Correlation Approach to Identify Coding Regions in DNA Sequences

S. M. Ossadnik,* S. V. Buldyrev,* A. L. Goldberger,‡ S. Havlin,*§ R.N. Mantegna,* C.-K. Peng,*‡ M. Simons,‡¶ and H. E. Stanley*

*Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachussetts 02215 USA, ‡Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, Massachussetts 02215 USA, §Department of Physics, Bar Ilan University, Ramat Gan, Israel, and ¶Department of Biology, Massachussetts Institute of Technology, Cambridge, Massachussetts 02139 USA

ABSTRACT Recently, it was observed that noncoding regions of DNA sequences possess long-range power-law correlations, whereas coding regions typically display only short-range correlations. We develop an algorithm based on this finding that enables investigators to perform a statistical analysis on long DNA sequences to locate possible coding regions. The algorithm is particularly successful in predicting the location of lengthy coding regions. For example, for the complete genome of yeast chromosome III (315,344 nucleotides), at least 82% of the predictions correspond to putative coding regions; the algorithm correctly identified all coding regions larger than 3000 nucleotides, 92% of coding regions between 2000 and 3000 nucleotides long, and 79% of coding regions between 1000 and 2000 nucleotides. The predictive ability of this new algorithm supports the claim that there is a fundamental difference in the correlation property between coding and noncoding sequences. This algorithm, which is not species-dependent, can be implemented with other techniques for rapidly and accurately locating relatively long coding regions in genomic sequences.

INTRODUCTION

One of the major problems facing researchers working with long genomic DNA sequences is the need for a rapid and accurate method of identifying coding regions. Currently, a typical search for a coding region involves scanning the DNA sequence for the presence of an open reading frame (longer than a certain arbitrarily defined length) for both orientations and for all possible frame-shift positions. The identified open reading frames are then searched for canonical intron splice sites and for the existence of cDNA or protein matches by using appropriate data bases. These methods are labor-intensive and require considerable operator participation. In contrast, an ideal technique would be fast and accurate and require only minimal operator input.

Recently, a multiple sensor neural network approach was developed by Uberbacher and Mural (1991) to locate protein-coding regions. Their approach involves calculating the values of a group of seven sensor algorithms over a window of 99 consecutive bp. A neural network training procedure is then performed on a training set of human DNA sequences for optimizing the weights of the different sensor algorithms. This approach has been used to detect coding regions in human DNA with good predictive power. However, because most of those sensor algorithms are species-sensitive, the parameters need to be adjusted for other organisms (especially nonmammalian DNA sequences). Therefore, an algorithm based on a more general principle that can be applied across the entire phylogenetic spectrum without modification would be desirable.

Received for publication 22 February 1994 and in final form 18 April 1994.

Address reprint requests to H. Eugene Stanley, Center for Polymer Studies, Department of Physics, 590 Commonwealth Avenue, Boston, MA 02215. Tel.: 617-353-2617; Fax: 617-353-3783; E-mail: hes@buphyk.bu.edu.

0006-3495/94/07/64/07 $2.00

We have developed such a tool for rapid identification of DNA coding elements based on our observation of the existence of power-law long-range correlations in noncoding, but *not* in coding, sequences (Peng et al., 1992). The key general concept underlying this new technique, which we call the "coding sequence finder" (CSF) algorithm, is to "drag" an observation box along the DNA sequence and to measure continuously the "signal" from a device that quantifies the degree of power-law long-range correlation. Noncoding regions are typically characterized by a correlation that is long-range in that it decays not exponentially but, rather, as a power law. On the other hand, coding regions typically display only short-range correlations, which decay exponentially.

We test the CSF algorithm on a variety of long DNA sequences, including the recently sequenced Yeast III chromosome, which comprises 315,344 bp (Oliver et al., 1992). The algorithm is found to work well when the coding regions are moderately large (over 1000 bp in length). We also confirm its accuracy on long, artificially generated "control" sequences comprised of known coding and known noncoding sub-sequences.

POWER-LAW LONG-RANGE CORRELATIONS

To quantify the correlation properties of a DNA sequence, it is convenient to introduce a graphical or "landscape" representation, termed a *DNA walk* (Peng et al., 1992). For the conventional one-dimensional random walk model (Montroll and Shlesinger, 1984), a walker moves either "up" [$u(i) = +1$] or "down" [$u(i) = -1$] one unit length for each step i of the walk. For the case of an *uncorrelated* walk, the direction of each step is independent of the previous steps. For the case of a *correlated* random walk, the direction of each step depends on the history ("memory") of the walker.

One possible choice for the DNA walk can be defined as follows: the walker steps "up" [$u(i) = +1$] if a pyrimidine (C or T) occurs at position a linear distance i along the DNA chain, whereas the walker steps "down" [$u(i) = -1$] if a purine (A or G) occurs at position i. Other definitions are discussed in Discussion and Summary. A key question is if such a walk displays only short-range correlations (as in an n-step Markov chain) or power-law long-range correlations (as in critical phenomena and other scale-free "fractal" phenomena).

The DNA walk provides a graphical representation for any DNA sequence and permits the degree of correlation in the base pair sequence to be visualized directly. To quantify this correlation, one calculates the "net displacement," $y(\ell)$, of the walker after ℓ steps, which is the sum of the unit steps $u(i)$ for each step i. Thus, $y(\ell) \equiv \sum_{i=1}^{\ell} u(i)$.

One difficulty in analyzing DNA sequences by random walk method is that DNA sequences are highly heterogeneous. Thus, the problem of how to distinguish "patchiness" from truly fractal (scale-invariant) type of behavior needs to be addressed (Karlin and Brendel, 1993). In Peng et al. (1992), a "min-max" method was proposed to take into account the nucleotide heterogeneity and changes in strand bias. A potential drawback of this method is that it requires the investigator to judge how many local maxima and minima of a landscape to utilize in the analysis. Recently, we presented a new method: *"detrended fluctuation analysis"* (*DFA*), that is independent of investigator input and permits the detection of power-law long-range correlations embedded in a patchy landscape, and also avoids the spurious detection of apparent power-law long-range correlations that are an artifact of nucleotide patchiness (Peng et al., 1994).

The DFA method is carried out as follows: first, we divide the entire sequence of length N into N/ℓ nonoverlapping boxes, each containing ℓ nucleotides, and define the "local trend" in each box to be the ordinate of a linear least-squares fit for the DNA walk displacement in that box. Next we define the "detrended walk," denoted by $y_\ell(n)$, as the difference between the original walk $y(n)$ and the local trend. We calculate the variance about the local trend for each box and calculate the average of these variances over all the boxes of size ℓ, denoted $F_d^2(\ell)$. Thus,

$$F_d^2(\ell) \equiv \frac{1}{N} \sum_{n=1}^{N} y_\ell^2(n). \tag{1}$$

It was shown (Peng et al., 1994) that the calculation of $F_d(\ell)$ can clearly distinguish two different types of behavior: (i) $F_d(\ell) \sim \ell^{1/2}$ for patchy but otherwise uncorrelated (or only short-range correlated) sequences, and (ii)

$$F_d(\ell) \sim \ell^{\alpha} \tag{2}$$

with $\alpha \neq 1/2$, if there is no characteristic length for the correlations.

Typical data for $F_d(\ell)$ are linear on double logarithmic plots, confirming that, indeed, $F_d(\ell) \sim \ell^{\alpha}$. A least-squares fit of such data produces a straight line with slope α. It was observed that for coding sequences, $\alpha \approx 1/2$, whereas for noncoding sequences, α is substantially larger than 1/2 (Peng et al., 1992, 1994).

CODING SEQUENCE FINDER (CSF) ALGORITHM

The focus of the CSF algorithm is the calculation of the correlation exponent α for different sub-regions of the DNA sequence. If α, measured from a sub-region, is close to 0.5 it means that this sub-region is more likely to belong to the coding part of the sequence, which is in accord with our finding that the coding sequences do *not* have power-law long-range correlations. If, on the other hand, the value of α for a region is much larger than 0.5, then this region is more likely to belong to the noncoding part of the sequence.

Note, however, that α cannot be calculated for a single nucleotide. Instead, the exponent α, defined by the behavior of the fluctuation $F_d(\ell)$, can be calculated only for a subsequence of nucleotides with length $w \gg \ell$.

Therefore, we have devised the following 6-step procedure:

Step 1. Calculate $F_d(\ell)$ for the subsequence (window of size w) from nucleotide $n - w/2$ to nucleotide $n + w/2$, for a continuous sequence of positions n ranging from the first nucleotide ($n = w/2$) to the last ($n = N - w/2$), where N is the total number of bp.

Step 2. Construct a log-log plot of $F_d(\ell)$ vs. ℓ. The exponent $\alpha \equiv \alpha(n)$ is estimated from the slope of the plot. To calculate the slope, we make a linear regression fit for the data in the range from ℓ_1 to ℓ_2. Thus, the local value of $\alpha(n)$ is a function of window size w and fitting range $[\ell_1, \ell_2]$.

Step 3. Select an appropriate window size w and fitting range $[\ell_1, \ell_2]$. The lower cutoff value ℓ_1 is chosen such that α is not affected by the short-range (Markovian) correlations. Although we prefer to have very large ℓ_2, we must take ℓ_2 much smaller than w, because the error of estimation of α rapidly increases when ℓ_2 approaches w. The ratio w/ℓ_2 represents the number of statistically independent measurements by which the value $F_d(\ell)$ is obtained. The error of α is, therefore, inversely proportional to the square root of this ratio. Indeed, we have shown rigorously (Peng et al., 1993) that the SD σ of the value of α can be calculated by the formula

$$\sigma = C\sqrt{\ell_2/w}, \tag{3}$$

where C is a coefficient that is close to 0.1. Our selection criterion for w and ℓ_2 is that the SD or "error" σ must be much smaller than the difference of α values between coding and noncoding sequences, i.e., the signal-to-noise ratio must be as large as possible.

Our unpublished observations, based on sampling over a wide range of phylogenetic spectrum, reveal that the average value of α for coding regions obtained by DFA for the fitting range $\ell_1 = 10$, $\ell_2 = 100$ is 0.51, whereas for noncoding regions it is 0.59. Therefore, we choose $w \geq 10\ell_2$, which from (3) gives $\sigma \leq 0.03$, an error considerably smaller than the excursions in α between coding and noncoding regions, $0.59 - 0.51 = 0.08$.

Furthermore, there is a trade-off in our choice of parameters: by increasing the window size and the fitting range, one *increases* the accuracy of the value of α but *decreases* the accuracy of locating this value along the sequence.

Step 4. Smooth out the resulting function $\alpha(n)$. The function $\alpha(n)$ is a rather irregular oscillatory function with many minima and maxima. Two factors contribute to this irregular spatial fluctuation: (i) alternating coding and noncoding regions have different exponent α (this is the "signal" that we want); and (ii) the error in estimating α from a finite subsequence (this is the "noise" that we do not want). Therefore, our goal is not to smooth arbitrarily but, rather, only to smooth in such a fashion as to minimize the effect of (ii). The two effects are distinguishable, because the fluctuations that are more likely caused by the noise are "high-frequency" compared with the fluctuations caused by alternation of coding and noncoding regions. For this reason, a simple low-pass filter (Press et al., 1991) is quite effective. Alternatively, we may simply average together $\alpha(n)$ for several nearby values of n. Our preliminary calculations show that both smoothing procedures give similar results.

Step 5. Compare the $\alpha(n)$ function with locations of known coding regions. The smoothed function $\alpha(n)$ usually has minima of about 0.5, which correspond to the local absence of power-law long-range correlations (see Fig. 1). Indeed, comparing the function $\alpha(n)$ for the sequence of yeast chromosome III (for which many of the coding regions are known), we can see that minima of $\alpha(n)$ correspond remark-

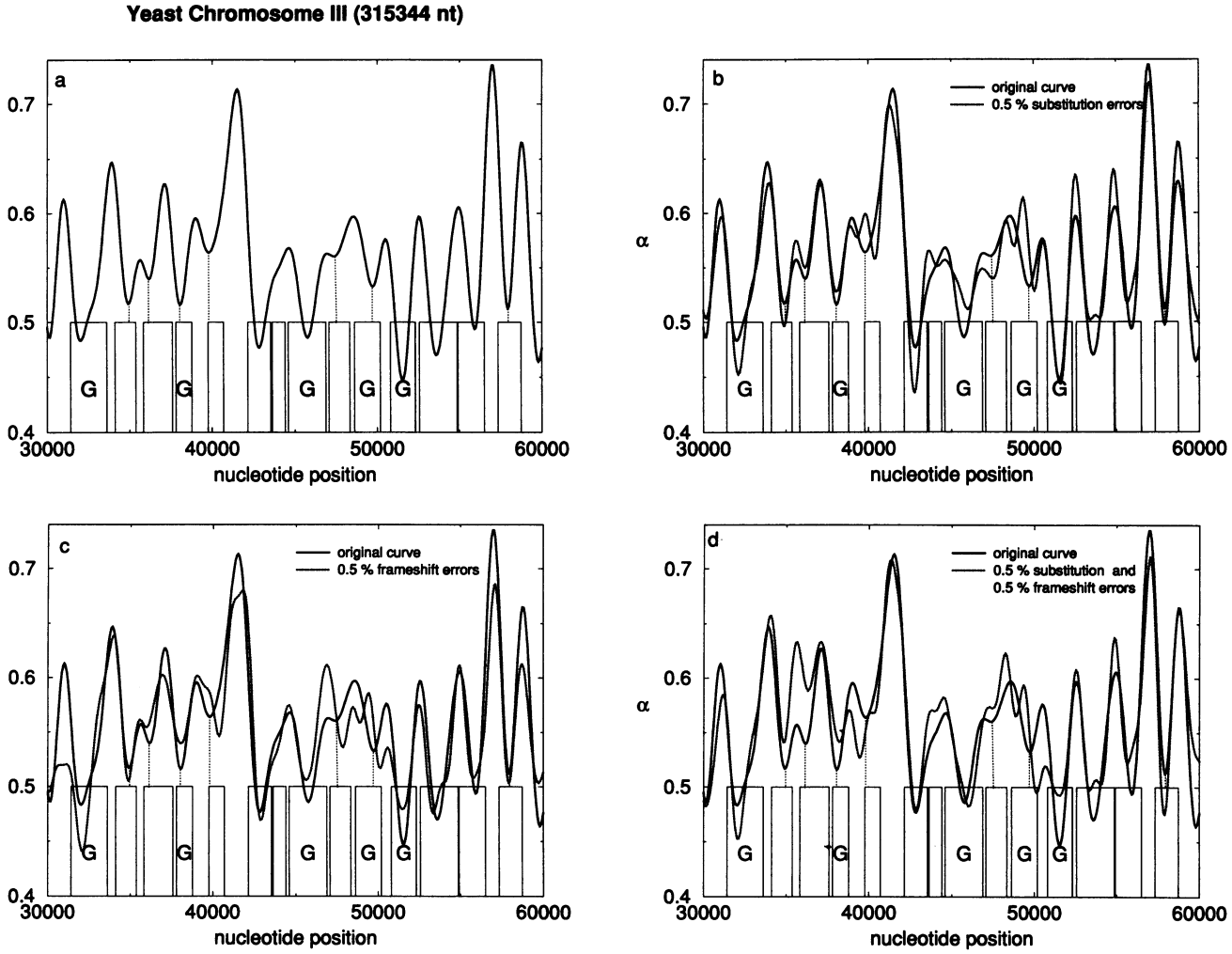

FIGURE 1 (*a*) Analysis of section of yeast chromosome III using the sliding box CSF algorithm. The value of the long-range correlation exponent α is shown as a function of position along the DNA chain. In this figure, the results for about 10% of the DNA are shown (from base pair 30,000 to base pair 60,000). Shown as vertical bars are the putative genes and open reading frames; denoted by the letter "G" are those genes that have been more firmly identified (March 1993 version of *GenBank*). Note that the local value of α typically displays minima where genes are suspected, whereas between the genes α displays maxima. This behavior corresponds to the fact that the DNA sequences of coding regions lack power-law long-range correlations ($\alpha = 0.5$ in the idealized limit), whereas the DNA sequences in between coding regions possess power-law long-range correlations ($\alpha \approx 0.6$). Parameter values: $w = 800$, $\ell_1 = 8$, $\ell_2 = 64$. (*b*) The solid curve is the same as in part *a*, whereas the dotted curve is the same analysis applied after 0.5% of the bp have in the same sequence been randomly mutated. (*c*) The solid curve is the same as in part *a*, whereas the dotted curve is the same analysis applied after 0.5% of the bp in the same sequence have been randomly shifted to a randomly-chosen position. (*d*) The solid curve is the same as in part *a*, whereas the dotted curve is the same analysis applied after both the operations of parts *b* and *c* have been carried out.

ably well to the positions of putative coding regions (identified genes or open reading frames), whereas intergenomic sequences usually correspond to the local maxima of $\alpha(n)$.

Step 6. The above procedure (steps 1–5) outlines the basic CSF algorithm. Step 6 demonstrates how the CSF algorithm can be combined with other local criteria for more precise identification of coding sequences. For example, for yeast chromosome III, where the coding regions are typically uninterrupted by introns, one can incorporate information about open reading frames and stop codons. A reading frame is one of three possible ways of dividing sequences on each of two complementary DNA strands into subsequent codons. An open reading frame is a reading frame without the stop signals TAG, TGA, and TAA. Therefore, to predict the actual boundaries of coding regions from our calculated function $\alpha(n)$, we can carry out the following additional procedure:

(a) Find all local minima of $\alpha(n)$.

(b) Identify the six open reading frames (three on each of the complementary DNA strands).

(c) Define the longest open reading frame corresponding to a minimum value of α. *Long open reading frames without power-law long-range correlations are very likely to be actual coding regions.*

EVALUATION OF THE CSF ALGORITHM

Test for yeast chromosome III

To characterize quantitatively the goodness of our algorithm, we consider the relative positions of local minima, maxima, and the boundaries of coding regions.

For example, the outcome of the CSF algorithm (steps 1–5) for the test case of yeast chromosome III, using the parameter choices $w = 800$, $\ell_1 = 16$, $\ell_2 = 64$ can be characterized by the following table:

- Total number of putative coding regions known from work of others (Oliver et al., 1992): 218.
- Fraction of the 315,344 bp belonging to putative coding regions: $p = 0.66$.
- Number of minima in $\alpha(n)$: 176.
- Number of such minima belonging to putative coding regions (true positives): 138.
- Number of false positives: 38.

Thus, of 176 minima, all but 38 correspond to putative coding regions. A key statistical test of the CSF algorithm is to demonstrate that the apparently striking agreement between the putative coding regions and the dips in $\alpha(n)$ is not simply a result of random coincidence. Therefore, we assume the contrary, i.e., that the dips are occurring at random. Then, because there are 176 minima in our $\alpha(n)$ plot, $176 \times p = 176 \times (0.66) = 116$ of the minima should lie *inside* putative coding regions, and $176 \times (1 - p) = 176 \times (0.34) = 60$ of the minima should lie *outside* putative coding regions. The SD for the above estimation (assuming that these 176 minima are occurring at random) is given by the formula $\sigma = \sqrt{176 \times p \times (1 - p)}$. Hence, in the present case, we would expect $\sigma = \sqrt{176 \times 0.66 \times 0.34} = 6.3$. The actual number of false positives is 38, three SD smaller than the expected value 60. The probability of obtaining this result if the minima did not correspond to the coding regions, therefore, is the chance of finding a signal 3 SD from the expected value, or 0.0014.

The combination of the CSF algorithm (based on global criteria of power-law long-range correlations) with local criteria (stop codons; see step 6 above) is very successful in identifying the precise boundaries of long coding regions. It enables us to identify correctly in the yeast chromosome III 100% of the putative coding regions with more than 3000 nucleotides, 92% of coding regions with between 2000 and 3000 nucleotides long and 79% of coding regions between 1000 and 2000 nucleotides long.

The "false positives" identified by the CSF algorithm might actually indicate the presence of former coding material, such as pseudo-genes, jumping genes, and retroviral inserts. For example, for yeast chromosome III, we found a clear minimum in $\alpha(n)$ near the position $n = 149200$, a region that is known to contain primarily noncoding sequences. We submitted the sequence from nucleotide position 149120 to nucleotide position 149401 to the experimental GENINFO BLAST (Altschul et al., 1990) network at the National Center for Biotechnology Information, which indicated a remarkably high similarity score of the submitted sequence to the jumping gene known as retroelement Ty4-476.

We measured "false negatives" by focusing on a subset of all open reading frames, those of more than 1000 bp. We define false negatives to be the absence of an unambiguous pronounced minimum in $\alpha(n)$. We find that the CSF algorithm fails to locate only 12% (3/25) of the known genes, but fails to locate 27% (16/60) of the open reading frames that are not known to be genes. Thus, the CSF algorithm is much more successful on the known genes than on putative genes.

Robustness of the CSF algorithm with respect to sequencing errors

We performed three tests designed to test the robustness of the CSF algorithm with respect to sequencing errors, by intentionally "mutating" a small fraction of the bp: 0.5%. The CSF algorithm would be most useful if it could still be able to identify most of the coding regions. Fig. 1 *b* shows the same region as shown in Fig. 1 *a* of the paper. The solid curve is identical to the solid curve in Fig. 1 *a*, namely, the CSF algorithm result for the long-range correlation exponent α as a function of position in yeast chromosome III. The dotted curve is the same analysis applied to an "artificially mutated" yeast chromosome III in which 0.5% of the bp were randomly substituted by a different bp. There is almost no difference between the two curves whatsoever.

Another test is to randomly shift the reading frame, as shown in Fig. 1 *c*. The solid curve is the same as in Fig. 1. The dotted curve is the same analysis applied to a mutated chromosome III, in which 0.5% of the bp were randomly cut out from one place and inserted in another randomly selected

locus of the chromosome in such a way as to shift the reading frame. There is almost no change whatsoever.

Finally, we show in Fig. 1 *d* what happens when *both* operations of Fig. 1, *b* and *c*, have been carried out, so that in fact twice as many (1%) of the bp are mutated in total.

Parameter optimization

We note that the accuracy of the algorithm depends importantly on the window size parameter. Fig. 2 shows the dependence on window size of the fraction of minima that occur in coding regions for yeast chromosome III. Reducing the window size increases the number of true positives (the sensitivity of the algorithm) but increases the number of "false positives" and "false negatives." This is the reason the algorithm in its present form is challenged for finding genes in mammalian sequences, which are highly fragmented by introns. Although the average size of an exon in mammalian DNA is only about 186 bp (Watson et al., 1992) (close to the lower threshold of applicability of our algorithm), there are many mammalian exons larger than the average that our method should detect readily.

We have also studied how the fluctuations of the DNA landscapes created by other rules of mapping can be used for detecting coding regions as well as various *two-dimensional* DNA walks (Berthelsen et al., 1992). In the generalized definition of a one-dimensional DNA walk, one can assign different values S_A, S_C, S_G, or S_T to an increment of the ith step $u(i)$ depending on which nucleotide A, C, G, or T occurs on the ith place. For example, we studied correlations of one nucleotide with itself; in this case, one can assign $u(i) = +1$ if nucleotide A occurs on the ith place and $u(i) = -1$ otherwise (in case of C, G, or T). Similarly, we studied correlations of pairs of nucleotides, such as the purine-pyrimidine rule we used above. Except for the definition of $u(i)$, the rest of the analysis remains the same as for the original purine-pyrimidine rule. Our calculations show that the original binary purine-pyrimidine rule (Peng et al., 1992) is the most robust one for detecting coding regions.

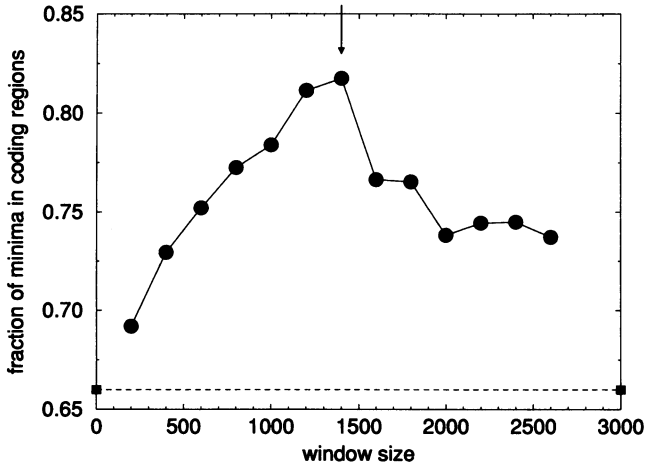

FIGURE 2 Dependence of sensitivity of CSF algorithm on window size, w, for yeast chromosome III. Sensitivity here is defined as the percentage of the minima of α that lie within putative coding regions (see Fig. 1). Window size is defined in the text (Step 1 of the algorithm). The solid circles show the results for the yeast chromosome III. For window sizes less than 600, the fitting range was chosen to be $\ell_1 = 8$, $\ell_2 = 32$. For window sizes larger or equal 600, we chose $\ell_2 = 64$. The vertical arrow indicates the optimal nucleotide window size. The dashed line is the expected sensitivity of the algorithm based on random coincidence (see text).

Test of the CSF algorithm on other long genomic sequences

We also applied the CSF algorithm (steps 1–5) to four additional long genomic sequences and observed comparable predictability as for the yeast chromosome III. The sequences were: liverwort marchantia polymorpha chloroplast genome (GenBank name: CHMPXX, 121024 bp; 59% coding region, CSF sensitivity = 74% with window size $w = 1000$); tobacco chloroplast genome DNA (CHNTXX, 155844 bp; 53% coding regions, CSF sensitivity = 72% with window size $w = 1200$); rice complete chloroplast genome (CHOSXX, 134525 bp; 50% coding regions, CSF sensitivity = 68% with window size $w = 2200$); and Epstein-Barr virus (EBV genome, 172281 bp; 71% coding, CSF sensitivity = 90% with window size $w = 1400$).

Test of the CSF algorithm on control sequences

Not all "coding regions" for the yeast chromosome III and other genomic sequences we tested are confirmed (in fact, they are termed "putative coding regions" (Oliver et al., 1992)). To obtain additional evidence about the reliability of the CSF algorithm, we analyzed "control sequences" that contain only firmly identified coding and noncoding regions. To this end, we have selected from GenBank 40 known coding sequences (including exons and cDNA sequences) and 39 known noncoding sequences (including introns and intergenomic sequences). These samples (total length 80,000 bp) represent a wide phylogenetic spectrum (including sequences from human, chicken, tobacco, bacterial, and viral DNA). The selection criteria for these sequences are: (i) they are all of length greater than 500 bp, and (ii) the percentage of coding and noncoding material approximates that of yeast chromosome III.

Next we "assembled" an artificial nucleotide sequence ("Type I controls") by randomly splicing together coding and noncoding sequences (in an alternating fashion) from the two sample pools. We then applied the CSF algorithm to this control sequence and computed the number of minima inside and outside the known coding regions. We found that for window size $w = 800$, almost 90% of the minima coincided with coding regions. The percentage of correct identifications decreased to 60% with increases or decreases in w, which is comparable with the results obtained for the actual yeast chromosome III sequence (Fig. 3).

This test confirms that for coding and noncoding sequences of length larger than 500 bp, the CSF algorithm is highly accurate. It also illustrates another generic feature of

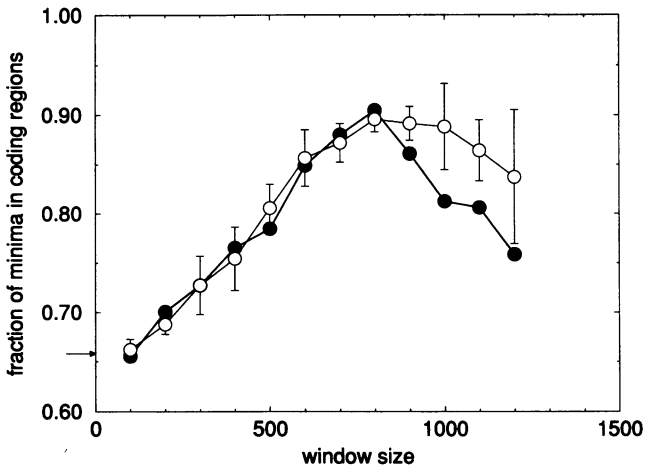

FIGURE 3 Dependence of sensitivity of CSF algorithm on window size, w, for control sequences. Sensitivity is defined in Fig. 2 caption. The solid circles show the results for a Type I control sequence, and the open circles show the averaged results for three Type II control sequences (see text). The error bars show the SD. The horizontal arrow on the left indicates the percentage of coding regions in the Type I control sequence (66%). We started our analysis for window size 100 and increased the window size up to 1200. For larger window sizes, the total number of minima decreases (down to 15), thus the statistical error increases. For small window sizes (~100), the signal is very noisy, so that the detection rate is about the value expected for a random signal, i.e., 66%. With increasing window sizes, the sensitivity of the CSF algorithm increases. The maximum sensitivity of the CSF algorithm for detecting coding regions (90%) for both classes of control sequences is obtained with window size 800.

the CSF algorithm, i.e., that it can, in principle, be applied to DNA sequences of very different organisms because the underlying mechanism for detecting coding sequences is the same.

Finally, we tested the CSF algorithm on a second type of control sequences ("Type II controls") constructed as follows: for the artificial chromosome sequence described above, we replaced each coding part by an uncorrelated computer-generated sequence of random letters A, C, G, and T. We also replaced each noncoding sequence by a computer-generated sequence of letters with "built-in" power-law long-range correlations having correlation parameter α of 0.62, using the method in Peng et al. (1993). We then calculated the percentage of correct positives for several independent realizations of such a sequence, and we computed the SE of this value. The result shows 90% sensitivity for $w = 800$.

When we compare Fig. 3 with Fig. 2, we find that our highest sensitivity for the artificial chromosome sequence is 90%, whereas for the yeast chromosome III sequence, we achieved a sensitivity of 82% with optimum parameter selection. It is not surprising that the results for the artificial chromosome sequence are somewhat better: (i) we eliminated the problem of the putative coding regions, and (ii) we only considered coding and noncoding regions larger than 500 nucleotides.

We note that the value of the exponent α measured over typically $<10^2$ bp is also highly correlated with certain other previously described quantities such as the length distribution of tandem nucleotide repeats (Uberbacher and Mural, 1991). This is not surprising because the tandem repeats of length more than 10 may contribute to the value of α calculated for these small fitting ranges. However, tandem repeats by themselves do not fully account for the power-law long-range correlation we observed. Furthermore, because power-law long-range correlations with $\alpha > 0.5$ generate a type of *persistence* (one nucleotide is more likely to be followed by another of the same class), tandem repeats are more likely to be found in correlated rather than uncorrelated sequences.

DISCUSSION AND SUMMARY

The results of the CSF analysis presented in this study are of interest for two primary reasons:

First, these results provide the most compelling evidence to date confirming the claim that noncoding sequences typically possess long-range power law correlations whereas coding sequences do not. The initial report (Peng et al., 1992) describing long-range (scale-invariant) correlations only in noncoding DNA sequences generated contradictory responses (Karlin and Brendel, 1993; Li and Kaneko, 1992; Munson et al., 1992; Grosberg et al., 1993; Nee, 1992; Chatzidimitriou-Dreismann and Larhammar, 1993; Voss, 1992; Prabhu and Claverie, 1992). Although some reports supported this finding (Li and Kaneko, 1992; Munson et al., 1992; Grosberg et al., 1993), it has also been challenged on two fronts: (i) by those claiming that *no* DNA sequences possess power-law long-range correlations (Karlin and Brendel, 1993; Nee, 1992; Chatzidimitriou-Dreismann and Larhammar, 1993; Prabhu and Claverie, 1992), and (ii) by those claiming that introns and exons both contain power-law long-range correlations (Voss, 1992). The data presented above and graphically displayed in Figs. 1–3 unambiguously confirm that there is a systematic correspondence between lower values of the scaling exponent α and coding sequences, and between higher values of α and noncoding sequences. Furthermore, these results apply in a statistically significant way both to the entire yeast chromosome III and other long genomic sequences as well as to control sequences constructed by alternating known coding and noncoding sequences of variable lengths. These findings, along with a recent re-analysis of the patchiness of DNA sequences (Peng et al., 1994), disprove the contention of Karlin and Brendel (1993) that power-law long-range correlations are simply an artifact of the heterogeneous (mosaic) structure of DNA. Furthermore, the results of the CSF analysis contradict Voss' (1992) report that power-law long-range correlations are found in both coding and noncoding sequences.

We also note the recent study by Prabhu and Claverie (1992) claiming that their analysis of the putative *coding* regions of the yeast chromosome III produced a *wide range of exponent values*, some larger than 0.5. Thus, they too failed to find statistical difference, based on the correlation exponent, between coding and noncoding regions. In contrast, the CSF analysis does demonstrate statistically sig-

nificant agreement between dips in $\alpha(n)$ and the presence of putative coding regions for yeast chromosome III. This apparent discrepancy results from the fact that Prabhu and and Claverie (1992) as well as Karlin and Brendel (1993) and Voss (1992) did not account for the presence of long regions of strand bias coding sequences. As recently reported (Peng et al., 1994), the detrended fluctuation analysis (DFA) method (used in the CSF algorithm) can successfully distinguish true power-law long-range correlations (e.g., those in noncoding sequences) from spurious correlations due to DNA "patchiness."

The fact that the value of α for coding regions is close to that of random uncorrelated sequences might be relevant to the theory of protein folding and is consistent with the recent claim of Shakhnovich and Gutin (1990) that the amino acid sequences of biologically active proteins are also statistically similar to uncorrelated random sequences. In contrast, the correlated properties of noncoding sequences might be related to evolutionary mechanisms involving nucleotide deletion and insertion (Buldyrev et al., 1993a, b).

Second, we show how the new algorithm based on these biologic differences in correlation properties can be used to screen long DNA sequences to identify coding and noncoding regions. The CSF algorithm is able to detect relatively long coding regions with a high degree of reliability. Its advantages include speed, simplicity of use, and operator independence. Furthermore, because it is based primarily on *global* statistical measurements, it is not affected by the particular species examined or by sequencing errors. Its major limitations relate to the requirement for a relatively large window (>800 bp) and the inability to precisely locate intron/exon boundaries. Given these limitations, the optimal application of the CSF algorithm might be to rapidly scan large genomic sequences, to identify any potential coding sites, and then to apply standard coding-sequence finding tools to a detailed analysis of these selected areas. Indeed, identification of even a single putative exon would imply the nearby location of a gene that can then be searched for with conventional techniques.

The CSF algorithm is particularly attractive because it can be applied to sequences from organisms across the phylogenetic spectrum. Furthermore, because it is based on a *global* statistical measurement, it is not affected by local mutation or lab sequencing errors. On the other hand, its global statistical nature, as emphasized above, limits its ability to locate precisely the boundaries of coding and noncoding regions. Therefore, in its present form, CSF can be used in concert with other algorithms (see "Step 6" in our procedure) that apply local property measurements.

We wish to thank C. R. Cantor, C. DeLisi, and T. M. Sanders for valuable discussions, and an anonymous referee for having suggested the calculations that led to Fig. 1 *b–d*.

Partial support was provided to C.-K. Peng by National Institutes of Health/National Institutes of Mental Health, to A. L. Goldberger by the G. Harold and Leila Y. Mathers Charitable Foundation, the National Heart, Lung and Blood Institute, and the National Aeronautics and Space Administration, to M. Simons by the American Heart Association, and to S. V. Buldyrev, S. Havlin, R. N. Mantegna, S. M. Ossadnik, and H. E. Stanley by the National Science Foundation and Office of Naval Research.

REFERENCES

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

Berthelsen, C. L., J. A. Glazier, and M. H. Skolnick. 1992. Global fractal dimension of human DNA sequences treated as pseudorandom walks. *Phys. Rev. A.* 45:8902–8913.

Buldyrev, S. V., A. L. Goldberger, S. Havlin, C.-K. Peng, H. E. Stanley, M. H. R. Stanley, and M. Simons. 1993a. Fractal landscapes and molecular evolution: modeling the myosin heavy chain gene family. *Biophys. J.* 65:2673–2679.

Buldyrev, S. V., A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley. 1993b. Generalized Lévy walk model for DNA nucleotide sequences. *Phys. Rev. E.* 47:4514–4523.

Chatzidimitriou-Dreismann, C. A., and D. Larhammar. 1993. Long-range correlations in DNA. *Nature.* 361:212–213.

Grosberg, A. Yu., Y. Rabin, S. Havlin, and A. Nir. 1993. Self-similarity in DNA structure. *Europhys. Lett.* 23:373–378.

Karlin, S., and V. Brendel. 1993. Patchiness and correlations in DNA sequences. *Science.* 259:677–680.

Li, W., and K. Kaneko 1992. Long-range correlations and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. *Europhys. Lett.* 17: 655–658.

Montroll, E. W., and M. F. Shlesinger. 1984. The wonderful world of random walks. *In* Nonequilibrium Phenomena II: From Stochastics To Hydrodynamics. J. L. Lebowitz and E. W. Montroll, editors. North-Holland, Amsterdam. 1–121.

Munson, P. J., R. C. Taylor, and G. S. Michaels. 1992. Long range DNA correlations extend over entire chromosome. *Nature.* 360:636–636.

Nee, S. 1992. Uncorrelated DNA walks. *Nature.* 357:450–450.

Oliver, S. G. et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature.* 357:38–46.

Peng, C.-K., S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley. 1992. Long-range correlations in nucleotide sequences. *Nature.* 356:168–170.

Peng, C.-K., S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, and H. E. Stanley. 1993. Finite size effects on long-range correlations: implications for analyzing DNA sequences. *Phys. Rev. E.* 47:3730–3733.

Peng, C.-K., S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. 1994. On the mosaic organization of DNA sequences. *Phys. Rev. E.* 49:1685–1689.

Prabhu, V. V., and J.-M. Claverie. 1992. Correlations in intronless DNA. *Nature.* 359:782–782.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1991. Numerical Recipes in C: The Art of Scientific Computing. Cambridge University Press, Cambridge.

Shakhnovich, E. I., and A. M. Gutin. 1990. Implication of thermodynamics of protein folding for evolution of primary sequences. *Nature.* 346: 773–775.

Uberbacher, E. C., and R. J. Mural. 1991. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA.* 88:11261–11265.

Voss, R. 1992. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys. Rev. Lett.* 68:3805–3808.

Watson, J. D., M. Gilman, J. Witkowski, and M. Zoller. 1992. Recombinant DNA. Scientific American Books, New York.