Building Innovative Representations of DNA Sequences to Facilitate Gene Finding

Jianbo Gao, *University of Florida*

Yinhe Cao, *BioSieve*

Yan Qi, *Johns Hopkins University*

Jing Hu, *University of Florida*

Indices that can discriminate DNA sequences' coding and noncoding regions are crucial elements of a successful gene identification algorithm. Multiscale analysis of various species' genome sequences facilitates construction of novel codon indices.

Finding genes is one of the most important tasks in genome research. The computational approaches based on comparative search and on Markov or hidden Markov models require considerable knowledge of the genome sequence under investigation. To succeed, a gene-finding algorithm must incorporate good indices that can discriminate coding and noncoding regions accurately.[1,2] While a number of good codon indices have been proposed, most involve the period-3 (P3) feature of coding sequences. P3's uniqueness is due to the fact that three nucleotide bases encode an amino acid and that the usage of the nucleotide bases at the three positions is highly biased. This feature can show up in a number of ways. For example, if you map a DNA sequence of length N to a numerical sequence and take the Fourier transform, typically you'll see a strong peak at or around $N/3$ in the Fourier transform's magnitude if the sequence is a coding one. However, such a peak is either weak or nonexistent if the sequence is noncoding.

Each codon index captures certain but not all features of a DNA sequence in the protein-coding region. You'll get the strongest signal when you combine multiple complementary sources of information,[3] by either integrating many codon indices to improve accuracy or building better representations of DNA sequences to devise more accurate codon indices. Here, we describe some fruitful work in the latter area.

A novel codon index based on recurrence time

This method tries to represent a genomic DNA sequence hierarchically by quantifying repeating patterns in a genome sequence in such a way that it properly characterizes the sequence's P3 feature and its entropy. This method is about 7 percent more accurate than that based on the Fourier transform method for quantifying the P3 feature.

Let's first define terms. We denote a sequence under study by $S = b_1 b_2 b_3 \cdots b_N$, where N is the sequence length and each b_i ($i = 1, \cdots, N$) is a nucleotide base. For instance, if S_1 = ACGAAAAACGATTTTAAA, then $N = 18$, $b_1 = \mathrm{A}$, $b_2 = \mathrm{C}$, $\cdots$, and $b_{18} = \mathrm{A}$. Next, we group consecutive nucleotide bases of window size w and call that a *word* of size w. Using maximal overlapping sliding windows, we obtain $n = N - w + 1$ such words.

We associate these words with the positions of the original DNA sequence from 1 to n; that is, $W_i = b_i b_{i+1} \cdots b_{i+w-1}$ is a word of size w associated with the position i along the DNA sequence. Two words are considered equal if all their corresponding bases match. That is, $W_i = W_j$ if and only if $b_{i+k} = b_{j+k}$, where $k = 0, \cdots, w - 1$. Also, $S[u \rightarrow v] = b_u b_{u+1} \cdots b_v$ denotes a subsequence of S from position u to v.

Mathematically, the *recurrence time* $T(i)$ for a position i along the DNA sequence is the smallest $j - i$ such that $j > i$ and $W_j = W_i$. If no such j exists, no repeat exists for the word W_i after position i in the sequence S, and we indicate such a situation by $T(i) = -1$.

Take S_1 as an example. If $w = 3$, then $n = 16$, and its recurrence time series $T(i)$ is

1541-1672/05/$20.00 © 2005 IEEE
Published by the IEEE Computer Society

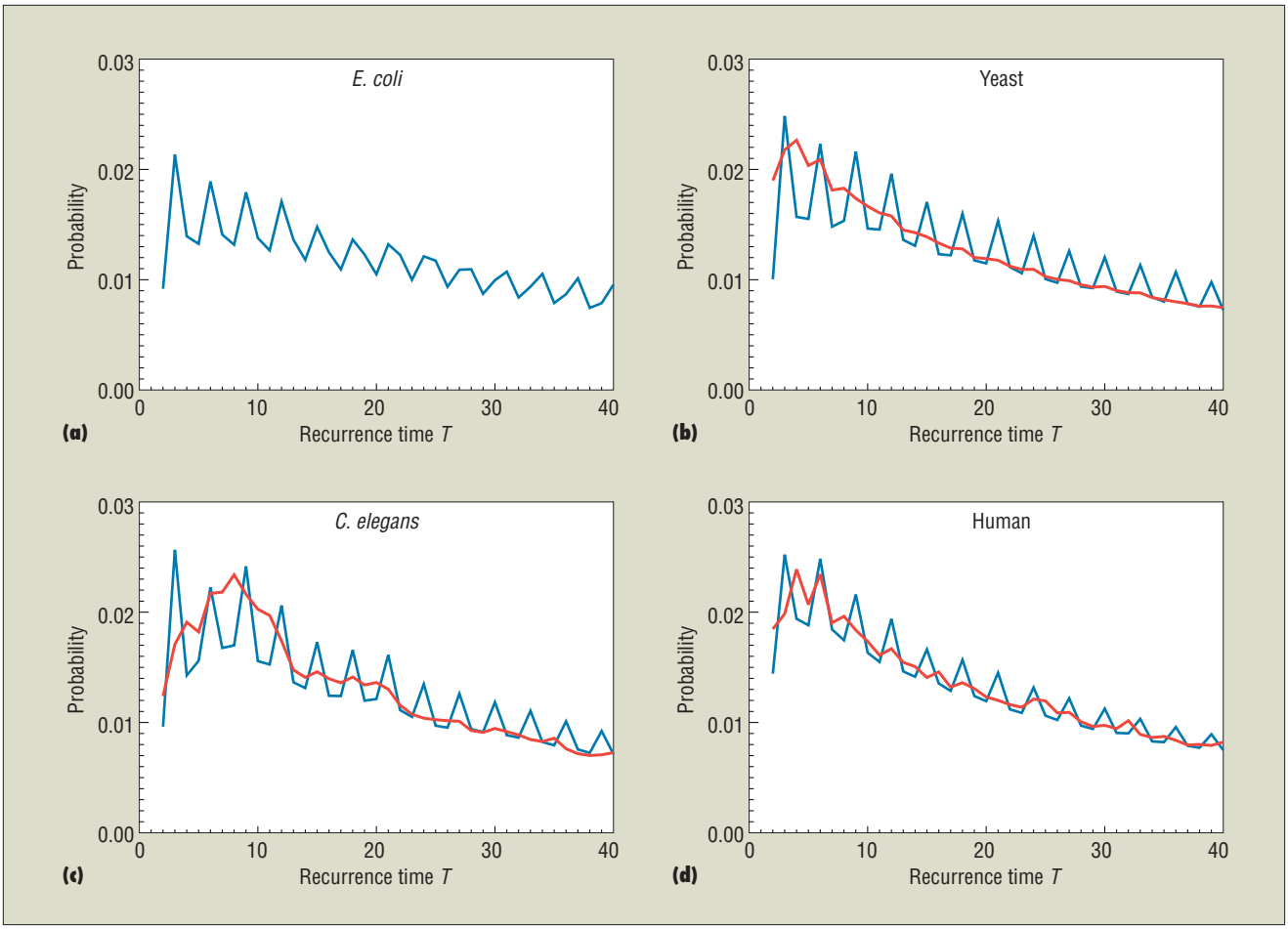

Figure 1. The probability distribution curves computed from the genomes of (a) *E. coli*, (b) yeast, (c) *C. elegans*, and (d) humans. The blue and red curves are for coding and noncoding sequences, respectively. The window size *w* is 3 in all the computations. We obtained similar results when *w* = 4 or 5.

7, 7, –1, 1, 1, 10, –1, –1, –1, –1, –1, 1, –1, –1, –1, –1

Discarding all the –1 terms from the $T(i)$ sequence, we get a recurrence time $T(i)$ series of six nontrivial recurrence times:

7, 7, 1, 1, 10, 1

As another example, consider a sequence of $(A)_l$, which represents a consecutive sequence of A's of length l. Such a sequence contributes to $T = 1$ a total of $l - w$ counts. Other single base repeats similarly contribute to $T = 1$, while a sequence such as $(AC)_l$ contributes a total of $2l - w$ counts to $T = 2$.

To understand why recurrence time is a hierarchical representation of a DNA sequence, imagine that we've computed the recurrence time for each position of the DNA sequence. The recurrence time for many different positions might be the same. Treating the recurrence time as a pointer, we can group those different positions or regions of the DNA sequence together. Such a strategy contrasts with conventional sequence analysis, where we compare a sequence against a specific pattern. This hierarchical representation of DNA sequences has enabled us to develop an effective method for finding all repeat-related features from a genomic DNA sequence and to study mutations, insertions, and deletions.[4]

Why does recurrence time characterize both the entropy and the P3 feature of a DNA sequence? Regarding entropy, the Ornstein-Weiss theorem[5] states that the entropy rate per symbol is given by the logarithm of the recurrence time divided by the word length, as the word length goes to infinity. In fact, we can compute the entire integer-order Renyi entropy spectrum from the recurrence time distribution

$$\psi(T) = P\{T(i) = T\} = \sum_{i=1}^{m} p_i^2 \cdot [1 - p_i]^{(T-1)} \quad (T \geq 1) \quad (1)$$

where p_i is the probability for word W_i to occur. From equation 1, we can obtain the Renyi entropy, defined by

$$H_q(W) = \frac{1}{1-q} \log\left(\sum_{i=1}^{m} p_i^q\right)$$

where q is real and $H_1(W)$ gives the Shannon entropy. Expanding $\psi(T)$ in equation 1 using binomial expansion, we get

$$\psi(T) = \sum_{i=0}^{T-1} (-1)^i \frac{(T-1)!}{i!(T-1-i)!} e^{-(1+i)H_{2+i}}$$

For example, if we take $T = 1, 2, 3, \ldots$, we then have

$$\psi(1) = e^{-H_2}$$
$$\psi(2) = e^{-H_2} - e^{-2H_3}$$
$$\psi(3) = e^{-H_2} - 2e^{-2H_3} + e^{-3H_4}$$

and so on. Hence, we can find all integer-order H_q, $q \geq 2$, by computing $\psi(T)$. We can find the Renyi entropy spectrum of order $q < 2$ by computing moments of the recurrence time T and $\log T$.

To illustrate how recurrence time can capture a DNA sequence's P3 feature, we've shown in figure 1 the probability distributions for the recurrence times ≤ 40 for the genome

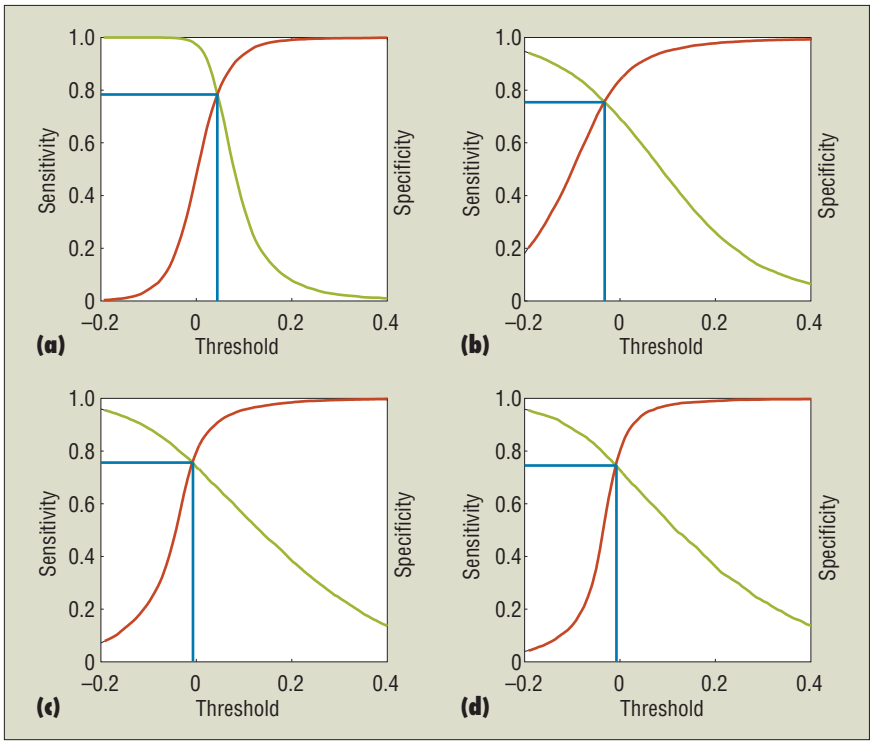

Figure 2. The specificity and sensitivity curves for the RTI index evaluated on (a) all 16 yeast chromosomes, (b) chromosome 3 of the *C. elegans* genome, (c) chromosome 19 of the human genome, and (d) chromosome 22 of the human genome.

Related Web Sites

You can obtain the genomes we studied from the following URLs:

***E. coli*:** www.genome.wisc.edu/sequencing/k12.htm
Yeast: ftp://genome-ftp.stanford.edu/pub/yeast/data_download
***C. elegans*:** www.sanger.ac.uk/Projects/Celegans
Human: www.ncbi.nlm.nih.gov/genome/guide/human

sequences of the *E. coli*, yeast, *C. elegans*, and human species. The blue and red curves are for the coding and noncoding regions, respectively. (Because of the low percentage of the *E. coli* genome's noncoding regions, we didn't compute the red curve.) The blue curves all have well-defined peaks at recurrence times of 3, 6, 9, …. On the basis of this feature, we can define a simple Recurrence Time Index codon as

$$\mathrm{RTI} = \sum_{i=1}^{m} \left[2p(3i) - p(3i+1) - p(3i+2) \right]$$

where $p(i)$ is the probability for the recurrence time $T = i$ calculated for a coding or noncoding sequence of length n, and m is a cutoff parameter typically chosen not to be larger than 20 so that short sequences can be studied. When you use RTI as your codon index, you can use the sliding-window technique, with $3m$ being the sliding window's length.

To evaluate RTI's accuracy in identifying protein-coding regions, we studied long genome sequences of yeast, *C. elegans*, and humans (see the "Related Web Sites" sidebar). We obtained information on the coding and noncoding sequences' exact locations from the annotations to these genomes. For yeast, our sample pool comprised two sets of DNA segments: the coding set (fully coding regions or *exons*) contained 4,125 verified open reading frames (ORFs), and the noncoding set contained 5,993 segments (fully noncoding regions or *introns*). For *C. elegans* chromosome 3, with 2,904 genes, our sample pool consisted of 18,469 exons and 16,421 introns. For human chromosome 19, with 2,453 genes, the sample pool consisted of 19,337 exons and 10,785 introns. For human chromosome 22, with only 831 genes, the sample pool consisted of only 6,618 exons and 4,068 introns. Figure 2 shows the specificity and sensitivity curves for the RTI index evaluated on all 16 yeast chromosomes, chromosome 3 of *C. elegans*, chromosome 19 of the human genome, and chromosome 22 of the human genome. The red curves are the cumulative distribution functions for RTI for the noncoding regions, and the green curves are the complementary cumulative distribution functions for the coding regions. The two curves' intercept point indicates the accuracy of the codon index.

Our accuracy for the yeast genome was 78 percent, about 7 percent higher than with the Fourier transform method for characterizing the P3 feature (discussed in detail later), and only slightly lower for the *C. elegans* and human genomes. This strongly suggests that the method is largely species-independent. This might reflect the well-known fact that entropies for genomes of different organisms are close. So, we expect that this codon index will be more advantageous than conventional ways of characterizing the P3 feature when it's applied to genomes of advanced organisms such as humans.

Characterizing DNA sequences' randomness and structure

Although recurrence time can characterize certain random aspects of a DNA sequence using entropy, it focuses more on the P3 feature. Because our codon index based on recurrence time is 7 percent more accurate in characterizing the yeast genome's P3 feature than the one based on Fourier transform, we wondered if we could develop more accurate indices by more actively characterizing a DNA sequence's random aspects besides its P3 feature. The answer is yes.

As we pointed out earlier, entropy might not sharply characterize differences among genomes of different organisms. What else can characterize DNA sequences' randomness? One promising scheme is the *random fractal theory*. We'll illustrate the idea through *detrended fluctuation analysis*.[6]

DFA

Intuitively speaking, a fractal means one part is similar to another part or to the whole, so it doesn't possess any well-defined scale.[7] Because P3 implies a specific scale of three, fractals and P3 are incompatible properties. However, we can characterize them simultaneously by describing the breaking of the fractal scaling at the specific scale of three.

First, to facilitate the fractal analysis of DNA sequences, we map a DNA sequence to a numerical sequence using the mapping rule[8]

C or T at position $n \rightarrow u(n) = +1$;
A or G at position $n \rightarrow u(n) = -1$

We then generate a DNA walk by forming the partial summation

$$y(n) = \sum_{i=1}^{n} u(i), \quad n = 1,2,3,\ldots$$

We then use DFA as follows:

1. Divide a given DNA walk of length N into $\lfloor N/l \rfloor$ nonoverlapping segments (where the notation $\lfloor x \rfloor$ denotes the largest integer not greater than x that contains l nucleotides).
2. Define the local trend in each segment to be the ordinate of a linear least-squares fit to the DNA walk in that segment.
3. Compute the detrended walk, denoted by $y_l(n)$, as the difference between the original walk $y(n)$ and the local trend.

The following scaling behavior (that is, fractal property) has been found for many DNA walks studied:

$$F_d(l) = \left\langle \sum_{i=1}^{l} y_l(i)^2 \right\rangle^{1/2} \sim l^H$$

where the angle brackets denote the ensemble average of all the segments, and $F_d(l)$ is the average variance over all segments. The exponent H is often called the Hurst parameter. When $H = 0.5$, the DNA walk resembles a standard random walk. When $H > 0.5$, the DNA walk possesses long-range correlations. Statistically speaking, a noncoding region is often more likely to possess long-range-correlation properties.[9] This feature enabled S. Ossadnik and colleagues to develop a coding sequence finder for genomes with long noncoding regions.[10]

To illustrate the idea, figure 3 shows a representative log-log plot of $F_d(l)$ versus l for a coding and a noncoding sequence of yeast chromosome I (the first of 16 chromosomes). The two sequences are of lengths $N = 1{,}742$ and $N = 3{,}598$, respectively. We choose l to increment with base $r = 2$, and the fitting range with the best scaling property is $[l_0, l_1] = [2^2, 2^8]$ for both figures 3a and 3b. Within this range, the Hurst parameters (H) are 0.54 and 0.62, respectively. The nice scaling law in $[l_0, l_1]$ indicates that DNA sequences are fractals. The observation that the Hurst parameters in noncoding regions are larger than those in coding regions is quite typical when the sequence under study is fairly long.

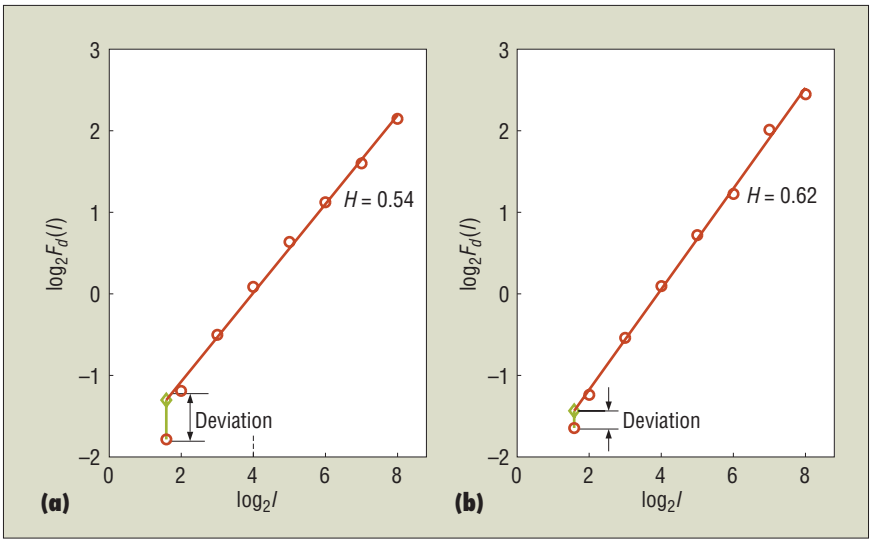

Figure 3. Representative period-3 fractal deviation for (a) coding and (b) noncoding regions in yeast chromosome I.

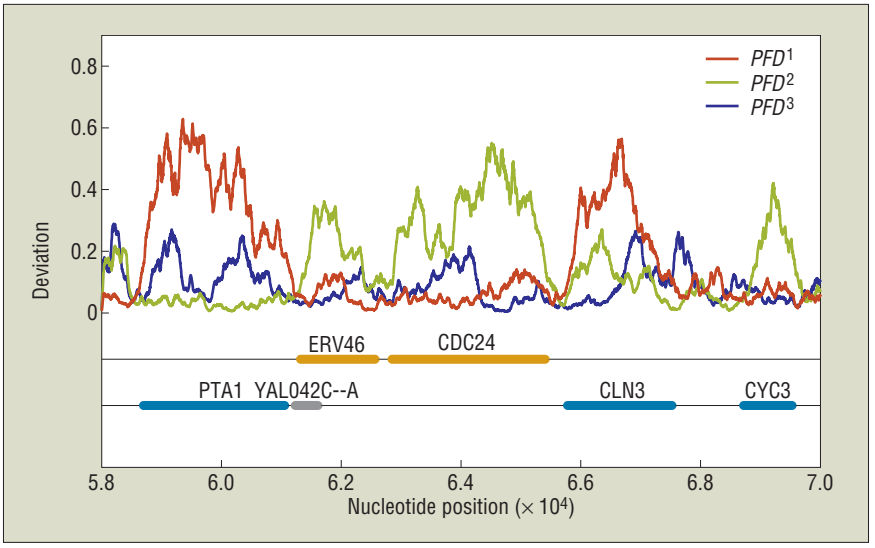

Figure 4. The reading-frame-specific fractal deviation curves (PFD^i, i = 1, 2, 3) for a DNA segment in yeast chromosome I (from nucleotide 58,000 to 70,000). The sliding-window size is w = 512. The colored bars on the two horizontal lines are the open reading frames on the two strands of the chromosome (the positive strand is on top, the reverse strand on the bottom). The orange and blue bars represent verified ORFs; the gray bar represents a dubious ORF.

How do we characterize the scaling break at the scale of three? After we fit a straight line to the points calculated through DFA (see figure 3), we calculate how far the computed point at $l = 3$ deviates from the fitted straight line. As the figure shows, this *period-3 fractal deviation* (*PFD*) is larger for coding sequences. In fact, this is a typical feature, more salient than the long-range-correlation property in noncoding sequences.

However, this feature isn't as simple as it seems. You have to consider the reading frames: when the coding segment starts with the gene-containing reading frame (the first nucleotide of a codon), the P3 feature collides with the DFA technique at the scale of $l = 3$ and results in a large *PFD*. When the segment starts with an incorrect reading frame, the P3 feature can't be captured, and the deviation value is small. In this case, the coding sequence behaves just like a noncoding sequence. This complexity is actually a great asset, since the method offers an interesting way to identify ORFs. To illustrate this, fig-

Table 1. Finding genes using fractal deviation (FD), P3 based on Fourier transform (FT), and the Hurst parameter (*H*) using detrended fluctuation analysis.

Coding / noncoding	Sensitivity / specificity (%)		
(n_1, N_1) / (n_2, N_2)	FD (w = 64)*	FT (w = 63)*	H (w = 64)*
(1, 4,125) / (1, 5,993)	82.6	70.9	43.9
(256, 4,067) / (256, 4,164)	84.3	71.5	45.2
(512, 3,756) / (512, 1,939)	87.3	71.2	45.5
(1,026, 2,674) / (512, 1,939)	89.2	72.0	45.1
(1,026, 2,674) / (1,026, 638)	92.4	71.1	43.7

* w = word size

ure 4 shows three reading-frame-specific fractal deviation curves, where we've applied a fifth-order moving-average filter. Note that the three colored curves (in green, red, and blue) generally don't overlap with one another. This is a necessary condition for the three reading frames to be separable. Also, you can see that in coding regions, both in the positive and the reverse strands, typically one of the three PFD^i curves displays a large value and separates considerably from the other two curves. And finally, in noncoding regions, the three PFD^i curves are mixed—meaning that the three reading frames are more or less equivalent and inseparable.

Index based on fractal deviation

On the basis of these observations, we devised a simple codon index. The algorithm works as follows. We employ a sliding-window technique to systematically calculate *PFD* along a DNA sequence. We then partition *PFD* into three subsets, one for each reading frame. Then, we define the codon index

$$\mathrm{FD} = \frac{1}{\lfloor M/3 \rfloor} \times \sum_{m=1}^{\lfloor M/3 \rfloor} \max\left(PFD^1(m), PFD^2(m), PFD^3(m)\right)$$

in which the maximum operation is carried out at each nucleotide position.

To evaluate the accuracy of FD as a codon measure, we analyze all 16 yeast chromosomes. Our sample pool comprises two sets of DNA segments: the coding set containing 4,125 verified exons and the noncoding set containing 5,993 introns. We extract different subsets of coding and noncoding segments from this sample pool according to the sequences' lengths. For any cutoff value n_1 and n_2, let C denote the (coding) set of ORFs whose lengths are greater than n_1, and let NC denote the set of noncoding segments whose lengths are greater than n_2. C and NC contain N_1 and N_2 elements, respectively. Table 1 lists FD's accuracy for several configurations of (n_1, n_2). For comparison, we also list the accuracy based on P3 (obtained using Fourier transform) and the Hurst parameter (obtained using DFA) alone. Clearly, the proposed codon index is much more accurate than either the P3 or the fractal method alone. It's also interesting to note that, in the yeast genome case, this codon index is more accurate than the recurrence-time-based method discussed earlier. This suggests that random fractal theory is indeed more effective in characterizing the randomness of DNA sequences.

We've developed two novel DNA codon indices, one based on recurrence time and one based on fractal deviation. Because both work well on short DNA sequences, they both hold the promise of being integrated into and thus improving existing gene identification algorithms.

Recently, we integrated the FD-based codon index with the P3 feature using Fisher linear discriminate analysis. The combined index achieves even higher accuracy than in the work reported here. We will continue to work on incorporating our indices into existing gene-finding algorithms so that their training can be simplified and accuracy improved.

Simultaneous characterization of complex

The Authors

Jianbo Gao is an assistant professor of electrical and computer engineering at the University of Florida. His research involves nanocomputing and fault-tolerant computing, nonlinear-time-series analysis, and bioinformatics. He has developed tools for studying a wide range of real-world signal-processing problems. He received his PhD in electrical engineering from the University of California, Los Angeles. He is a member of Sigma Xi and the IEEE. Contact him at the Dept. of Electrical & Computer Eng., NEB 427, Univ. of Florida, Gainesville, FL 32611; gao@ece.ufl.edu.

Yinhe Cao is the founder of and a bioinformatics application software architect at BioSieve, which builds bioinformatics software tools for analyzing microarray data and genome and protein sequence data. His current research interests include inferring various bionetworks such as protein-protein interaction networks and gene regulatory networks from microarray data. He received his PhD in mechanical engineering from the University of Missouri–Rolla. Contact him at BioSieve, 1026 Springfield Dr., Campbell, CA 95008; contact@biosieve.com.

Yan Qi is a PhD candidate in biomedical engineering at Johns Hopkins University. Her research interests include computational biology and bioinformatics, especially probabilistic regulatory motif discovery and functional genomics. She received her MSc in electrical and computer engineering from the University of Florida, Gainesville. Contact her at 6908 Bonnie Ridge Dr., Apt. 202, Baltimore, MD 21209; yanqi@jhu.edu.

Jing Hu is a PhD candidate in electrical and computer engineering at the University of Florida, Gainesville. Her research interests include signal processing, chaos and fractal time series analysis, biological-data analysis, and bioinformatics. She received her MEng in electrical engineering from Huazhong University of Science and Technology. Contact her at 366 Maguire Village, Apt. 7, Gainesville, FL 32603; jinghu@ufl.edu.

data's randomness and structure represents a new paradigm of thinking. Under such a strategy, we can use theories developed in mathematics, physics, and engineering synergistically rather than singly to characterize different facets of complex data, so that the data can be fully characterized and pattern discovery can be most effectively carried out. These approaches might also be considered new and effective ways to characterize data complexity, so they may become an important new element of complexity theory. In particular, we can use the recurrence time method to expedite database search and improve data compression schemes for large data sets—noticing that it identifies dominant features first before compressing data. In contrast, the classic Lempel-Ziv algorithm processes a file sequentially. Such an operation might break dominant features into many pieces, thus increasing entropy and reducing compression efficiency. With the recurrence time method, even directories can be compressed hierarchically. ■

Acknowledgments

Parts of this article were presented at the 2005 Information Sciences and Systems Conference.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.

References

1. M.Q. Zhang, "Computational Prediction of Eukaryotic Protein-Coding Genes," *Nature Reviews Genetics*, vol. 3, Sept. 2002, pp. 698–709.
2. C. Mathé et al., "Current Methods of Gene Prediction, Their Strengths and Weaknesses," *Nucleic Acids Res.*, vol. 30, no. 19, 2002, pp. 4103–4117.
3. M. Snyder and M. Gerstein, "Genomics—Defining Genes in the Genomics Era," *Science*, vol. 300, no. 5617, 2003, pp. 258–260.
4. Y.H. Cao et al., "Recurrence Time Statistics: Versatile Tools for Genomic DNA Sequence Analysis," *J. Bioinformatics and Computational Biology*, vol. 3, no. 4, 2005, pp. 677–696.
5. D. Ornstein and B. Weiss, *IEEE Trans. Information Theory*, vol. 48, no. 6, 2002, p. 1694.
6. C.K. Peng et al., "Mosaic Organization of DNA Nucleotides," *Physical Rev. E*, vol. 49, no. 2, 1994, pp. 1685–1689.
7. B.B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, 1982.
8. C.-K. Peng et al., "Long-Range Correlations in Nucleotide Sequences," *Nature*, vol. 356, no. 6365, 1992, pp. 168–170.
9. H.E. Stanley et al., "Scaling Features of Noncoding DNA," *Physica A*, vol. 273, no. 1, 1999, pp. 1–18.
10. S.M. Ossadnik et al., "Correlation Approach to Identify Coding Regions in DNA Sequences," *Biophysical J.*, vol. 67, no. 1, 1994, pp. 64–70.