Journal of Bioinformatics and Computational Biology
Vol. 3, No. 3 (2005) 677–696
© Imperial College Press

Imperial College Press
www.icpress.co.uk

RECURRENCE TIME STATISTICS: VERSATILE TOOLS FOR GENOMIC DNA SEQUENCE ANALYSIS

YINHE CAO

Biosieve, 1026 Springfield Drive
Campbell, CA 95008, USA
contact@biosieve.com

WEN-WEN TUNG

Department of Earth and Atmospheric Sciences, Purdue University
West Lafayette, IN 47907, USA
wwtung@purdue.edu

J. B. GAO

Department of Electrical and Computer Engineering
University of Florida, Gainesville, FL 32611, USA
gao@ece.ufl.edu

YAN QI

Department of Biomedical Engineering
Johns Hopkins University, Baltimore, MD 21205, USA
yanqi@bme.jhu.edu

Received 25 August 2004
Revised 5 November 2004
Accepted 10 December 2004

With the completion of the human and a few model organisms' genomes, and with the genomes of many other organisms waiting to be sequenced, it has become increasingly important to develop faster computational tools which are capable of easily identifying the structures and extracting features from DNA sequences. One of the more important structures in a DNA sequence is repeat-related. Often they have to be masked before protein coding regions along a DNA sequence are to be identified or redundant expressed sequence tags (ESTs) are to be sequenced. Here we report a novel recurrence time-based method for sequence analysis. The method can conveniently study all kinds of periodicity and exhaustively find all repeat-related features from a genomic DNA sequence. An efficient codon index is also derived from the recurrence time statistics, which has the salient features of being largely species-independent and working well on very short sequences. Efficient codon indices are key elements of successful gene finding algorithms, and are particularly useful for determining whether a suspected EST belongs to a coding or non-coding region. We illustrate the power of the method by studying the genomes of *E. coli*, the yeast *S. cervisivae*, the nematode worm *C. elegans*, and the human, *Homo sapiens*. Our method requires approximately $6 \cdot N$ byte memory and a computational time of $N \log N$ to extract all the repeat-related and periodic or quasi-periodic features from a sequence of length N without any prior knowledge on the consensus sequence of those features, hence enables us to carry out sequence analysis on the whole genomic scale by a PC.

Keywords: Genomic DNA sequence; repeated-related structures; coding region identification; recurrence time statistics.

1. Introduction

The structure of human genome and genomes of other organisms is very complicated. With the completion of many different types of genomes, especially the human genome, one of the grand challenges for the future genomics research is to comprehensively identify the structural and functional components encoded in a genome.[1] Outstanding structural components include all kinds of repeat-related structures,[2,3] and periodicity and quasi-periodicity, such as period-3, which is considered to reflect codon usage,[4] and period 10–11, which may be due to the alternation of hydrophobic and hydrophilic amino acids[5] and DNA bending.[6] Extracting and understanding these structural components will greatly facilitate the identification of functional components encoded in a genome, and the study of the evolutionary variations across species and the mechanisms underlying those variations. Equally or even more important, repeat-related features often have to be masked before protein coding regions along a DNA sequence are to be identified or redundant expressed sequence tags (ESTs) are to be sequenced.

More important than finding repeat-related structures in a genome is the identification of genes and other functional units along a DNA sequence. In order to be successful, a gene finding algorithm has to incorporate good indices for the protein coding regions. A few representative indices are the Codon Bias Index (CBI),[7] the Codon Adaptation Index (CAI),[8,9] the period-3 feature of nucleotide sequence in the coding regions[10–13] and the recently proposed YZ score.[14] Each index captures certain but not all features of a DNA sequence. The strongest signal can only be obtained when one combines multiple different sources of information.[15] In order to improve the accuracy and simplify the training of existing coding-region or gene identification algorithms (see the recent review articles[16,17] and the references therein), and to facilitate the development of new gene recognition algorithms, it would be highly desirable to find new codon indices.

Over the past decades, sequence alignment and database search[18–27] have played a significant role in molecular biology, and extensive research in algorithms has resulted in a few basic software tools such as FASTA[28,29] and BLAST.[30–33] Although these tools have been routinely used in many different types of researches, finding biologically significant information with these tools is far from trivial, for the following reasons: (i) The results of sequence alignment and database search strongly depend on some model-based score function.[19–21] However, the model may not be general enough to be appropriate to the biological problem under study. For example, a widely used affine gap cost model[21] assumes that insertions and deletions are exponentially less likely to occur as their length increases. Nevertheless, long insertions and deletions may occur as a single event, such as insertion of a transposon element. (ii) The dynamic programming approach to the maximization of the score function, although mathematically sound, requires a computational time at least proportional to the product of the length of the two sequences being compared. This makes the approach less feasible for sequence comparison on the whole genomic scale. (iii) Assessment of the statistical significance of an alignment is hard to make.[34–40] All theoretically tractable score statistics are based upon certain probability models about the sequence under study. However, those models may not capture interesting sequence segments such as repeat structures inherent in natural DNA sequences. For example, it is a common experience for a user of BLAST that for some input sequence, the output from BLAST may be very noisy: many matches with very high score may only hit on low complexity regions and are not biologically interesting, while biologically significant units such as binding sites, promoter regions, or expression regulation signals, due to their low scores, do not have a chance to show up as the output.[26,27]

Here, we propose a simple recurrence time based method, which has a number of distinct features: (i) It does not utilize a score function in general and does not penalize gaps in particular. This makes it most convenient to find out large blocks of insertions or deletions. (ii) Computationally it is very efficient: with a computational time proportional to $N \log N$, where N is the size of the sequence, and a memory of $6N$, it can exhaust all repeat-related and periodic or quasi-periodic structures. This feature allows us to carry out genome analysis on the entire genomic scale by a PC. (iii) It is model free in the sense that it does not make any assumption about the sequences under study. Instead, it builds a representation of the sequence in such a way that all interesting sequence units are automatically extracted. (iv) The method defines an efficient codon index, which is largely species-independent and works well on very short sequences. This feature makes the method especially appealing for the study of short ESTs. Below, we shall illustrate the power of the method by extracting outstanding structures including insertion sequences (ISs), rRNA clusters, repeat genes, simple sequence repeats (SSRs), transposons, and gene and genome segmental duplications such as inter-chromosomal duplication from genome sequences. We shall also discuss the usefulness of the method for the study of the evolutionary variations across species by carefully studying mutations, insertions and deletions. Finally, we shall evaluate the effectiveness of the recurrence time-based codon index by studying all of the 16 yeast chromosomes.

2. Databases and Methods

2.1. *Databases*

We have studied the DNA sequence data from the following four species: (a) *E. coli*,[41] (b) the yeast *S. cervisivae*,[42] (c) the nematode worm *C. elegans*,[43] and (d) the human, *Homo sapiens*.[2,44] These DNA sequence data are available from the following *URLs* respectively:

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

Except the *E. coli* genome, the other three contain gaps, since they are not yet completely sequenced. Those gaps are deleted before we compute the recurrence times from them. For the yeast *S. cervisivae*, the sequences of chromosome 1 to chromosome 16 are joined together into a single sequence in that order.

2.2. *Basic idea of recurrence time statistics*

Notations. Let us denote a sequence we want to study by $S = b_1 b_2 b_3 \cdots b_N$, where N is the length of the sequence, $b_i,\ i = 1, \ldots, N$, are nucleotide bases. For instance, if we take

$$S_1 = \text{ACGAAAAACGATTTTAAA},$$

then $N = 18, b_1 = \text{A}, b_2 = \text{C}, \ldots, b_{18} = \text{A}$. Next we group a consecutive nucleotide bases of window size w together and call that a word of size w. Using maximal overlapping sliding window, we then obtain $n = N - w + 1$ such words. We associate these words with the positions of the original DNA sequence from 1 to n, i.e., $W_i = b_i b_{i+1} \cdots b_{i+w-1}$ is a word of size w associated with the position i along the DNA sequence. Two words are considered equal if all of their corresponding bases match. That is, $W_i = W_j$, if and only if $b_{i+k} = b_{j+k},\ k = 0, \ldots, w-1$. $S[u \to v] = b_u b_{u+1} \cdots b_v$ will denote a subsequence of S from position u to v.

Recurrence time. The recurrence time $T(i)$ of position i for a DNA sequence S is a discretized version of the recurrence times of the second type for dynamical systems introduced recently by Gao.[45–48] It is defined as follows.

Definition: 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, then there is no repeat for the word W_i after position i in the sequence S, and we indicate such a situation by $T(i) = -1$.

To analyze the $T(i)$ sequence, we first filter out all those $T(i) = -1$, then denote the remaining positive integer sequence by $R(k),\ k = 1, \ldots, m$. The length of such a sequence is comparable to the size of a genome sequence, hence, is very long. We assume that the probability distribution functions for both $R(k)$ and $\log_{10} R(k)$ sequence can be reliably estimated based on such a sequence (e.g., by forming suitable histograms). When the word size is not too large, this is indeed the case. These two probability distribution functions, when they are well defined, are what we mean by the recurrence time statistics. The reason that we also work on $\log_{10} R(k)$ is that the largest $R(k)$ computed from a genomic DNA sequence can be extremely long, hence, it may be difficult to visualize the distribution for $R(k)$ in linear scale.

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

$$7, 7, -1, 1, 1, 1, 10, -1, -1, -1, -1, -1, 1, -1, -1, -1, -1$$

Discarding all the -1 terms from the $T(i)$ sequence, we then get the following recurrence time $R(i)$ series:

$$7, 7, 1, 1, 10, 1$$

where $m = 6$. The motivation for introducing the above definition is that the recurrence time sequence $T(i),\ i = 1, \ldots, n$, for a DNA sequence and a completely random sequence will be very different, and that by exploiting this difference, we would be able to exhaustively identify most of the interesting features contained in a DNA sequence.

2.3. *Recurrence time statistics for completely random (pseudo-DNA) sequences*

In order to characterize the difference between a DNA sequence and a completely random sequence in terms of the recurrence times, we study a completely random sequence first. We have the following interesting theorem.

Theorem: Given a sequence of independent words $W_i,\ i = 1, \ldots, n$, where there are a total of K distinct words, each occurs with probability $p = 1/K$, the probability that the recurrence time $T(i)$ being $T \geq 1$ is given by

$$P\{T(i) = T\} \propto [n - T] \cdot p \cdot [1 - p]^{(T-1)} \quad (1 \leq T < n). \tag{1}$$

Proof. It suffices to note that the probability for an arbitrary word W_i, where i is from the positions 1 to $n - T$, to repeat exactly after $T \geq 1$ positions is given by the geometrical distribution, $p \cdot [1 - p]^{(T-1)}$. Since there are a total of $n - T$ such positions or words, while each position along the sequence from 1 to $n - T$ has the same role, the total probability is then proportional to the summation of $n - T$ terms of $p \cdot [1 - p]^{(T-1)}$. This completes the proof. □

How can the above theorem be applied to a DNA sequence? If we assume the four chemical bases A, C, T and G to occur completely randomly along a (pseudo) DNA sequence, then there are a total of 4^w words of length w. If each word occurs equally likely, then $p = 4^{-w}$. However, due to overlapping of adjacent words, the above theorem is not applicable when $T \leq w - 1$. This problem is quite minor, since we consider T up to the scale of the genome sequence. Hence, the probability for a word to repeat exactly after $T \geq w$ locations is given by Eq. (1), while the distribution for the log-recurrence time $\log_{10} R(k)$ is given by

$$f(t) = C \cdot T \cdot [n - T] \cdot p \cdot [1 - p]^{(T-1)}, \quad (0 \leq t < \log_{10} n), \tag{2}$$

where $T = 10^t$, and C is a normalization constant. To prove Eq. (2), it suffices to note that $p(T)dT = f(t)dt$.

2.4. *Recurrence time statistics for DNA sequences*

2.4.1. *Recurrence time statistics and a novel codon index*

We have plotted in Fig. 1 the probability density functions (pdfs) of log recurrence time, i.e., $\log_{10} R(i)$, for the DNA sequence data from the four species. The dotted curves in Fig. 1 are computed according to Eq. (2) and represent those of completely random sequences with their length and the word size chosen to analyze them the same as those of the DNA sequences. The word sizes used are 12, 15, 16, 15 for Fig. 1 (a) to Fig. 1 (d) respectively. We observe two interesting features: (i) the pdfs for the genome sequences are very different from those for the random sequences, as signaled by the many sharp peaks in the curves of the pdfs for the genome sequences; and (ii) the degree of this difference varies vastly among the four genomes studied. In fact, Eq. (2) fairly well describes the background distribution for the $\log_{10} R(i)$ sequence for *E. coli*, but very poorly describes that for the chromosome 16 of the human. This may be due to biased base compositions in eukaryotic genome sequences. Each sharp peak in Fig. 1 may actually represents many sharp peaks if we plot the pdf for the $R(i)$ sequence instead of that for the $\log_{10} R(i)$ sequence. This is because with logarithmic scale, a whole interval of $R(i)$ will be lumped together. It is important to emphasize that all the sharp peaks indicate distinct features of a genome sequence. To appreciate this better, we have plotted in Fig. 2 the pdf for the $R(i)$ sequence computed from the yeast genome sequence. We clearly observe that even if only a very small portion of the pdf curve (up to $R = 300$) has been plotted out, it already contains many sharp peaks. It is interesting to note that the corresponding $R(i)$ sequence for a completely random (pseudo DNA) sequence is a smooth curve very close to the x-axis, as depicted by the dotted curve in the figure.

To understand why the sharp peaks in Figs. 1 and 2 indicate distinct features of a genome sequence, let us take an example. A sequence of $(A)_l$, which represents a consecutive sequence of A's of length l, contributes to a peak at $R = 1$, if l is larger than the word length w. In fact, when $l > w$, $(A)_l$ contributes a total of $l - w$ counts to $R = 1$. Other single base repeats similarly contribute to $R = 1$. As another example, we note that a sequence such as $(AC)_l$ contributes to $R = 2$ a total of $2l - w$ counts.

We are now ready to propose a novel recurrence time based codon index. This index is based on the period-3 feature. To appreciate the idea, we have shown in Fig. 3 the probability distributions for the recurrence times not greater than 40,

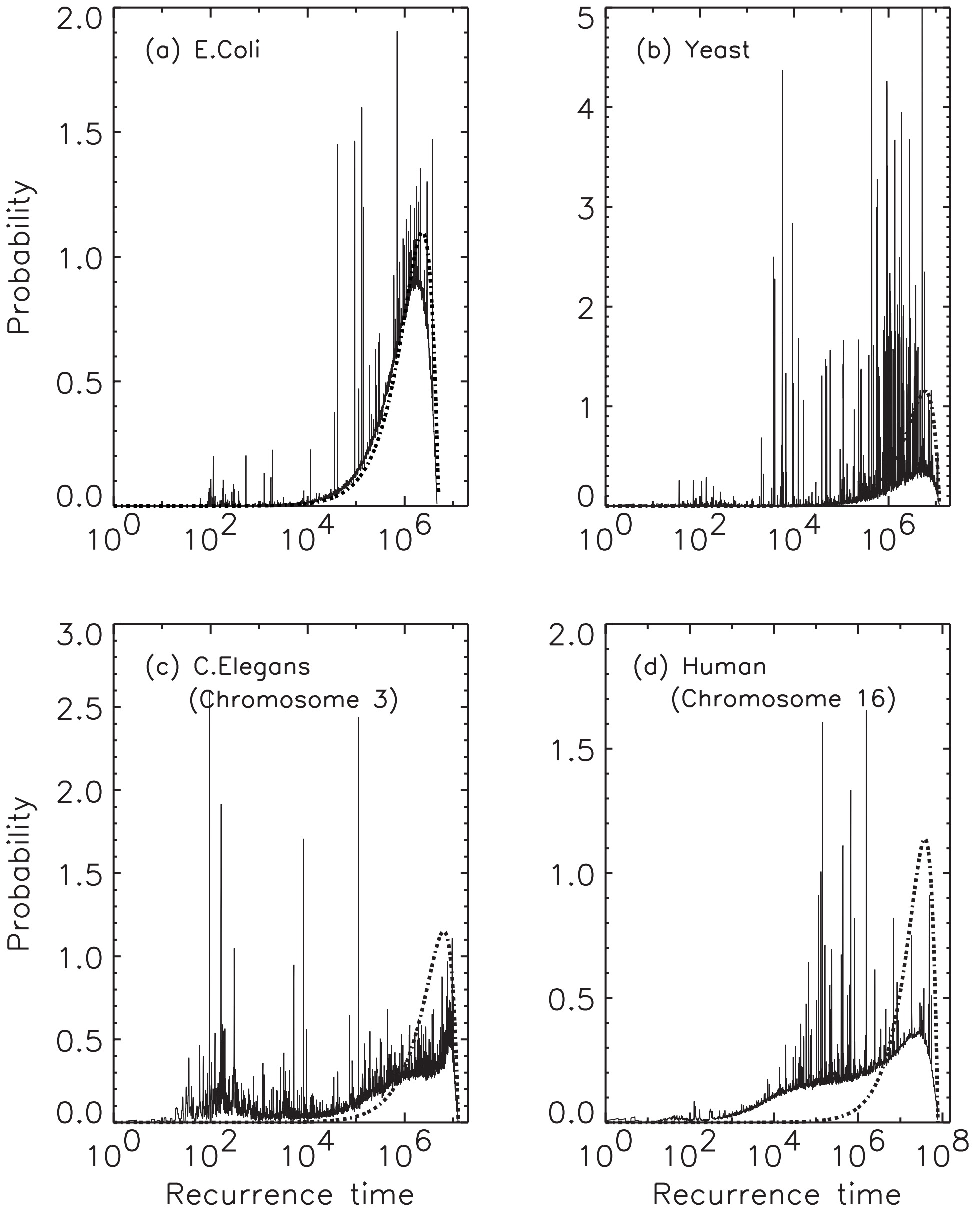

Fig. 1. The probability density function (pdf) for the recurrence time $\log_{10} R(i)$ sequence computed from the DNA sequence of (a) *E. coli*, (b) the yeast *S. cervisivae*, (c) chromosome 3 of the nematode worm *C. elegans*, and (d) chromosome 16 of the human. Dotted curves are computed from Eq. (2) and represent the situation where the four bases A, C, T, and G occur completely randomly with equal probability.

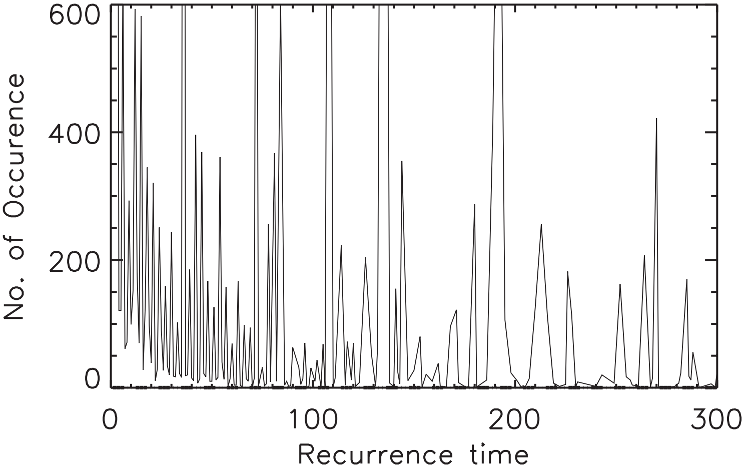

Fig. 2. A portion of the number of occurrence for the recurrence time $R(i)$ sequence for the Yeast genome. The red line close to the x-axis is generated from Eq. (1).

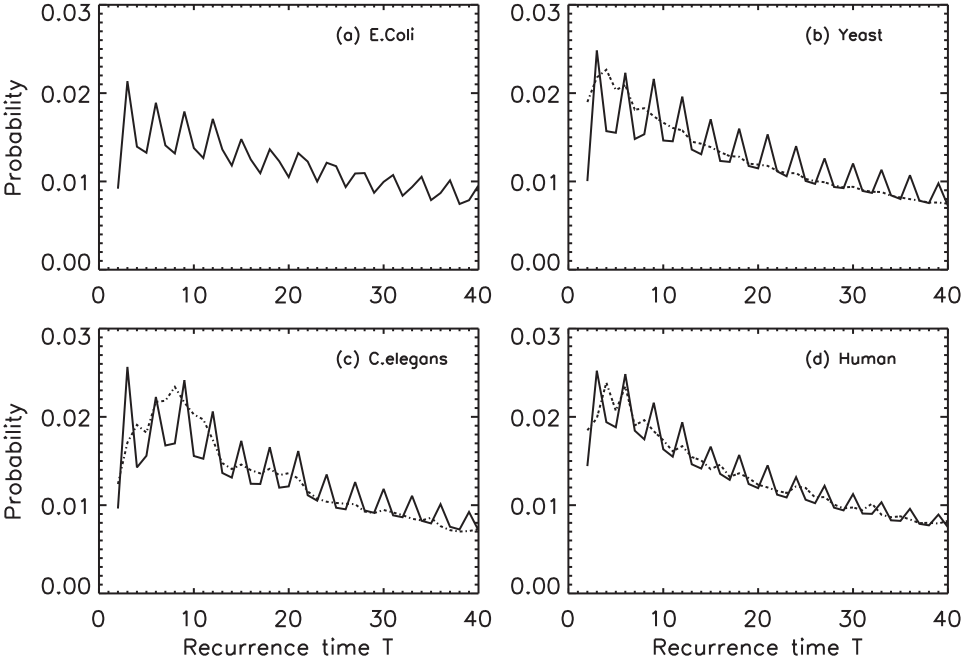

Fig. 3. The probability distribution curves computed from the genomes of four organisms studied. The red and black curves are for non-coding and coding sequences, respectively. The window size w is 3 in all the computations.

for the genome sequences of four species, *E. coli*, yeast, *C. elegans*, and the human. The black curves are for the coding regions. The dotted curves in Fig. 2(b–d) are for the non-coding regions. Due to the low percentage of non-coding regions in the *E. coli* genome, such a curve is not computed. We observe that the black curves all have very well defined peaks at recurrence times of 3, 6, 9, etc. Also note that the black curves are very similar among the four different species. Such period-3 feature can be conveniently used to define a codon index, which we shall denote by RT_{p3}:

$$RT_{p3} = \sum_{i=1}^{m} [2p(3i) - p(3i+1) - p(3i+2)], \tag{3}$$

where $p(i)$ is the probability for the recurrence time $T = i$ calculated for a coding or non-coding sequence, n is the number of bases of the coding/non-coding sequence, and m is a cutoff parameter typically chosen not to be larger than 20 so that very short sequences can be studied.

Before we go ahead to evaluate the efficiency of RT_{p3} as a codon index, let us focus on how we can exhaustively find all the repeat-related structures by tracing the peaks in Fig. 1 back to the DNA sequence. This can be easily done.

2.4.2. *Computation of exact repeat elements from recurrence times*

Let $T(i)$ be the recurrence time for position i of a DNA sequence S, where $i = 1, 2, \ldots, n$. For each particular value $T (1 \leq T < n)$ of the recurrence time, we build a list of indices $L(T : S)$ by linearly scanning $T(i)$ from $i = 1$ to $i = n$ and adding i to $L(T : S)$ whenever $T(i) = T$. Denote the index set $L(T : S)$ by

$$L(T : S) = \{i_1, i_2, \ldots, i_C\},$$

where $T(i_k) = T,\ k = 1, 2, \ldots, C,\ \ i_k < i_{k+1}$ for $k = 1, 2, \ldots, C - 1$, and C is the count of the occurrence of T in the recurrence time series $T(i)$. If we take S_1 as an example, then

$$L(1 : S_1) = \{4, 5, 12\},\ L(7 : S_1) = \{1, 2\},\ L(10 : S_1) = \{6\}.$$

When the count C is larger than 1, we define the gap between two consecutive indices of $L(T : S)$ as:

$$g_k = i_{k+1} - i_k \quad (k = 1, 2, \ldots, C - 1)$$

Let $g^* = w$. What happens when all $g_k \leq g^*$? In this situation, the sequence segment $S_r = S[i_1 \to i_C - i_1 + w - 1]$ is an exact repeat with period T. To see this, we note that for each term i_k in $L(T : S)$, we have $W_{i_k} = W_{i_k+T}$, or more explicitly, $b_{i_k+j} = b_{i_k+T+j}$ for $j = 0, 1, \ldots, w - 1$, and $k = 1, 2, \ldots, C$. Hence, we can concatenate b_{i_2} at $b_{i_1+g_1}$ to combine the repeats starting from b_{i_1} and b_{i_2}. More concretely, for $k = 1$, we have $b_{i_1+j} = b_{i_1+T+j}$, $j = 0, 1, \ldots, w - 1$; similarly, for $k = 2$, we have $b_{i_2+j} = b_{i_2+T+j}$, $j = 0, 1, \ldots, w-1$. Noting $g_1 \leq w$ means $i_2 - i_1 \leq w$, or $i_2 \leq i_1 + w$, we have $b_{i_1+j} = b_{i_1+T+j}$ for $j = 0, 1, \ldots, i_2 - i_1 + w - 1$. Continuing this procedure till $k = C$, we have $b_{i_1+j} = b_{i_1+j+T}$ for $j = 0, 1, \ldots, i_C - i_1 + w - 1$. Hence, the sequence $S_r = S[i_1 \rightarrow i_C - i_1 + w - 1]$ is an exact repeat with period T with its length $l = i_C - i_1 + w$. If $T < l$, S_r is then a tandem repeat.

In general, some gaps g_k may be larger than $g^* = w$. When there are P such gaps, we can decompose $L(T : S)$ into $P + 1$ subsets, such that within each subset all of the gaps are not larger than g^*. Then following the procedure detailed in the last paragraph, we see that each subset represents an exact repeat of period T. Taking S_1 as an example again, then from $L(7 : S_1) = \{1, 2\}$ we get an exact repeat ACGA of period 7, and from $L(1 : S_1) = \{4, 5, 12\}$, we get two exact repeats of period 1: the first one is AAAAA which is a simple sequence repeat, the other is TTTT, which is also a simple sequence repeat.

Before we move on to evaluating the efficiency of RT_{p3} as a codon index, we emphasize that the procedure outlined here makes the recurrence time method largely independent of the word size w: any feature with length longer than w can be re-combined. For simplicity, we shall call this a **re-combination algorithm**. This algorithm, together with the features related to mutations, deletions, and insertions, which are to be discussed shortly, makes the recurrence time-based method especially convenient for identifying horizontally transfered genes.

2.4.3. *Single Nucleotide Mutation and Single Nucleotide Polymorphism (SNP)*

Suppose we have two exactly repeating sequence segments, S_{lead} and S_{lag}, where the subscript lead and lag mean S_{lead} appears earlier than S_{lag} in a genome. In the simplest case, each word constructed from segments of S_{lead} has the same period T. In general, however, a few words constructed from segments of S_{lead} may have smaller recurrence times, due to the possibility that those words may find their copies in between S_{lead} and S_{lag}. Now suppose one nucleotide somewhere within S_{lead} is mutated. Since the mutated nucleotide appears in a consecutive w words, each of length w, we see that for those words, the period T will have to be different than T. This means if we plot out the recurrence time vs. the sequence position curve, then we should observe a gap of length w in an otherwise almost constant (T) curve. When this is the case, we can suspect that there may be a single nucleotide mutation at the end of the gap. If the gap corresponds to recurrence times larger than T or equal to -1 (meaning no repeats), then we can conclude that there is a single nucleotide mutation at the end of the gap. This is actually a sufficient condition, since it excludes the possibility that a few words may have copies in between S_{lead} and S_{lag}. An example is shown in Fig. 4, where we observe three single nucleotide mutations which correspond to gaps from positions 220,357 to 220,371, 220,474 to 220,488, and 220,609 to 220,623, respectively. Very interestingly, Fig. 4 actually also shows two small gaps of smaller recurrence times, which means a few words indeed have recurrence times smaller than T.

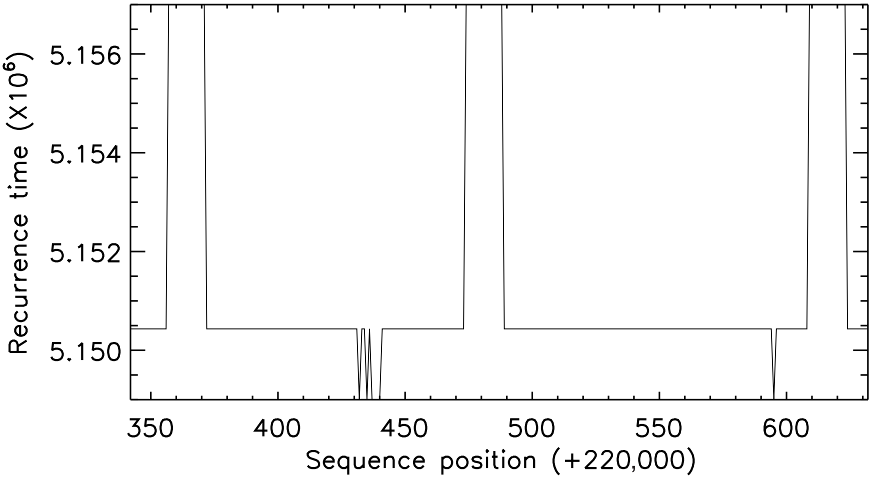

Fig. 4. Recurrence time vs. sequence position for the yeast *S. cervisivae* genome, showing effects of single nucleotide mutations.

The study of single nucleotide mutation is most relevant to the study of Single Nucleotide Polymorphism (SNP), where DNA sequence variations occur when a single nucleotide (A, T, C, or G) in the genome sequence among different populations is changed, possibly due to evolution. It is clear that if we concatenate two genome sequences for different subjects together, then we can treat SNP as a special type of single nucleotide mutation.

2.4.4. *Insertion/deletion and relations between repeat sequences of different periods*

Suppose we have a sequence segment starting from the position i_a. What happens if we insert a sequence of length, say, a few thousand bases, in the middle of that segment, then let the segment with insertion to repeat somewhere in the genome? Equivalently, the original sequence segment can be considered a result of deletion from the longer (i.e., with insertion) sequence segment. This interesting situation is revealed by a jump in the recurrence time vs. sequence position plot, with the height of the jump being the length of the insertion sequence, as we explain below.

Let $T(i)$ denote the recurrence time for the word at the position i of the sequence S. Suppose

$$\begin{aligned} &i_a < i_b, \quad T(i_a) = T(i_{a+1}) = \cdots = T(i_b) = T_1 > 0, \quad \text{and} \\ &\quad i_b < i_c \le i_b + w < i_d, \quad T(i_c) = T(i_{c+1}) = \cdots = T(i_d) = T_2 > T_1. \end{aligned}$$

Then the sequence segment

$$S_a = S[i_a \to (i_d + w - 1)]$$

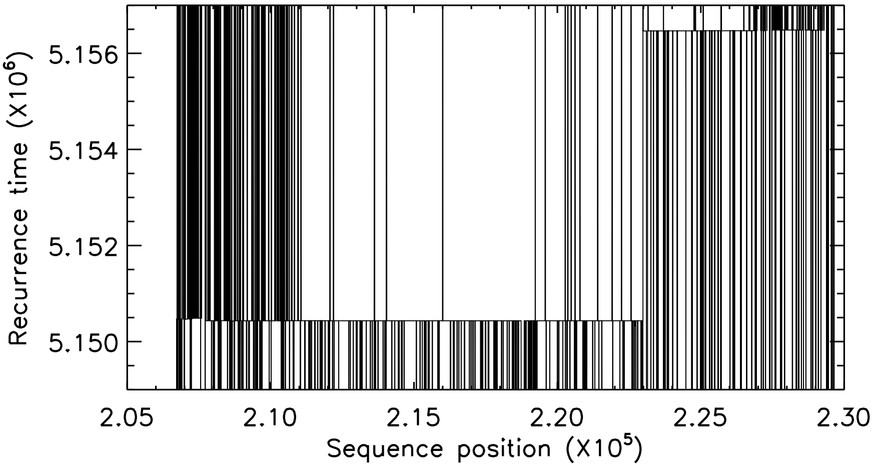

Fig. 5. Recurrence time vs. sequence position for the yeast *S. cervisivae* genome, showing a block of inter-chromosome duplication, with mutations at various locations, and insertion/deletion of a **Ty** element.

can be considered the result of deleting the sequence

$$S_{deletion} = S[(i_c + T_1 - 1) \rightarrow (i_c + T_2 - 1)]$$

from the sequence segment

$$S_b = S[(i_a + T_1) \rightarrow (i_d + w - 1 + T_2)].$$

Equivalently, S_b is the result of inserting the sequence $S_{deletion}$ into S_a right before the position i_c. Note that the condition of $i_c \leq i_b + w$ comes from the fact that the boundary always affects a consecutive w words, each of length w. When the first w bases of the deleted sequence segment do not have their copies at the positions starting from i_c, we have $i_c = i_b + w$. Otherwise, we have in-equalities. To make this discussion more concrete, we have plotted in Fig. 5 a recurrence time vs. sequence position curve. A blow-up of the figure is shown in Fig. 6. Clearly we observe a jump with height around 6000, which is the difference between the two recurrence times T_1 and T_2. We note that the gap in Fig. 6 has a width smaller than $w = 15$, which corresponds to the condition of $i_c < i_b + w$. Also note numerous deviations of the recurrence times from either T_1 or T_2. This is due to single or multiple nucleotide mutations, or due to the fact that some words have copies long before the sequence segment S_b.

3. Results and Discussion

In this section, we present examples of repeat-related structures extracted by the proposed method and evaluate the efficiency of the codon index.

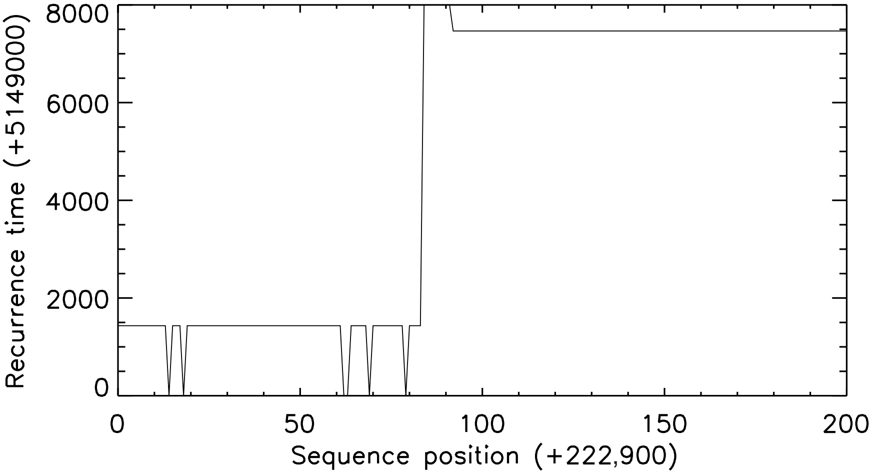

Fig. 6. A blowup of Fig. 4 around the jump of height of about 6000, indicating the location of insertion.

3.1. *Extraction of repeat-related structures*

We now present examples of structures which can be found by tracing the peaks in Fig. 1 back to the genome sequences. These structures include insertion sequences (ISs), rRNA clusters, repeat genes, simple sequence repeats (SSRs), transposons, and gene and genome segmental duplications such as inter-chromosomal duplication. We shall illustrate most of these structures using the yeast *S. cervisivae* as an example.

We first study SSRs. SSRs are perfect or slightly imperfect tandem repeats of particular k-mers. They have been extremely important in human genetic studies, because they show a high degree of length polymorphism in human population owing to frequent slippage by DNA polymerase during replication.[2] Any tandem repeat of k-mers, disregarding its exact content, will contribute to the count of occurrence of period $T = k$ in recurrence time statistics, hence can be easily found by following the peak of $T = k$ in Fig. 1. As an example, we note that there are 39 sequence segments contributing to $k = 13$. Three of them are CCACACCCA CACA, GGTGTGTGGGTGT, and TACCGACGAGGCT. Note that by Fig. 1(a), we can conclude that *E . coli* has very few SSRs.

One of the more striking features of the yeast *S. cervisivae* genome is that it contains many copies of transposon yeast (**Ty**) elements. Each **Ty** element is about 6.3 kb long, with the last 330 bp at each end constituting direct repeats, called δ. Those direct end repeats are responsible for the peaks around 5500 in Fig. 1(b), which enable us to find all of those **Ty** elements on both strands of the genome. As two examples, we mention that the transposon Ty3-1 on the Watson strand of chromosome 7 starts at the position 707,196 and ends at 712,546, and has a period of 5011. The transposon Ty1-1 on the Crick strand of chromosome 1 starts at the position 166,162 and ends at 160,238, and has a period of 5588.

Gene duplication is an important source of evolutionary novelty. Many duplicate genes have been found in the yeast *S. cervisivae* genome, and they often seem to be phenotypically redundant.[49–51] Any gene duplication will contribute to one of the sharp peaks in Fig. 1(b). As an example, we note that a gene (standard name MCH2, systematic name YKL221W), which is on chromosome 6 starting from the position 6931, is repeated on chromosome 13, starting from the position 7749. We have listed some of the duplicate genes in Table 1. Other duplications that we have found but not shown in Table 1 include rRNA clusters, tRNA gene pairs, and a snRNA pair.

Genome segmental duplications consist of large blocks that have been copied from one region of the genome to another. They have been found among genomes of many species including the yeast *S. cervisivae*,[49] and the *Homo sapiens*.[2,52] In fact, they contribute to some of the sharpest peaks in Fig. 1. An example of such segmental duplications is the inter-chromosomal duplication corresponding to the peak at $T = 5,150,433$ in Fig. 1(b), which has been shown in Fig. 5. We observe that duplication is between chromosome 1 and 8, and the copy on chromosome 1 is about 21.5 kb long. Between the two copies of that duplication, there are mutations involving single or multiple nucleotides at various locations, and an insertion of a **Ty** element about 6000 bp long at the location corresponding to the jump of that period value at chromosome 1.

3.2. *Evaluation of the novel codon index*

In order to evaluate the effectiveness of the RT_{p3} as a codon index, we study all of the 16 yeast chromosomes. Our sample pool is comprised of two sets of DNA

Table 1. Examples of gene duplication in the yeast *S. cervisivae*.

Type	Standard name	Systematic name	Chr. No. A	Start position	Chr. No. B	Start position position
Inter-chromosome repeat genes	MCH2	YKL221W	6	6931	13	7749
	SSA1	YAL005C	1	140183	12	96234
	RPL1B	YGL135W	7	254659	16	135804
	RPL11B	YGR085C	7	648400	16	731244
	EFT2	YDR385W	4	1243218	15	575097
	SSB1	YDL229W	4	44225	14	252217
	RPS6B	YBR181C	2	591670	16	377286
	TEF2	YPR080W	2	477629	16	700590
	YRF1-6	YNL339C	14	91	15	1091284
	YRF1-5	YLR467W	12	1071651	15	1084613
Intra-chromosome repeat genes	UBI4	YLL039C	12	64158	12	64614
	DOG2	YHR043C	8	192806	8	194069
	PHO3	YBR092C	2	427693	2	429543
	ALD3	YMR169C	13	599685	13	601895
	HXT7	YDR342C	4	1160726	4	1164075
	ASP3-3	YLR158C	12	468812	12	472463
	ASP3-3	YLR158C	12	481899	12	485551

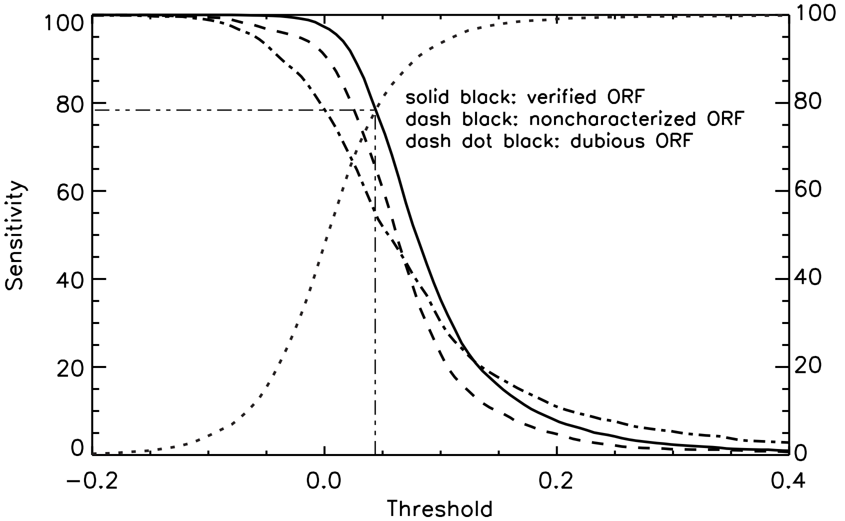

Fig. 7. The specificity and sensitivity curves for the RT_{p3} index evaluated on all of the 16 yeast chromosomes.

segments: the coding set (fully coding regions or exons), which contains 4125 verified ORFs, 1626 uncharacterized ORFs, and 812 dubious ORFs, and the non-coding set, which contains 5993 segments (fully non-coding regions or introns). Some of these coding and non-coding segments are very short. Regardless of their length, each segment is counted as one when calculating the sensitivity and specificity curves. Figure 7 shows the specificity and sensitivity curves for all of the 16 yeast chromosomes, where the dotted curve is the cumulative distribution function for RT_{p3} for the non-coding regions, and the black curves are the complementary cumulative distribution function for the coding regions, where for clarity, we have computed such distributions for verified ORFs, uncharacterized ORFs, and dubious ORFs, separately. To understand the meaning of such curves, let us focus on the intersection of the solid black curve and the dotted curve. When we choose RT_{p3_0} as a threshold value, then a coding sequence with 78% probability is characterized as coding sequence, while a non-coding sequence with 78% probability is also taken as a non-coding sequence. As expected, this percentage is lower for uncharacterized and dubious ORFs. It is interesting to note that the percentage of accuracy calculated on human genomes is around 74%, close to 78%. Because of this (see also Fig. 3), we conclude that the method is largely species-independent. It is also important to emphasize that the threshold value for defining the accuracy is also fairly species-independent.

It is interesting to note that the period-3 feature is often quantified by performing the Fourier spectral analysis on fairly long DNA sequences. In order to make such analysis applicable to sequences as short as 162 bases, recently a lengthen-shuffle algorithm has been proposed.[11] Fourier spectral analysis together with the lengthen-shuffle algorithm gives about 69% of sensitivity and specificity when evaluated on a prokaryote genome, the *V. cholerae* chromosome I, and about 61% when evaluated on eukaryotic genomes.[12] It is clear that the RT_{p3} index is more accurate. Other features of the recurrence time-based method are: (i) DNA sequences as short as 40 bases can be very well studied. Noting that an expressed sequence tag (EST) is usually very short and that little may be known about the genome to which the EST belongs, this feature, together with the species-independent one, makes the method particularly useful for determining whether a suspected EST belongs to a coding or non-coding region. (ii) The method directly works on the DNA sequence. In contrast, numerical sequences have to be obtained by certain mapping rules in order to use the Fourier spectral analysis based methods. (iii) The measure is somewhat complementary to other well-known codon indices such as codon adaptation index.

4. Discussion

In this paper, we have proposed a simple recurrence time-based method for DNA sequence analysis, and shown that the method can conveniently exhaust all repeat-related structures of length greater than an arbitrarily chosen small word of size w in a genome. We have also shown that the method is very convenient for the study of mutations, insertions and deletions, hence, it holds great potential for the study of evolutionary variations across species and the mechanisms underlying it. By characterizing the peaks at multiples of 3, we have defined a very efficient codon index which is largely species independent and works well on very short sequences. We emphasize that one of the more appealing features of RT_{p3} as a codon index is that no *a priori* knowledge about the sequence is used. Hence, the method will be especially convenient for the study of genome sequences that very little is known. This is the case, for example, when a genome sequence is to be sequenced by a few small research groups by studying expressed sequence tags (ESTs).

While the accuracy of 78% for the yeast genome is already satisfactory, we note that it is possible to improve this percentage by designing other indices from the recurrence times. Readers interested in this issue are encouraged to contact the authors for the raw recurrence time data.

Why are the recurrence time statistics so useful for genomic sequence analysis? The answer may lie at the following interesting facts: (i) The recurrence time is the basic period for periodic phenomena. (ii) The recurrence time is related to the fractal dimension of nonlinear dynamical systems.[45] (iii) For ergodic random sequences/fields, the well-known Ornstein–Weiss theorem[53] states that $\log T/w,\ w \to \infty$, gives the Shannon entropy. Hence, the recurrence time is related to the concept of entropy. (iv) In our analysis, no assumptions about stationarity, ergodicity, and so on, have been made. This makes the method applicable to different types of genome sequences.

Where else may the recurrence time statistics be useful? While an exhaustive list is impossible, we surmise they will be useful in the following areas: (i) The recurrence time statistics may provide useful new information for constructing Markov/Hidden Markov models for finding genes in a genome sequence. (ii) The recurrence time statistics may also be very useful in motif discovery and protein domain analysis. (iii) Our analysis implies that the information content of a long genome sequence may be considerably lower than the values reported so far, if one takes into account repeating sequences separated by very large distances along a genome sequence.

Acknowledgments

The Authors have benefited considerably by participating workshops organized by the Institute of Pure and Applied Mathematics at UCLA.

References

1. Collins FS, Green ED, Guttmacher AE, Guyer MS, A vision for the future of genomics research, *Nature* **422**(6934):835–847, 2003.
2. International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, *Nature* **409**:860–921, 2001.
3. Jurka J, Repeats in genomic DNA: Mining and meaning, *Curr Opin Struct Biol* **8**: 333–337, 1998.
4. Guigó R, DNA Composition, Codon Usage and Exon Prediction, in Bishop MJ (ed.) *Genetics Databases*, Academic Press, San Diego, CA, pp. 53–80, 1999.
5. Herzel H, Weiss D, Trifonov EN, 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding, *Bioinformatics* **15**(3):187–193, 1999.
6. Fukushima A, Ikemura T, Kinouchi M *et al.*, Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis, *Gene* **300**(1–2):203–211, 2002.
7. Bennetzen JL, Hall BD, Codon selection in yeast, *J Biol Chem* **257**:3026–3031, 1982.
8. Sharp PM, Li W-H, The codon adaptation index–a measure of directional synonymous codon usage bias, and its potential applications, *Nucleic Acids Res* **15**:1281–1295, 1987.
9. Jansen R, Bussemaker HJ, Gerstein M, Revisiting the codon adaptation index from a whole-genome perspective: Analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models, *Nucleic Acids Res* **31**: 2242–2251, 2003.
10. Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R, Prediction of probable genes by Fourier analysis of genomic sequences, *Comput Appl Biosci* **13**:263–270, 1997.
11. Yan M, Lin ZS, Zhang CT, A new Fourier transform approach for protein coding measure based on the format of the Z curve, *Bioinformatics* **14**:685–690, 1998.
12. Issac B, Singh H, Kaur H, Locating probable genes using Fourier transform approach, *Bioinformatics* **18**:196–197, 2002.
13. Kotlar D, Lavner Y, Gene prediction by spectral rotation measure: A new method for identifying protein-coding regions, *Genome Res* **13**:1930–1937, 2003.
14. Zhang CT, Wang J, Recognition of protein coding genes in the yeast genome at better than 95% accuracy based an the Z curve, *Nucleic Acids Res* **28**:2804–2814, 2000.
15. Snyder M, Gerstein M, Genomics — Defining genes in the genomics era, *Science* **300**:258–260, 2003.

16. Fickett JW, Guigó R, Computational gene identification, in Swindell S, Miller R, Myers G, (eds.), *Internet for the Molecular Biologist*, Horizon Scientific Press, Wymondham, UK, pp. 73–100, 1996.
17. Zhang MQ, Computational prediction of eukaryotic protein-coding genes, *Nat Rev Genet* **3**:698–709, 2002.
18. Needleman SB, Wunsch C, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J Mol Biol* **48**:443–453, 1970.
19. Smith TF, Waterman MS, Identification of common molecular subsequences, *J Mol Biol* **147**:195–197, 1981.
20. Fitch WM, Smith TF, Optimal sequence alignments, *Proc Natl Acad Sci* **80**:1382–1386, 1983.
21. Altschul SF, Erickson BW, Optimal sequence alignment using affine gap costs, *Bull Math Biol* **48**:603–616, 1986.
22. Pearson WR, Comparison of methods for searching protein sequence databases, *Prot Sci* **4**:1145–1160, 1995.
23. Delcher AL, Kasif S *et al.*, Alignment of whole genomes, *Nucl Acids Res* **27**:2369–2376, 1999.
24. Delcher AL, Phillippy A, Carlton J, Salzberg SL, Fast algorithems for large-scale genome alignment and comparison, *Nucl Acids Res* **30**:2478–2483, 2002.
25. Henikoff S, Henikoff JG, Performance evaluation of amino acid substitution matrices, *Proteins* **17**:49–61, 1993.
26. Jurka J, Klonowski P, Dagman V, Pelton P, CENSOR-A program for identification and elimination of repetitive elements from DNA sequences, *Comput Chem* **20**: 119–122, 1996.
27. Smit AFA, Origin of interspersed repeats in the humna genome, *Curr Opin Genet Devl* **6**:743–749, 1996.
28. Lipman DJ, Pearson WR, Rapid and sensitive protein similarity searches, *Science* **227**:1435–1441, 1985.
29. Pearson WR, Lipman DJ, Improved tools for biological sequence comparison, *Proc Natl Acad Sci* **85**:2444–2448, 1988.
30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, Basic local alignment search tool, *J Mol Biol* **215**:403–410, 1990.
31. Altschul SF, Boguski MS, Gish W, Wootton JC, Issues in searching molecular sequence databases, *Nature Genet* **6**:119–129, 1994.
32. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res* **25**:3389–3402, 1997.
33. Schäffer AA, Aravind L *et al.*, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, *Nucl Acids Res* **29**:2994–3005, 2001.
34. Lippert RA, Huang HY, Waterman MS, Distributional regimes for the number of k-word matches between two random sequences, *Proc Natl Acad Sci* **99**: 13980–13989, 2002.
35. Karlin S, Altschul SF, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, *Proc Natl Acad Sci* **87**: 2264–2268, 1990.
36. Waterman MS, Vingron M, Rapid and accurate estimates of statistical significance for sequence database searches, *Proc Natl Acad Sci* **91**:4625–4628, 1994.
37. Waterman MS, Vingron M, Sequence comparison significance and Poisson approximation, *Stat Sci* **9**:367–381, 1994.

38. Smith TF, Waterman MS, Burks C, The statistical distribution of nucleic acid similarities, *Nucleic Acids Res* **13**:645–656, 1985.
39. Altschul SF, Gish W, Local alignment statistics, *Meth Enzymol* **266**:460–480, 1996.
40. Reich JG, Drabsch H, Daumler A, On the statistical assessment of similarities in DNA sequences, *Nucl Acids Res* **12**:5529–5543, 1984.
41. Blattner FR *et al.*, The complete genome sequence of Escherichia coli K-12, *Science* **277**:1453–1474, 1997.
42. Mewes HW *et al.*, Overview of the yeast genome, *Nature* **387**:7–8, 1997.
43. The C. elegans Sequencing Consortium, Genome Sequence of the Nematode Caenorhabditis elegans-A Platform for Investigating Biology, *Science* **282**:2012–2018, 1998.
44. The Celera Genomics Sequencing Team, The sequence of the human genome, *Science* **291**:1304–1351, 2001.
45. Gao JB, Recurrence time statistics for chaotic systems and their applicaitons, *Phys Rev Lett* **83**:3178–3181, 1999.
46. Gao JB, Cai HQ, On the structures and quantification of recurrence plots, *Phys Lett A* **270**:75–87, 2000.
47. Gao JG, Detecting nonstationarity and state transitions in a time series, *Phys Rev E* **63**:066202, 2001.
48. Gao JB, Cao YH, Gu LY, Harris JG, Principe JC, Detection of weak transitions in signal dynamics using recurrence time statistics, *Phys Lett A* **317**:64–72, 2003.
49. Wolfe KH, Shields DC, Molecular evidence for an ancient duplication of the entire yeast genome, *Nature* **387**:708–13, 1997.
50. Seoighe C, Wolfe KH, Updated map of duplicated regions in the yeast genome, *Gene* **1**:253–261, 1999.
51. Glaever G *et al.*, Functional profiling of the *Saccharomyces cervisivae* genome, *Nature* **418**:387–391, 2002.
52. Brendan J *et al.*, Genome duplications and other features in 12 Mb of DNA sequence from human chromosome 16p and 16q, *Genomics* **60**:295–308, 1999.
53. Ornstein D, Weiss B, Entropy and recurrence rates for stationary random fields, *IEEE Trans Inform Theory* **48**:1694, 2002.

Dr. Yinhe Cao received his Ph.D from University of Missouri-Rolla in 1996. After being Unix System's Architect and Senior Software Engineer at various companies in the area of data security and data communications for a few years, he founded BioSieve a California — based company. At BioSieve, he has developed Expression-Sieve — a microarray data analysis and visualization package. Currently, Dr. Cao serves as a Bioinformatics Application Software Architect at BioSieve, focusing on delivering effective and easy-to-use bioinformatics software tools for the analysis of microarray data, and genome & protein sequence data. He is an expert of software engineering and novel algorithm development. His current research interests include inferring various bio-networks such as protein-protein interaction network and gene regulatory networks from microarray data.

Dr. Wen-wen Tung is an assistant professor in the Department of Earth and Atmospheric Sciences at Purdue University. Her research is mainly focused on multiscale convective systems in the atmosphere. Drawing from her extensive experiences in cloud and climate systems diagnosis and modeling, she also masters complex and massive data processing and analysis techniques and is inspired to involve in interdisciplinary research.

Dr. Jianbo Gao received his Ph.D from the University of California, Los Angeles in 2000. After working at EE of UCLA for another one and half years, he joined ECE of the University of Florida as an assistant professor. He has been working in a number of fields including genomics, nano-computing, nonlinear dynamical systems, and signal processing using random fractal and chaos theory. He has developed a number of novel signal processing tools and found applications in as diverse fields as biology, finance, engineering, and physical sciences. Currently, he has been focusing on the study of bio-molecular networks and systems biology.

Yan Qi received her B.Sc. degree in Electrical Engineering from Zhejiang University, Hangzhou, China, in 2002 and the M.Sc. degree in Electrical and Computer Engineering from the University of Florida, Gainesville in 2003. She is currently working towards the Ph.D degree in Biomedical Engineering at the Johns Hopkins University, Baltimore. Her research interests include fault-tolerant computation, nanoelectronic system architectures, statistical modeling and network dynamics analysis in computational biology and bioinformatics.