A Novel Statistical Approach to Identify Coding Regions in DNA Sequences

S. M. Ossadnik[1], S. V. Buldyrev[1], A. L. Goldberger[2],
S. Havlin[1,3], R.N. Mantegna[1], C.-K. Peng[1,2], M. Simons[2,4] & H. E. Stanley[1]

[1]Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA

[2]Cardiovascular Division, Harvard Medical School, Beth Israel Hospital, Boston, MA 02215, USA

[3]Department of Physics, Bar Ilan University, Ramat Gan, ISRAEL

[4]Department of Biology, MIT, Cambridge, MA 02139, USA

ABSTRACT

Recently it was observed that non-coding regions of DNA sequences possess long-range power-law correlations, while 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 sequence to locate possible coding regions. The algorithm is particularly successful in predicting the location of coding regions $>$ 500–1000 base pairs (bp) in length. For example, for the complete genome of yeast chromosome III (315,344 bp), 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 confirms the claim that there is a fundamental difference in the correlation property between coding and non-coding 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.

I. 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, accurate and require only minimal operator input.

Recently, a multiple sensor neural network approach was developed by Uberbacher and Mural [1] to locate protein-coding regions. Their approach involves calculating the values of a group of seven sensor algorithms over a window of 99 consecutive base pairs. 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, due to the fact that most of those sensor algorithms are species-sensitive, the parameters need to be adjusted for other organisms (especially non-mammalian DNA sequences). Therefore, an algorithm based on a more general principle which can be applied across the entire phylogenetic spectrum without modification would be desirable.

We have developed such a tool for rapid identification of DNA coding elements based on our observation of the existence of long-range correlations in non-coding but *not* in coding sequences [2]. 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 long-range correlation. Non-coding 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 [3]. The algorithm is found to work well when the coding region is moderately large (over 1000 bp in length). We also confirm its accuracy on long artificially generated "control" sequences comprised of known coding and known non-coding sub-sequences.

II. Long-Range Correlations

In order to quantify the correlation properties of a DNA sequence, it is convenient to introduce a graphical or "landscape" representation, termed a *DNA walk* [2]. For the conventional one-dimensional random walk model [4], 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, while the walker steps "down" $[u(i) = -1]$ if a purine (A or G) occurs at position i. Other definitions are discussed below. A key question is whether such a walk displays only short-range correlations (as in an n-step Markov chain) or 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 directly visualized. 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-invariance) type of behavior needs to be addressed [5]. In Ref. [2], 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 long-range correlations embedded in a patchy landscape, and also avoids the spurious detection of apparent long-range correlations that are an artifact of nucleotide patchiness [6].

The DFA method is carried out as follows: First, we divide the entire sequence of length N into N/ℓ non-overlapping 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

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

It was shown [6] 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, $\alpha \approx 1/2$, while for non-coding sequences, α is substantially larger than 1/2 [2,6].

III. 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 in accord with our finding that the coding sequences do *not* have 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 non-coding part of the sequence.

Note, however, 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 $w >> \ell$.

Therefore, we have devised the following 7-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 base pairs.

Step 2. Construct a log-log plot of $F_d(\ell)$ versus ℓ. The exponent $\alpha \equiv \alpha(n)$ is estimated from the slope of the plot. In order 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 ℓ_1, ℓ_2.

Step 3. Select an appropriate window size w and fitting range ℓ_1, ℓ_2. The lower fitting range 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 [7] that the standard deviation 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 standard deviation or "error" σ must be much smaller than the difference of α values between coding and non-coding 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, while for non-coding 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 non-coding 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 non-coding regions have different exponent α (this is the "signal" that we want) ; and (ii) the error in estimating α from a finite sub-sequence (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, since the fluctuations that are more likely due to the noise are "high-frequency" compared to the fluctuations due to alternation of coding and non-coding regions. For this reason, a simple low-pass filter [8] is quite effective. Alternatively, we may simply average together $\alpha(n)$ for several nearby values of n. Our preliminary calculations show that both averaging 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 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 remarkably well to the positions of putative coding regions (identified genes or open reading frames), while intergenomic sequences usually correspond to the local maxima of $\alpha(n)$.

Step 6. In order to quantitatively characterize 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 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 [3]:	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, since 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 standard deviation 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 standard deviations smaller than the expected value 60. The probability of obtaining this result if the minima did not correspond to the coding regions is therefore the chance of finding a signal 3 standard deviations from the expected value, or 0.0014.

The above procedure (step 1 to 6) outlined our CSF algorithm. However, to demonstrate that the CSF algorithm can be combined with other local criteria for more precise identification of coding sequences, we include the following optional step 7 as an example.

Step 7. For yeast chromosome III, most coding sequences contain a stop codon. Therefore, to actually predict the boundaries of coding regions from our calculated function $\alpha(n)$, we carry out the following procedure:

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

(b) Define the largest of all 6 open reading frames, that contain each minimum.* *Large open reading frames without long-range correlations are very likely to be actual coding regions.*

The combination of the CSF algorithm (global criteria of long-range correlations) with some local criteria (such as the stop codons) is very successful. It enables us to correctly identify in the yeast chromosome III 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.

We also note that some of the "false positives" may actually indicate the presence of former coding material, such as pseudo-genes, jumping genes, 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 non-coding sequences. We submitted the sequence from nucleotide position 149120 to nucleotide position 149401 to the experimental GENINFO BLAST [9] 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.

IV. Optimization of the CSF algorithm

We have systematically studied how varying the window size, fitting range and smoothing parameter affect the accuracy of the algorithm (see Fig. 2). Note that 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 why 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 [10], (close to the lower threshold of applicability of our algorithm), there are many exons larger than the average which our method should detect readily.

We 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 [11]. 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 i^{th} step $u(i)$ depending on which nucleotide A, C, G, or T occurs on the i^{th} place. For example we can study correlations of one nucleotide with itself, in this case, one can assign $u(i) = +1$ if nucleotide A occurs on the $i^{\rm th}$ place and $u(i) = -1$ otherwise (in case of C, G, or T). Similarly, we can study 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 [2] is the most robust one for detecting coding regions.

V. Test of the CSF Algorithm on Other Long Genomic Sequences

We also applied the CSF algorithm 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% of this sequence code and among them 74% were located by CSF with window size $w = 1000$); tobacco chloroplast genome DNA (CHNTXX, 155844 bp; 53% coding regions, 72% were located by CSF with window size $w = 1200$); rice complete chloroplast genome (CHOSXX, 134525 bp; 50% coding regions, 68% were located by CSF with window size $w =$); and Epstein-Barr virus (EBV genome, 172281 bp; 71% coding regions, 90% were located by CSF with window size $w = 1400$).

VI. Test of the CSF Algorithm on Control Sequences

[*] A reading frame is one of three possible ways of dividing sequences on each of two DNA strands into subsequent codons. An open reading frame is a reading frame without the stop signals TAG, TGA, and TAA.

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" [3]). In order to obtain additional evidence about the reliability of the CSF algorithm, we analyzed "control sequences" that contain only firmly identified coding and non-coding regions. To this end, we have selected from GenBank 40 known coding sequences (including exons and cDNA sequences) and 39 known non-coding 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 non-coding material approximates that of yeast chromosome III.

Next we "assembled" an artificial nucleotide sequence ("Type I controls") by randomly splicing together coding and non-coding 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, comparable to the results obtained for the actual yeast chromosome sequence (Fig. 3).

This test confirms that for coding and non-coding sequences of length larger than 500 bp, the CSF algorithm is highly accurate. It also illustrates another generic feature of the CSF algorithm, i.e., it can in principle be applied to DNA sequences of very different organisms since 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 non-coding sequence by an computer-generated sequence of letters with "built-in" long-range correlations having correlation parameter α of 0.62, using the method in Ref. [7]. We then calculated the percentage of correct positives for several independent realizations of such a sequence and we computed the standard error of this value. The result shows 90% of correct positives for $w = 800$.

When we compare Fig. 3 to Fig. 2, we find that our highest sensitivity for the artificial chromosome sequence is 91%, whereas for the yeast chromosome III sequence we achieved a sensitivity of 82% with optimum parameters selection. It is not surprising that the results for the artificial chromosome sequence are better: (i) We eliminated the problem of the putative coding regions. (ii) We only consider coding and non-coding regions larger than 500 nucleotides.

The difference of the optimum window size, i.e., for our artificial chromosome sequence the maximum at window size 800 and for the yeast chromosome III sequence at window size 1500, is actually related to the length distribution of the coding and non-coding sequences for the two different examples we studied. As we noticed that the average length of coding sequences in yeast chromosome III is much longer than that in our control sequences.

VII. Discussion and Summary

The results of the CSF analysis are of interest for two primary reasons:

First, these results are notable because they provide the most compelling evidence to date confirm the claim that non-coding sequences typically possess long-range power law correlations while coding sequences do not. The initial report [2] describing long-range (scale-invariant) correlations only in non-coding DNA sequences generated contradicting responses [5,12–18]. While some reports supported this finding [12–14], it has also been challenged on two fronts: (i) by those claiming that *no* DNA sequences possess long-range correlations [5,15,16,18], and (ii) by those claiming that introns and exons both contain long-range correlations [17]. 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 non-coding sequences. Furthermore, these results apply in a statistically significant way both to the entire yeast III chromosome as well as to control sequences constructed by alternating known coding and non-coding sequences of variable lengths. These findings, along with a recent re-analysis of the patchiness of DNA sequences [6], disprove the contention of Karlin and Brendel [5] that long-range correlations are simply an artifact of the heterogeneous (mosaic) structure of DNA. Furthermore, the results of the CSF analysis contradict Voss's [17] report that long-range correlations are found in both coding and non-coding sequences.

We also note the recent study by Prabhu and Claverie [18] 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 non-coding regions. In contrast, our CSF analysis does demonstrate statistically significant 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 Ref. [18] as well as Refs. [5] and [17] did not account for the patchy nature of coding sequences which typically contain long regions of strand bias. As recently reported [6], the detrended fluctuation analysis (DFA) method (used in the CSF algorithm) can successfully distinguish true long-range correlations (e.g. those in non-coding sequences) from spurious correlations due to DNA "patchiness".

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 non-coding 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, since 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 (greater than 800 bp) and the inability to precisely locate intron/exon boundaries. Given these limitations, the optimal application of the CSF algorithm may be to rapidly scan large genomic sequences, to identify any potential coding sites, and then to apply standard coding-sequence finding tools to an analysis of the 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, since 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 precisely locate the boundaries of coding and non-coding regions. Therefore, in its present form, CSF can be used in concert with other algorithms (see Step **7** in our procedure) that apply local properties measurements.

Finally, we find that the value of the exponent α measured over relatively short range (e.g., $< 10^2$ bp) is also highly correlated with certain other previously described quantities such as the length distribution of tandem nucleotide repeats [ref]. This is not surprising since 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 long-range correlation we observed. Furthermore, since 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.

1 E.C. Uberbacher, R.J. Mural: Proc. Natl. Acad. Sci. USA **88**, 11261 (1991).

2 C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley: Nature **356**, 168 (1992).

3 S.G. Oliver et al.: Nature **357**, 38 (1992)

4 E.W. Montroll, M.F. Shlesinger: "The Wonderful World of Random Walks" in: *Nonequilibrium Phenomena II. From Stochastics to Hydrodynamics*, ed. by J.L. Lebowitz, E.W. Montroll (North-Holland, Amsterdam 1984), pp. 1–121

5 S. Karlin, V. Brendel: Science **259**, 677 (1993)

6 C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, A.L. Goldberger: Phys. Rev. E **49**, xxx (1994)

7 C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, M. Simons, H.E. Stanley: Phys. Rev. E **47**, 3730 (1993)

8 W. H. Press, B. P. Flannery, S. A. Teukolsky, W.T. Vetterling, *Numerical Recipes in C - The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1991)

9 S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic Local Alignment Search Tool," *J. Mol. Biol.* **215**, 403–410 (1990).

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

11 C.L. Berthelsen, J.A. Glazier, M.H. Skolnick: Phys. Rev. A **45**, 8902 (1992)

12 W. Li, K. Kaneko: Europhys. Lett. **17**, 655 (1992)

13 P.J. Munson, R.C. Taylor, G.S. Michaels: Nature **360**, 636 (1992)

14 A.Yu. Grosberg, Y. Rabin, S. Havlin, A. Neer: Europhys. Lett. **23**, 373 (1993)

15 S. Nee: Nature **357**, 450 (1992)

16 C.A. Chatzidimitriou-Dreismann, D. Larhammar: Nature **361**, 212 (1993)

17 R. Voss: Phys. Rev. Lett. **68**, 3805 (1992)

18 V.V. Prabhu, J.-M. Claverie: Nature **357**, 782 (1992)

Fig. 1: Analysis of section of yeast chromosome III using the sliding box *Coding Sequence Finder* "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, while between the genes α displays **maxima**. This behavior corresponds to the fact that the DNA sequences of coding regions lack long-range correlations ($\alpha = 0.5$ in the idealized limit), while the DNA sequences in between coding regions possess long-range correlations ($\alpha \approx 0.6$).

Fig. 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. The vertical arrow shows the optimal nucleotide window size. The open circles show the results for a randomized sequence. The randomized sequence was produced by shuffling nucleotides in each coding and non-coding region separately (without shuffling across different regions), thus preserving the coding and non-coding structure but destroy the correlation inside each sub-sequence. For the randomized sequence, the fraction of minima inside

coding regions are close to 0.66, the value we would expect by random coincidence. The significant higher accuracy for locating coding sequences for yeast chromosome III indicates that our results for the yeast chromosome III sequence are far from being just random. For window sizes less than 600 the fitting range was chosen to be $l_1 = 8$, $l_2 = 32$. For window sizes larger or equal 600, we choose $l_2 = 64$.

Fig. 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 (see text), the open circles show the averaged results for three Type II control sequences. The error bars show the standard deviation. The horizontal arrow on the left indicates the fraction of coding length 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 ($\sim$ 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 fraction of minima lying inside the coding regions increases. The maximum sensitivity of the CSF algorithm for detecting coding regions (90%) is obtained with window size 800bp. From window sizes 100 to 800 the solid and the open circles are in good agreement. For window sizes larger than 800, the results for the Type I control sequence are not as good as those for the Type II control sequences.