Published online 12 September 2012

Nucleic Acids Research, 2013, Vol. 41, No. 1 ***e17***
doi:10.1093/nar/gks721

Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm

Matko Glunčić[1,*] and Vladimir Paar[1,2,*]

[1]Faculty of Science, University of Zagreb, Bijenička 32 and [2]Croatian Academy of Sciences and Arts, Zrinski trg 11, 10000 Zagreb, Croatia

Received April 23, 2012; Revised June 19, 2012; Accepted July 5, 2012

ABSTRACT

The main feature of global repeat map (GRM) algorithm (www.hazu.hr/grm/software/win/grm2012.exe) is its ability to identify a broad variety of repeats of unbounded length that can be arbitrarily distant in sequences as large as human chromosomes. The efficacy is due to the use of complete set of a *K*-string ensemble which enables a new method of direct mapping of symbolic DNA sequence into frequency domain, with straightforward identification of repeats as peaks in GRM diagram. In this way, we obtain very fast, efficient and highly automatized repeat finding tool. The method is robust to substitutions and insertions/deletions, as well as to various complexities of the sequence pattern. We present several case studies of GRM use, in order to illustrate its capabilities: identification of α-satellite tandem repeats and higher order repeats (HORs), identification of Alu dispersed repeats and of Alu tandems, identification of Period 3 pattern in exons, implementation of 'magnifying glass' effect, identification of complex HOR pattern, identification of inter-tandem transitional dispersed repeat sequences and identification of long segmental duplications. GRM algorithm is convenient for use, in particular, in cases of large repeat units, of highly mutated and/or complex repeats, and of global repeat maps for large genomic sequences (chromosomes and genomes).

INTRODUCTION

Repetitive DNA sequences are of increasing importance because of their regulatory role and the role as one of principal factors for evolutionary development (1–16). Therefore, their identification and analysis are currently of substantial interest.

Eukaryotic genomes are characterized and often dominated by repetitive sequences. More than 50% of the human genome is made up of repeats (17). Different types of repeats in DNA can be classified as tandem repeats, dispersed repeats, equidistant repeat copies separated by spacers, segmental duplications and complex repeats (18–28). According to the length of repeat units, tandem repeats can be classified as microsatellites (1 to ~6 bp), minisatellites (~6 to ~100 bp), satellites (~100 bp to ~2 kb) and macrosatellites (>~2 kb); the thresholds between different classes of satellites may vary in different authors. Repeats are mostly approximate, containing nucleotide substitutions, insertions and/or deletions with respect to consensus.

There is a vast range of algorithms for repeat detection (partial survey of list of references on algorithms from plethora of repeat finding methods is given in Supplementary Table S1). For reviews of repeat algorithms, see for example (29–42). A broad scope of various algorithms reveal both the complexity of challenges posed by this enormous task and, in spite of significant advances achieved so far, there still remains limitations and high potential for improving efficiency.

Some of repeat finding algorithms are designed specifically for identification of tandem repeats or of dispersed repeats, or of both. Of particular interest are *ab initio* programs that in the repeat identification process do not rely upon previously known repeats.

In order to extract the mathematical and statistical information embedded in symbolic DNA sequences, one can use the powerful analysis tools developed in traditional signal processing, such as Fast Fourier transform, wavelet transform and correlation function technique. In that case one must map the symbolic elements into numerical values. Numerous mappings of symbolic DNA into numerical sequences have been proposed, for example (43–52). However, there appears

*To whom correspondence should be addressed. Tel: +385 1 4605663; Fax: +385 1 4680321; Email: matko@phy.hr
Correspondence may also be addressed to Vladimir Paar. Tel: +385 1 4680321; Fax: +385 1 4680321; Email: paar@hazu.hr

a question whether some result is an inherent property of symbolic data or just an artefact of numerical mapping.

Tandem repeat detection algorithms can be broadly assigned to two main categories: flexible statistical string matching algorithms [e.g. (30,32,42,53–60)] and signal processing algorithms [e.g. (45,61–74)]. As pointed out (42), the most often used tandem repeat detection algorithms are Tandem Repeat Finder (TRF) (30) and Spectral Repeat Finder (SRF) (69). The TRF is based on string matching and the use of a probabilistic model of tandem repeats with statistically based recognition criteria (30). Some ideas incorporated in TRF have been used in earlier homology detection program BLAST (75), but the goals and methods differ (30). The signal processing algorithm SRF maps a given genomic sequence into numerical sequence which is analysed by discrete Fourier transform (69).

Various problems related to the use of available repeat finding algorithms have been pointed out and partly investigated, for example (30,34–39,42,76). Some of the relevant results and discussions from these references could be summarized as follows.

The TRF algorithm can have problems with repeats containing sizeable substitutions and/or insertions/deletions, with repeats having very large repeat units (>2 kb), or giving several possible repeat structures of the same sequence. On the other hand, the SRF algorithm can produce numerical artefacts and poor resolution of spectral analyses.

The problem of repeat finding and analysis is very complex and has many facets, primarily because of approximate nature of repeats (different algorithms can exhibit different degree of robustness for various types of repeats) and because of uncertainties introduced by arbitrarily defined threshold (regions detected as repeats are those whose alignment is higher than a given threshold—larger threshold sizes can prevent detection of more tandem repeats). For example, different algorithms applied to a given genomic sequence with higher order structures can sizeably differ in what is detectable repeat structure, as primary repeat or higher order repeat (HOR). In general, significant differences may appear among different algorithms; some approaches identify repeats that are missed by some other approaches; some algorithms can detect more divergent repeats than the other. As for recognition of tandem repeats, it was suggested that the average number of tandem repeats found in a set of random sequences could serve as a value for background noise produced by chance occurrence of tandem repeats. For algorithms with user-defined parameters, the choice of parameter values can influence the results for identified repeats. Also, a major drawback of many computational algorithms, when running against very long sequences, can be that they produce a large amount of cumbersome results, which require a painstaking interpretation. It was pointed out that in addition to the use of better and faster algorithms for repeat finding, the use of combined results from several different algorithms holds promise.

Investigations comparing different repeat finding algorithms have been made for some sequences [e.g. (34–39,76,77)]. For example, testing five commonly used repeat finding programs (TRF, Sputnik, Mreps, RepeatMasker and STAR) across several eukaryotic genomes, it was found major divergence in the repeats detected, depending on the program, and more significantly, depending on parameter setting selected (34). A meta-analysis on published distribution in yeast showed divergence of up to several orders of magnitude in the frequency of microsatellite motifs reported among seven studies. A bias depending on the algorithm employed was found, mainly in number of repeats detected, size classes identified and length distribution (36). Each method has its own limitations, revealing a need for the development of newer methods to overcome more limitations (77).

The increased availability of sequenced genomes as well as the increasing recognition of biological importance of repetitive elements motivates the development of more sensitive and effective algorithms for *ab initio* repeat discovery and characterization. In light of profound variation in performance of currently available *ab initio* repeat finders as well as the situation that each of many available algorithms has some advantages and disadvantages, there remains substantial room for improvement and development in algorithms for detection and characterization of novel repeats. It was pointed out that there are many ways how the computational identification and characterization of repeat sequences could be improved by creation of more efficient, sensitive, selective and faster algorithms, as well as by combination of results from several different algorithms. Taking all these aspects into account, in spite of significant progress made so far, the repeat finding programs are still a challenge.

In previous application of initial restricted version of GRM algorithm (restricted to 100 kb fragment lengths), we identified repeats in human chromosomes 1 and Y (Build assemblies). We reproduced previously known repeats, and in addition we discovered a dozen of novel repeats (78,79). In this article we extend the software of GRM algorithm (new source code www.hazu.hr/grm/tools.html#grm2012) to be able to treat as large genomic sequences as hundreds of megabases and we formulate this model in a general framework incorporating crucial elements from both the digital signal processing and the string matching approaches. As will be presented and discussed in this article, the novelty of GRM approach is a direct mapping of symbolic DNA sequence into frequency domain using complete *K*-string ensemble instead of statistically adjusted individual *K*-strings optimized locally. In this way, we show that the GRM provides a straightforward identification of DNA repeats using frequency domain, but avoiding mapping of symbolic DNA sequence into numerical sequence, and uses *K*-string matching, but avoiding statistical methods of locally optimizing individual *K*-strings (Figure 1). We also present a set of case studies of various types of repeats in human genome in order to demonstrate efficacy and robustness of GRM.

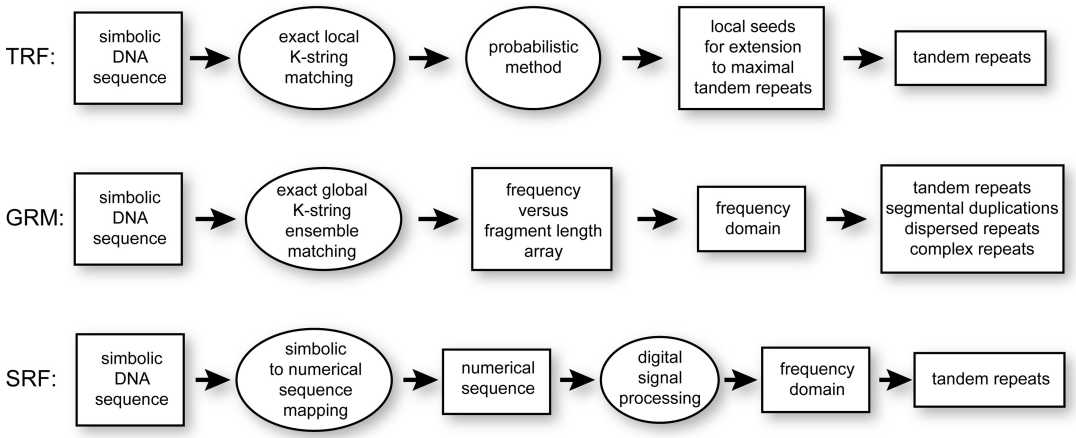

Figure 1. Basic scheme of GRM in comparison to basic schemes of TRF (30) and SRF (69).

MATERIALS AND METHODS

Single *K*-string spacing and frequency arrays for a given genomic sequence

Concerning the encoded information, the DNA sequence can be treated as 1D symbolic sequence made of four letters A, C, G and T, representing the 4 nt. A *K*-string (also called *k*-word, *k*-mer, *k*-tuple, seed or key string) is a sequence of *K* nucleotides, i.e. a symbolic sequence of *K* elements from four-letter alphabet {A, C, G, T} (30,38,39,81–84):

$$s_K(j) = \alpha_1(K,j)\alpha_2(K,j)\ldots\alpha_K(K,j), \quad j = 1, 2, \ldots, 4^K,$$

where $\alpha_i(K,j)$ stands for a nucleotide (A, C, G or T) at the *i*th position in the *j*th *K*-string. There are 4^K different *K*-strings. The set of all *K*-strings $s_K(j)$ will be called the *K*-string ensemble, denoted E_K. The ordering of *K*-strings in E_K is chosen arbitrarily and is of no significance for mapping.

In a given genomic sequence of length *L*, all exact matches are determined separately for each single *K*-string $s_K(j)$, i.e. for each *j* from $j = 1$ to $j = 4^K$ (schematically shown in Figure 2), by sliding the window across the sequence in steps of 1 nt. This starting point of GRM is analogous as in standard string matching approaches (30).

Using for example an ensemble of 4^8 *K*-strings corresponding to $K = 8$, we align with each *K*-string consecutively sub-sequences from Position 1 to 8, from 2 to 9, from 3 to 10, etc., recording for each *K*-string the start position each time when an 8-bp sub-sequence from genomic sequence and the 8-bp *K*-string match.

Denote the start positions of exact matches of each *K*-string $s_K(j)$,(i.e. for a fixed *j*) by:

$$\{x_K(j)\} = [x_K(j)]_1, [x_K(j)]_2, [x_K(j)]_3, \ldots [x_K(j)]_n, [x_K(j)]_{n+1}, \ldots \quad j = 1, 2, \ldots, 4^K.$$

For each fixed *j*, the first match of $s_K(j)$ starts at a position denoted $[x_K(j)]_1$, the second match of $s_K(j)$ at a position denoted $[x_K(j)]_2$, etc. In this way, for each *K*-string we obtain a sequence $\{x_K(j)\}$ of start positions of the corresponding exact matches.

A spacing between start position of the *n*th match and $(n+1)$th match is

$$[d_K(j)]_n = [x_K(j)]_{n+1} - [x_K(j)]_n, \quad n = 1, 2, 3, \ldots$$

While in the standard heuristic approach the distances between *K*-string matches are treated by statistical criteria and are used for estimate of best *K*-string (30), in GRM we use these distances for transformation of symbolic DNA sequence into frequency domain. To this end let us consider an array of all spacings $[d_K(j)]_n$, corresponding to the *j*th *K*-string $s_K(j)$ from *K*-string ensemble E_K:

$$\{d_K(j)\} = [d_K(j)]_1, [d_K(j)]_2, [d_K(j)]_3, \ldots \quad j = 1, 2, \ldots, 4^K.$$

Such spacing array for a single *K*-string will be referred to as a single-*K*-string spacing array [called key string distance array in (78,79,83,85,86)].

An example of a single-*K*-string spacing array for three-string GGC and an illustrative genomic sequence is shown in Table 1 and Figure 2.

Each spacing $[d_K(j)]_n$ in a single-*K*-string array $\{d_K(j)\}$ for a fixed *j* is equal to an integer number of base pairs. Let us denote by $[f_K(j)]^{(1)}$ the number (i.e. frequency) of appearance of all single-*K*-string spacings $[d_K(j)]_n$ that are equal to 1 bp; by $[f_K(j)]^{(2)}$ the frequency of all single-*K*-string spacings $[d_K(j)]_n$ equal to 2 bp, ... by $[f_K(j)]^{(v)}$ the frequency of all single-*K*-string spacings $[d_K(j)]_n$ equal to *v* bp, ... In this way, for each spacing length we obtain the corresponding frequency of appearance by counting how many times this spacing length appears in the spacing length sequence. A single-*K*-string frequency array is constructed for each of the 4^K *K*-strings:

$$\{f_K(j)\} = [f_K(j)]^{(1)}, [f_K(j)]^{(2)}, \ldots [f_K(j)]^{(v)}, \ldots \quad j = 1,2, \ldots, 4^K.$$

K-string ensemble frequency array for a given genomic sequence

In the next step we superpose all 4^Ksingle-*K*-string frequency arrays of all *K*-strings from the *K*-string

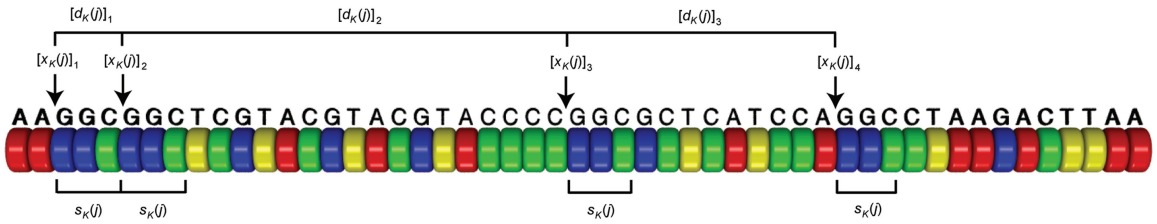

Figure 2. Illustration of exact matches of a single *K*-string and the corresponding spacings. The single *K*-string GGC is assigned to $j = 42$ within the three-string ensemble $\{s_K(j), j = 1,2,3,...64\}$ of all 4^3 possible *3*-strings of the length 3: $s_K(j) = s_3(42) = GGC$.

Table 1. Match positions and spacings corresponding to schematic genomic sequence from Figure 2 using single-three-string s_3 = GGC

$[x_3(42)]_1 = 3$	$[d_3(42)]_1 = [x_3(42)]_2 - [x_3(42)]_1 = 6 - 3 = 3$
$[x_3(42)]_2 = 6$	$[d_3(42)]_2 = [x_3(42)]_3 - [x_3(42)]_2 = 26 - 6 = 20$
$[x_3(42)]_3 = 26$	$[d_3(42)]_3 = [x_3(42)]_4 - [x_3(42)]_3 = 38 - 26 = 12$
$[x_3(42)]_4 = 38$	
	$\{d_K(42)\} = 3,20,12$

ensemble. In this way we obtain the *K*-string ensemble frequency array of a given genomic sequence:

$$\{f_{K(E)}\} = \sum_{j=1}^{N}[f_K(j)]^{(1)},\ \sum_{j=1}^{N}[f_K(j)]^{(2)},\ \ldots \sum_{j=1}^{N}[f_K(j)]^{(\nu)},$$

where $N = 4^K$.

This frequency array, computed for a given genomic sequence using the *K*-string ensemble E_K, provides a discrete frequency spectrum versus spacing length *m*:

$\sum_{j=1}^{N}[f_K(j)]^{(1)}$is the frequency of superposed spacings of length 1 bp,
$\sum_{j=1}^{N}[f_K(j)]^{(2)}$is the frequency of superposed spacings of length 2 bp,
$\sum_{j=1}^{N}[f_K(j)]^{(\nu)}$is the frequency of superposed spacings of length ν bp.

In summary regarding frequency array, each *K*-string from E_K ensemble is moving in steps of 1 nt along the genomic sequence. Thus, we align successively each *K*-string to sub-sequences from Position 1 to *K*, from 2 to $K+1$, from 3 to $K+2$,... in a given genomic sequence, recording start position each time when a *K*-bp sub-sequence and a *K*-string fully match. In this way, for each *K*-string from the E_K ensemble we obtain a sequence of key string start positions within genomic sequence. In this way the genomic sequence is for each *K*-string separately segmented into fragments. Each fragment is equal to spacing between start positions of the two neighbouring matches of the same *K*-string. Such spacing is referred to as fragment length. For each fragment length and each *K*-string, the corresponding frequency of appearance is determined by counting how many times this fragment length appears in the fragment length sequence along a given genomic sequence. Superposing frequencies of appearance of fragment lengths for all *K*-strings in the ensemble, we obtain the frequencies of fragment lengths 1 bp, 2 bp, 3 bp,... This set of frequencies represents the map of genomic sequence in the frequency domain.

Global repeat map diagram in frequency domain as repeat finder

Using *K*-string ensemble E_K, we obtain the mapping $\{f_{K(E)}\}$ of a given symbolic sequence (genomic sequence) into frequency domain. Diagrammatic presentation of this frequency dependence is referred to as global repeat map (GRM).

Using the *K*-string ensemble approach in GRM, we compute efficiently the frequency map of genomic sequences as large as human chromosomes using our computer code grm2012 with $K = 8$, which gives results corresponding to the previous GRM computer program from (78,79) that was previously not fully automatized and restricted to maximum fragment lengths of 100 kb, i.e. the maximum distance between start positions of neighbouring repeat copies of 100 kb.

Here we develop the computer program grm2012 which is highly automatized and extended to fragment lengths as large as hundreds of mega base pairs, which is of particular importance for identification of segmental duplications. (Initial version of GRM (78,79) (source code grm2011) was restricted to fragment lengths up to 100 kb.)

With regard to biological significance of frequency peaks in GRM diagram, from the construction method it is apparent that the corresponding fragment length (spacing length) at position of each frequency peak is equal to a distance between starts of the neighbouring repeat copies. We find three possible GRM situations:

(1) In the case of dispersed repeats, the fragment length is equal to a distance between start positions of dispersed repeat copies. Such repeats that can be identified using GRM algorithm are, for example, LINEs, SINEs and segmental duplications. GRM can quickly detect segmental duplications at distances as large as hundreds of mega base pairs. We also found that GRM detects each Alu copy, thanks to the internal structure of Alu, which consists of two approximately homologous sub-sequences separated by a short spacing, and also detects each tandem of Alus.
(2) In the case of tandem repeats, a fragment length obviously corresponds to the repeat unit length, equal

to distance between starts of neighbouring repeat copies. In this way, the GRM algorithm provides a very efficient tool to detect tandem repeats and HORs.

(3) In the case of complex repeats, such as intertwined repeats, or dispersed repeats with regular spacings and/or sizeably distorted pattern, several fragment lengths can form a complex geometrical pattern in the corresponding GRM diagram, which reveals the underlying repeat pattern.

Identification and analysis of repeats corresponding to each GRM peak

Using the computer code grm2012 in the first step ('identification step') selecting significant pronounced peaks, we identify significant GRM fragment lengths (i.e. repeat unit lengths of tandem repeats and/or spacings between dispersed repeat copies). In the second step ('analysis step'), for each significant fragment length (corresponding to a GRM peak), using grm2012 code we determine the corresponding repeat sequences and its positions, the consensus repeat unit and divergence of repeat copies with respect to each other and to consensus. The basic point is the use of dominant *K*-string associated with a GRM peak (fragment length). The dominant *K*-string for a particular fragment length m is the one $s_K(j)$ among the 4^K *K*-strings in the E_K ensemble with highest frequency for this fragment length: $max\left(\left[f_K(j)^{(m)}\right]\right)$. In computer program grm2012 it is selected automatically from the stored set of frequencies $\left\{f_{K(j)}\right\}$ for each fragment length in the first step of GRM peak identification. To each GRM peak its particular dominant *K*-string corresponds, determined automatically in this way.

RESULTS AND DISCUSSION

Case study: GRM diagram for human chromosome 7 and identification of α-satellite tandem repeats and HORs

α-Satellite DNA, which consists of tandem repetitions of ~171 bp repeat unit (called a monomer), is the major constituent of primate centromeres. In humans, a large fraction of α-satellite monomers is arranged into HORs) Individual human α-satellite monomers mutually diverge by 20–40%, while the sequence divergence between HOR copies is <5% and often even <2% (19,59,81,85–92). As an illustration of GRM identification of α-satellite monomers and HORs, we present the GRM diagram for human chromosome 7 using the 8-bp *K*-string ensemble (Figure 3). A strong peak at fragment length ~171 bp corresponds to α-satellite monomer repeat unit of ~171 bp. Peaks at its multiples (~2 × 171 bp, ~3 × 171 bp, ~4 × 171 bp, ...), decreasing with increasing multiple order, correspond to tandem repeats of ~171 bp α-satellite monomers. In addition to this multiple pattern, there is a strong peak at 2734 bp, corresponding to consensus HOR length. This peak reveals higher order structure of α-satellite organization: 16 (2734 bp/171 bp ≈ 16) tandemly arranged α-satellite monomers, which mutually diverge by 20–40%, are arranged into more homogenous second-order units. The high homogeneity of second order units, as well as relatively high heterogeneity of primary repeat units, is reflected in GRM diagram with a characteristic pattern of HOR-signature.

Case study: identification of Alu dispersed repeats and 2-Alu tandems in human chromosome 7

The most abundant dispersed repeats in humans and other primates are Alu elements. The human genome contains >1.1 million dispersed Alu elements. Alu elements have the highest copy number of all of the human mobile elements. They occur at an average of one every 3–6 kb, but distribution within the human genome is not uniform (9).The structure of each Alu element is bi-partite: it is a dimer of two approximately similar monomers; the 3′-monomer contains an additional 31-bp A-rich insertion relative to the similar 5′-monomer (Figure 4a) (9,92). Monomers are separated by a middle A-rich region that contains the sequence A_5TACA_6. Thus the body of an Alu element is ~282 bp long. Due to similarity between two monomers constituting an Alu element, the distance between the starts of parts M1 and M1′ is 135 bp, which gives rise to the GRM peak at 135 bp, while the distance between the starts of parts M2 and M2′ is 135 bp + length of insertion I = 166 bp, which gives rise to the GRM peak at 166 bp (Figure 5a and b). In this way, the two GRM peaks, at 135 and at 166 bp, are signature of dispersed Alus (Figure 5b). Each Alu element in the genomic sequence under study contributes to these peaks. Thus, the height of peaks at 135 and 166 bp increases with increasing number of Alu copies in genomic sequence. From the scheme in Figure 4a, it is seen that the incomplete Alu fragments <135 bp cannot be presented as GRM peaks, because the shortest distance between two similar Alu sub-segments is >135 bp.

The basic Alu dimer is flanked by a poly-A tail at its 3′-end. This poly-A tail can be a perfect A repeat of variable lengths, from a few bases up to a hundred bases long, occasionally interspersed with other bases (93,94). Sequence composition of *Alu* poly-A tail constructs contains, for example, A10, A20, A30, A40, A50, ... The total length of each Alu-sequence is a sum of a basic *Alu* dimer of ~282 bp and the widely scattered length D_A of the 3′ poly-A tail (Figure 4a). This sum is referred to as ~300 bp.

An additional GRM peak appears in the case when two Alu elements form a tandem in a genomic sequence (Figure 5a and c). The distance between the starts of the first and of the second Alu element is $282 + D_A(1)$, where $D_A(1)$ denotes the length of poly-A tail in the first Alu element (on the left hand side). As the length of A-tail differs in a wide range among Alus, for many different pairs of Alus within a chromosome the distance between the two Alu elements within a tandem can differ by dozens of base pairs and therefore the GRM peak corresponding to a 2-Alu tandem for a whole human chromosome 7 sequence is broadened over dozens of base pairs above the 282-bp fragment length (Figure 5a, and its magnified segment, Figure 6). A broad GRM peak is seen in the interval of A-tail lengths from ~4 to ~50 bp (towards still

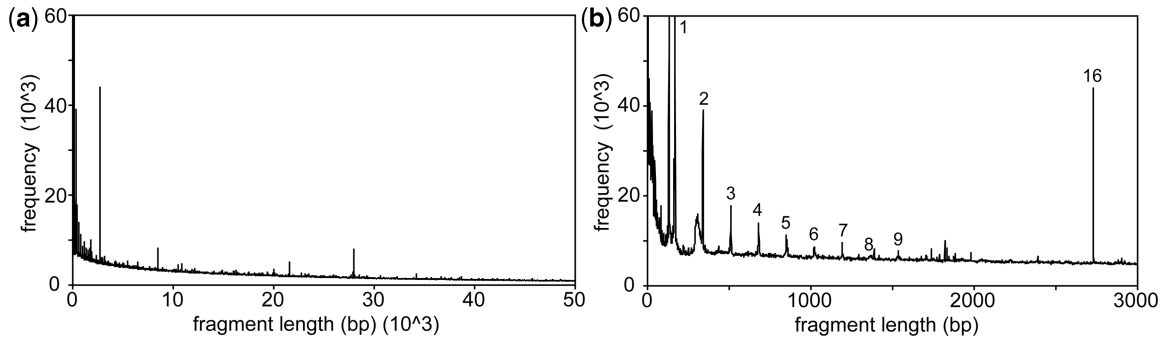

Figure 3. GRM diagram for human chromosome 7 (Build 37.3 assembly). (**a**) fragment lengths up to 50 kb, (**b**) fragment lengths up to 3 kb. In magnified section of GRM diagram (b) (interval 0–3000 bp fragment lengths) the peaks corresponding to fragment lengths at approximate multiples $n \times 171$ bp of α-satellite primary repeat unit are denoted by integers n. Pronounced peak at $n = 16$ corresponds to HOR consensus length of 2734 bp.

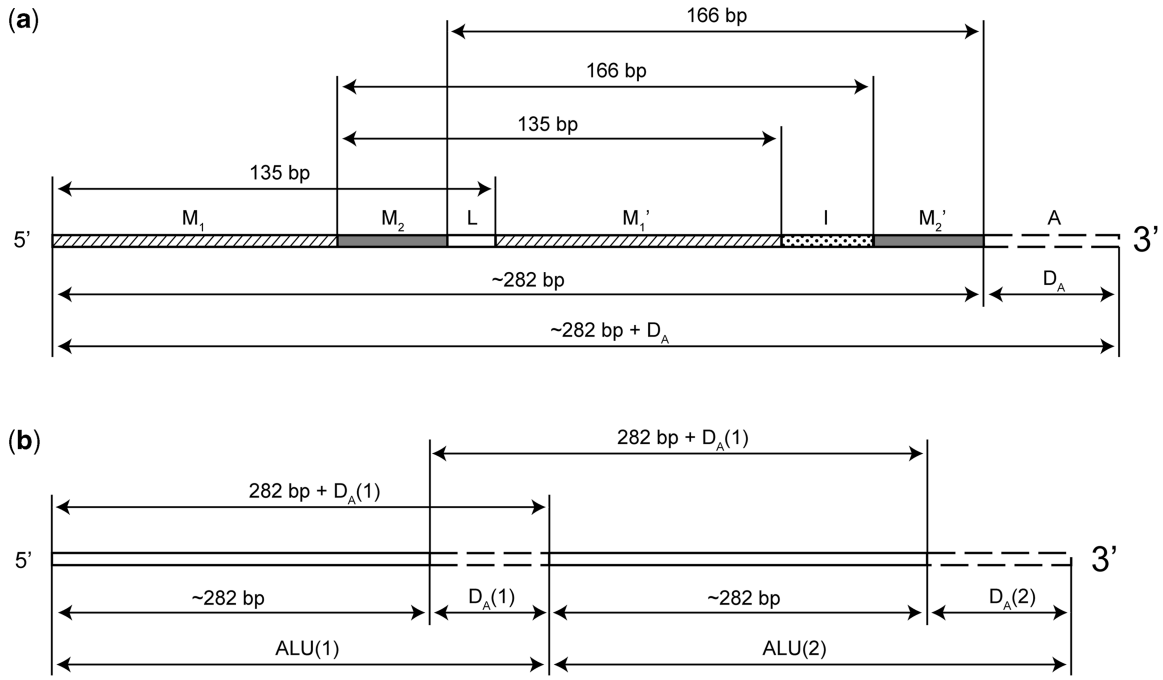

Figure 4. Schematic presentation of the origin of GRM peaks corresponding to dispersed Alu elements and to tandems of Alu elements. (**a**) Origin of the 135 bp and 166 bp GRM peaks from each Alu element. The left constituent monomer of Alu is divided into two segments, M1 and M2. The segments denoted M1′ and M2′ in the right monomer are similar to the segments M1 and M2 in the left monomer, respectively. In between the segments M1′ and M2′ in the right monomer there is a 31-bp insertion, denoted I. Between the left and right monomer there is a short A-rich linker denoted L. (**b**) Origin of the ~300-bp GRM peak from each tandem of two Alu elements. The first Alu element, denoted Alu(1), consists of the ~282-bp main part (two constituent monomers and a linker) and of A-tail. The lengths of A-tail in the first and second Alu element are denoted $D_A(1)$ and $D_A(2)$, respectively. The distance between the starts of the first and second Alu element is $282 + D_A(1)$. As the length of tail differs in a wide interval of several dozens of bp, the distance between the two Alu elements in different dimers can differ by dozens of bp and therefore the peak corresponding to a two-Alu tandem is broad.

higher fragment lengths the tail of this peak is shadowed by the emerging α-satellite peak $\sim 2 \times 171$ bp = 340 bp). The maximum height of this A-tail peak is at $D_A = 28$ bp (frequency ~16 k), with an estimated half-width of ~50 bp (frequency ~8 k) extending in the interval of fragment lengths from ~285 to ~335 bp. The width of fragment length interval exceeding three-fourth of maximum frequency, i.e. $0.75 \times 16 = 12$ bp is ~ 24 bp, from $D_A \sim 15$ bp to $D_A \sim 39$ bp (Figure 6). This GRM analysis for A-tail length distribution for the whole human chromosome 7 is in accordance with direct results obtained for a selective sample of Alus from chromosome 7: analysed Alu old subfamilies S, J and young Ya5 had a distribution of A-tail lengths with a mean size of 21 and 26 bp, respectively (21 ± 8 and 26 ± 9, respectively). For Alu J, S: minimum length 1 bp, maximum length 43 bp; for Alu Ya5: minimum length 2 bp, maximum length 59 bp (93).

Case study: period 3 distinguishing human coding and non-coding regions

The 3-nt periodicity in coding sequences was evidenced as a sharp peak at frequency $f = 1/3$ in the corresponding

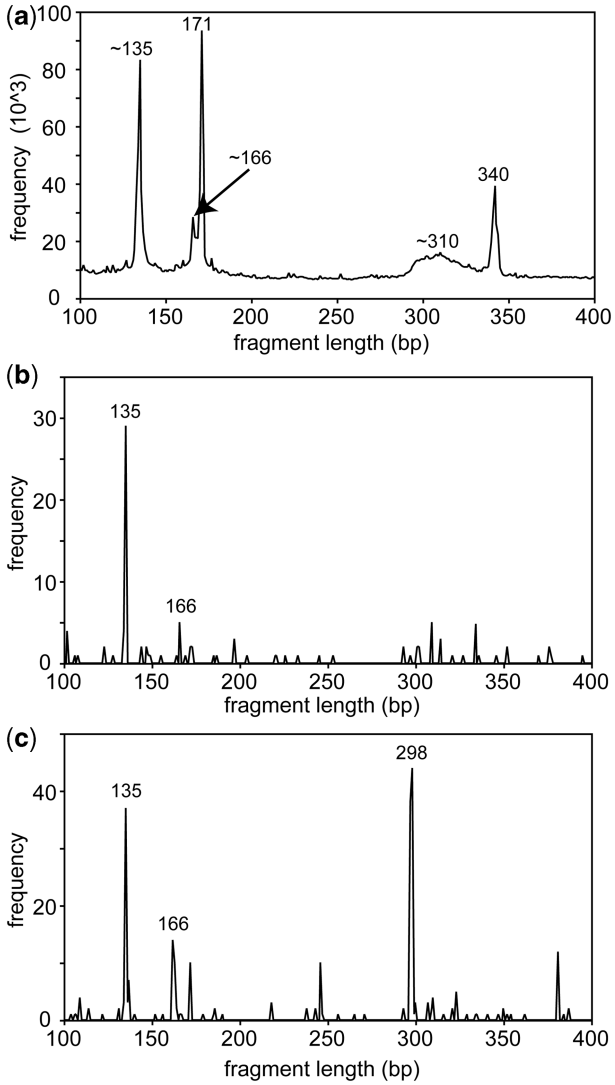

Figure 5. GRM peaks corresponding to single Alus and tandems of two Alus in human chromosome 7 in the interval of fragment lengths from 100 to 400 bp. (**a**) GRM diagram is computed for the whole sequence of human chromosome 7 from Build 37.3 assembly. In this fragment length interval we see peaks at ~135 bp and ~166 bp corresponding to all dispersed single Alu copies in chromosome 7. A broad peak at ~310 bp corresponds to all tandems consisting of two Alu copies. Pronounced peaks at 171 and 340 bp correspond to the length of α-satellite primary repeat unit and twice the length of primary unit due to HOR structure. (**b**) GRM diagram computed for a 5-kb segment of chromosome 7 (interval from position 2 765 000 to 2 770 000 in contig NT_007741.14) encompassing one Alu element. In the interval from 100 to 400 bp only two significant peaks appear, at ~135 bp and ~166 bp, corresponding to the Alu element. (**c**) GRM diagram computed for a 5-kb segment of chromosome 7 (interval from position 2 620 000 to 2 625 000 in contig NT_007741.14) encompassing one tandem of two Alu elements. In the interval from 100 to 400 bp only three significant peaks appear, at ~135 bp and ~166 bp, corresponding to each Alu element separately, and a narrow peak at ~298 bp, corresponding to the tandem of two Alu elements.

Fourier spectrum, while in non-coding sequences such peak is missing (45,61,63,69,95). Here we investigate GRM results for the case study of coding sequences, in comparison to non-coding sequences (Figure 7). As we are interested in repeat units of 3 bp, the best GRM resolution

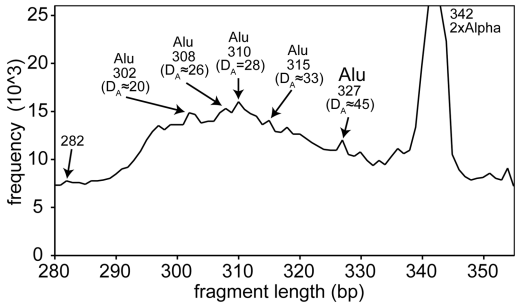

Figure 6. Magnified segment from Figure 5a containing interval of fragment lengths with GRM peak of two-Alu tandems. Maximum of the peak is at 310 bp. Therefrom we determine the corresponding A-tail length as 310 – 282 bp = 28 bp. Peaks corresponding to the A-tail lengths 26 bp and 45 bp are also shown. To the right of this broad peak, above the fragment length of ~332 bp (i.e. above the A-tail length of ~52 bp), the broad peak is shadowed by pronounced GRM peak corresponding to to twice the α-satellite monomer length with maximum at 340 bp.

is obtained for the string length $K = 3$. (The results for $K = 4$ are quite similar.) From GRM diagrams we find pronounced GRM peaks at 3 bp and their multiples in most exons, while in non-coding sequences (i.e. in ~99% of human genome), the peak at 3-bp period is much smaller or absent. Thus, in GRM diagrams for whole human chromosomes (where non-coding sequences strongly dominate) the frequency corresponding to Period 3 is about three times smaller than the frequency corresponding to Period 4, while in the coding sequences it is the opposite. In non-coding regions, a pronounced GRM periodicity is mostly 6. These GRM results are in accordance with the pattern of previous results obtained by using heuristic algorithms (96). We note that in digital signal processing, such as Fourier transform, the pronounced Period 3 peak appears due to hidden periodicity as defined in (64), while in GRM the equidistant peaks at periods $3n$ ($n = 1, 2, 3, \ldots$) (Period 3 and their multiples) are over-represented due to specific equidistant arrangements of three strings.

'Magnifying glass' for GRM

Significant repeats in a given genomic sequence are easily recognizable as pronounced peaks in GRM diagrams. For less pronounced repeats that can be submerged in the background of noise, we can use computational 'magnifying glasses' for their identification, thus increasing the resolution power of the method.

In GRM algorithm, the first 'magnifying glass' consists simply in magnifying segments of fragment lengths in GRM diagram, but without performing any additional GRM computation for genomic sequence.

The second 'magnifying glass' consists in performing additional GRM computations for smaller sub-segments of genomic sequence. The accuracy of algorithm is improved if we divide very long DNA sequences into smaller sub-regions, which decreases the background of noise and thus increases the accuracy of predictions, i.e. smaller GRM peaks become visible above the background of noise. For example, if a DNA sequence is 10 Mb, we

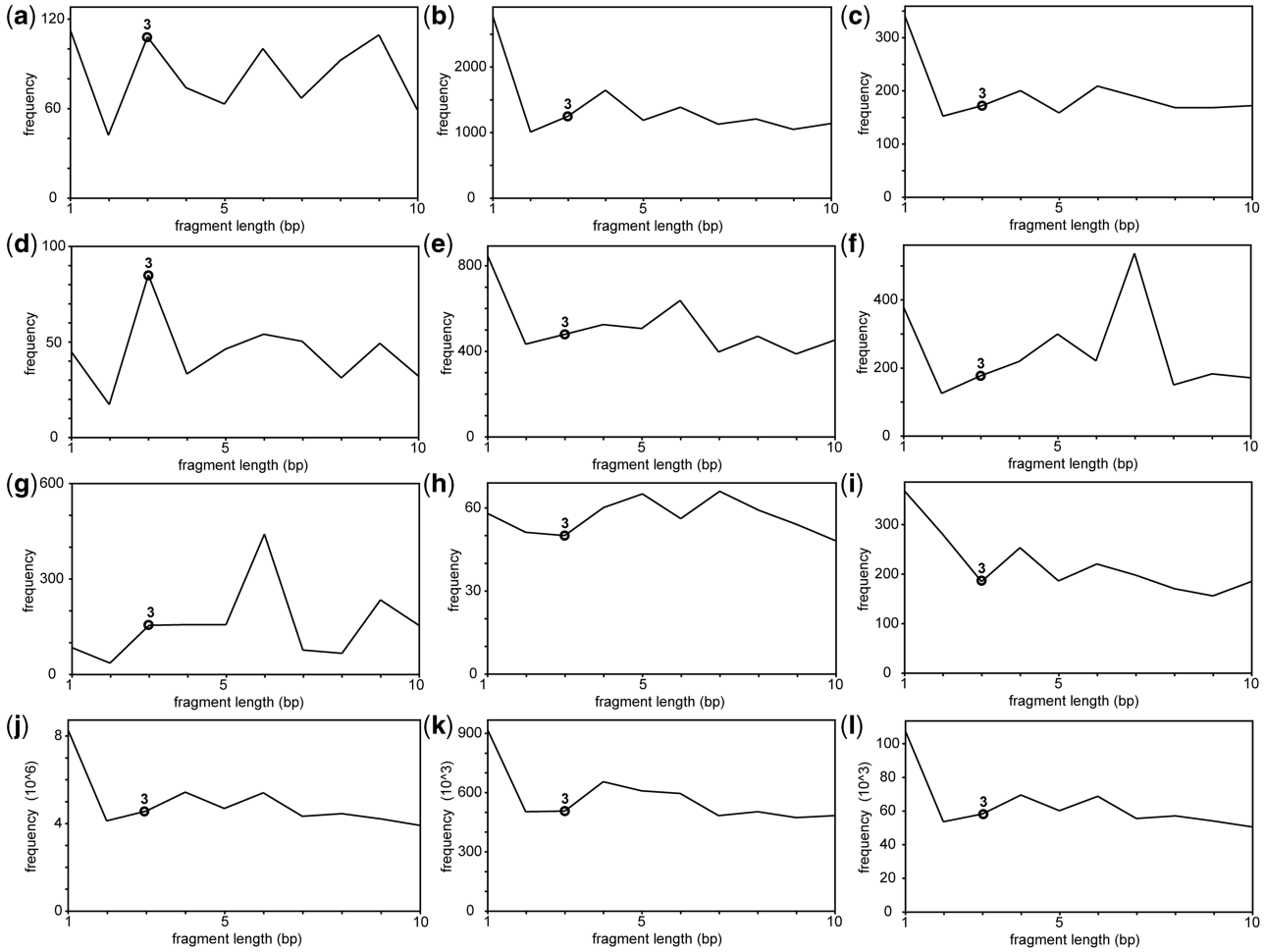

Figure 7. Illustration of low fragment length segment (1–10 bp) of GRM diagrams for some human exons, introns, chromosomes and human genome (Build assembly). GRM diagrams at $K = 3$ are shown for: (**a**) exons from DAZ1 gene in Y chromosome, (**b**) introns from DAZ1 gene in Y chromosome, (**c**) 9 kb intergenic segment starting at position 1 kb distant from DAZ1 gene, (**d**) exons from Tubulin Tyrosine Ligase 10 gene in chromosome 1, (**e**) introns from Tubulin Tyrosine Ligase 10 gene in chromosome 1, (**f**) 9 kb intergenic segment starting at position 1-kb distant from Tubulin Tyrosine Ligase 10 gene in chromosome 1, (**g**) exons from hornerin gene in chromosome 1, (**h**) introns from hornerin gene in chromosome 1, (**i**) 9-kb intergenic segment starting at position 1-kb distant from hornerin gene in chromosome 1, (**j**) chromosome 1, (**k**) chromosome Y and (**l**) whole human genome.

can improve the accuracy of repeat identification by dividing it in sub-sequences of 1 Mb. In this way we exclude from GRM diagram for each sub-segment the peaks due to repeats lying outside of this sub-segment and in this way we reduce the density of peaks, making peaks corresponding to repeats in this sub-segment more visible. This procedure is implemented into current source code (grm2012). Identified repeats, which are stored in associated *.txt file, are result of this 'magnifying glass' technique.

The third 'magnifying glass' consists of performing additional GRM computation for randomized sequence of the same nucleotide composition as in the real sequence and in discarding all GRM peaks that lie below the threshold defined by heights of GRM peaks for randomized sequence.

We find that the background in GRM diagrams is mostly small, thus allowing identification of most significant GRM peaks (i.e. repeats) without the need for increase of resolution.

We note that the general problem of 'magnifying glass' was pointed out previously in another framework, as for example in the case of mreps algorithm with a resolution parameter playing a role of 'magnifying glass' (57). In the case of REPuter algorithm the user is also able to 'zoom in' on the details of particular repetitive regions (97).

Case study: reduction of background noise for human chromosome Y

As a case study for reduction of random background noise from counting results, we compute the GRM diagram for human chromosome Y (Build 37.3 assembly). The GRM diagram computed for fragment lengths up to 25 kb is shown in Figure 8a. For comparison, we compute GRM diagrams for a set of 100 randomized sequences of the same nucleotide composition as in human chromosome Y and their mean GRM diagram (Figure 8b and c). This mean of GRM diagram of randomized sequences is used as a background noise produced by chance occurrence. For every fragment length from GRM diagram of

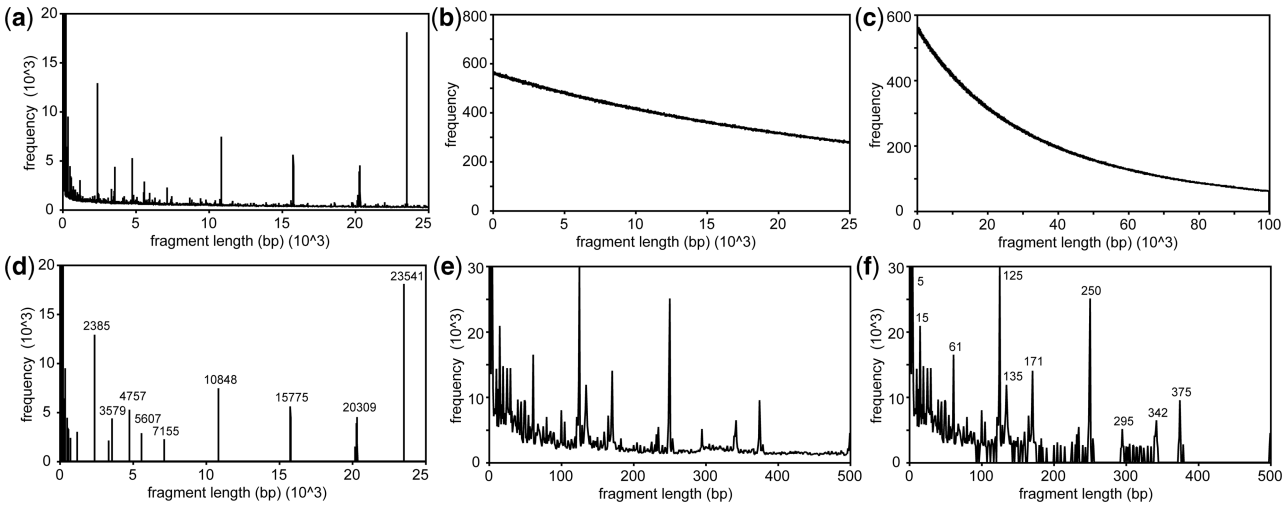

Figure 8. (**a**) GRM diagram for human chromosome Y with fragment lengths up to 25 kb. Some of pronounced peaks were discussed in (78). (**b**) GRM diagram of mean of one hundred of randomized sequences of the same nucleotide composition and fragment length interval as for Y chromosome in (a). (**c**) Mean randomized GRM diagram from (b) extended to a larger fragment length interval (up to 100 kb). (**d**) GRM diagram from (a) with reduced background. Frequencies lying below the threshold defined by the quadruple mean randomized background from (b) are set to null. (**e**) Magnified section of GRM diagram from (a) in the fragment length interval up to 500 bp. (**f**) Magnified section of GRM diagram with reduced background from (d) in fragment length interval up to 500 bp.

human chromosome Y, we discard each frequency that is smaller than quadruple frequency of mean randomized GRM diagram at that fragment length, by assigning the value 0 to that frequency. Using this procedure we obtain the GRM diagram with reduced background (Figure 8d) and the peaks in this diagram are taken as significant. Some magnified sections of GRM diagrams are presented in Figure 8e and f. Increased mean GRM diagram (Figure 8c) reveals exponential correlation between background noise frequency and fragment length. Correlation between frequency and small fragment lengths in regular GRM diagram (magnified section, Figure 8e) is also exponential. However, for small fragment lengths, the rate of exponential decay in mean randomized GRM diagram is much smaller than in regular GRM diagram, and as a consequence, all small fragment length frequencies in regular GRM diagram survive quadruple threshold test (Figure 8f). This is, on the other hand, a consequence of intrinsic characteristic of DNA sequence; the shorter a sequence, the mechanism of its multiplication is more efficient.

Case study: increasing resolution by GRM computations for smaller sub-segments

As pointed out, the resolution power of GRM computation depends on the length of section of genomic sequence computed in a single run. Fragmenting genomic sequence into smaller sub-sections for GRM computations, the level of background noise is rapidly decreasing. Already for genomic segments of as much as 1 Mb the level of noise in computing GRM diagram is sufficiently low so that we can clearly identify GRM peaks corresponding significant repeats, even to those substantially mutated and/or of low copy number. We find that the level of noise in computed GRM diagrams is much lower than in spectra obtained by Fast Fourier Transform.

As a case study for sensitivity of resolution power on length of genomic sequence in GRM computation, we present the GRM diagram computed for a sequence of a complete long contig NT_032977.9 (length 91.3 Mb) from human chromosome 1 (Build 37.3) (Figure 9a). As seen, in that case the resolution power is not sufficiently high in GRM calculation to recognize a significant peak at fragment length 902 bp. However, by separating this contig into smaller segments of 10 Mb and performing GRM computation of each of them, we identify a significant GRM peak at the fragment length of 902 bp (Figure 9b). This GRM peak becomes even more pronounced by further reducing the length of genomic sequence for GRM computation to 1 Mb (Figure 9c).

GRM identification of complex repeats

Case study: complex HOR pattern based on ~2.4 kb monomer in human Y chromosome

For human chromosome Y (Build 37.3) we compute the GRM diagram in the interval of fragment lengths from 2 to 8 kb (Figure 10a). In that fragment length interval the seven most pronounced GRM peaks are (in order of decreasing frequencies) at 2385, 4757, 3579, 5607 and 7155 bp, respectively. Three of these peaks are nearly equidistant approximate multiples of the basic length l_1:

2385 bp = l_1, 4757 $\approx$ 2 l_1, and 7155 $\approx$ 3 l_1.

This indicates that the lengths of these three peaks correspond to a tandem with basic repeat unit ~2.4 kb, and to the corresponding 2-mer and 3-mer HORs (78). This straightforward prediction of GRM diagram is investigated here as a detailed case study.

According to the general outlay of the GRM method, first we determine which contig gives the main contribution to these three GRM peaks; we find that it is NT_011903.12. Its GRM diagram in the interval from 2 to 8 kb fragment length is displayed in Figure 10b. In this

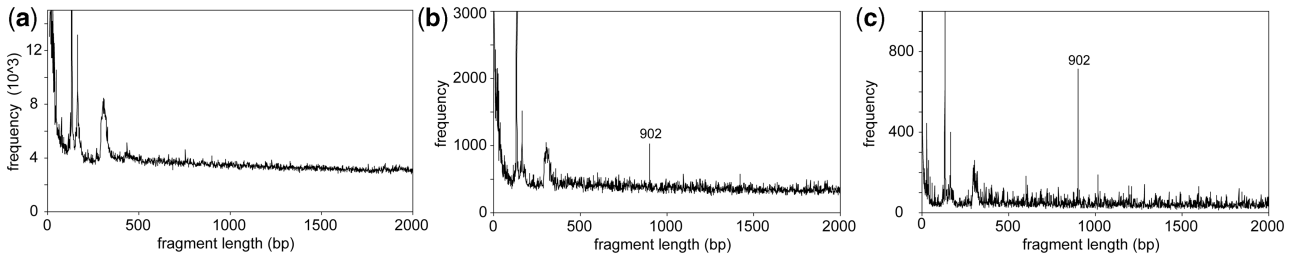

Figure 9. GRM diagram for long contig NT_032977.9 (length 91.3 Mb) from human chromosome 1 computed with increasing resolution. GRM diagram for: (**a**) the whole contig NT_032977.9; (**b**) a 10 Mb sub-segment of NT_032977.9 containing location of tandem array based on the 902-bp repeat unit and (**c**) 1 Mb sub-segment encompassing the location of tandem array based on the 902-bp repeat unit.

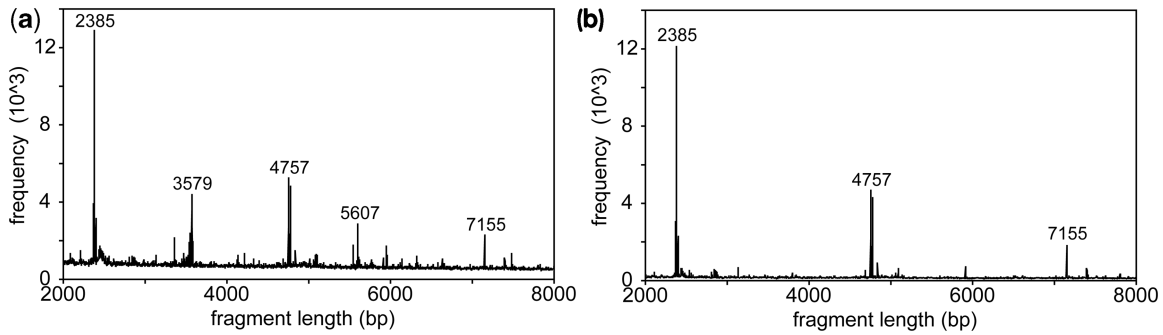

Figure 10. GRM diagram in the fragment length interval from 2 to 8 kb for: (**a**) human chromosome Y (Build 37.3 assembly) and (**b**) contig NT_011903.12 from human chromosome Y (Build 37.3 assembly). Fragment lengths given correspond to the most pronounced GRM peaks.

GRM diagram there are only three significant peaks in the interval from 2 to 8 kb, at 2385 bp, 4757 bp and 7155 bp, corresponding to the equidistant lengths based on ~2.4 kb monomer. The dominant key string for the 2385-bp fragment length is TATTTTTA. The corresponding GRM peaks are associated with monomers of close-lying lengths 2385 bp, 2384 bp, 2376 bp, 2377 bp, 2380 bp, 2404 bp etc., centred at 2385 bp (monomer of highest frequency). On the basis of mutual divergence these monomers can be classified into five different families, taking the 3% upper limit of divergence between monomers of the same family. For each family the corresponding consensus length is determined. These five monomer families are denoted as m01 (2446 bp), m02 (2404 bp), m03 (2376 bp), m04 (2375 bp) and m05 (2385 bp) (consensus lengths in parentheses). The monomer family with highest frequency of appearance is m05. For each of five monomer families we determine the corresponding consensus sequence. Consensus sequences of different families differ mutually by ~13%.

In order to demonstrate that the ~2.4-kb monomer sequences are primary repeat units and that they are without any hidden internal repeat structure, we compute their GRM diagrams showing that they are without any peak. As an illustration, in Figure 11a we present the GRM diagram for 2404-bp monomer which has no peaks.

Using BLAST (75) we have checked that there are no other sub-sequences in NT_011903.12 which are homologous (within the 3% divergence limit) to consensus sequences of five monomer types m01–m05. The scheme of all monomer copies in NT_011903.12 having divergence with respect to consensus <3% is displayed in Table 2 and Figure 12. As seen, these monomers are organized into six tandem arrays. The tandem arrays III and VI are formed solely from m05 monomers. These are highly homologous tandem arrays and therefore give the main contribution to frequency of the 2385 bp GRM peak. Thus, the m05 monomers form a separate family of monomeric type. Tandem arrays I and IV contain reverse complement monomer sequences. Tandem arrays II contain two copies of 2-mer and two copies of 3-mer HORs. Each of tandem arrays IV and V contains three copies of 2-mers, and tandem array I contains two copies of 2-mers. The 2-mer copies are obtained by deleting one monomer, m02 or m04, from the 3-mer m02m03m04 HOR. Each of tandem arrays of direct orientation containing HOR copies starts with m01 and ends with m04, and for reverse complement the orientation is in reverse order.

Let us now investigate the GRM peak at 7155 bp from Figure 10b. We find computationally that the dominant *K*-string for fragment length 7155 bp is GTAAATTT. Performing *K*-string segmentation we obtain three dispersed 7155-bp copies, with start positions within the NT_011903.12 contig at ~1.02, ~1.44 and ~2.65 Mb, respectively. For each of these three genomic sequences we compute the corresponding GRM diagrams. The first and third 7155-bp genomic sequences have no pronounced GRM peaks, i.e. they have no internal repeat structure. On the other hand, the GRM diagram for 7155-bp genomic sequence at the position ~1.44 Mb is characterized by two pronounced GRM peaks, at ~2.4 and at $\sim 2 \times 2.4$ kb, which are centred at 2404 and

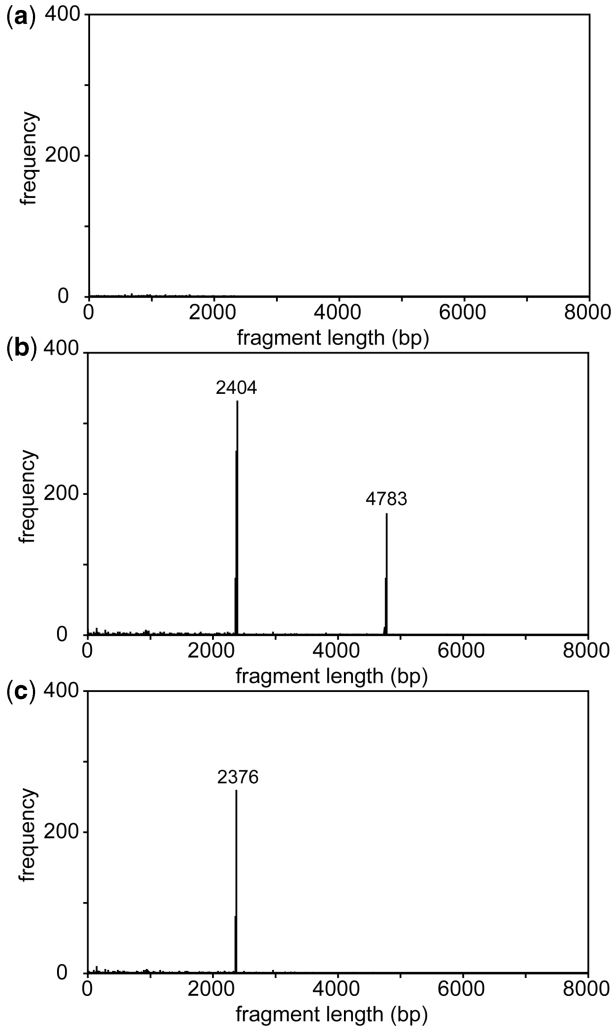

Figure 11. GRM diagrams for human chromosome Y showing internal repeat substructure of: (**a**) 2404-bp monomer consensus sequence, (**b**) 7155-bp sub-sequence at position ~1.44 Mb in NT_011903.12, revealing a 3-mer HOR consensus structure and (**c**) 4757 bp consensus length revealing a 2-mer HOR consensus structure.

4783 bp, respectively (Figure 11b). Applying *K*-string segmentation to the contig NT_011903.12 using AAATATT T *K*-string for which the 2404-bp fragments have the largest frequency, we find at the position of 7155 bp a 3-mer HOR copy m02 m03 m04. Hence, the 7155-bp genomic sequence can be presented as a tandem of three monomer consensus sequences, m02 m03 m04 (average divergence of only 0.2%).

Let us comment on the origin of 7155-bp frequency peak in the GRM diagram of NT_011903.12. In tandem array of monomers at the position of two 7155-bp copies we find the monomer array m02 m03 m04 m02 m03 m04. Divergence between any two monomers from different monomer families is ~11%, while the average divergence between monomers of the same family is <1%. Therefore, the fragments generated by *K*-string segmentation extend largely from each nucleotide position in the first m02 monomer to the corresponding nucleotide position in the second m02 monomer, and analogously from the first m03

Table 2. Tandem arrays of ~2.4 kb monomers in NT_011903 classified into monomer families m01, m02. m03, m04, m05 and HORs

Monomer type	Length (bp)	Start	Orientation	Div (%)	HOR	Tandem No.
m04	2375	1 346 649	R	0.08		
m03	2376	1 349 023	R	0.21		
m02	2406	1 351 399	R	1.04	m03m02	
m03	2376	1 353 802	R	0.13		
m02	2404	1 356 178	R	0.75	m03m02	
m01	2447	1 358 582	R	0.12		I
m01	2446	1 425 263	D	0.00		
m02	2404	1 427 709	D	0.67		
m03	2376	1 430 113	D	0.13	m02m03	
m02	2406	1 432 489	D	0.87		
m03	2376	1 434 891	D	0.08	m02m03	
m02	2407	1 437 267	D	0.96		
m03	2376	1 439 671	D	0.08		
m04	2376	1 442 047	D	0.13	m02m03m04	
m02	2404	1 444 422	D	0.62		
m03	2376	1 446 826	D	0.08		
m04	2376	1 449 202	D	0.04	m02m03m04	II
m05	2385	1 458 984	D	0.21		
m05	2385	1 461 368	D	0.08		
m05	2385	1 463 753	D	0.21		
m05	2385	1 466 137	D	0.17		
m05	2385	1 468 521	D	0.17		
m05	2385	1 470 905	D	0.13	m05 tandem	III
m04	2375	2 977 988	R	0.04		
m03	2382	2 980 363	R	2.60		
m04	2375	2 982 745	R	0.04	m03m04	
m03	2382	2 985 120	R	2.60		
m04	2375	2 987 502	R	0.08	m03m04	
m03	2376	2 989 876	R	0.17		
m02	2404	2 992 252	R	0.79	m03m02	
m01	2446	2 994 656	R	0.08		IV
m01	2446	3 050 498	D	0.12		
m02	2404	3 052 943	D	0.79		
m03	2376	3 055 347	D	0.08	m02m03	
m02	2407	3 057 723	D	1.08		
m03	2376	3 060 127	D	0.04	m02m03	
m04	2375	3 062 503	D	0.04		
m03	2382	3 064 878	D	2.60	m04m03	
m04	2375	3 067 260	D	0.04		V
m05	2385	3 077 041	D	0.42		
m05	2385	3079426	D	0.08		
m05	2385	3 081 811	D	0.25		
m05	2385	3 084 195	D	0.17		
m05	2385	3 086 580	D	0.29	m05 tandem	VI

Encircled: monomers organized into 2-mer and 3-mer HORs. Column 1: monomer family classification; Column 2: monomer length (bp); Column 3: start position within NT_011903.12; Column 4: direct (D) or reverse complement (R); Column 5: divergence with respect to consensus (%); Column 6: *n*mer (if not monomeric); Column 7: array No. Monomers with divergence < 3% are included in table.

Figure 12. Schematic presentation of 2-mer and 3-mer HOR copies from Figure 11, based on ~2.4-bp monomers in human chromosome Y. For description see the text.

to the second m03, and from the first m04 to the second m04 monomer. Thus, all fragment lengths arising from six-monomer sub-sequence correspond to the combined length of three constituent monomers m02, m03 and m04.

In general, if we have four tandem monomers, where the first and fourth monomers are from the same family while the second and third monomers are from two different families, with a sizeable mutual divergence, most of the corresponding GRM fragment lengths will be equal to the sum of three monomer lengths, generating in GRM diagram a peak at that length.

Finally, for repeats associated with the 4757 peak in GRM diagram of NT_011903.12 the corresponding dominant key string is TTTTGTTA. The GRM diagram for this 4757-bp consensus length shows just one pronounced peak, corresponding to a ~2.4-kb monomer length (Figure 11c). This reveals a 2-mer structure based on ~2.4-kb monomer. This is further substantiated by expressing the 4757-bp consensus sequence in terms of monomer consensus sequences: the 4757-bp consensus sequence at contig position ~2.98 Mb can be very nearly presented as reverse complement of a two-monomer sequence m03 m04, in accordance with Table 2.

Similarly, we find some other cases of tandem sub-sequences characterized by every-other-monomer similarity:

m03 m02 m03 m02 (start position at 1 349 023bp),
m02 m03 m02 m03 m02 m03 (start position at 1 427 709bp),
m03 m04 m03 m04 (start position at 2 980 363bp) and
m02 m03 m02 m03 (start position at 3 052 957bp).

In each of these cases every second monomer belongs to the same family, contributing to the frequency of fragment lengths of $\sim 2 \times 2.4$ kb, i.e. to the 4757-bp peak. We note that the contribution from m02 m03 m04 m03 is smaller. It is mainly due to the frequency of fragment length from the first m03 to the second m03, while the frequency of fragment length from m02 to m04 is smaller because m02 and m04monomers are mutually more divergent.

Case study: inter-tandem transitional dispersed repeat sequences

In the spacings between the tandem sequences I–VI from Table 2 we identify peculiar insertions which contain dispersed monomers m01, m02, m05 and three types of additional larger monomers. Each of them is without any internal repeat substructure, which is seen from the absence of peaks in their GRM diagrams.

In front of tandem I we find a sequence of three monomers m05 m05 m7406. Here, m7406 denotes a new dispersed monomer of 7406 bp, which is located between the monomer m05 in front of region I and the first monomer m04 from the region I. Between the tandems II and III we find a spacing of ~9.8 kb that contains a 7407-bp monomer (denoted m7407). Within the large 1.5-Mb spacing between tandems III and IV, we find immediately in front of tandem IV the monomers m05 m7406. In the 9.8-kb spacing between the tandems V and VI we find again a monomer m7407. All these four ~7.4-kb monomers, appearing within spacings, are highly homologous (divergence <1%) and without any internal repeat structure.

Within the 64.5-kb spacing between the tandems I and II we find an array of monomers m8531 m01 m8535 m01 m42547 (summary length 64 505bp), where m8531 denotes a monomer of 8531 bp (similar to m8535), and m42547 a monomer of 42 547bp. In the 53.5-kb spacing between the tandems IV and V we find an array of monomers m42549 m02 m8556 (summary length 53 509bp). All three ~ 8.5-kb monomers are highly homologous (divergence <1%). Similarly, the two ~42.5-kb sequences are also highly homologous (divergence <1%).

Extension of GRM to very large fragment lengths (hundreds of megabases)

In our preliminary version of GRM algorithm (78,79) the computation of GRM diagrams was possible for fragment lengths up to 100 kb. In this way computation has covered dispersed repeat copies at mutual distance <100 kb. We found that the K-string ensemble $K = 2^3$ is suited to obtain significant GRM peaks in that fragment length interval. This interval covers practically all tandem repeats and Alus. Also it covers other dispersed repeats and segmental duplications with distance between copies <100 kb.

However, in the cases of segmental duplications repeat copies could appear at much larger spacings. Segmental duplications are portions of DNA present at two or more locations in the genome that satisfy the minimum requirement of 90% sequence identity and are >1 kb in length (22,98,99). Spacing between these copies can be very large. For this reason, we extend here the GRM algorithm to enable computation of fragment lengths to much larger spacings, up to hundreds of megabases, and to much longer copy lengths (as large as megabases). In general, with increase of K-string length the interval of accessible fragment lengths increases, i.e. segmental duplication copies and other dispersed copies can be detected at larger spacings.

The present GRM algorithm (grm2012) is based computationally on a method for K-string lengths $K = 2^n$ (n integer), in which case the mapping is performed

using operators on bits only. To identify very distant and long segmental duplications, at spacings of 100 Mb or more, we use grm2012 choosing longer *K*-strings ensembles, for example $K = 2^5$ or 2^6. For shorter *K*-string lengths, like $K = 8$, very long copies cannot be identified because very long fragments are being masked by a background generated by short fragments. In addition, our new procedure increases the speed of computation by additional hundreds of times. We have tested that this improved method applied to the $K = 8$ case sizeably decreases computation time, but does not influence results of our previous version of GRM. At the same time, this method enables extension of computed fragment length much above the 100-kb limit, to hundreds of megabases.

Case study: GRM identification of 0.6-Mb segmental duplications at 1-Mb spacing in human chromosome Y

Here we show a case study for identification of segmental duplications in contig NT_011903.12 of chromosome Y. Due to small divergence between the copies and substantial length segmental duplications show up strong peaks in GRM diagram. Using our grm2012 computer code for $K = 2^5$we compute GRM diagram up to fragment length of 3 Mb (Figure 13). In this GRM diagram we find a pronounced peak at the fragment length 1 633 744bp, in accordance with spacings of ~1.6 Mb between highly homologous sub-sequences from Figure 14. In that fragment length interval a dominant group of peaks appears at ~1.6 Mb; the most prominent is the GRM peak at the fragment length 1 633 744bp. GRM diagram indicates that all peaks around ~1.6 Mb are due to a group of several connected sub-sequences (total length of group is <1.6 Mb) which is repeated at a head-to-head distance of ~1.6 Mb.

An alternative approach to identify repeats using a single-*K*-string GRM is based on similarity of length arrays in different regions of genomic sequence. The GRM length array of NT_011903.12 for a 8-bp single-*K*-string TATTTTTA shows a mosaic repeat pattern in the length array (Table 3), which is being repeated in a different region within this contig, at distance of ~1.6 Mb (Table 3), revealing segmental duplications within this contig. The largest segmental duplications in NT_011903.12, of the length ~0.6 Mb each (denoted here as regions I and I′, respectively), lie at a mutual distance of ~1.6 Mb. The start positions of each of these two regions (944 621 and 2 578 383, respectively) correspond to positions of the onset of repeat segments within the length array, as illustrated in Table 3. Analogously, at the end of each of the two repeat regions the repeat sub-sequences of the two similar length arrays abruptly diverge into completely different patterns.

A difference of ~0.02 Mb in the length of two duplicons is due to deletions of some sub-segments. The largest deletion is at position ~2 999 615 in the region I′ with respect to the region I. Precise value of position of the onset of periodic pattern of length array depends on the *K*-string used, since the precision of determined position is sensitive to the length of fragments near the point of the onset. In general, the smaller is the average fragment length, i.e. the smaller the key string, the more precise will be determination of region boundaries. A more detailed scheme with mosaic pattern of segmental duplications in NT_011903.12 is shown in Figure 14.

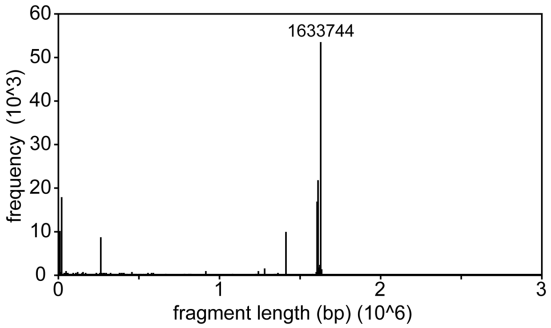

Figure 13. GRM diagram of the contig NT_011903.12 in chromosome Y (grm2012.exe at $K = 2^5$). Frequencies are shown up to the fragment length of 3 Mb revealing a strong peak of segmental duplication at the spacing of ~1.6 Mb.

CONCLUSION

Although extensive and impressive work has been done in order to identify approximate repeats from sequenced DNA data, there are still inherent limitations with available approaches. We present here the framework of GRM (application, help and test run at www.hazu.hr/grm/tools.html#grm2012) which is based on novel concept of direct mapping of symbolic DNA sequence into frequency domain. The concept of GRM is compared with standard approaches of digital signal processing, which uses mapping of symbolic DNA into numerical sequence, and of statistically based methods, which use locally optimized *K*-string mapping. We demonstrate the ability of GRM to detect various types of repeats (tandem repeats, segmental duplications, dispersed repeats, equidistant repeat copies separated by different spacers, complex repeats, Period-3 coding repeats) that could have extensive mutational changes (substitutions, deletions and insertions). GRM is effective, robust, fast and provides a global survey of repeats. Using a desktop computer it can identify within minutes repeats in sequences as long as human chromosomes.

A threshold of noise above which the GRM peaks are detectable in GRM diagram depends on the length of genomic sequence used in GRM computation. If we perform GRM computation for complete human chromosome sequences (genomic sequences as large as ~200 Mb) we identify GRM peaks of the 'first sight' survey of most pronounced repeats (tandem repeat arrays, HOR arrays, regularly or irregularly dispersed repeats, segmental duplications). For smaller but still substantial sub-sequences, of ~1–5 Mb, the noise background becomes lower and we can identify peaks corresponding to even shorter and/or more mutated (substitutions and insertions/indels) repeats than in the initial whole-chromosome GRM computation. Performing GRM computations for even smaller sub-sequences, of <~0.5 Mb length, the noise background is so much reduced that we can identify weak GRM peaks corresponding to very short tandem arrays and/or with

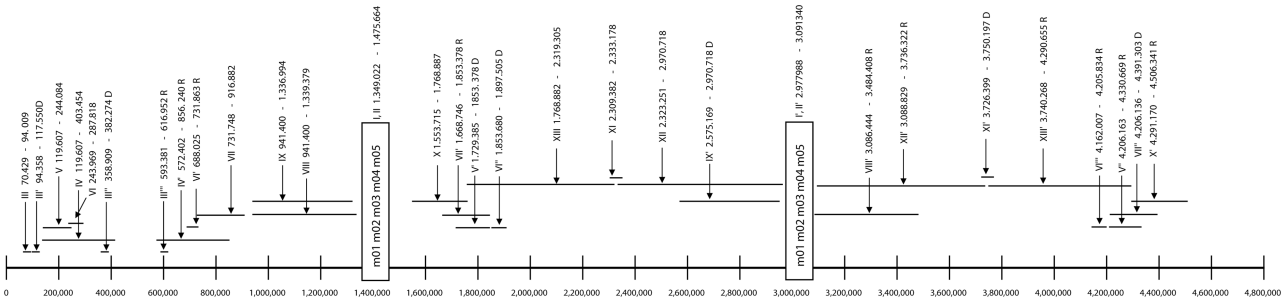

Figure 14. Scheme of the structure of segmental duplications within NT_011903.12 in chromosome Y. Horizontal axis represents position along the contig. Start and end positions of each duplicated region are determined as the onset and end of repeat pattern in *K*-string length arrays. A dominant 8-bp single *K*-string corresponding to the 2385-bp peak is AAATATTT. Heavy lines denote pairs of nearly identical sub-regions, denoted by n and n', respectively. Positions of duplicated sub-regions (in megabases): (1) 0.94–1.55, (1′) 2.58–3.16; (2) 0.12–0.28, (2′) 1.73–1.89; (3) 0.69–0.92, (3′) 4.17–4.39; (4) 0.28–0.40, (4′) 0.58–0.69 (reversed length array of 4); (5) 0.074–0.088, (5′) 0.097–0.111; (6) 1.56–2.97, (6′) 3.09–4.50 (reversed length array of 6); (7) 2.978–3.013; (7′) 3.023–3.069) (reversed length array of 7).

Table 3. Starting segments of two similar length arrays in NT_011903.12, obtained by GRM segmentation *K*-string TATTTTTA

region I	region I′
17	17
2521	2521
10 248	10 257
137	137
9489	9499
22 271	22 264
3977	3977
1561	1561
635	639
861	861
...	...

These length arrays correspond to two highly homologous genomic sub-sequences at starts of regions denoted I and I′ (each of ∼0.6 Mb length) at a mutual distance of ∼1.6 Mb. This pattern is an easily recognizable *K*-string signature of segmental duplications.

substantial substitutions and indels. Thus, the length of a genomic sub-sequence for calculating GRM diagram acts as a threshold for identifying less pronounced repeats. Current version of GRM application (grm2012.exe) removes almost all background noise during the process of repeats identification for a given genomic sequence segmented into 1-Mb sub-sequences. Thus, the GRM algorithm is an efficient tool to de-noise predictions for approximate and complex repeats.

In general, if we are interested to identify repeat units in a fragment length interval from 8 bp to 100 kb, it is convenient to use GRM algorithm with complete ensemble of all 8-bp *K*-strings (grm2012.exe with K-string length parameter $K = 8$). The value $K = 8$ is referred to as the standard GRM*K* value. If the *K*-string length is not specified, the code takes automatically the default length of 8 bp.

If we are interested in short repeat units, <8 bp, it is more convenient to use shorter *K*-string; then we use as the value $K = 3$. However, most of results are similar as obtained for the standard $K = 8$ input. On the other hand, in the cases when we are interested to identify very distant and very long segmental duplications (as large as megabases), longer *K*-strings are preferable, and we use $K = 2^5$ or $K = 2^6$.

In the standard string matching algorithms, like TRF (30) and BLAST (75), the optimal strings for analysis are determined locally by using statistical methods, so the chosen strings can differ from segment to segment within genomic sequence. On the other hand, in standard signal processing algorithms, like SRF (69), the intermediate symbolic to numerical sequence mapping can introduce some ambiguities or artefacts, and a significant background noise. In GRM we use consistently a complete ensemble of all *K*-strings to map directly the whole DNA sequence into frequency domain, which reveals repeats without need to use statistical methods or to involve ambiguities of symbolic to numerical sequence mapping.

The useful features of GRM algorithm are

- to embody some advantages of both TRF and SRF approaches, while avoiding some difficulties of heuristics (as present in TRF) and of symbolic to numerical mapping (as present in SRF),
- to be a highly automatized program appropriate for identification of different classes of repeats (tandem repeats, HORs, dispersed or equidistantly spaced repeats and complex repeats),
- to be capable of detecting repeats with any period,
- to provide a highly robust algorithm with respect to mutations (substitutions, insertions and deletions),
- to provide fast computations,
- to avoid need for users to optimize model parameters,
- to provide an overview (map) of repetitive structure for the whole sequence,
- to identify and locate repeats *ab initio*, without need of any prior knowledge of repeats.

Available algorithms have some advantages and some disadvantages, as pointed out, for example in (36–39,42,72), and a systematic comparison of efficacy and applicability of different algorithms is still a challenge. It was pointed out that what is needed are systematic comparisons of efficacy and selectivity of different algorithms

on the same, suitably selected standard set of test cases. This standard testing set should include different repeat patterns: tandem repeats, segmental duplication, dispersed repeats, equidistantly spaced repeats and complex repeats, including specified test substitutions, deletions and insertions. On this ground it should be possible to select optimal algorithms or combination of algorithms for systematic search of repeats in ever growing body of genomic sequences. As for comparison of algorithms, an opinion was expressed (76) that, following a common fairness practice, a thorough comparative benchmark should be best systematically performed by a third party, on a hopefully accepted set of test problems, taking in due account possible presence of different tunable input parameters, that can drastically affect the performance. It would be advisable that the authors of each approach test the sensitivity and efficacy of own algorithm analyzing complex genomic patterns looking for new repeats.

AVAILABILITY

The computer application (Windows and Linux versions) and test run are freely accessible from http://www.hazu.hr/grm//tools.html. For additional information, if needed, contact matko@phy.hr

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1.

FUNDING

Funding for open access charge: Croatian Ministry of Science, Education and Sport.

Conflict of interest statement. None declared.

REFERENCES

1. Jacob,F. and Monod,J. (1961) Genetic regulatory mechanisms in synthesis of proteins. *J. Mol. Biol.*, **3**, 318–356.
2. Britten,R.J. and Davidson,E.H. (1969) Gene regulation for higher cells—a theory. *Science*, **165**, 349–357.
3. Britten,R.J. and Davidson,E.H. (1971) Repetitive and nonrepetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Quart. Rev. Biol.*, **46**, 111–138.
4. Tautz,D., Trick,M. and Dover,G.A. (1986) Cryptic simplicity in DNA is a major source of genetic variation. *Nature*, **322**, 652–656.
5. Wessler,S.R. (1997) Transposable elements and the evolution of gene expression. *Exp. Biol.*, **1039**, 115–122.
6. Dorer,D.R. and Henikoff,S. (1994) Expansions of transgene repeats cause heterochromatin formation and gene silencing in Drosophila. *Cell*, **77**, 993–1002.
7. Nakamura,Y., Koyama,K. and Matsushima,M. (1998) VNTR (variable number tandem repeat) sequences as transcriptional, translational, or functional regulators. *J. Hum. Genet.*, **43**, 149–152.
8. Pennacchio,L.A. and Rubin,E.M. (2001) Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.*, **2**, 100–109.
9. Batzer,M.A. and Deininger,P.L. (2002) Alu repeats and human genomic diversity. *Nature Genet.*, **3**, 370–379.
10. Gelfand,Y., Rodriguez,A. and Benson,G. (2007) TRDB – the tandem repeats database. *Nucleic Acids Res.*, **35**, D80–D87.
11. King,D.C., Taylor,J., Zhang,Y., Cheng,Y., Lawson,H.A., Martin,J., Chiaromonte,F., Miller,W. and Hardison,R.C. (2007) Finding cis-regulatory elements using comparative genomics: Some lessons from ENCODE data. *Genome Res.*, **17**, 775–786.
12. Visel,A., Akiyama,J.A., Shoukry,M., Afzal,V., Rubin,E.M. and Pennacchio,L.A. (2008) Functional autonomy of distant-acting human enhancers. *Genomics*, **93**, 509–513.
13. Mercer,T.R., Dinger,M.E. and Mattick,J.S. (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–159.
14. Garfield,D.A. and Wray,G.A. (2010) The evolution of gene regulatory interactions. *BioScience*, **60**, 15–23.
15. Gemayel,R., Vinces,M.D., Legendre,M. and Verstrepen,K.J. (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.*, **44**, 445–477.
16. Noonan,J.P. and McCallion,A.S. (2010) Genomics of long-range regulatory elements. *Annu. Rev. Genomics Hum.Genet.*, **11**, 1–23.
17. 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.
18. Charlesworth,B., Sniegowski,P. and Stephan,W. (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, **371**, 215–220.
19. Warburton,P.E. and Willard,H.F. (1996) Evolution of centromeric alpha satellite DNA: molecular organization within and between human and primate chromosomes. In: Jackson,M., Strachan,T. and Dover,G. (eds), *Human Genome Evolution*. BIOS Scientific, Oxford, pp. 121–145.
20. Toth,G., Gaspari,Z. and Jurka,J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, **10**, 967–981.
21. Vergnaud,G. and Denoeud,F. (2000) Minisatellites: mutability and genome architecture. *Genome Res.*, **10**, 899–907.
22. Bailey,J.A., Yavor,A.M., Massa,H.F., Trask,B.J. and Eichler,E.E. (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.*, **11**, 1005–1017.
23. Ellegren,H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435–445.
24. Ames,D., Murphy,N., Helentjaris,T., Sun,N. and Chandler,V. (2008) Comparative analyses of human single- and multilocus tandem repeats. *Genetics*, **179**, 1693–1704.
25. Mayer,C., Leese,F. and Tollrian,R. (2010) Genome-wide analysis of tandem repeats in Daphnia pulex—a comparative approach. *BMC Genomics*, **11**, 277.
26. Tremblay,D.C., Alexander,G., Moseley,S. and Chadwick,B.P. (2010) Expression, tandem repeat copy number variation and stability of four macrosatellite arrays in the human genome. *BMC Genomics*, **11**, 632.
27. McLaughlin,C.R. and Chadwick,B.P. (2011) Characterization of DXZ4 conservation in primates implies important functional roles for CTCF binding, array expression and tandem repeat organization on the X chromosome. *Genome Biol.*, **12**, R37.
28. Tremblay,D.C., Moseley,S. and Chadwick,B.P. (2011) Variatian in array size, monomer composition and expression of the macrosatellite DXZ4. *PLoS One*, **6**, e18969.
29. Roy,A., Raychaudhury,C. and Nandy,A. (1998) Novel techniques of graphical representation and analysis of DNA sequences – a review. *J. Biosci.*, **23**, 55–71.
30. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
31. Chakravarthy,N., Spanias,A., Iasemidis,L.D. and Tsakalis,K. (2004) Autoregressive modeling and feature analysis of DNA sequences. *EURASIP J. Appl. Sign. Process.*, **1**, 13–28.
32. Krishnan,A. and Tang,F. (2004) Exhaustive whole genome tandem repeat search. *Bioinformatics*, **20**, 2702–2710.
33. Nandy,A., Harle,M. and Basak,S.C. (2006) Mathematical descriptors of DNA sequences: development and applications. *ARKIVOC*, **9**, 211–238.
34. Leclercq,S., Rivals,E. and Jarne,P. (2007) Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinformatics*, **8**, 125.

35. Sharma,P.C., Grover,A. and Kahl,G. (2007) Mininig microsatellites in eukaryotic genomes. *Trends Biotechnol.*, **25**, 490–498.
36. Merkel,A. and Gammel,N. (2008) Detecting short tandem repeats from genome data: opening the software black box. *Brief. Bioinformatics*, **9**, 355–366.
37. Richard,G.F., Kerrest,A. and Dujon,B. (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.*, **72**, 686–727.
38. Saha,S., Bridges,S., Magbanua,Z.V. and Peterson,D.G. (2008) Computational approaches and tools used in identification of dispersed repetitive DNA sequences. *J. Trop. Plant Biol.*, **1**, 85–96.
39. Saha,S., Bridges,S., Magbanua,Z.V. and Peterson,D.G. (2008) Empirical comparison of ab initio repeat finding programs. *Nucleic Acids Res.*, **36**, 2284–2294.
40. Arniker,S.B. and Kwan,H.K. (2009) Graphical representation of DNA sequences. *Proceedings of IEEE International Conference Electro/Information Technology*. Windsor. Ontario, Canada, pp. 311–314.
41. Lorenzo-Ginori,J.V., Rodriguez-Fuentes,A., Abalo,R.G. and Rodrigues,R.S. (2009) Digital signal processing in the analysis of genomic sequences. *Curr. Bioinformatics*, **4**, 28–40.
42. Zhou,H.X., Du,L.P. and Yan,H. (2009) Detection of tandem repeats in DNA sequences based on parametric spectral estimation. *IEEE Trans. Inform. Technol. Biomed.*, **13**, 747–755.
43. Silverman,B.D. and Linsker,R. (1986) A measure of DNA periodicity. *J. Theor. Biol.*, **118**, 295–300.
44. Li,W. and Kaneko,K. (1992) Long-range correlation and partial 1/f spectrum in a noncoding DNA sequence. *Europhys.Lett.*, **17**, 655.
45. Voss,R.F. (1992) Evolution of long-range correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.*, **68**, 3805–3808.
46. Li,W., Marr,T.G. and Kaneko,K. (1994) Understanding long-range correlations in DNA sequences. *Physica D*, **75**, 392–416.
47. Buldyrev,S.V., Goldberger,A.L., Havlin,S., Mantegna,R.N., Matsa,M.E., Peng,C.K., Simons,M. and Stanley,H.E. (1995) Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis: *Phys. Rev. E*, **51**, 5084–5091.
48. Anastassiou,D. (2001) Genomic signal processing. *Sign.Process. Mag. IEEE*, **8**, 8–20.
49. Cristea,P. (2002) Conversion of nucleotides sequences into genomic signals. *J. Cell Mol. Med.*, **6**, 279–303.
50. Wang,W. and Johnson,D.H. (2002) Computing linear transforms of symbolic signals. *IEEE Trans. Sign.Process.*, **50**, 628–634.
51. Rushdi,A. and Tuqan,J. (2006) Gene identification using the Z-curve representation. *Proceedings of IEEE International Confference on Acoustics, Speech and Signal Processing*, ICASSP II Toulouse, France, pp. 1024–1027.
52. Wang,L. and Schonfeld,D. (2009) Mapping equivalence for symbolic sequences: theory and applications. *IEEE Trans. Sign. Process.*, **57**, 4895–4905.
53. Benson,G. (1995) A space efficient algorithm for finding the best nonoverlapping alignment score. *Theor. Comput. Sci.*, **145**, 357–369.
54. Kannan,S.K. and Myers,E.W. (1996) An algorithm for locating nonoverlapping regions of maximum alignment score. *SIAM J. Comput.*, **25**, 648–662.
55. Sagot,M.F. and Myers,E.W. (1998) Identifying satellites and periodic repetitions in biological sequences. *J. Comput. Biol.*, **5**, 539–553.
56. Hauth,A.M. and Joseph,D.A. (2002) Beyond tandem repeats: complex pattern structures and distant regions of similarity. *Bioinformatics*, **18**, S31–S37.
57. Kolpakov,R., Bana,G. and Kucherov,G. (2003) Mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.*, **31**, 3672–3678.
58. Delgrange,O. and Rivals,E. (2004) STAR: an algorithm to search for tandem approximate repeats. *Bioinformatics*, **20**, 2812–2820.
59. Warburton,P.E., Hasson,D., Guillem,F., Lescale,C., Jin,X. and Abrusan,G. (2008) Analysis of the largest tandemly repeated DNA families in the human genome. *BMC Genomics*, **9**, 533.
60. Sokol,D., Benson,G. and Tojeira,J. (2007) Tandem repeats over the edit distance. *Bioinformatics*, **23**, e30–e35.
61. Chechetkin,V.R. and Turygin,A.Y. (1995) Search of hidden periodicities in DNA sequences. *J. Theor. Biol.*, **175**, 477–497.
62. Herzel,H. and Grosse,I. (1995) Measuring correlations in symbol sequences. *Physica A*, **216**, 518–542.
63. Tiwari,S., Ramachandran,S., Bhattacharya,A., Bhattacharya,S. and Ramaswami,R. (1997) Prediction of probable genes by Fourier analysis of genomic sequences. *Comp. Appl. Biosci.*, **13**, 263–270.
64. Trifonov,E.N. (1998) 3-, 10.5-, 200- and 400-base periodicities in genome sequences. *Physica A*, **249**, 511–516.
65. Anastassiou,D. (2000) Frequency-domain analysis of biomolecular sequences. *Bioinformatics*, **16**, 1073–1081.
66. Fukushima,A., Ikemura,T., Kinouchi,M., Oshima,T., Kudo,Y., Mori,H. and Kanaya,S. (2002) Periodicity in prokaryotic and eukaryotic genomes identified by power spectrum analysis. *Gene*, **300**, 203–211.
67. Cristea,P.D. (2003) Large scale features in DNA genomic signals. *Sign. Process.*, **83**, 871–888.
68. Tran,T.T., EmmanueleII,V.A. and Zhou,G.T. (2004) Techniques for detecting approximate tandem repeats in DNA. *Proc. IEEE Internat. Conf. Acoust., Speech, Sign. Process*, **5**, 449–452.
69. Sharma,D., Isaac,B., Raghava,G.P.S. and Ramaswamy,R. (2004) Spectral repeat finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics*, **20**, 1405–1412.
70. Vaidyanathan,P.P. and Yoon,B.J. (2004) The role of signal-processing concepts in genomics and proteomics. *J. Franklin Inst.*, **341**, 111–135.
71. Berryman,M.J., Allison,A., Wilkinson,C.R. and Abbott,D. (2005) Review of signal processing in genetics. *Fluctuation Noise Lett.*, **5**, R13–R35.
72. Gupta,R., Sarthi,D., Mittal,A. and Singh,K. (2007) A novel signal processing measure to identify exact and inexact tandem repeat patterns in DNA sequences. *EURASIP J. Bioinform. Syst. Biol.*, **1**, 43596.
73. Akhtar,M., Epps,J. and Ambikairajah,E. (2008) Signal processing in sequence analysis: advances in eukaryotic gene prediction. *IEEE J. Selected Topics Sign. Process.*, **2**, 310–321.
74. Chechetkin,V.R. (2011) Spectral sum rules and search for periodicities in DNA sequences. *Phys. Lett. A*, **375**, 1729–1732.
75. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
76. Parisi,V., De Fonzo,V. and Aluffi-Pentini,F. (2003) STRING: finding tandem repeats in DNA sequences. *Bioinformatics*, **19**, 1733–1738.
77. Poddar,A., Chandra,N., Ganapathiraju,M., Sekar,K., Klein-Seetharaman,J., Reddy,R. and Balakrishnan,N. (2007) Evolutionary insights from suffix array-based genome sequence analysis. *J. Biosci.*, **32**, 871–881.
78. Paar,V., Glunčić,M., Basar,I., Rosandić,M., Paar,P. and Cvitković,M. (2011) Large tandem, higher order repeats and regularly dispersed repeat units contribute substantially to divergence between human and chimpanzee Y chromosomes. *J. Mol. Evol.*, **72**, 34–55.
79. Paar,V., Glunčić,M., Rosandić,M., Basar,I. and Vlahović,I. (2011) Intragene higher order repeats in neuroblastoma breakpoint family genes distinguish humans from chimpanzees.*Mol. Biol. Evol.*, **28**, 1877–1892.
80. Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes – a genomic signature. *Trends Genet.*, **11**, 283–290.
81. Benson,G. and Waterman,M. (1994) A method for fast database search for all k-nucleotide repeats. *Nucleic Acids Res.*, **22**, 4828–4836.
82. Hampson,S., Kibler,D. and Baldi,P. (2002) Distribution patterns of over-represented k-mers in non-coding yeast DNA. *Bioinformatics*, **18**, 513–528.
83. Rosandić,M., Paar,V. and Basar,I. (2003) Key-string segmentation algorithm and higher-order repeat 16mer (54 copies) in human alpha satellite DNA in chromosome 7. *J. Theor. Biol.*, **221**, 29–37.

84. Qi,J., Wang,B. and Hao,B.L. (2004) Whole proteome prokaryote phylogeny without sequence alignmenmt: a K-string composition approach. *J. Mol. Evol.*, **58**, 1–11.
85. Rosandić,M., Paar,V., Basar,I., Glunčić,M., Pavin,N. and Pilaš,I. (2006) CENP-B box and pJα sequence distribution in human alpha satellite higher-order repeats (HOR). *Chromosome Res.*, **14**, 735–753.
86. Paar,V., Pavin,N., Rosandić,M., Glunčić,M., Basar,I., Pezer,R. and Durajlija Žinić,S. (2005) ColorHOR – novel graphical algorithm for fast scan of alpha satellite higher-order repeats and HOR annotation for GenBank sequence of human genome. *Bioinformatics*, **21**, 846–852.
87. Waye,J.S., England,S.B. and Willard,H.F. (1987) Genomic organization of alpha satellite DNA on human chromosome 7: evidence for two distinct alphoid domains on a single chromosome. *Mol. Cell Biol.*, **7**, 349–356.
88. Tyler-Smith,C. and Brown,W.R.A. (1987) Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J. Mol. Biol.*, **195**, 457–470.
89. Rudd,M.K. and Willard,H.F. (2004) Analysis of the centromeric regions of the human genome assembly. *Trends Genet.*, **20**, 529–533.
90. Lee,C., Wevrick,R., Fisher,R.B., Ferguson-Smith,M.A. and Lin,C.C. (1997) Human centromeric DNAs. *Hum. Genet.*, **100**, 291–304.
91. Alexandrov,I., Kazakov,A., Tumeneva,I., Shepelev,V. and Yurov,Y. (2001) Alpha-satellite DNA of primates: old and new families. *Chromosoma*, **110**, 253–266.
92. Alkan,C., Ventura,M., Archidiacono,N., Rocchi,M., Sahinalp,S.C. and Eichler,E.E. (2007) Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. *PLoS Comput. Biol.*, **3**, 1807–1818.
93. Mighell,A.J., Markham,A.F. and Robinson,P.A. (1997) Alu sequences. *FEBS Lett.*, **417**, 1–5.
94. Roy-Engel,A.M., Salem,A.H., Oyeniran,O.O., Deininger,L., Hedges,D.J., Kilroy,G.E., Batzer,M.A. and Deininger,P.L. (2002) Active Alu element "A-tails": size does matter. *Genome Res.*, **12**, 1333–1344.
95. Comeaux,M.S., Roy-Engel,A.M., Hedges,D.J. and Deininger,P.L. (2009) Diverse cis factors controlling Alu retrotransposition: what causes Alu elements to die? *Genome Res*, **19**, 545–555.
96. Paar,V., Pavin,N., Basar,I., Rosandić,M., Glunčić,M. and Paar,N. (2008) Hierarchical structure of cascade of primary and secondary periodicities in Fourier power spectrum of alphoid higher order repeats. *BMC Bioinformatics*, **9**, 466.
97. Boeva,V., Regnier,M., Papatsenko,D. and Makeev,V. (2006) Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics*, **22**, 676–684.
98. 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.
99. Eichler,E.E. (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet.*, **11**, 661–669.
100. Bailey,J.A. and Eichler,E.E. (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat. Rev. Genet.*, **7**, 552–564.