BIOINFORMATICS

Vol. 20 no. 9 2004, pages 1405–1412
DOI: 10.1093/bioinformatics/bth103

Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation

Deepak Sharma[1], Biju Issac[2], G. P. S. Raghava[2] and R. Ramaswamy[3,4,]*

[1]Department of Biotechnology, All India Institute of Medical Sciences, New Delhi 110029, India; [2]Institute of Microbial Technology, Chandigarh 160036, India; [3]School of Physical Sciences and [4]School of Information Technology, Jawaharlal Nehru University, New Delhi 110067, India

Received on June 4, 2003; revised on August 31, 2003; accepted on November 30, 2003
Advance Access publication February 19, 2004

ABSTRACT

Motivation: Repetitive DNA sequences, besides having a variety of regulatory functions, are one of the principal causes of genomic instability. Understanding their origin and evolution is of fundamental importance for genome studies. The identification of repeats and their units helps in deducing the intra-genomic dynamics as an important feature of comparative genomics. A major difficulty in identification of repeats arises from the fact that the repeat units can be either exact or imperfect, in tandem or dispersed, and of unspecified length.

Results: The Spectral Repeat Finder program circumvents these problems by using a discrete Fourier transformation to identify significant periodicities present in a sequence. The specific regions of the sequence that contribute to a given periodicity are located through a sliding window analysis, and an exact search method is then used to find the repetitive units. Efficient and complete detection of repeats is provided together with interactive and detailed visualization of the spectral analysis of input sequence. We demonstrate the utility of our method with various examples that contain previously unannotated repeats. A Web server has been developed for convenient access to the automated program.

Availability: The Web server is available at http://www.imtech.res.in/raghava/srf and http://www2.imtech.res.in/raghava/srf

Contact: r.ramaswamy@mail.jnu.ac.in

INTRODUCTION

The genomes of all eukaryotes contain repetitive elements of varying lengths that can occupy a significant fraction of the total DNA content. More than 50% of the human genome is thought to consist of repeats of various types (Lander *et al*., 2001). Since repetitive DNA sequences are presumed to be important in a number of regulatory functions (Tautz *et al*., 1986; Pardue *et al*., 1987; Bucher and Yagil, 1991; Burge *et al*., 1992; Lu *et al*., 1993; Kundu and Rao, 1999) and are one of the principal causes of genomic instability, their identification within genomic DNA sequences is of considerable importance and interest.

There are two major groups of repeats in eukaryotic genomes: tandem repeats that are usually confined to specific chromosomal regions and repeats interspersed with genomic DNA mainly represented by inactive (pseudogenes) copies of historically or contemporarily active transposable elements (Strachan and Read, 1999). Tandem repeats are grouped into three major subclasses: satellites, minisatellites and microsatellites (Strachan and Read, 1999). Satellite repeats are composed of very long tandem arrays of 5–171 bp repeat units usually present at centromeres. Minisatellites consists of tandem repeats of short units of about 6–64 bp located near the telomeres, while microsatellite repeats are highly repetitive sequences consisting of 1–4 bp repeated up to 50 times as tandem arrays dispersed throughout all the chromosomes. Likewise, interspersed repeats can also be subgrouped into five types: Short Interspersed Nuclear Elements (SINEs), Long Interspersed Nuclear Elements (LINEs), Long Terminal Repeats (LTRs), DNA transposons and others (Smit, 1996).

Exact repeats are multiple copies of a given unit that may be present in tandem or separated by distances of more than 1 bp (i.e. dispersed repeats). When exact repeats in DNA have mutated excessively, both tandem and dispersed repeats become difficult to locate, and one of the major challenges in repeat detection is the location of such latent or hidden repeats. Owing to the lack of observable periodicity, these can be difficult to identify using traditional programs.

Both exact and heuristic methods for locating repeats have been developed in the past (Kurtz *et al*., 2000). Typically, exact methods first formally define a model of a repeat and then locate all regions in a given sequence that satisfy

*To whom correspondence should be addressed.

this definition. A simple a priori method that finds repeats without prior knowledge is a dot plot in which a plot of sequence against itself is created, as in Dotter (Sonnhammer and Durbin, 1995). Heuristic methods are useful in dealing with large sequences but are not guaranteed to find all possible repeats. However, these are of considerable utility: for instance, the program RepeatMasker (Smit and Green, unpublished data) helps to remove or mask repetitive DNA in genomic sequences since these tend to confuse sequence analysis programs. The program defines a repeat as a substring that occurs 'very often' in a genome and then employs an exact or approximate string matching of the given sequence using a dictionary of known repeat sequences. A number of programs such as FORRepeats (Lefebvre *et al*., 2003) and REPuter (Kurtz *et al*., 2001) have been developed for identification of repeats on a genomic scale. Similarly, programs such as Tandem Repeat Finder (TRF; Benson, 1999), Tandyman (http://www.stdgen.lanl.gov/tandyman/index.html, unpublished data), Sputnik (http://abajian.net/sputnik, unpublished data), Mreps (Kolpakov and Kucherov, 1999, 2001) and TROLL (Castelo *et al*., 2002) are designed to find tandem repeats from a DNA sequence.

In the present paper, we implement a Fourier transform (FT) method in order to locate and identify repetitive DNA and its constituent units. This method, which we have named Spectral Repeat Finder (SRF), is useful for detecting tandem as well as dispersed repeats, and can also locate latent periodicities within genomic DNA. There has been extensive use of correlation functions and the related Fourier spectra to analyse DNA sequences (Arques and Michel, 1987, 1990; Pasquier *et al*., 1998; Herzel *et al*., 1999), including the use of power spectra to locate protein-coding regions using the three-base periodic property of coding genes (Tiwari *et al*., 1997; Yan *et al*., 1998; Issac *et al*., 2002). Spectral techniques have not, so far, been extensively used to detect repetitive sequences even though it would seem that they are ideally suited for detecting periodic patterns. The method we propose here first identifies the length of the potential repeat unit present in any DNA sequence by evaluating the power spectrum. Subsequently, the sequence is scanned at particular individual frequencies to locate the approximate region(s) where the repeat units are contained. Potential seed patterns from these regions are then used to identify repeats through an exact method.

We have developed a Web server (http://www.imtech.res.in/raghava/srf) that implements the method. Here, we present results from applications of SRF on a number of examples. These include unannotated DNA sequences as well as cases of existing annotations, where we show that our method reveals hitherto unknown repeat units in addition to the annotated repeats. SRF extends the previously developed GeneScan algorithm (Tiwari *et al*., 1997) to distinguish the repeat regions. The server gives users information such as the initial seed pattern that is searched for in the predicted region, the Fourier spectrum from both the initial FT and the window-based FT, the score for each pattern and a colored display of patterns and their positions along the DNA sequence.

METHODS

Algorithm

Consider a DNA sequence of length n,

$$\alpha_1\alpha_2\alpha_3\cdots\alpha_n,$$

where α_i is the base (A, T, G or C) at position i. The structure of symbolic correlations within the sequence is most simply explored through correlation functions of the type

$$C_{\alpha\beta}(r) = \langle U_\alpha(i)\, U_\beta(i+r)\rangle \quad \alpha,\beta \in \{A,T,G,C\}, \qquad (1)$$

where $U_\alpha(i) = 1$ if α_i, the symbol at position i, is α and 0 otherwise. $U_\alpha(i)$ is termed the 'indicator function' or 'projection operator' (Voss, 1992; Tiwari *et al*., 1997) and $\langle\cdots\rangle$ denotes an average over the string [note that $C_{\alpha\beta}(r)$ is not normalized]. In the past several years, the self-correlation,

$$C(r) = \sum_\alpha C_{\alpha\alpha}(r), \qquad (2)$$

has been most commonly studied. If there is a non-trivial N-base correlation such as is caused by a repetitive unit, the correlation function has a near recurrence at this period, namely, $C(r) \approx C(N+r)$ (Herzel *et al*., 1999). It is computationally simpler to consider the Fourier transform of $C(r)$, which is an entirely equivalent method of examining correlations. One commonly used definition for the power spectrum (Silverman and Linsker, 1986; Li *et al*., 1994) is

$$S(f) = \sum_\alpha \frac{1}{n^2}\left|\sum_{j=1}^{n} U_\alpha(j)\mathrm{e}^{2\pi i f j}\right|^2. \qquad (3)$$

A peak at frequency $f = 1/N$ in the spectrum indicates that the correlation function has N-base periodicity; see Li (1997) for a detailed discussion of the Fourier spectrum and its connection with the correlation function. It should be pointed out that while a N-base correlation does not necessarily imply the presence of a repeat unit of length N, the converse is true: the existence of repeats of length N will give rise to a peak at frequency $f = 1/N$. [In this context, recall that the three-base periodicity in coding sequences gives a sharp peak in the power spectrum at frequency 1/3 (Fickett, 1982; Tiwari *et al*., 1997), though, of course, coding regions are not composed of three-base repeats.]

Our study of the power spectrum of a number of sequences suggests a simple method for discovering hidden periodicities. This applies in both instances, for tandem and dispersed repeats, as well as when they are imperfect or 'noisy'. For a sequence string of length n, periodicities of the order of $n/2$ can in principle show up in the power spectrum, which is

computed at frequencies $f = k/2n,\ k = 1, 2, \ldots, n$. We first identify those frequencies, f_i, such that

$$S(f_i)/\bar{S} > T, \tag{4}$$

where the spectral average

$$\bar{S} = \frac{1}{n}\left(1 + \frac{1}{n} - \sum_{\alpha} \rho_{\alpha}^2\right) \tag{5}$$

and ρ_α is the frequency of nucleotide α in the sequence.

The significance of any spectral line should be assessed with respect to the spectral average, namely through its signal-to-noise (S/N) ratio. Peaks with S/N greater than 2 typically begin to be significant, but from a number of studies, we find that since random fluctuations can frequently drive a spectral line to this level, a more prudent criterion would require a higher threshold (T); this threshold is similar to the discriminator used in the gene finding method (Tiwari *et al*., 1997) and is quite insensitive to window and/or sequence length. We find from a number of simulations (Fig. 4 and the related discussion) that a threshold of S/N = 4 is useful (Tiwari *et al*., 1997; Ramaswamy and Ramachandran, 1999; Aggarwal and Ramaswamy, 2002) and this makes it possible to locate repeats, which constitute as little as 2% of the total sequence (Table 2).

Having found candidate repeat lengths $N_i = 1/f_i$, the S/N ratio of the spectral peak at frequency f_i is computed in a sliding window along the sequence. As for coding regions (Tiwari *et al*., 1997), it is simple to identify the regions containing the repeats as those where the S/N ratio exceeds the threshold. Since the length of the repeat ($1/f_i$) and the region containing the repeats are both completely specified, the actual repeats can be easily identified by testing possible $N_i$mers in the region for the occurrence of multiple copies. This can be done by exact enumeration or even by a heuristic local alignment method.

We summarize the SRF algorithm in the following steps:

Step 1: Input a DNA sequence of length n.

Step 2: Compute the power spectrum, $S(f)$, and the spectral average, $\bar{S}$, of the entire sequence.

Step 3: Identify all peaks with $S(f_i)/\bar{S} > T$ (the threshold, here chosen to be 4). For each frequency f_i so identified, there are potential repeats of length $N_i = 1/f_i$.

Step 4: For each of these, compute $P_m(j) = S(f_i)/\bar{S}$ in a sliding window of length m centered on position j in the sequence. Regions containing a repeat of length N_i can be identified directly as those where $P_m(j)$ exceeds the threshold.

Step 5: Since both the repeat length, N_i, and its location are known, an exact method (step 6) is used to identify the repeat units.

Step 6: Consider all $N_i$mers in the repeat region and identify those occurring most frequently by local alignment. This automatically makes it possible to allow for insertions and deletions to any desired level.

In principle, sequences of any size can be analyzed through the above approach, but there are practical limitations arising from the fact that the power spectrum tends to get crowded, and therefore assessing the significance of a peak can be difficult. Therefore we split long sequences (>15 kb) in overlapping segments of length 10 kb; each of these segments is then analyzed individually for the presence of repetitive units. Results from the analysis of *Plasmodium* sequences of varying lengths (>50 kb) are available at http://www.imtech.res.in/raghava/srf/supl/longseq.html. The time complexity of the present algorithm is $O(n^2)$; efforts are under way to improve the time performance through incorporation of a heuristic local alignment technique such as BLAST (Altschul *et al*., 1997) or FASTA (Pearson and Lipman, 1988).

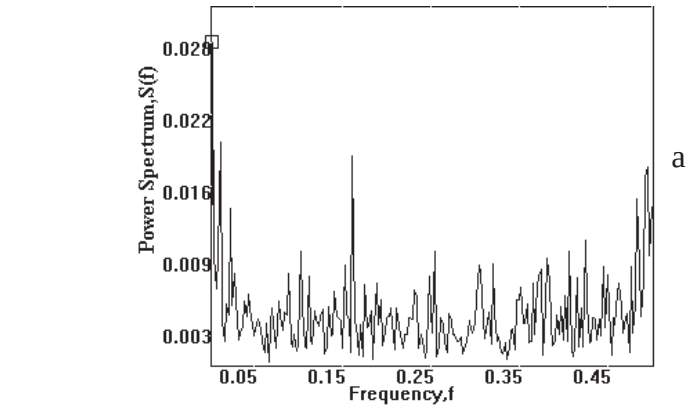
Tandem repeats give the strongest signals, namely very high S/N ratios of the corresponding spectral peaks. This signal degrades both with dispersion and with imperfection, as can be expected since this affects the correlation function, $C(r)$ (Fig. 4). We illustrate the utility of the present technique with some representative examples in the next section. In all cases we use the threshold $T = 4$. The choice of window length and slide length is made heuristically, but typically we use m a multiple of N_i, and slide lengths of 1 nt.

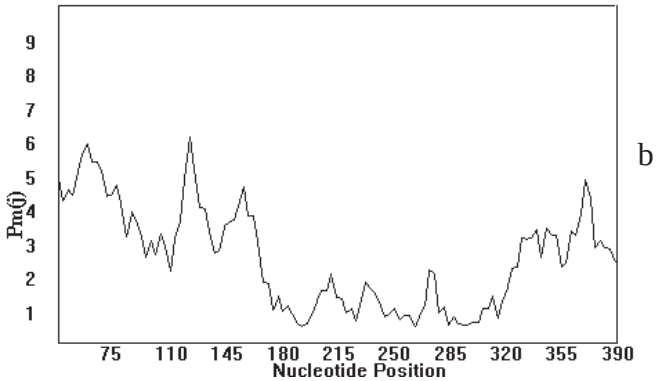
Implementation

To evaluate the ability of SRF to find repeats, we analyzed a number of annotated sequences that are known to contain repeats. We present here two examples of previously annotated sequences and demonstrate the ability of SRF to detect additional and previously unknown latent periodicities of potential significance.

Consider the microsatellite sequence M96445, annotated for the presence of dinucleotide repeats. The power spectrum shown in Figure 1a however suggests that there may be, in addition, unannotated hexamer repeats that give rise to the peak at $f = 1/6$. The peaks are of comparable height, indicating that the two repeats are in about the same proportion. A sliding window analysis at this latter frequency (Fig. 1b) suggests that the hexamer repeats are located between 1 and 160 bp; detailed analysis of the sequence by SRF (Fig. 1c) revealed the presence of imperfect copies of the 6mer GGCTGT.

As is well known, many naturally occurring sequences are composed of a complex pattern of repeats, many of which are not conserved. Existing algorithms often have difficulty in this regard: e.g. in the above-mentioned case, TRF located only the dinucleotide repeats in the sequence (analysis was done using default parameters). Similarly, microsatellite





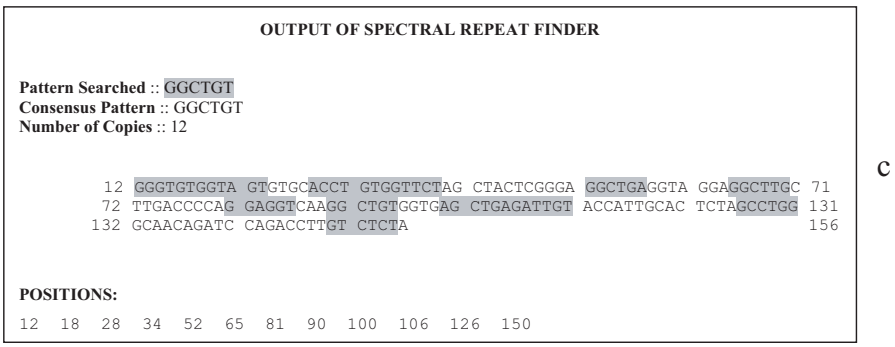


Fig. 1. (**a**) Power spectrum of the microsatellite sequence M96445. The major peaks are at $f = 1/2$ and 1/6. (**b**) Sliding window analysis at scanning frequency $f = 1/6$ and a window length of 96. (**c**) Detailed result of the SRF analysis showing the exact locations of the dispersed hexamer repeats (shaded) that give rise to the spectral feature at $f = 1/6$ in (a).

sequence M65145 shows two peaks at $f = 1/2$ and 1/11 (Fig. 2a). It is clear from the sliding window analysis of this sequence, at frequency $f = 1/2$ (Fig. 2b) and at frequency $f = 1/11$ (Fig. 2c), that dinucleotide repeats occur between positions 800 and 900 bp (GenBank annotation is between 860 and 900 bp), while the 11mer repeats are located between positions 100 and 500 bp (unannotated). SRF analysis of the sequence reveals that the 11 base correlation corresponds to a dispersed imperfect copy of the 11mer TGACTTTGGGG (Table 1). The method of Pasquier *et al*. (1998) and REPuter (Kurtz *et al*., 2001) did not locate either the hexamer or the 11mer repeats in the above examples. However, it would be interesting to point out that REPuter was able to find repeats of varying lengths in the regions mentioned in Table 1; clustering methods can then be used to align these 'specific' repeats to extract the hidden periodicity. The identification of the tandem dinucleotide repeats and the dispersed as well as imperfect hexamer/11mer repeats clearly substantiates the advantage of our method in locating hidden periodic patterns.

When the repeats are well conserved and in tandem, it frequently happens that the power spectrum has

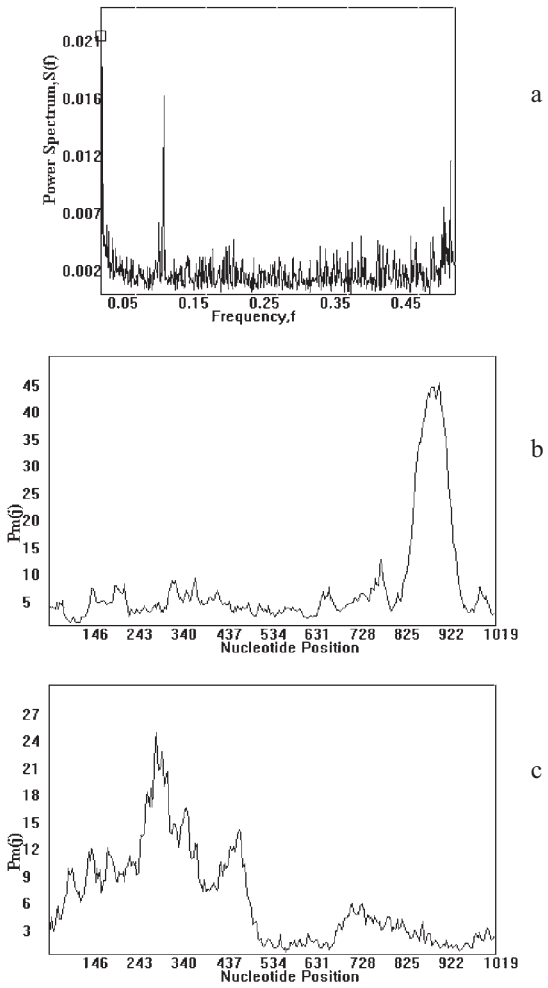


Fig. 2. (**a**) Power spectrum of the microsatellite sequence M65145. The major peaks are at $f = 1/2$ and 1/11. (**b**) Sliding window analysis at scanning frequency $f = 1/2$ and a window length of 100. The doublet repeats are located between positions 800 and 900 bp. (**c**) Sliding window analysis at scanning frequency $f = 1/11$ and a window length of 99. The 11mer repeats are located between positions 100 and 500 bp.

Table 1. 11mer repeats in the microsatellite sequence M65145

Nucleotide position	Sequence										
131–141	T	G	A	C	C	T	T	T	G	G	G
157–167	T	G	A	C	C	T	T	G	G	G	G
256–266	T	G	A	C	T	T	T	A	G	G	G
300–310	T	T	T	C	T	T	T	G	G	G	G
322–332	T	G	A	C	T	T	T	G	G	G	G
346–356	T	G	A	T	T	T	T	G	A	G	G
411–421	T	G	A	C	T	T	T	G	A	A	G
458–468	T	G	A	C	T	C	T	G	G	G	G
634–644	T	G	G	C	T	T	G	G	G	G	G
738–748	T	G	T	C	T	C	T	G	G	G	G
Consensus sequence	T	G	A	C	T	T	T	G	G	G	G

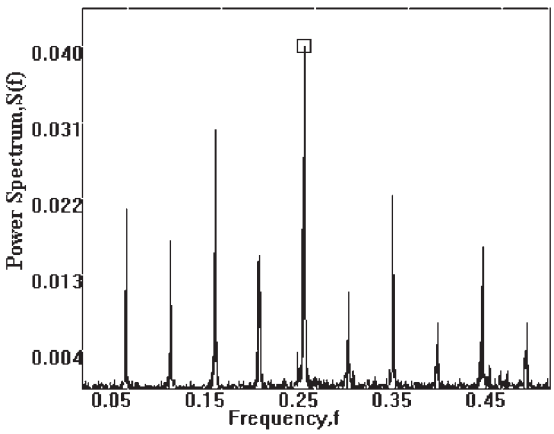


Fig. 3. Power spectrum of AE001365 (region 6001–8000 bp) of Chromosome 2 of *P.falciparum*. The individual peaks are overtones of the basic frequency $f = 1/21$, which is the left-most peak.

Table 2. Partial results[a] of the repeats found in two of the sections of *P. falciparum* by Spectral Repeat Finder and comparison with TRF results

Accession no.	Program	Period size	Indices	Copy no.
AE001381	TRF[b]	24	9619–9978	14.5
	SRF	10	8064–10 756	60
		11[c]	4330–5003	12[d]
		12	1–1114	13[d]
		16	1440–2738	26
		21[c]	1767–5542	30
		23[c]	8659–10 921	14
		24	5765–11 850	30
		32[c]	10 191–12 779	11
AE001383	TRF	—	—	—
	SRF	10	3614–7665	66
		10	8521–12 389	79
		11[c]	1–1372	22[d]
		12	13 558–14 186	19[d]
		16	12 939–13 802	10[d]
		18	9271–11 343	11[d]
		50[c]	519–4664	12

[a]Complete and detailed results are available at http://www.imtech.res.in/raghava/srf/supl/table2.html. Only repeats of length ≥10, copy number ≥10 and with percentage matches ≥75 have been shown.
[b]A program to find tandem repeats only.
[c]These are entirely dispersed repeats, with no occurrence of repeat units in tandem.
[d]Constitutes <2% of the length of the sequence.

a very characteristic fingerprint consisting of the basic spectral peak at frequency $1/N$ and the overtones at k/N, $k = 2, 3, \ldots, (N/2)$. In such cases, the left-most peak of the spectrum represents the actual length of the repeats. An example of such a spectrum, deriving from Chromosome 2 of *Plasmodium falciparum* (AE001365, 6001–8000 bp), having 21 bp repeats, is shown in Figure 3 (detailed results are available at http://www.imtech.res.in/raghava/srf/supl/AE001365.html). Given in Table 2

Table 3. 23mer repeats present in AE001381 of *P. falciparum*

Nucleotide position	Sequence																						
9322–9344	T	C	C	A	A	G	A	T	A	A	T	A	T	T	T	T	C	T	T	C	T	T	C
9544–9566	T	C	A	T	T	T	A	A	A	A	A	A	T	T	T	T	T	T	T	T	T	T	C
9646–9668	T	T	A	T	G	T	A	A	A	A	T	A	T	T	T	T	C	C	T	C	C	T	C
9670–9692	T	T	A	T	A	T	A	A	A	A	T	A	T	T	T	T	C	T	T	C	T	T	C
9697–9719	T	T	A	T	G	T	A	A	A	A	T	A	T	T	T	T	C	C	T	C	T	T	C
9721–9743	T	T	A	T	G	T	A	A	A	A	T	G	T	T	T	T	C	C	T	C	T	T	C
9748–9770	T	T	A	T	G	T	A	A	A	A	T	A	T	T	T	T	C	C	T	C	C	T	C
9772–9794	T	T	A	T	A	T	A	A	A	A	T	G	T	T	T	T	C	C	T	C	T	T	C
9799–9821	T	T	A	T	G	T	A	A	A	A	T	A	T	T	T	T	C	C	T	C	C	T	C
9823–9845	T	T	A	T	A	T	A	A	A	A	T	A	T	T	T	T	C	C	T	C	T	T	C
9850–9872	T	T	A	T	G	T	A	A	A	A	T	A	T	T	T	T	C	T	T	C	T	T	C
9898–9920	T	T	A	T	G	T	A	A	A	A	T	A	T	G	T	T	C	C	T	C	T	T	C
9922–9944	T	T	A	T	G	T	A	A	A	A	T	A	T	T	T	T	C	C	T	C	T	T	C
9946–9968	T	T	A	T	T	T	A	T	A	A	T	A	T	T	T	T	C	C	C	C	T	T	C
Consensus sequence	T	T	A	T	G	T	A	A	A	A	T	A	T	T	T	T	C	C	T	C	T	T	C

are results from our analysis of two other sections of Chromosome 2, and a detailed comparison is made with the predictions of TRF. One of the dispersed repeats identified in Table 2 by our technique is shown explicitly in Table 3.

DISCUSSION

The location of repetitive DNA segments is currently a problem of considerable importance. We have presented here a spectral method, SRF, which has the potential to identify and locate direct repeats present within a sequence. While the significance of micro- and minisatellites is known to some extent, the biological significance of latent periodicities that show up as significant peaks in the power spectrum of a DNA sequence is not apparent at present. However, the large- and small-scale structure of genomic DNA is only beginning to be understood, and it is likely that these weakly periodic structures have some importance, either functionally or from an evolutionary point of view.

Although it would seem a very natural technique to apply for the analysis of repetitive sequences, spectral analysis has not so far been used extensively in the problem of repeat detection. Some earlier studies have indeed used the power spectrum and the correlation function in order to identify periodicities within DNA sequences, but a systematic use of such measures to locate repeats has been lacking.

One practical problem in identifying repetitive DNA is that the repeated elements can vary in length due to insertions and deletions, and even when they are of the same length, they can differ considerably due to substitutions. String-matching based algorithms for finding repeats require the specification of error tolerances. By making the match conditions more stringent, only exact or near exact repeats are located. With more relaxed tolerance levels, more approximate repeats can be found, but their significance is doubtful.

In contrast, a correlation based technique, or equivalently a power spectral technique, averages information over the entire DNA sequence. Such methods operate in a different manner from string matching and are thus sensitive to latent periodicities. The method that we have presented here, SRF, extends the previously suggested algorithm, GeneScan (Tiwari *et al*., 1997), for coding sequence identification (which is based on detecting the three-base periodicity that is characteristic of protein-coding DNA sequences). Like GeneScan, SRF is an *ab initio* technique since no prior assumptions need to be made regarding the repeat length, its fidelity or whether the repeats are in tandem or not. A DNA sequence is converted into a digital signal whose power spectrum is computed via the discrete Fourier transform. Repeats show up as peaks of high intensity in the power spectrum, but this varies with the relative number of repeats. A large number of exact tandem repeats give rise to a very strong signal; the quality of the signal degrades with dispersion as well as when the repeats lose fidelity either through substitutions or through insertions and deletions that change the length of the repeat unit (Fig. 4). We have also found from simulations that the relationship between the percentage of repeats present and signal strength (or peak height) is almost quadratic for exact tandem repeats. In these simulations, exact repeats were tandemly inserted, in artificially constructed random DNA sequences (the power spectra of which were essentially Gaussian white noise). Similar experiments with artificially constructed sequences containing completely dispersed or imperfect tandem repeats all suggest that a spectral technique can successfully identify such repeats even when they constitute as little as 2% of the entire sequence.

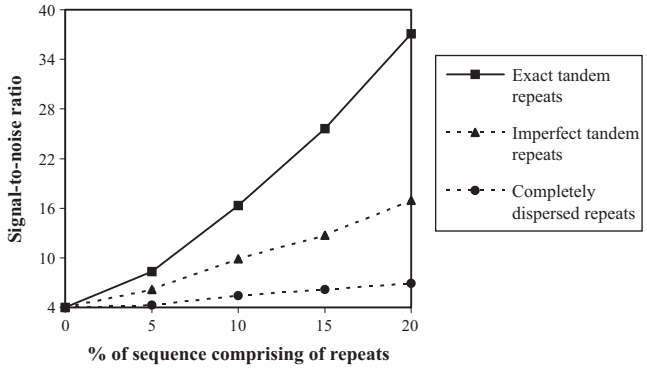


Fig. 4. Signal-to-noise ratio versus percentage of repeats present in the sequence for artificially constructed sequences containing 5 bp units as tandem repeats (filled squares), completely dispersed repeats (filled circles) and imperfect tandem repeats with point mutations and indels (filled triangles). Results have been averaged over several realizations and represent typical variations. As can be seen, the signal strength degrades with both imperfection (here, $P_{\mathrm{mut}} = 0.18$ per site and $P_{\mathrm{indel}} = 0.02$ per site) and dispersion.

ACKNOWLEDGEMENT

We have greatly benefited from discussion and collaboration with Alok Bhattacharya, Sudha Bhattacharya and Andrew Lynn during initial stages of this work. R.R. is partially supported by a grant from DBT, Government of India. D.S. is thankful to J.S. Tyagi for permission to carry out this work and to CSIR for a Senior Research Fellowship. B.I. and G.P.S.R. acknowledge the financial and infrastructural contribution made by DBT and CSIR.

REFERENCES

Aggarwal,G. and Ramaswamy,R. (2002) *Ab initio* gene identification: prokaryote genome annotation with GeneScan and Glimmer. *J. Biosci.*, **27**, 7–14.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Arques,D.G. and Michel,C.J. (1987) Periodicities in introns. *Nucleic Acids Res.*, **15**, 7581–7592.

Arques,D.G. and Michel,C.J. (1990) Periodicities in coding and noncoding regions of the genes. *J. Theor. Biol.*, **143**, 307–318.

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Bucher,P. and Yagil,G. (1991) Occurrence of oligopurine. oligopyrimidine tracts in eukaryotic and prokaryotic genes. *DNA Seq.*, **1**, 157–172.

Burge,C., Campbell,A.M. and Karlin,S. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl Acad. Sci.*, USA, **89**, 1358–1362.

Castelo,A.T., Martins,W. and Gao,G.R. (2002) TROLL—Tandem Repeat Occurrence Locator. *Bioinformatics*, **18**, 634–636.

Fickett,J.W. (1982) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.*, **10**, 5303–5318.

Herzel,H., Weiss,O. and Trifonov,E.N. (1999) 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics*, **15**, 187–193.

Issac,B., Singh,H., Kaur,H. and Raghava,G.P.S. (2002) Locating probable genes using Fourier transform approach. *Bionformatics*, **18**, 196–197.

Kolpakov,R. and Kucherov,G. (1999) Finding maximal repetitions in a word in linear time. *Symposium on Foundations of Computer Science (FOCS)*, New York, USA, pp. 596–604.

Kolpakov,R. and Kucherov,G. (2001) Finding approximate repetitions under Hamming Distance. *9th European Symposium on Algorithms (ESA)*, Arhus, Denmark, Lecture Notes in Computer Science, **2161**, pp. 170–181.

Kundu,T.K. and Rao,M.R. (1999) CpG islands in chromatin organization and gene expression. *J. Biochem.*, **125**, 217–222.

Kurtz,S., Ohlebusch,E., Schleiermacher,C., Stoye,J. and Giegerich,R. (2000) Computation and visualization of degenerate repeats in complete genomes. *Proceedings of the International Conference on Intelligent Systems For Molecular Biology*, AAAI Press, Menlo Park, CA, pp. 228–238.

Kurtz,S., Choudhuri,J.V., Ohlebusch,E., Schleiermacher,C., Stoye,J. and Giegerich,R. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.*, **29**, 4633–4642.

Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Dolye,M., FitzHugh,W. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lefebvre,A., Lecroq,T., Dauchel,H. and Alexandre,J. (2003) FORRepeats: detects repeats on entire chromosomes and between genomes. *Bioinformatics*, **19**, 319–326.

Li,W. (1997) The study of correlation structures of DNA sequences: a critical review. *Comput. Chem.*, **21**, 257–271.

Li,W., Marr,T.G. and Kaneko,K. (1994) Understanding long-range correlations in DNA sequences. *Physica D*, **75**, 392–416.

Lu,Q., Wallrath,L.L., Granok,H. and Elgin,S.C. (1993) $(CT)_n$ $(GA)_n$ repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the *Drosophila hsp26* gene. *Mol. Cell Biol.*, **13**, 2802–2814.

Pardue,M.L., Lowenhaupt,K., Rich,A. and Nordheim,A. (1987) $(dC\text{-}dA)_n$ $(dG\text{-}dT)_n$ sequences have evolutionarily conserved chromosomal locations in *Drosophila* with implications for roles in chromosome structure and function. *EMBO J.*, **6**, 1781–1789.

Pasquier,C.M., Promponas,V.I., Varvayannis,N.J. and Hamodrakas,S.J. (1998) A Web server to locate periodicities in a sequence. *Bioinformatics*, **14**, 749–750.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci.*, USA, **85**, 2444–2448.

Ramaswamy,R. and Ramachandran,S. (1999) Gene identification in bacterial and organellar genomes using GeneScan. *Comput. Chem.*, **23**, 165–174.

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

Smit,A.F. (1996) The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.*, **6**, 743–748.

Sonnhammer,E.L. and Durbin,R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1–GC10.

Strachan,T. and Read,A.P. (1999) *Human Molecular Genetics*, 2nd edn. John Wiley and Sons (Asia) Pte Ltd, pp. 139–168.

Tautz,D., Trick,M. and Dover,G.A. (1986) Cryptic simplicity in DNA is a major source of genetic variation. *Nature*, **322**, 652–656.

Tiwari,S., Ramachandran,S., Bhattacharya,A., Bhattacharya,S. and Ramaswamy,R. (1997) Prediction of probable genes by Fourier analysis of genomic sequences. *Comput. Appl. Biosci.*, **13**, 263–270.

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

Yan,M., Lin,Z.-S. and Zhang,C.-T. (1998) A new Fourier transform approach for protein coding measure based on the format of the Z curve. *Bioinformatics*, **14**, 685–690.