Nucleic Acids Research, 1993, Vol. 21, No. 12 **2837–2844**

Estimation of protein coding density in a corpus of DNA sequence data

James W.Fickett and Roderic Guigó
Theoretical Biology and Biophysics Group and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Received March 29, 1993; Revised and Accepted May 25, 1993

ABSTRACT

A number of experimental methods have been reported for estimating the number of genes in a genome, or the closely related *coding density* of a genome, defined as the fraction of base pairs in codons. Recently, DNA sequence data representative of the genome as a whole have become available for several organisms, making the problem of estimating coding density amenable to sequence analytic methods. Estimates of coding density for a single genome vary widely, so that methods with characterized error bounds have become increasingly desirable. We present a method to estimate the protein coding density in a corpus of DNA sequence data, in which a 'coding statistic' is calculated for a large number of windows of the sequence under study, and the distribution of the statistic is decomposed into two normal distributions, assumed to be the distributions of the coding statistic in the coding and noncoding fractions of the sequence windows. The accuracy of the method is evaluated using known data and application is made to the yeast chromosome III sequence and to *C.elegans* cosmid sequences. It can also be applied to fragmentary data, for example a collection of short sequences determined in the course of STS mapping.

INTRODUCTION

Fundamental knowledge about an organism includes an estimate of the number of genes in its genome—one measure of the overall complexity of the organism—and an estimate of the closely related coding density (defined as the fraction of base pairs that are in codons). The latter is a basic aspect of genome structure, related to the intriguing question of the prevalence of 'junk' or 'selfish' DNA [Orgel and Crick, 1980]. An estimation of coding density has important practical consequences as well, for example in deciding whether more information will be gained by sequencing cDNAs or genomic DNA.

Current estimates of coding density for most eukaryotic organisms are given only in rather wide ranges. For example, Clark *et al.* [1988] estimate that there are roughly 3500 essential genes in *Caenorhabditis elegans*, giving both reasons to think this estimate may be too high as well as reasons that indicate it may be minimal. Combined with the results of Park and Horvitz [1986], which suggest that half of the genes in *C.elegans* may be inessential, this gives an estimate of roughly 7000 genes. However Waterston *et al.* [1992] estimate that the true number of genes in *C.elegans* may be closer to 15000.

Kaback, Angerer, and Davidson [1979] showed that roughly 50–60% of the yeast genome is transcribed in roughly 5000 transcripts, under laboratory conditions. (A transcript density of 50–60% corresponds to a coding density of less than 50%, since an mRNA contains untranslated regions.) If all genes were distributed uniformly over the genome, this estimate would give about 120 genes on chromosome III. However Yoshikawa and Isono [1990] found 156 transcripts from chromosome III, and Oliver *et al.* [1992], equating genes with open reading frames of length at least 100 amino acids on the chromosome III sequence, find 182 genes, giving a coding density estimate of 67% and an estimate of about 8000 genes in the whole organism.

The analyses of Waterston *et al.*, Sulston *et al.*, and Oliver *et al.* were made possible because an important new source of data has recently become available. Whereas most sequences in the current databases are from highly expressed genes, sequence is now becoming available which is a much less biased sample of the genome. In some cases this means a very long stretch of DNA encompassing many genes, as in the case of the recent determination of yeast chromosome III [Oliver *et al.* 1992], in others it means a large number of short sequences, randomly selected from the genome in the course of determining STSs for genome mapping [Olson *et al.* 1989].

Current methods to determine coding density, both experimental and computational, rely on counting genes. The experimental methods typically give low estimates because, under the experimental conditions chosen, not all genes are required or expressed. A major difficulty with the computational methods applied to date is that current gene recognition methods have rather large, and sometimes uncharacterized, error rates. Thus Oliver *et al.* simply count open reading frames exceeding a certain size and the studies of Waterston *et al.* and Sulston *et al.* depend on the gene recognition program Genefinder. In neither case is it easy to evaluate the accuracy of the predicted number of genes.

Here, we present a method to estimate coding density in a corpus of sequence data—either long stretches of sequence data or a large number of short sequences—that does not rely on identifying genes. Indeed, it is possible to define on windows of sequence a number of simple measures, or coding statistics, which are indicative of protein coding function (reviewed in

[Fickett and Tung 1992]). And while such statistics have a large random component and a large variance when observed on individual windows, the overall distribution of such a statistic, when observed on a large set of windows, is closely correlated with global coding density (Fig. 1).

One simple approach to the problem considered here would be (1) to infer, using sequence data of known coding density from public databases, a model of the relationship between coding density and an ensemble property of the coding statistic distribution, as for example the linear regression shown in Fig. 1, and (2) use such a model to predict the coding density of the new sequence data under study. However, because the data in the public sequence databases is a very biased sample of the genome, extrapolation from the database to the genome may not be justifiable.

This problem may be surmountable, but here we pursue an alternative method which does not rely on first establishing a model of the relationship between coding density and a coding statistic in previously characterized sequences, but rather depends exclusively on the distribution of the coding statistic on the corpus of sequence data under study. In the method detailed below, the sequence data under study are first partitioned into a set of fixed-size windows, and the chosen coding statistic is calculated on each. Then the distribution of the statistic is decomposed into two normal distributions that are assumed to correspond to the distribution of the statistic in the noncoding and coding fractions of the sequence data.

In what follows we first introduce the Max Codon Usage coding statistic, or MCU, for which we have generally observed reasonably gaussian behavior. We describe in detail the method used to decompose the distribution of the coding statistic, using the yeast genomic sequences from GenBank to explicitly illustrate it. We then evaluate the accuracy of the method in large sets of characterized genomic sequences from five distantly related genomic organisms, and describe two applications. In the first, the coding density of yeast chromosome III is estimated by decomposing the distribution of the MCU distribution. In the second, we estimate the coding density of a collection of sequenced cosmids from *C.elegans*. In this case, however, since we strongly suspect that the MCU statistic does not have a normal distribution, we estimate the coding density by decomposing the distribution of a different coding statistic. Finally, we discuss the applicability of the method to other genomes, and its limitations.

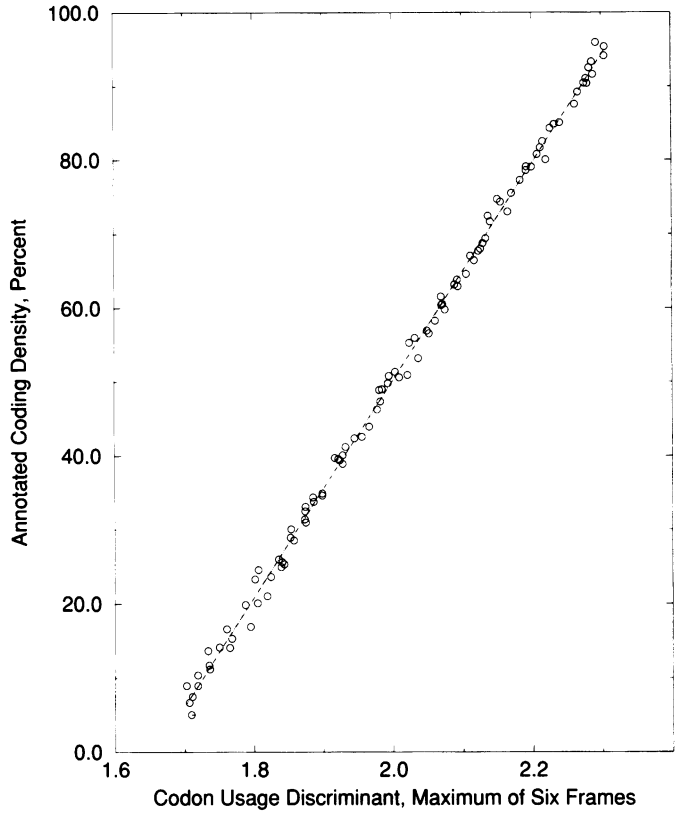

Figure 1. Correlation between the mean of a coding statistic and the coding density calculated from GenBank annotation. For each density d, d =0,..., 100, a set of yeast 240 bp sequence windows was made by selecting windows at random from the GenBank reference set (see Methods) with probabilities chosen to obtain an expected overall percentage d of coding windows, and an expected set size of 315 kb. Each such set is represented as a point in the figure, the abscissa giving the average of the coding statistic (maximum of a codon usage discriminant function over the six frames of the window—see Methods) over the set, and the ordinate giving the coding density computed from GenBank annotation.

MATERIALS AND METHODS

Sequence data

Nucleotide sequence data were taken from the GenBank™/EMBL/DDBJ international collection, accessed via the on-line relational GenBank database [Cinkosky *et al.* 1991] and the EMBL on-line service [Higgins *et al.* 1992].

Reference sets of annotated sequences released on or before 30 June 1992 were extracted from GenBank for eukaryotic organisms with more than one megabase of available genomic sequence. (This date was chosen to avoid large amounts of unannotated 'raw' sequence; the date makes the data essentially equivalent to that available in GenBank release 71). The yeast reference set excluded the chromosome III sequence. Successive, non-overlapping, 240 basepair windows were taken from each genomic sequence. Windows with ambiguous bases were discarded. Positions annotated as 'CDS' in any of the six frames were considered coding positions. All other positions were considered non-coding. The organisms considered and the total number of windows of genomic sequence extracted from GenBank (by means of a query on division and locus name prefix, as shown): Human (PRI/HUM), 29696 windows; Mouse (ROD/MUS), 13733; Yeast (PLN/YSC), 13233; Drosophila (INV/DRO), 8725; Rat (ROD/RAT), 8723.

The yeast chromosome III sequence (accession number X59720, totalling 1313 240 bp windows) was retrieved from the EMBL on-line service. *C.elegans* genomic sequences were extracted from GenBank on 15 Dec 1992, and two separate data sets were generated. One contained six recently published cosmid sequences (accession numbers L07143, L07144, M77697, Z11115, Z11126 and Z11505; cf. [Sulston *et al.* 1992]), totalling 906 240bp windows, and another, the reference set, containing 2443 windows, was made up of the remaining genomic sequences in GenBank.

The MCU coding statistic

For each organism, a function to discriminate protein coding from noncoding DNA was derived using linear discriminant analysis (e.g. [Manly 1986]) on the windows in the GenBank reference set and their reverse complements. At this stage, for the

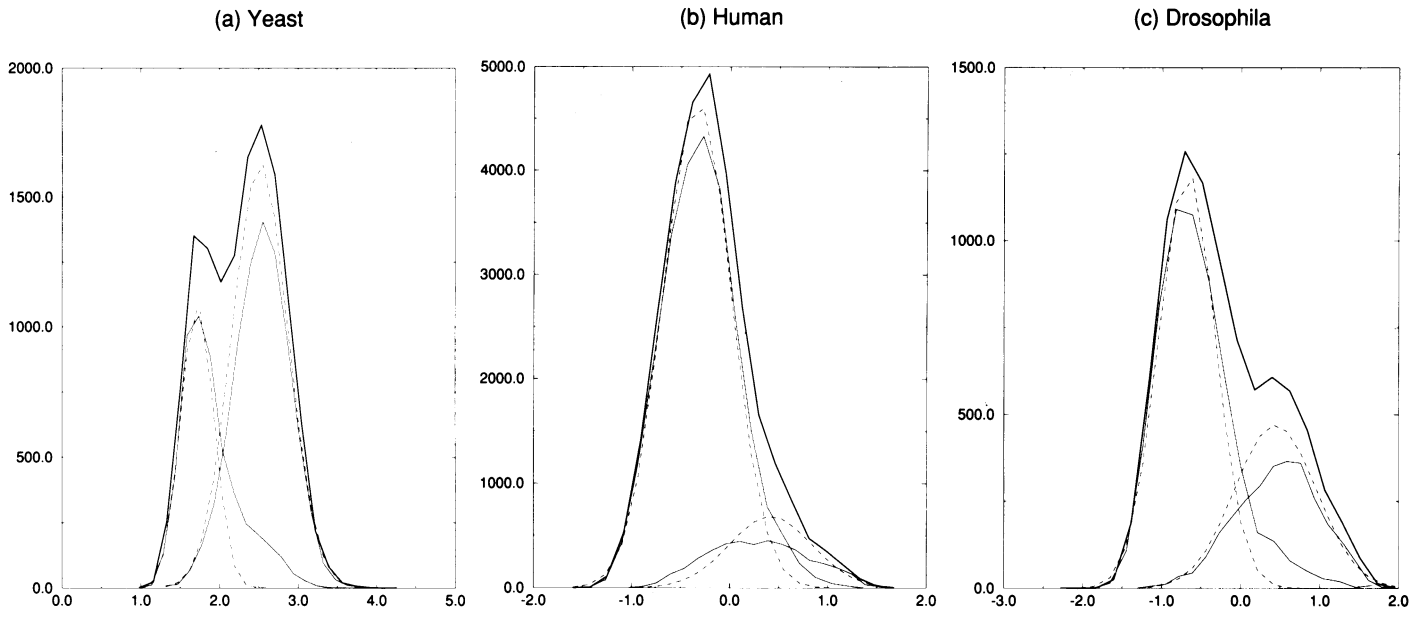

Figure 2. Global MCU distribution (thick line) in all windows, and component distributions on coding and non-coding windows—estimated (dashed lines) and based on GenBank annotation (thin lines)—for some of the reference data sets studied (the cases of mouse and rat are very similar to the case of human). The values of the parameters for the estimated distributions appear (for all five test organisms) in Table 1 below.

discriminant analysis, only fully coding and fully noncoding windows were used. Because phase-specific coding measures are generally more accurate than those which are only region-specific [Fickett and Tung 1992], the discrimination was between windows that were both coding and in phase, on the one hand, and windows that were either coding and out of phase, or noncoding, on the other hand. The basis for discrimination was codon usage, this being a simple but quite accurate measure of coding function [ibid.]. Thus a codon usage vector C was calculated for each window, and a discriminant vector D was determined by linear discriminant analysis. D was replaced by $-$D, if necessary, to make the average value of C·D higher on coding than noncoding windows. The coding statistic we used in the primary analysis, which of course must not be phase specific, was then defined as the maximum of D·C/(# of codons) over the six reading frames of a window. For brevity we will refer to this statistic as MCU (Max Codon Usage) in the rest of the paper. MCU is simple and seems to have close to gaussian behavior in many cases, but of course the linear discriminant function can be derived from sequence properties other than codon usage (see the estimation of the coding density of the *C.elegans* cosmids below.)

Decomposition of the MCU distribution

The distribution of MCU for all windows in the GenBank yeast reference set is shown in Fig. 2.a. The individual MCU distributions for the sets of coding (>50% of coding bases) and non-coding windows (≤50%) are also shown in Fig. 2.a. Note that the peaks of these individual distributions coincide with the peaks of the compound bimodal distribution. The method that we have designed to estimate the coding density of genomes relies on inferring such individual underlying distributions from the compound distribution. In our approach, an additional assumption is initially made: that the individual MCU distributions for coding and non-coding windows are normal. In several cases, for example that of yeast, this assumption appears to be violated to a significant extent by the long tail characterizing the MCU distribution for non-coding windows (see Fig. 2.a). However, we think it likely that this is largely an artifact of the database annotation—the reasons are discussed below.

The problem of estimating the parameters of a mixture of distributions was first approached by Pearson [1894] who attempted to estimate the parameters of a mixture of two normal populations by equating the first five moments with their sample values. Since then, alternative methods have been developed (for a review, see [Redner and Walker, 1984]). Here, we have used a maximum likelihood method. Briefly, the maximum likelihood estimate of parameters associated with a sample of observations can be defined as that choice of parameters which maximizes the probability of the sample. The probability of the sample, written as a function of the parameters to be chosen, is called the *likelihood function*. In particular, we have used an iterative procedure for numerically approximating maximum-likelihood estimates of parameters in mixture distributions, known as the EM (Expectation Maximization) algorithm, which was formalized in a more general context by Dempster *et al.*, [1977]. An extensive review of the EM algorithm can be found in [Redner and Walker 1984].

We have used a version of the EM algorithm initially proposed by Hasselblad [1966, 1969] for mixtures of normal distributions (see also [Equihua 1988]). If f_j denotes the density function of the jth ($j = 1,\ldots, k$) component normal distribution, then the density function of the mixture is given by

$$g = \sum_{j=1}^{k} \alpha_j f_j \qquad (1)$$

for some α_j, with $0 < \alpha_j < 1$, and $\Sigma\alpha_j = 1$. The log-likelihood function is then

$$L = \sum_{i=1}^{m} n_i \ln g_i \qquad (2)$$

where m is the number of sample values, n_i is the frequency of x_i (the ith sample value), g_i is the value of g at x_i,. Let μ_j denote the mean of the jth component of the distribution and θ_j the variance. Intuitively, the EM algorithm proceeds in the following way: First, initial estimates are made of α_j, μ_j, and θ_j. Next, the following two steps are repeated iteratively. (1) each observation is assigned to the component distribution to which it has highest probability of belonging. (2) new estimates of α_j, μ_j, and θ_j are obtained from the distributions resulting from this assignment. This procedure leads to solving the following set of iterative equations simultaneously for $j = 1,\ldots, k-1$.

$$\alpha_j^{(r+1)} = \frac{\alpha_j^{(r)}}{n} \sum_{i=1}^{m} \frac{n_i f_{ji}^{(r)}}{g_i^{(r)}} \quad (3)$$

$$\mu_j^{(r+1)} = \left[\sum_{i=1}^{m} \frac{n_i x_i f_{ji}^{(r)}}{g_i^{(r)}}\right] \Big/ \left[\sum_{i=1}^{m} \frac{n_i f_{ji}^{(r)}}{g_i^{(r)}}\right] \quad (4)$$

$$\theta_j^{(r+1)} = \left[\sum_{i=1}^{m} \frac{n_i (x_i - \mu_j^{(r)})^2 f_{ji}^{(r)}}{g_i^{(r)}}\right] \Big/ \left[\sum_{i=1}^{m} \frac{n_i f_{ji}^{(r)}}{g_i^{(r)}}\right] \quad (5)$$

where $n = \Sigma n_i$.

The log-likelihood function L is computed at each iteration, and the procedure continues until an iteration is reached such that the absolute difference of two consecutive log-likelihoods is smaller than a specified value. Although initial estimates for α_j, μ_j, and θ_j need to be provided, the procedure has been proved to converge irrespective of their value [Hasselblad 1966, 1969]; [Dempster *et al.*, 1977]. To evaluate the goodness of the fit for the observed data to the estimated mixture, the likelihood ratio test statistic, G, is computed. The statistic is given by

$$G = 2(\Gamma_1 - \Gamma_0) \quad (6)$$

where Γ_1 and Γ_0 are the log-likelihoods computed under the alternative and null hypothesis (which assumes the mixture of distributions), respectively. In practice, G is computed as

$$G = 2(\Sigma\, n_i \ln n_i - \Sigma\, n_i \ln n\, g_i) \quad (7)$$

G approximates a χ^2 with $m-1-[ck+(k-1)]$, where c is the number of parameters of the distributions f_j ($j = 1,\ldots, k$). That is, $ck+(k-1)$ is the total number of parameters to be estimated. In our case $c=2$ and $k=2$, so the degrees of freedom are $m-6$.

We have written a program in C implementing such a procedure, after the ALGOL code of Agha and Ibrahim [1983].

The input data must be a frequency distribution; therefore the observations of a continuous variable—such as MCU—need to be grouped in equal-sized intervals. We have observed that the number of intervals, if chosen between reasonable boundaries—from 10 to 40—has little effect on the final estimations. Following Sokal and Rohlf [1981] we have usually grouped the MCU distributions in 20 intervals.

Estimation of the coding density

In practice, to estimate the coding density of anonymous sequence data we proceed in the following way: the sequence data under study is decomposed in a number of fixed length windows. The value of the MCU statistic is calculated in each window, and the global MCU distribution obtained. After discretizing it in twenty intervals, the EM algorithm is used to decompose this distribution in two component normal distributions, which are assumed to be the MCU distributions in coding windows (more than 50% coding nucleotides) and in non-coding windows (50% or less coding nucleotides). From them, the fraction of coding windows is estimated. Such a fraction of coding window is taken as an estimate of the coding density. (It is easy to see that under the very reasonable hypothesis that the distribution of the proportion of coding bases in partially coding windows is uniform, the expected value of the fraction of coding windows is in fact the fraction of coding bases. In practice, we have found that the difference between the two fractions is ordinarily less than 1%).

All of our software, written in C for the Sun Sparcstation, is available upon request.

RESULTS

Evaluation of the method

We have obtained the EM estimates of the MCU distribution in coding and non-coding windows for the five reference data sets. The parameter estimates are shown in Table 1 and the distributions in Figure 2 (the rat and mouse cases are similar to the human, and are not shown.) Note that the EM estimation of the underlying MCU distributions for coding and non-coding windows are close to those derived from the GenBank annotation, even though the decomposition is much less obvious for the metazoans than for yeast. Despite this relatively good agreement, the EM decomposition of the MCU distribution results in an apparent systematic overprediction of the proportion of coding windows, of magnitude $+0.08 \pm 0.06$. The apparent overprediction is the consequence of the long tail to the right characterizing the MCU distribution for windows annotated as non-coding, that causes this distribution to depart significantly from normality. (Observe, in Table 1, the large values of the log-likelihood ratio obtained for the gaussian fitting of the MCU distribution for non-coding windows, G0, when compared with the log-likelihood ratios obtained for the fitting of coding windows, G1.) Then, when the global MCU distribution is forced into two normals, an important fraction of the windows annotated as non-coding (but with high MCU) is put in the distribution of coding windows, thus resulting in the observed overprediction of its proportion.

Although it can not be ruled out that the departure from normality characterizing the distribution of non-coding windows is intrinsic to the MCU behavior (due, for example, to the existence of sequences which are noncoding but resemble coding sequence, such as pseudogenes, or certain repetitive elements), it is also possible that such a departure is only apparent; the result at least partially of the incompleteness in the annotation of the public databases. It is well known that the use of 'CDS' in GenBank is conservative, often not identifying a likely coding sequence as 'CDS' unless there is experimental evidence. Thus, a number of windows annotated as non-coding, but with high MCU value are likely to be coding; in such a case, the long tail observed for the MCU distribution on non-coding windows would be just an artifact, and the EM estimates of the coding density that we have obtained would be closer to the actual values than they appear to be. In support of this possibility consider the case of yeast, where splicing is rare and long ORFs usually correspond to genes. We have detected around 400 open reading frames in the GenBank yeast sequences longer than 300bp starting with an ATG and ending in an stop codon that are not annotated. It has been suggested in relation with chromosome III (Oliver *et al.*,

Table 1. EM estimates of the coding and non-coding MCU distributions

organism		Proportion coding	Proportion no coding	Mean coding	Mean no coding	Standard Deviation coding	Standard Deviation no coding	G	G1	G0
Human	annot.	0.15	0.85	0.297	−0.294	0.500	0.386			
	estimated	0.17	0.83	0.420	−0.336	0.422	0.346	134	153	576
Mouse	annot.	0.17	0.83	0.259	−0.086	0.282	0.230			
	estimated	0.27	0.73	0.231	−0.122	0.322	0.177	107	140	1659
Yeast	annot.	0.61	0.39	2.536	1.894	0.354	0.375			
	estimated	0.71	0.29	2.515	1.718	0.355	0.208	32	42	1094
Drosoph.	annot.	0.33	0.67	0.465	−0.565	0.527	0.484			
	estimated	0.41	0.59	0.442	−0.680	0.514	0.357	86	58	768
Rat	annot.	0.14	0.86	0.235	−0.251	0.340	0.267			
	estimated	0.28	0.72	0.117	−0.301	0.364	0.217	28	30	438

Parameters of the MCU distributions on coding and non-coding windows estimated from the compound MCU distribution. EM estimates and values based on GenBank annotation are provided for the proportion of coding windows and for the mean and standard deviation of the MCU distributions on coding and non-coding windows. G is the value of the log-likelihood ratio for the global MCU distribution to fit the two component estimated normal distributions. The corresponding value of χ^2 with 14 degrees of freedom and a level of significance of 0.01 is 29.14. G1 and G0 are the values of the log-likelihood ratio for the actual MCU distribution in coding and non-coding windows respectively to fit a normal distribution. The corresponding value of the χ^2 with 17 degrees of freedom and a level of significance of 0.01 is 33.41.

1992) that most such ORFs would correspond to protein-coding genes. If these unannotated ORFs were to be counted as coding sequence, the coding density of the GenBank yeast sequences would be at least 0.68, very close to the EM estimation.

The log-likelihood ratio of equation (7) gives a measure of how well the sum of the estimated normal distributions fits the empirical distribution. Values larger than those of the corresponding chi-square distribution indicate that the hypothesis of two underlying normal distributions is unlikely; and in such a case the estimates obtained may be unreliable. We have usually found (table 1) rather large log-likelihood ratios for the EM estimates obtained in the GenBank reference sets. This is not surprising, as lack of gaussian behavior should be expected on GenBank, where essentially the same sequence is often reported multiple times, so that particular MCU values are overrepresented. (Note the much lower log-likelihood values in the yeast chromosome III and *C.elegans* cosmid cases below, where the sampling bias should be very low.) The EM method is in principle valid only when the log-likelihood ratio is low enough to indicate a good fit to two normal distributions. However since we find empirically that the estimates of coding density on known data are close to what we deduce from other sources, it seems that the EM algorithm is quite robust in the face of some deviation from normality.

We have built empirical confidence intervals for the EM estimates by systematically measuring their deviation from the annotated coding density on a large number of human sequence data samples. To minimize the additional error introduced by the incompleteness in the annotation of the database, only sequence entries containing the keyword 'CDS' were considered, since the presence of unannotated coding regions in entries containing 'CDS' is less likely. No action was taken, however, to eliminate the bias resulting from unequal representation of homologous gene families. Specifically, we have generated 1000 random samples of a given number of 240 bp windows from GenBank human genomic entries containing the keyword 'CDS', and for each of them, we have obtained the EM decomposition of the MCU distribution and calculated the difference between the estimated and annotated values. Then, we have determined the smallest interval containing 90% of the observed differences. Results obtained for samples of different number of windows (thus, for different amounts of sequence data) are shown in Table 2. For about one megabase of sequence data, we can obtain an estimation of the density within a ±6% error. For smaller data sets, the accuracy obviously decreases. Note incidentally than when only GenBank entries containing the keyword 'CDS' are considered, the previously observed overprediction of the coding density becomes a slight underprediction, strongly suggesting the existence of unannotated coding regions in GenBank.

Table 2. Variation of Estimates with Different Sample Sizes

Windows in Sample	Bases in Sample	Estimate-Annotation, Percent
3750	900,000	−2.45±5.75
1983	476,000	−2.30±7.70
1067	256,000	−2.65±9.95
415	100,800	−1.70±14.00

For each sample size, 1000 random sequence samples were obtained from the GenBank human genomic entries containing the keyword 'CDS'. For each of the samples, the EM decomposition of the MCU distribution was obtained and the difference between the estimated and annotated values of coding density was calculated. Then, the smallest interval containing 90% of the observed differences was determined (the third column in the table.)

Estimation of the coding density of yeast chromosome III

We have applied the EM algorithm to decompose the MCU distribution derived from the recently published sequence of the yeast chromosome III. The global MCU distribution on these windows and the EM estimates of the component MCU distributions for coding and non-coding windows are shown in Fig. 3. The EM decomposition results in an estimated proportion of coding windows of 0.71. The value of log-likelihood ratio is low, 21.1, indicating that the global MCU distribution fits very well two underlying normal distributions.

Again in the case of yeast we constructed an empirical confidence interval by generating 1000 random samples from the GenBank yeast sequence data, each roughly the size of chromosome III. For each one of these sets, we have calculated the difference between the estimated coding density and the coding density derived from the GenBank annotation. We have found that such a difference stays basically constant. The interval defined by the predicted density −0.11±0.05 contains the value

derived from GenBank in approximately 90% of the samples. Perhaps the interval is smaller than in the corresponding human case because codon usage is more consistent in the simpler organism, or because the window length chosen is more appropriate for yeast genes than for human (see discussion.)

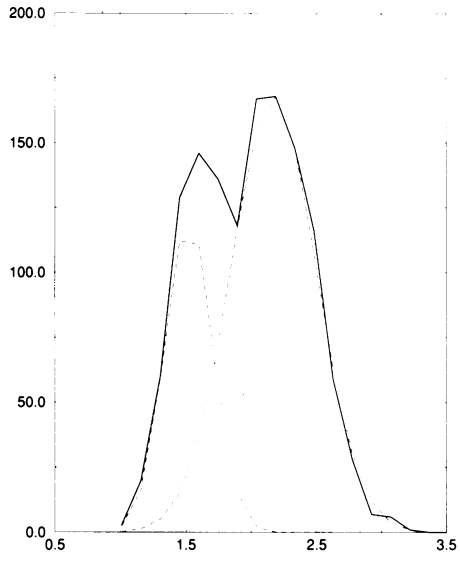

Figure 3. Global MCU distribution (thick line) on all windows extracted from yeast chromosome III, and decomposition into two component normal distributions (dashed lines), using the EM algorithm. The value of the log-likelihood ratio for the global distribution to fit the two normal distributions is 21.1, indicating a good fit. Assuming such component distributions to be the MCU distributions on coding and non-coding windows, 71% of all windows extracted from yeast chromosome III would be coding.

Assuming that the systematic bias is due to the incompleteness of GenBank annotation, our prediction of coding density in yeast is then 0.71 ± 0.06. For comparison, the coding density of yeast chromosome III derived from the Oliver *et al.* (1992) hypothetical annotation, which assumes that every open reading frame longer than 300 bp starting with an ATG and ending with a stop codon corresponds to a protein coding gene, is 0.67.

Estimation of the coding density of *C.elegans* cosmids

We have also used the EM algorithm to estimate the coding density of six recently published *C.elegans* cosmids [Sulston *et al.* 1992]. We have first used the reference set (genomic sequence without cosmids) to study the behavior of the MCU statistic, and the performance of EM algorithm. The coding density of the reference set derived from the GenBank annotation is 0.37. In figure 4.a we have plotted the global MCU distribution on this set, and the underlying distributions for coding and non-coding windows based on the GenBank annotation. As it is possible to see, the MCU distribution for coding windows departs obviously from normality, being clearly bimodal. The EM algorithm can not be applied to infer such a distribution. (Indeed, if applied the estimated coding density obtained is 0.21.) Such an observed distribution could indicate the existence of two classes of genes with very different codon usage. In figure 4.b, we have plotted the distribution of a coding statistic, DIA, derived from di-amino acid counts, which is less sensitive to details of codon usage. Similarly to the MCU, the DIA statistic is a linear combination of 400 di-amino acid pair counts, obtained by discriminant analysis, and also computed in successive non-overlapping windows. The actual DIA distribution for coding windows is much closer to normal than that of the MCU. The EM estimate of the coding density derived from the DIA distribution is 0.47, which is within the limits of the overprediction that we have systematically obtained.

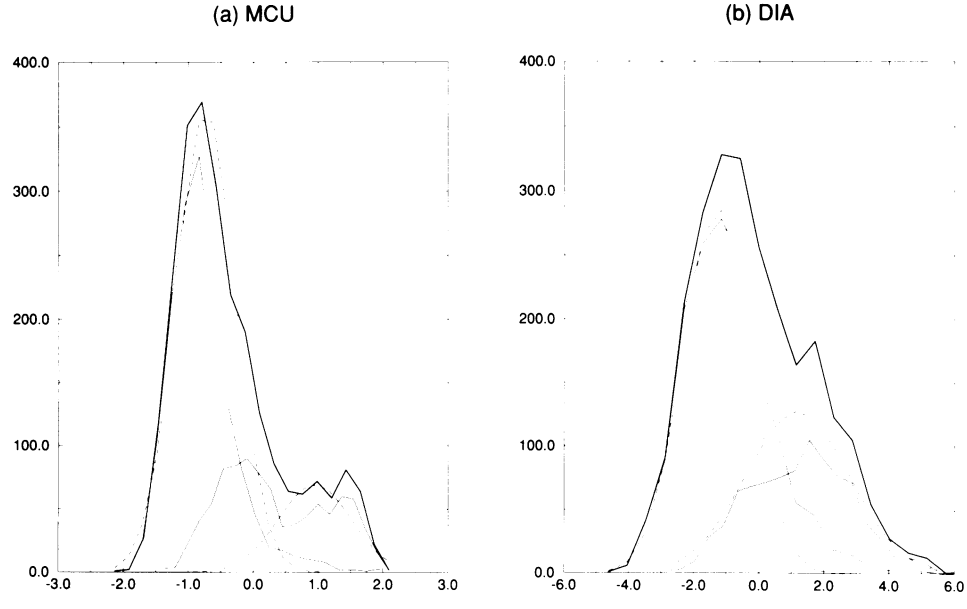

Figure 4. a) Global MCU distribution (thick line) on all sequence windows in the *C.elegans* reference set (GenBank genomic sequences without cosmids), and MCU distributions for coding and non-coding windows based on GenBank annotation (thin lines). The MCU distribution in coding windows departs obviously from normality and the EM algorithm can not be used to infer such a distribution. **b)** Distribution of the DIA coding statistic (thick line) on all sequence windows in the *C.elegans* reference set, and DIA distributions for coding and non-coding windows. Note that the DIA distribution for coding windows is much closer to normal than the MCU distribution (the log-likelihood ratio is 28.2). The corresponding EM estimates are plotted using dashed lines. The resulting estimated proportion of coding windows is 0.47, versus a value of 0.37 derived from (probably incomplete) GenBank annotation.

Therefore we have chosen to use the DIA statistic, instead of the MCU, in an attempt to predict the coding density of the *C.elegans* cosmids. The EM decomposition of the DIA distribution in the *C.elegans* cosmids results in an estimated proportion of coding windows of 0.23. If the empirical confidence interval derived from the human data may serve as a guide here, the 90% confidence interval on this prediction is 0.26 ± 0.10. The value of the log-likelihood ratio is 41.4. For comparison with our estimated value, the tentative, and in part computer-derived, annotation supplied to GenBank with the cosmids gives a density of 0.29.

DISCUSSION

It is of considerable interest to estimate the overall coding density of a genome, as this is one important estimate of the overall complexity of the organism. In addition, determining 'a priori' the coding density of anonymous sequence data in a particular region may be beneficial. It may, for example, influence the approach to further identify the potential genes contained in the sequence, which may be different for sequences with predicted high coding density than for sequences with predicted low coding density. It may also help to establish priorities when in the near future large amounts of sequence data be routinely obtained, since it may be desired to study in more detail first those sequences likely to be rich in genes. In addition, to the extent that the sequence data available is representative of a whole genome, the estimate obtained can be taken as an estimate of the whole genome coding density, thus possibly suggesting the more appropriate strategy to follow towards the complete inventory of an organism's genes.

We have described a method to estimate the coding density of a corpus of sequence data. The method relies on the decomposition of the global distribution of a given coding statistic on the sequence data in two component normal distributions, which are assumed to be the distributions of the coding statistic on the coding and non-coding fractions of the sequence data. The method is based thus on the assumption that the coding statistic employed behaves in a gaussian manner on each of these fractions. We have used an EM algorithm to obtain maximum likelihood estimates for the parameters of the underlying distributions in the hypothetical mixture, and a log-likelihood ratio to evaluate the fit between the estimated mixture and the observed data.

As coding statistic, we have generally used a linear combination of the codon usage vector—the MCU statistic. We have observed that the MCU coding statistic shows stronger correlation with coding density than other simple coding statistics, and that it exhibits a good gaussian behavior. However, we have data indicating that such normal behavior of the MCU statistic may not be universal. In such a case, the EM algorithm can still be applied, provided that an alternative coding statistic with the appropriate normal behavior is found.

To obtain the distribution of the coding statistic, the sequence data needs to be decomposed in windows of fixed length. The choice of window length is somewhat arbitrary, within certain bounds. A longer window length contributes to precision, since the method requires that the correct frame be discernible from codon usage within the window. The window should probably be at least 100 bases for this purpose. On the other hand, windows occasionally span two exons in different phases; and in this case the codon usage counts are obviously less meaningful. For example, we estimate that about 0.3% of windows of length 240 span such introns in human sequences. In addition, a shorter window length minimizes the number of partially coding windows, which also contributes to precision. In that sense, the length we chose, 240, should be considered a compromise. It is quite possible that a longer window would give better results in yeast, where most genes are long open reading frames without interrupting introns, and that a shorter window would be more appropriate to human sequences, where most genes are constituted by rather short exons. For example, while 85% of the coding windows (>50% coding) for the yeast reference set are fully coding windows, only 40% of the them are so for the human reference set.

When tested in large corpus of sequence data (>1 Mb) from a number of distantly related eukaryotic organisms with very different coding densities, the EM decomposition of the MCU distribution resulted in consistent estimates. The estimates were systematically high when compared with the values of the coding density derived from the GenBank annotation. However, such a discrepancy may partially be the result of an underestimation of the actual coding density when derived from the GenBank annotation. The fact that there are sequences in GenBank that are clearly coding but not annotated as such, and that the overprediction is not observed when the EM estimates are obtained only from sequences in GenBank entries containing the keyword 'CDS'—where unannotated coding sequences are less likely to occur—supports such a possibility.

The values of the log-likelihood ratio obtained for the MCU distribution to fit the EM decomposition were often not consistent with the hypothesis of two underlying normals. Although the method requires, in principle, a good fit to gaussian distributions, as measured by this log-likelihood ratio, we have found that the method is in fact quite robust, and that whenever the coding statistic exhibits unimodal, approximately gaussian behavior, the EM estimates are quite reliable. Note, incidentally, that normality can not be expected in a data set as biased as GenBank. Lower values of the log-likelihood ratio (corresponding to more confidence in the hypothesis that the empirical distribution may be fit to a sum of two normal distributions) were found in the two cases where the sequence data were more representative of the genome as a whole, and thus more confidence may be placed in the estimates in this case.

In some cases, confidence intervals for the estimates have been empirically obtained for different sample sizes. Results obtained show that for human sequence data, an estimation of the density can be obtained within $\pm 6\%$ error, when 1 Mb of sequence data is available. For yeast sequences, such accuracy can be obtained with only 300 Kb of sequence data. Since the estimates for such distantly related organisms have been obtained using the same coding statistic and the same window length, it is likely that the precision can be increased by using specialized coding statistics, and by calibrating window length appropriate to the organism's characteristic gene structure. Thus, the robustness of the method to different coding statistics, window lengths, data sets used to calculate the discriminant vector, presence of ambiguous bases in the sequence, etc., has to be carefully analyzed. Preliminary results obtained in the five reference data sets when the discriminant vector was obtained from diaminoacid counts and dicodon counts, instead of codon counts, seem to indicate that the estimates are relatively independent to the choice of coding statistic (as far as the statistic exhibits the appropriate close-to-gaussian behavior.)

We have applied the method described here to two different cases. In the first case, although the actual density of the yeast

chromosome III sequence has not yet been experimentally determined, we believe we have obtained a very accurate estimate. In support of its reliability are first, the good agreement with the value derived from the independent analysis carried out by Oliver *et al.* (1992), and second, the excellent fit obtained for the global MCU distribution to two component normal distributions, together with the good gaussian behavior observed for the MCU statistic in the yeast sequences from GenBank. If the density estimated for yeast chromosome III is representative of the whole yeast genome, we should conclude that about $71 \pm 6\%$ of the yeast DNA is coding for proteins. This value strengthens the general conclusion that eukaryotic coding densities are considerably higher than expected. In the second case, to predict the coding density for the *C.elegans* cosmids we have not used the MCU statistic, since we have observed that its distribution in characterized genomic sequence data from *C.elegans* departs clearly from normality. Instead, we have used a linear discriminant function derived from a diaminoacid usage vector, for which we have observed close to normal behavior. The resulting EM estimate of coding density, 0.23, is lower than the independent estimate of 0.29 implied by the annotation supplied to GenBank with the cosmids (and based on Genefinder, a gene identification program), but is within the limits of variation (-0.03 ± 0.10) observed for human sequence data samples of similar size.

In summary, we have developed a method to estimate the coding density in a corpus of DNA sequence data. The method is not species-specific and can be used across taxonomic boundaries. It can be applied to both long stretches of contiguous DNA sequence or a large number of short sequences scattered over a genome. It requires only a coding statistic with reasonably gaussian behavior in windows of the sequence under study, and does not depend on extrapolating a model of the relationship between coding density and coding statistic inferred from characterized sequence data. Results obtained when tested in a number of distantly related eukaryotic organisms show that usually the method provides useful estimates of the coding density in sets of sequence data containing at least a few hundred kilobases.

ACKNOWLEDGMENTS

This work was performed under the auspices of the U.S. Department of Energy and the Los Alamos Center for Human Genome Studies, with funding by DOE/OHER grant ERFWPF116, and LANL grant X16K. We are grateful for helpful comments from C.Burks, M.J.Cinkosky, R.K.Moyzis, D.C.Torney, and the referees. We gratefully acknowledge also the use of software written by C.-S.Tung for [Fickett & Tung 1992].

REFERENCES

1. Agha, M. and Ibrahim, M.T. (1983) Appl. Stat., 33, 327.
2. Cinkosky, M.J., Fickett, J.W., Gilna,P., and Burks,C. (1991) Science, 252, 1273–1277.
3. Clark, D.V., Rogalski, R.M., Donati, L.M., and Baillie, D.L. (1988) Genetics, 119, 345–353.
4. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) J. Royal Stat. Soc. Ser. B., 39, 1–38
5. Fickett, J.W. and Tung, C.-S. (1992) Nucl. Acids Res., 24, 6441–6450.
6. Hasselblad, V. (1966) Technometrics, 8, 431.
7. Hasselblad, V. (1969) J. Am. Statist. Assoc., 64, 1459–1471.
8. Higgins,D.G., Fuchs,R., Stoehr,P.J., Cameron,G.N. (1992) Nucl. Acids Res., 20 supplement, 2071–2074.
9. Kaback, D.B., Angerer, L.M., and Davidson, N. (1979) Nucl. Acids Res., 6, 2499–2517.
10. Manly, B.F.J. (1986) Multivariate Statistical Methods. Chapman and Hall, London.
11. Oliver S.G., *et al.* (1992) Nature, 357, 38–46.
12. Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989) Science, 245, 1434–1435.
13. Orgel and Crick (1980) Nature, 284, 604–607.
14. Park, E.-C. and Horvitz, H.R. (1986) Genetics, 113, 821–852.
15. Pearson (1894) Phil. Trans. A., 185, 71–110.
16. Redner, R.A. and Walker, H.F. (1984) SIAM Review, 26, 195.
17. Sokal, R.R. and Rolf, F.J. (1981) Biometry. Freeman and Company, New York.
18. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R., and Waterston, R. (1992) Nature, 356, 37–41.
19. Waterston, R., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R., Green, P., Shownkeen, R., Halloran, N., Metzstein, M., hawkins, T., Wilson, R., Berks, M., Du, Z., Thomas, K., Thierry-Mieg, J., and Sulston, J. (1992) Nature Genetics, 1, 114–123.
20. Yoshikawa, A., and Isono, K. (1990) Yeast, 6, 383–401.