BIOINFORMATICS

Vol. 14 no. 8 1998
Pages 685-690

A new Fourier transform approach for protein coding measure based on the format of the Z curve

Ming Yan, Zhe-Suai Lin and Chun-Ting Zhang

Department of Physics, Tianjin University, Tianjin 300072, China

Received on December 27, 1997; revised on June 27, 1998; accepted on July 1, 1998

Abstract

Motivation: *At the core of most protein gene-finding algorithms are the coding measures used to make a decision on coding/non-coding. Of the protein coding measures, the Fourier measure is one of the most important. However, due to the limited length of the windows usually used, the accuracy of the measure is not satisfactory. This paper is devoted to improving the accuracy by lengthening the sequence to amplify the periodicity of 3 in the coding regions.*

Results: *A new algorithm is presented called the lengthen-shuffle Fourier transform algorithm. For the same window length, the percentage accuracy of the new algorithm is 6–7% higher than that of the ordinary Fourier transform algorithm. The resulting percentage accuracy (average of specificity and sensitivity) of the new measure is 84.9% for the window length 162 bp.*

Availability: *The program is available on request from C.-T. Zhang.*

Contact: *ctzhang@tju.edu.cn*

Introduction

Computer-aided protein gene finding in uncharacterized genomic DNA sequences is one of the most important issues of bioinformatics. The problem seems to be simple, although the algorithms may be complicated. For most prokaryotic DNA sequences, the problem is to determine which ORFs in a given sequence are really coding sequences coding for proteins. For eukaryotic DNA sequences, the problem is to determine how many exons and introns in a given sequence there are, and what are the exact boundaries between the exons and introns. In 1992, Fickett and Tung published a review paper which highlighted the progress of the gene-finding algorithms proposed over the past 13 years. The paper reviewed and synthesized the published algorithms, and compared them by a standardized benchmark. They pointed out that future algorithms should be based on Fourier, run, ORF and the in-phase hexamer measures. Based on these conclusions, other powerful gene recognition algorithms have been developed. For example, in GeneMark, the gene recognition algorithm used the fifth-order phased Markov chain model (Borodovsky *et al*., 1994). Here, utilization of the fifth-order phased Markov chain was based on the fact that the in-phase-hexamer statistics were thought of the most effect algorithm (Fickett and Tung, 1992).

Since then, great progress has been made. Probably the most important event that accompanied the development of computer-aided gene-finding studies in this period is the great advance of personal computers and the Internet, including the World Wide Web (WWW). A user can submit his (her) DNA sequence via the Internet to some address or URL of WWW to have the sequence analyzed and returned automatically. Furthermore, users may have many choices. For example, for an integrated gene identification task, they can choose FGENEH (human), GeneID (vertebrate), GeneParser (human), GenLang (dicots, *Drosophila* and vertebrates), GRAIL (human) and EcoParse (*Escherichia coli*), where the organisms suitable for the special algorithm concerned are denoted within parentheses. For only a coding region identification task, they can choose GeneMark (many individual species). The detailed e-mail address or WWW URL for each of the above network services are described in Table 1 of a recent review by Fickett (1996). Readers may refer to other reviews and papers for the relevant algorithm description (Mural *et al*., 1992; Borodovsky *et al*., 1994; Fickett, 1995; Gelfand, 1995; Guigo and Fickett, 1995; Claverie, 1996; Fickett and Guigo, 1996; Snyder and Stormo, 1996; etc.).

Table 1. Fisher discriminant vector **c** and the corresponding threshold t^a

Window length	c_1	c_2	c_3	t
63 bp	0.805	0.109	0.583	21.095
129 bp	0.735	0.092	0.672	12.212
162 bp	0.704	0.382	0.599	11.722

[a] The decision on coding/non-coding for each DNA fragment with the given length is performed by the criterion of $\mathbf{c}\cdot\mathbf{m} > t/\mathbf{c}\cdot\mathbf{m} < t$, where the measure vector **m** is defined by equation (4).

Although great progress in computer-aided gene recognition studies has been made, the situation is still far from being perfect. This may be reflected by the fact that no algorithm currently available can yield a 100% recognition accuracy in general cases. Furthermore, the parameters determined for an algorithm based on previously discovered sequences cannot be applied to identify genes on some recently discovered sequences with an accuracy as high as before (Fickett, 1996). In addition, although the genetic codes are universal for all organisms, the artificially invented computer algorithms are generally only applicable to one or several organisms. The reasons are still not clear. All of these indicate that the development of protein gene-finding algorithms is still in its early stage. There is much room for further improvement. As mentioned above, Fickett and Tung (1992) pointed out that the Fourier measure is one of the most important gene recognition algorithms. In a recent review, Fickett (1996) still addressed the importance of direct measure of periodicity of 3, 6 and 9 for a given DNA sequence to look for possible genes. However, due to the limited length (usually 100 bp or so) of the window used in the gene-finding process, the application of the Fourier measure is without impressive success. This paper is devoted to improving the ordinary Fourier measure currently available. A new algorithm called the lengthen-shuffle FFT algorithm is proposed. The resulting percentage accuracy (average of sensitivity and specificity) reaches 84.9% for a window length of 162 bp. It is hoped that the algorithm proposed here is useful to improve the accuracy of some existing gene-finding algorithms, as discussed later.

Algorithm

Format of Z curves

Consider a DNA sequence with N bases read from the 5-end to the 3-end. Beginning from the first base, inspect the sequence one base at a time. Let the number of steps be denoted by n, i.e. $n = 1, 2, \ldots, N$. In the nth step, count the cumulative numbers of the bases A, C, G and T, respectively, occurring in the subsequence from the first to the nth base in the DNA sequence inspected. Denote the four positive integers by A_n, C_n, G_n and T_n, respectively. The Z curve consists of a series of nodes P_n ($n = 1, 2, \ldots, N$), whose coordinates are denoted by x_n, y_n and z_n. It was shown that (Zhang and Zhang, 1994):

$$\begin{cases} x_n = 2(A_n + G_n) - n, \\ y_n = 2(A_n + C_n) - n, \quad n = 0,\ 1, \ldots,\ N \\ z_N = 2(A_n + T_n) - n, \end{cases} \tag{1}$$

where $A_0 = C_0 = G_0 = T_0 = 0$ and thus $x_0 = y_0 = z_0 = 0$. The connection of the nodes P_0 (i.e. the origin), $P_1, P_2, \ldots, P_N$ one by one by lines is defined as the Z curve of the DNA sequence inspected. It was demonstrated that the Z curve contains all the information in the DNA sequence, and vice versa; each can be reconstructed given the other. We then define:

$$\begin{cases} \Delta x_n = x_n - x_{n-1}, \\ \Delta y_n = y_n - y_{n-1}, \quad n = 1, 2,\ \ldots,\ N \\ \Delta z_n = z_n - z_{n-1}, \end{cases} \tag{2}$$

where Δx_n, Δy_n and Δz_n can only have the values of 1 or –1 (Zhang and Zhang, 1994). Δx_n is equal to 1 when the nth base is A or G (purine), or –1 when the nth base is C or T (pyrimidine); Δy_n is equal to 1 when the nth base is A or C (amino-type), or –1 when the nth base is G or T (keto-type). Similarly, Δz_n is equal to 1 when the nth base is A or T (weak hydrogen bond), or –1 when the nth base is G or C (strong hydrogen bond). Therefore, a DNA sequence can be decomposed into three series of digital signals, consisting of 1 or –1, each of which has clear biological meaning. The first series of digital signals Δx_n represents the distribution of the bases of the purine/pyrimidines along the DNA sequences. The second series Δy_n represents the distribution of the bases of the amino/keto types along the sequence. Similarly, the third series Δz_n represents the distribution of the bases of the strong/weak hydrogen bonds along the sequence (Zhang, 1997).

A lengthen-shuffle Fourier transform

It is well known that there exists an imperfect periodicity of 3 in protein coding sequences (Silverman and Linsker, 1986; Trifonov, 1987; Lio *et al*., 1994; etc.), which is the basis of our method to distinguish between coding and non-coding sequences. For a long sequence, say, longer than 1024 bp, it is easier to detect the periodicity by the FFT algorithm, but for a short sequence, say, shorter than 150 bp or even much shorter, a typical window size usually used, the periodicity of 3 cannot be easily detected by applying the FFT algorithm directly. To solve the problem, the relatively short DNA sequence is first lengthened by repeating the sequence K times, where K is an integer >1. For a sequence with 150 bp, for example, taking $K = 8$, we obtain a lengthened DNA sequence with a length of $8 \times 150 = 1200$ bp. Because the FFT algorithm needs data number to be 2^n (n is a positive integer), the sequence of the first 1024 (2^{10}) bp is used in the FFT program to detect the periodicity. Obviously, a bogus periodicity of 150 will be observed in the power spectrum of the FFT in the example case. To eliminate such a bogus periodicity, and at the same time keep the periodicity of 3 unchanged, the lengthened sequence is then shuffled M times with three consecutive bases as a unit. A typical value of M used here is 10 000.

As mentioned above, based on the format of the Z curve, any DNA sequence can be transformed into three series of digital signals, Δx_n, Δy_n and Δz_n, to which we can apply the FFT algorithm. The power spectrum for each digital series is calculated as follows:

$$P_c(f) = \frac{1}{N}\left|\sum_{n=1}^{N} \Delta c_n \exp[-i2\pi(f/N)n]\right|^2, \ f = 1,\ 2,\ \ldots,\ N \qquad (3)$$

where $P_c(f)$ is the power spectrum associated with Δc_n which represents Δx_n, Δy_n and Δz_n, respectively. It is well known that protein coding genes may exist in one of three possible phases of either strands of a DNA double helix. One advantage of the present method is that the coding measure for six phases can be explored simultaneously.

The detailed procedure of our method is described as follows.

1. Given a DNA sequence with any relatively short length, which should be a multiple of 3, lengthen the sequence by repeating the given sequence many times until the total length of the lengthened sequence is >1024 (2^{10}) bp. Then the first 1024 bp of the resulting sequence are used as the input of the FFT algorithm.
2. To eliminate the bogus periodicity due to the repeat procedure, and at the same time keep the periodicity of 3, shuffle the resulting sequence at least 10 000 times with three consecutive bases as a unit at each shuffling step.
3. Transform the shuffled sequence into three series of digital signals, Δx_n, Δy_n and Δz_n, according to equation (2).
4. Calculate the power spectrum for each of Δx_n, Δy_n and Δz_n to obtain the three numbers $P_x(N/3)$, $P_y(N/3)$ and $P_z(N/3)$ according to equation (3), where N = 1024. Note that 1024/3 is not an integer. To solve this minor problem, the maximum power spectra within the small interval (1024/3) ± 2 in the frequency axis f are taken as the values of $P_x(N/3)$, $P_y(N/3)$ and $P_z(N/3)$.

The benchmark to evaluate the algorithm

The standardized benchmark to evaluate the algorithms used by Fickett and Tung (1992) is used again here to evaluate the lengthen-shuffle FFT algorithm. For the reader's convenience, we describe the whole procedure briefly. For each window length, 1000 fragments of DNA sequences in fully coding regions or exons are prepared in advance. At the same time, 1000 fragments of DNA sequences of fully non-coding regions or introns are also prepared in advance. Each set of 1000 fragments is divided randomly into two equal parts, i.e. 500 are used as the training set and another 500 as the test set. Consequently, both the training and test sets consist of 1000 fragments; 500 are fully coding and another 500 are fully non-coding, respectively. Then the Fisher discriminant algorithm is used to distinguish between the coding and non-coding fragments. In our case, a three-dimensional (3D) space is spanned by the three numbers $P_x(N/3)$, $P_y(N/3)$ and $P_z(N/3)$, denoted by a 3D vector $\boldsymbol{m}$ hereafter. The vector $\boldsymbol{m}$ has three components m_1, m_2 and m_3, where:

$$m_1 = P_x(N/3),\ m_2 = P_y(N/3) \text{ and } m_3 = P_z(N/3) \qquad (4)$$

The Fisher linear discriminant equation in this case represents a plane in the 3D space, described by a vector **c** which has three components c_1, c_2 and c_3. The determination of **c** is simple. Denoted by **T** and **W**, the total covariance matrix and the within-population covariance matrix, respectively, we define $\mathbf{B} = \mathbf{T} - \mathbf{W}$. Using the data in the training set, we calculate **T**, **W** and **B** for each window length. The eigenvector associated with the maximum eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$ is the desired vector **c** (Mardia *et al.*, 1979). The vector **c** is not unique in the sense that **c** multiplied by a constant is still acceptable. Without losing generality, we choose the constant such that $|\mathbf{c}|^2 = 1$. We should point out that the within-population covariance matrix **W** is not singular in our case. So, utilization of the Penrose discriminant algorithm (Fickett and Tung, 1992) is not necessary. Based on the data in the training set, an appropriate threshold t for each window length is determined to make the coding/non-coding decision. The threshold t is uniquely determined by equalizing the sensitivity and specificity or, equivalently, by making the false-negative rate and the false-positive rate be identical. Once the vector **c** and the threshold t are obtained, the decision on coding/non-coding for each fragment in the test set is simply performed by the criterion of $\mathbf{c}\cdot\mathbf{m} > t/\mathbf{c}\cdot\mathbf{m} < t$. The evaluation of the lengthen-shuffle FFT algorithm is simply described by the percentage accuracy, which is the average of the sensitivity and specificity.

Results and discussion

The window lengths 63, 64, 128, 129 and 162 bp are studied here. The lengthen-shuffle FFT algorithm is applied to the window lengths 63, 129 and 162 bp, respectively. The ordinary FFT algorithm, but based on the format of the Z curve, is applied to the window lengths 64 and 128 bp, respectively. The DNA sequences are obtained from the human genome in the GenBank (Burks and Burks, 1988). The vector **c** and the threshold t for each of the window lengths 63, 129 and 162 bp are listed in Table 1. The false-negative rate, the false-positive rate and the percentage accuracy (average of the sensitivity and specificity) for the fragments in the test set for the window lengths 63, 64, 128, 129 and 162 are listed in Table 2. Note that the lengthen-shuffle FFT algorithm is applied only to the lengths 63, 129 and 162, and the ordinary FFT algorithm is applied only to the lengths 64 and 128. Both algorithms are based on the format of Z curves. We compare the false-negative rate, the false-positive rate and their percentage accuracy of the window lengths 64 with those of 63, 128 with 129. The false-negative rate is defined as the fraction of errors on the coding windows. The false-positive rate is defined as the fraction of errors on the non-coding windows. Consequently, the sensitivity and specificity are defined simply by (1 – false-negative rate) and (1 – false-positive rate), respectively. We find that the percentage accuracy (average of the sensitivity and specificity) of the lengthen-shuffle FFT algorithm is higher than that of the ordinary FFT algorithm. In the window length studied here, the accuracy of the new algorithm is 6–7% higher than that of the ordinary algorithm. Interestingly, the increase in accuracy is not only due to the decrease in the false-negative rate, indicating that the signal is amplified, but also due to the decrease in the false-positive rate, indicating that the noise is suppressed. This fact strongly implies that the lengthen-shuffle procedure really raises the ratio of signal/noise. Note that the databases of 63 and 64 bp are almost identical. In fact, we first choose a fragment of 64 bp as an element in the database of 64 bp. Deleting the 64th base from this fragment, we obtain a fragment of 63 bp, which is exactly the corresponding element in the database of 63 bp. A similar situation takes place between the databases of 128 and 129 bp. Therefore, the percentage accuracy of the lengthen-shuffle FFT algorithm is on average 6.5% higher than that of the direct (i.e. without the lengthen-shuffle procedure) FFT algorithm. Although 6.5% is not a high value, it might be useful to improve the accuracy of some existing gene recognition algorithms.

Table 2. The false-negative rate, the false-positive rate and the percentage accuracy for various window lengths and algorithms

Window length	64 bp[a]	63 bp[b]	128 bp[a]	129 bp[b]	162 bp[b]
False-negative rate[c]	0.318	0.260	0.226	0.134	0.108
False-positive rate[d]	0.296	0.240	0.316	0.258	0.194
Accuracy[e]	0.693	0.750	0.729	0.804	0.849

[a]Use the ordinary FFT algorithm, based on the format of the Z curve.
[b]Use the lengthen-shuffle FFT algorithm, based on the format of the Z curve.
[c]The false-negative rate is the fraction of errors on the coding windows.
[d]The false-positive rate is the fraction of errors on the non-coding windows.
[e]The percentage accuracy is the average of the sensitivity and specificity, i.e. the average of (1 – false-negative rate) and (1 – false-positive rate).

The periodicity of 3 in the coding regions was observed by many authors (Silverman and Linsker, 1986; Trifonov, 1987; Lio *et al*., 1994; etc.). Silverman and Linsker studied the overall patterns of periodicity in DNA sequences by the FFT algorithm. On the contrary, Lio *et al.* (1994) studied the periodicity of G + C in the third codon position. They first transformed the DNA sequence studied into S and W sequence, where S represents G or C and W represents A or T. Furthermore, S bases are coded as 1 and W bases are coded as –1. Accordingly, the DNA sequence studied was transformed into a series of 1 and –1. Based on this format, the periodicity of G + C in the third codon position was studied by the correlation function and FFT methods (Lio *et al*., 1994). Interestingly enough, the series they used is exactly the minus z component of the Z curve. Therefore, the periodicity they observed can be detected by the measure $m_3 = P_z(N/3)$ defined above. On the other hand, Trifonov found a G-non–G-N pattern in the coding regions, where N represents any base (Trifonov, 1987). Trifonov suggested that the pattern may be responsible for a reading frame correcting effect during the translation process. It was found early that the preferred codons are of the pattern RNY, where R and Y represent the purine and pyrimidine bases, respectively (Shepherd, 1984). Based on a graphic technique (Zhang and Zhang, 1991), we have observed that the predominant bases in the first codon position are purines. This finding is true for *E.coli* (Zhang and Chou, 1994), human (Zhang and Chou, 1993), HIV (Chou and Zhang, 1992) and many other species (data not yet published). Obviously, the above periodicity of 3 can be detected by the measure $m_1 = P_x(N/3)$ defined above. Compared with $m_1 = P_x(N/3)$ and $m_3 = P_z(N/3)$, $m_2 = P_y(N/3)$ seems to be less important for detecting the 3-periodicity in DNA sequences. Because the *y* component of the Z curve reflects only the distribution of the bases of amino/keto type along the sequence, it seems to us that the bases of amino/keto (M/K) type have less biological significance than those of purine/pyrimidine (R/Y) and strong H bond/weak H bond (S/W) types. This is also reflected by the fact that the magnitude of c_2 is generally far less than c_1 and c_3 (refer to Table 1). See the discussion below with respect to this point, too.

According to the theory of the Z curve (Zhang and Zhang, 1994; Zhang, 1997), any DNA sequence can be decomposed into RY, MK and SW sequence, corresponding to the *x*, *y* and *z* components of the Z curve, respectively. Consequently, various 3-periodicity of DNA sequences can be detected simultaneously by the coding measure vector $\mathbf{m}$ proposed here. The three components m_1, m_2 and m_3 of the vector $\mathbf{m}$ measure the 3-periodicity of bases of the purine/pyrimidine (R/Y), amino/keto (M/K) and strong H bond/weak H bond (S/W) types, respectively, in the DNA sequence studied. To compare the importance of the three measures m_1, m_2 and m_3 more clearly, we have performed the following test. Deleting one component from the 3D vector $\mathbf{m}$ each time, we obtain three 2D vectors. They are denoted by $\mathbf{m}_{12} = (m_1, m_2)$, $\mathbf{m}_{23} = (m_2, m_3)$ and $\mathbf{m}_{13} = (m_1, m_3)$, respectively. Replacing the 3D vector $\mathbf{m}$ by the 2D vectors $\mathbf{m}_{12}$, $\mathbf{m}_{23}$ and $\mathbf{m}_{13}$, respectively, we hope to see what will happen. The database of the window length 162 bp is used to test this idea. Repeating exactly the same lengthen-shuffle procedure and using the standardized evaluation benchmark, we obtain the percentage accuracy (average of the sensitivity and specificity) for each case, i.e. for $\mathbf{m}_{12}$, $\mathbf{m}_{23}$ and $\mathbf{m}_{13}$, respectively. The results are listed in Table 3. As we can see, the accuracy derived from $\mathbf{m}_{12}$ is worse than that from $\mathbf{m}_{23}$, and both are worse than that from $\mathbf{m}_{13}$, indicating that the 3-periodicity of bases of the amino/keto (M/K) types is less important than those of the purine/pyrimidine (R/Y) and strong H bond/weak H bond (S/W) types. It seems that the 3-periodicity of bases of the strong H bond/weak H bond (S/W) type is more important than that of purine/pyrimidine (R/Y) type. Consequently, the order of importance seems to be m_3, m_1 and m_2. Furthermore, the accuracy of all three 2D vectors is worse than that of the 3D vector $\mathbf{m}$, indicating that all of m_1, m_2 and m_3 have their respective contribution to the overall recognition accuracy, even including the component m_2.

Table 3. The percentage accuracy of various measures for window length 162 bp[a]

Measures[b]	$\mathbf{m}_{12}$	$\mathbf{m}_{23}$	$\mathbf{m}_{13}$	$\mathbf{m}$
Accuracy[c]	0.707	0.778	0.830	0.849

[a] Based on the lengthen-shuffle FFT algorithm.
[b] The various measure vectors are defined as $\mathbf{m} = (m_1, m_2, m_3)$, $\mathbf{m}_{12} = (m_1, m_2)$, $\mathbf{m}_{23} = (m_2, m_3)$ and $\mathbf{m}_{13} = (m_1, m_3)$, where m_1, m_2 and m_3 measure the 3-periodicity of RY, MK and SW sequences, respectively. See equation (4).
[c] Average of sensitivity and specificity.

Based on the above analysis, the importance of m_1 and m_3 reminds us to compare the results of the lengthen-shuffle FFT and the ordinary FFT algorithm schematically by using a 2D diagram. Let the x and y axes represent m_1 and m_3, respectively. The distribution of m_1 and m_3 can be displayed on the 2D coordinate plane. The databases of the window lengths 128 and 129 bp are used as examples. Consider the database of 128 bp first. Accordingly, 1000 coding points representing 1000 coding fragments (including 500 in the training set and another 500 in the test set) and 1000 non-coding points representing 1000 non-coding fragments (500 in the training set and another 500 in the test set) are distributed in Figure 1a. The coding points are denoted by open circles and the non-coding points by filled circles. Then consider the database of 129 bp. The corresponding distribution is shown in Figure 1b. Again, the coding points are denoted by open circles and the non-coding points by filled circles. Note that the database of 128 bp and the database of 129 bp are almost identical, as mentioned above. Compare Figure 1a and b. As we can see, the coding points diverge wider after the lengthen-shuffle procedure than before, indicating that the signal of 3-periodicity is amplified. At the same time, the distribution area of the non-coding points is relatively shrunk after the lengthen-shuffle procedure, indicating that the noise is suppressed. Consequently, the overlap between the two kinds of points is reduced and the ratio of the signal/noise is raised by the lengthen-shuffle procedure, as also quantitatively reflected by the data listed in Table 2. Besides, we can find other useful thing from Figure 1a and b. As we can see, the two kinds of points overlap severely. This is the reason why we cannot reach a 100% recognition accuracy. It seems that there is really no obvious 3-periodicity for most of the non-coding fragments and so too for a number of coding fragments. In other words, there is no obvious 3-periodicity for a considerable fraction of coding sequences. The reason is still not clear. This fact leads to the conclusion that a 100% recognition accuracy probably could not be reached based solely on the 3-periodicity detection.

One advantage of the present method is that the coding potential of six phases in a DNA double helix can be explored simultaneously. The method proposed here is of the ‘region-coding’ (Fickett and Tung, 1992). The result of our method has no apparent difference for three phases in a direct strand. It can be shown that the result obtained in the direct strand can be applied to the complementary strand directly. Therefore, the present method provides a tool to scan the double helix quickly to explore the coding potential. Related to the lengthen-shuffle procedure, the second advantage of the present method is that the result is quite insensitive to the

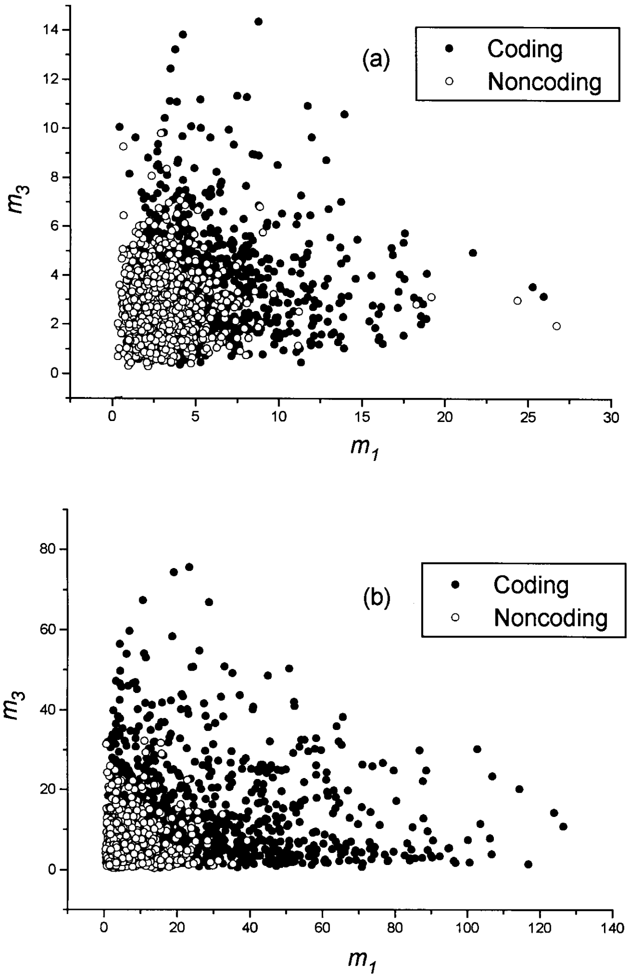

Fig. 1. The graph of m_3 versus m_1, where m_3 and m_1 measure the 3-periodicity of SW and RY sequences, respectively. (**a**) The data of (m_1,m_3) are derived from the ordinary FFT algorithm for the window length 128 bp and (**b**) from the lengthen-shuffle FFT algorithm for the window length 129 bp. Note that the database of 128 bp and that of 129 bp are almost identical. On each figure, there are 1000 coding points representing 1000 coding fragments (denoted by filled circles) and 1000 non-coding points representing 1000 non-coding fragments (denoted by open circles). Compare (a) with (b). Note that the coding points diverge wider after the lengthen-shuffle procedure than before, indicating that the signal of 3-periodicity is amplified. The distribution area of the non-coding points is relatively shrunk after the lengthen-shuffle procedure, indicating that the noise is suppressed. Consequently, the overlap between the two kinds of points is reduced and the ratio of the signal/noise is raised by the lengthen-shuffle procedure.

sequencing errors that are substitutions, but it may be very sensitive to frame-shift sequencing errors. Besides the above two advantages, there is a third possible advantage. Since the 3-periodicity is generally a coherent feature for most of the coding DNA sequences, it is expected that the method and its improved version could be applied to recently discovered sequences with an accuracy as high as to previously discovered sequences, which are used to derive the Fisher discriminant vector **c** and the threshold *t*.

For the window length 162 bp, the percentage accuracy of the lengthen-shuffle FFT algorithm reaches 84.9% (see Table 2), 4% higher than the corresponding value listed in Table 2 of Fickett and Tung (1992). Note that besides the 3-periodicity, 2-, 4-, 5-, 6-, 7-, 8- and 9-periodicity were also considered to obtain the accuracy 80.8% in Table 2 (Fickett and Tung, 1992). It is hoped that the accuracy of the lengthen-shuffle FFT algorithm could be improved further by taking other periodicity into account. The meaning of our work is not only of theoretical interest, but also for some realistic applications. By collaborating on the present algorithm with other existing algorithms, the accuracy of the joint algorithm would be increased more than without such collaboration. For example, there are some artificial intelligence approaches to the protein gene-finding problem, e.g. GRAIL used the neural network to recognize the coding regions or exons (Mural *et al*., 1992). By changing the form of the output result of the present algorithm appropriately, we suggest that the output of the present algorithm may served as an additional input of the neural network of GRAIL. Since the 3-periodicity is a universal feature for most coding sequences, as discussed above, it is hoped that the accuracy of the neural network approach, such as GRAIL (Mural *et al*., 1992), might be increased by collaborating with the method proposed here. The key challenge in eukaryotic gene finding is to recognize the splice sites, i.e. to find the boundaries between introns and exons. We are developing a new algorithm to tackle this important problem (e.g. Zhang *et al*., 1998). By collaborating on such an algorithm, we will develop a new integrated gene identification package based on the present algorithm and its improved version.

Acknowledgements

The present study was supported in part by the Pandeng Project of China and the 39570187 of the China Natural Science Foundation.

References

Borodovsky,M., Koonin,E.V. and Rudd,K.E. (1994) New genes in old sequence: a strategy for finding genes in the bacterial genome. *Trends Biochem. Sci*., **19**, 310–313.

Burks,H.S. and Burks,C. (1988) The Genbank sequence data bank. *Nucleic Acids Res*., **15**, 1861–1864.

Claverie,J.-M. (1996) Effective large-scale sequence similarity searches. *Methods Enzymol*., **266**, 212–227.

Chou,K.-C. and Zhang,C.-T. (1992) Diagrammatization of codon usage in 339 human immunodeficiency virus protein coding sequences and its biological implication. *AIDS Res. Human Retroviruses*, **8**, 1967–1976.

Fickett,J.W. (1995) ORFs and genes: how strong a connection? *J. Comp. Biol*., **2**, 117–123.

Fickett,J.W. (1996) Finding genes by computer: the state of the art. *Trends Genetics*, **12**, 316–320.

Fickett,J.W. and Guigo,R. (1996) In Computational gene identification, Swindell,S.R., Miller,R.R. and Myers,G. (eds), *Internet for the Molecular Biologist*. Horizon Scientific Press, Norfolk, UK, pp. 73–100.

Fickett,J.W. and Tung,C.-S. (1992) Assessment of protein coding measures. *Nucleic Acids Res*., **20**, 6441–6450.

Gelfand,M.S. (1995) Prediction of function in DNA sequence analysis. *J. Comp. Biol*., **2**, 87–115.

Guigo,R. and Fickett,J.W. (1995) Distinctive sequence features in protein coding, genic non-coding and intergenic himan DNA. *J. Mol. Biol*., **253**, 51–60.

Lio,P., Ruffo,S. and Buiatti,M. (1994) Third codon G+C periodicity as a possible signal for an internal selective constraint. *J. Theor. Biol*., **171**, 215–223.

Mardia,K.V., Kent,J.T. and Bibby,J.M. (1979) *Multivariate Analysis*. Academic Press, London.

Mural,R.J., Einstein,J.R., Guan,X., Mann,R.C. and Uberbacher,E.C. (1992) An artificial intelligence approach to DNA sequence feature recognition. *Trends Biotechnol*, **10**, 66–69.

Shepherd,J.C.W. (1984) Fossil remnants of a primeval genetic code in all forms of life? *Trends Biochem. Sci*., **1**, 8–10.

Silverman,B.D. and Linsker,R. (1986) Periodicity of DNA sequences. *J. Theor. Biol*., **118**, 295–300.

Snyder,E.E. and Stormo,G.D. (1996) In Bishop,M.J. and Rawlings,C.J. (eds), *DNA and Protein Sequence Analysis: A Practical Approach*. IRL Press, Oxford, pp. 209–224.

Trifonov,E.N. (1987) Translation framing code and frame-monitoring mechanism as suggested by analysis of mRNA and 16S rRNA nucleotide sequences. *J. Mol. Biol*., **194**, 643–652.

Zhang,C.-T. (1997) A symmetrical theory of DNA sequences and its applications. *J. Theor. Biol*., **187**, 297–306.

Zhang,C.-T. and Chou,K.-C. (1993) Graphic analysis of codon usage strategy in 1490 human protein coding sequences. *J. Protein Chem*., **12**, 329–335.

Zhang,C.-T. and Chou,K.-C. (1994) A graphic approach to analyzing codon usage in 1562 *E.coli* protein coding sequences. *J. Mol. Biol*., **238**, 1–8.

Zhang,C.-T. and Zhang,R. (1991) Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res*., **19**, 6313–6317.

Zhang,C.-T., Lin,Z.-S., Yan,M. and Zhang,R. (1998) A novel approach to distinguish between intron-containing and intronless genes based on the format of Z curves. *J. Theor. Biol*., in press.

Zhang,R. and Zhang,C.-T. (1994) Z curves, an intuitive tool for visualizing and analyzing the DNA sequences. *J. Biomol. Struct. Dyn*., **11**, 767–782.