and 90 degrees. Even if the second parameter is not so predictive, the first parameter causes genes to form a narrow ring on a torus. The third position seems to be a better predictor than the second for genomes of *E. coli, Mycoplasma* (not shown), and *M. jannaschii* (compare also the pairs of distributions: Fig. 2h and i, Fig. 3b and c, Fig. 3e and f). We also found that for mitochondrial genomes the third position seems to be a better predictor than the second one (data not shown). Note that using the third position as one of the parameters does not correspond to CAI or the method of McInerney (1997) because (1) mutation pressure exploits the transition mechanism, and the changes in [A-T] and [G-C] in the third positions of codons in the coding strand that we measured cannot result from transition but rather transversion, and (2) half of the substitutions in the third positions of the type purine<=>pyrimidine are not silent and cannot be subject to a simple translational selection.

A coding sequence can generate a noncoding ORF in a specific phase (Cebrat and Dudek, 1996). As two of the three stop codons (TAA,TAG) have the first two nucleotides in a palindromic relationship, they can

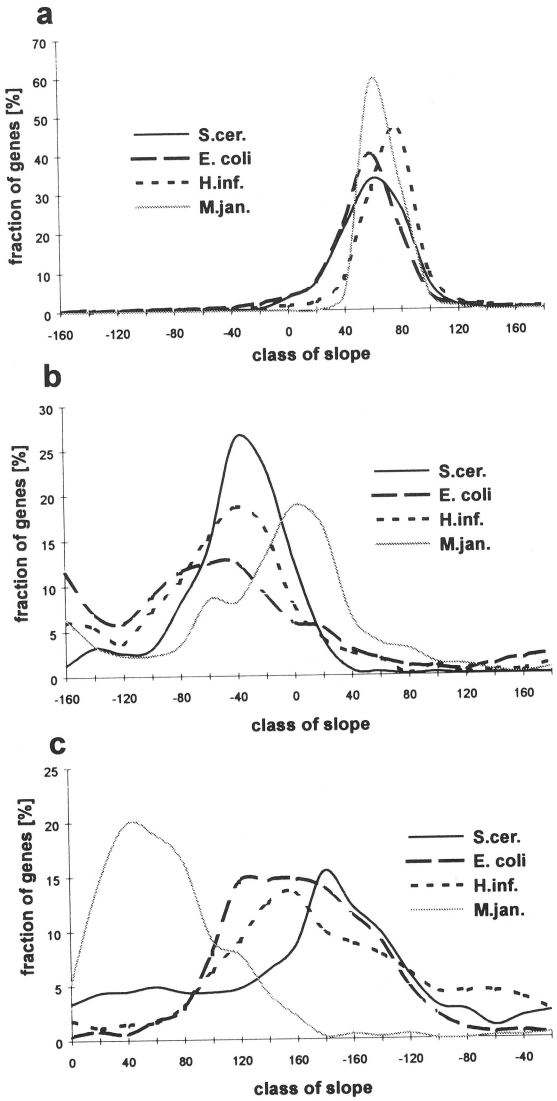

FIG. 5. Distribution of ORFs with known function into classes according to the values of angles of legs describing three positions in codons. **(a)** Angles of the first legs. **(b)** Angles of the second legs. **(c)** Angles of the third legs. The width of classes −20 degrees.

generate stops in the related phase of the opposite strand. By definition, coding sequences do not have stops in frame. The frequency of stops in the related phase of the opposite strand is lower, and the probability of the ORF appearing is higher. To find the coding ORF in a pair of overlapping ORFs, it is necessary to distinguish between the proper reading frame and the reading frame of the ORF generated by the coding sequence. Because the method described here shows differences in asymmetry for different positions in codons, it is simple to locate the coding frame. It can be seen readily in the cases of *Synechocystis* sp. and *M. thermotrophicum* (Fig. 3g–l). The first angles for the overwhelming fraction of ORFs of these genomes have a very narrow distribution, but for a specific class of ORFs, this parameter does not seem to be predictive (ORFs situated in the lower part of the plots). To check this, we selected all overlapping ORFs and divided this set (of overlapping ORFs) into the shorter ORFs in a pair and the longer ones. In Figure 3 g–l, we show the distributions for these three sets of ORFs. Assuming that in a pair of overlapping ORFs the longer one is coding (which should be true in most cases), we can conclude that the points distributed horizontally on plots g, i, j, and l in Figure 3 are not coding.

The method presented here seems to be universal, as all genomes, even those of viruses (Fig. 2j–l), show specific asymmetry in coding vs noncoding strands. There are also many other numerical parameters describing spiders that can be used for ORF discrimination. One of these parameters is the normalized length of spider legs.

ACKNOWLEDGMENTS

We thank Prof. A. Goffeau for many long discussions, for encouragement, for supplying the information on transmembrane protein coding ORFs, and for help in understanding the specific shift observed in the distribution of these ORFs. This work was supported by a KBN grant number 1016/S/IMi/97.

REFERENCES

BERTHELSEN, Ch. L., GLAZIER, J.A., and SKOLNICK, M.H. (1992). Global fractal dimension of human DNA sequences treated as pseudorandom walks. Phys Rev **A45,** 8902–8913.

BENETZEN, J.L., and HALL, B.D. (1982). Codon selection in yeast. J Biol Chem **257,** 3026–3031.

CASARI, G., de DRUVAR, A., SANDER, C., and SCHNEIDER, R. (1996). Bioinformatics and the discovery of gene function. Trends Genet **12,** 244–255.

CEBRAT, S., and DUDEK, M.R. (1996). Generation of overlapping reading frames. Trends Genet **12,** 12.

CEBRAT, S., DUDEK, M.R., and MACKIEWICZ, P. (1997a). The number of coding ORFs in *Saccharomyces cerevisiae* genome and the mystery of orphans. Yeast **13,** P189. (abstract). 18th Int Conf Yeast Genetics Mol Biol Stellenbosch, South Africa.

CEBRAT, S., DUDEK, M.R., and ROGOWSKA, A. (1997b). Asymmetry in nucleotide composition of sense and antisense strands as a parameter for discriminating open reading frames as protein coding sequences. J Appl Genetics **38,** 1–9.

DUJON, B. (1996). The yeast genome project: what did we learn? Trends Genet **12,** 263–270.

DUJON, B., and 106 co-authors. (1994). Complete DNA sequence of yeast chromosome XI. Nature **369,** 371–378.

FICKETT, J.W. (1996). Finding genes by computer: the state of the art. Trends Genet **12,** 316–320.

IKEMURA, T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. J Mol Biol **158,** 573–597.

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

McINERNEY, J.O. (1997). Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. Microb Comp Genom **2,** 89–97.

SHARP, P.M., and LI, W.-H. (1987). The codon adaptation index: a measure of directional synonymous codon usage bias and its potential applications. Nucleic Acids Res **15,** 1281–1295.

SHARP, P.M., STENICO, M., PEDE, J.F., and LLOYD, A.T. (1993). Codon usage: mutational bias, translational selection or both? Biochem Soc Trans **21,** 835–841.

SUEOKA, N. (1988). Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA **85,** 2653–2657.

Address reprint requests to:
Stanisław Cebrat
Institute of Microbiology
Wrocław University
ul. Przybyszewskiego 63/77
51-148 Wrocław
Poland
e-mail: cebrat@angband.microb.uni.wroc.pl.