CROATIAN MEDICAL JOURNAL
CMJ

44(4):386-406,2003

BASIC SCIENCES

Key-string Algorithm – Novel Approach to Computational Analysis of Repetitive Sequences in Human Centromeric DNA

Marija Rosandić, Vladimir Paar[1], Matko Glunčić[1], Ivan Basar[1], Nenad Pavin[1]

Division of Gastroenterology, Department of Medicine, Zagreb University Hospital Center; and [1]Faculty of Science, University of Zagreb, Zagreb, Croatia

Aim. To use a novel computational approach, Key-string Algorithm (KSA), for the identification and analysis of arbitrarily large repetitive sequences and higher-order repeats (HORs) in noncoding DNA. This approach is based on the use of key string that plays a role of an arbitrarily constructed "computer enzyme".

Method. A cluster of novel KSA-related methods was introduced and developed on the basis of a combination of computations on a very modest scale, by eye inspection and graphical display of results of analysis. Sequence analysis software was developed, containing seven programs for KSA-related analyses. This approach was demonstrated in the case study of alpha satellites and HORs in the human genetic sequence AC017075.8 (193277 bp) from the centromeric region of human chromosome 7. The KSA segmentation method was applied by using DCCGTTT, GTA, and TTTC key strings.

Results. Fifty-five copies of 2734-bp 16mer HORs were identified and investigated, and a start-string TTTTTTAAAAA was identified. The HOR-matrix was constructed and employed for graphical display of mutations. KSA identification of HORs in AC017075.8 was compared with that of RepeatMasker and Tandem Repeat Finder, which identified alpha monomers in AC017075.8, but not the HORs. On the basis of KSA study, the centromere folding was described as an effect of HORs and super-HORs (3×2734 bp) in AC017075.8. The following novel computational KSA-based methods, easy-to-use and intended for computational "pedestrians", were demonstrated: color-HOR diagram, KSA-divergence method, 171-bp subsequence-convergence diagram, and total frequency distribution of the key-string subsequence lengths. The results were supplemented by Fast Fourier Transform, employing a novel mapping of symbolic genomic sequence into a numerical sequence.

Conclusion. The KSA approach offers a simple and robust framework for a wide range of investigations of large repetitive sequences and HORs, involving a very modest scope of computations that can be carried out by using a PC. As the KSA method is HOR-oriented, the identification of HORs is even easier than the identification of underlying alpha monomer itself. This approach provides an easy identification of point mutations, insertions, and deletions, with respect to consensus. This may be useful in a wide range of investigations and applied in forensic medicine, medical diagnosis of malignant diseases, biological evolution, and paleontology.

Key words: *centromere; chromosomes, human pair 7; computational biology; DNA; DNA satellite; Fourier analysis; molecular sequence data; repetitive sequences, nucleic acid; point mutation*

The centromere is a specific structure present in chromosomes of all eukaryotes. It is essential for proper segregation of chromosomes during mitosis and meiosis. It is also crucial for the transfer of genetic information between cell and organismal generations (1). The centromere is the primary site of the formation of functional kinetochores, specialized DNA-protein complexes, which interact with spindle microtubules. Its essential function is to provide the point of attachment of the chromosome to the spindle microtubules and to ensure that the sister chromatides are held together until the end of the metaphase and that they efficiently segregate at anaphase.

The centromeric DNA of several lower eukaryotes has been completely defined. However, much less is known about the centromeric DNA of higher eukaryotes. There are two classes of localized centromeres: point and regional centromeres (2). The point centromere is characteristic for yeast chromosomes (*S. cerevisiae*) and contains 125 base pairs (bp), which corresponds with the diameter of a microtubule, indicating that only one microtubule attaches to each

centromere during mitosis or meiosis in that yeast cell (3,4).

The regional centromere is characteristic for all other eukaryotic organisms and encompasses DNA range in size from a few hundred kilobase pairs (kbp) to several megabase pairs (mbp) that anchor more microtubules (5,6). It is composed of arrays of heterochromatic, centromere-specific repeated DNA sequences. These arrays vary in number and relative position across different chromosomes and strains. In such a way, centromeres in multicellular eukaryotes are embedded in large heterochromatic blocks and need a specific DNA structure for proper function.

Human Alpha Satellite Family

Satellite DNAs are highly repetitive sequences with tandem arrangement located primarily in the centromeric heterochromatin (7). They were recognized through their differential buoyant density after centrifugation of genomic DNA in $CsSO_4$, which resulted in the formation of I, II, and III "satellite bands" in addition to the main band. Each satellite DNA family, consisting of a heterogeneous population of repetitive DNA sequences, is characterized by the length and primary structure of the basic repeat unit, called monomer. Alpha satellite DNA is tandemly repeated DNA family found at the centromeres of all examined primate chromosomes (8). The fundamental repeat units of alpha satellite DNA are divergent monomers of approximately 171-bp (169-172 bp), often organized in chromosome-specific higher-order-repeat (HOR) units (8-11).

The human alpha satellite family based on the approximately 171-bp monomer unit is the only satellite DNA family present in the centromeric regions of all human chromosomes (9-11).

Satellite families, including alpha satellite, are not homogenous but contain monomer variants that differ from the consensus sequence up to 40%. They are often organized into different tandem arrays, which constitute clearly definable HOR structures or alphoid subfamilies. At least 33 different alphoid subfamilies have been identified (12). Some of these subfamilies are specific for a single chromosome, whereas others are common to a small group of chromosomes. Certain chromosomes seem to have a single HOR within their centromeres, whereas others may contain several different HORs.

The human centromere DNA makes 3-5% of each chromosome and accordingly contains repeating monomeric units of approximately 171 bp. In human chromosome 7 a 2.7-kbp HOR unit consisting of 16 monomers was found previously (13). The centromeric DNA analysis of chromosome 7 is interesting because a change in a copy number of chromosome 7 (aneuploidy or polysomy) is commonly observed in solid tumors (prostate cancer, renal tumors, and bladder cancer) and hematological disorders (myelodisplastic syndrome and acute myeloid leukemia) (12,14).

Biochemical Analysis of Centromeric DNA

The physical maps of the centromeric regions have been made by using biochemical long range restriction analysis and Southern hybridization (13,15-18). Due to the large size of multiple blocks of distinct satellite DNAs, breakpoints close to the centromeres cannot always be obtained, and the activity of centromeres can be altered when they are removed from their natural context. They do not always indicate a precise order and might be inaccurate in some size estimates. Enzymes used in these studies are the so-called "frequent cutters", which digest frequently non-satellite DNA in flanking regions while leaving tandem arrays intact because of the lack of recognition sites within satellite sequences. A more precise biochemical method is the ultra-high resolution fluorescence in situ hybridization (FISH) (19,20).

Novel Segmentation Key-string Algorithm – Case Study of Genomic Sequence AC017075.8

In spite of many available algorithms for computational analyses of genomic sequences, there has been no efficient method to identify HORs and find mutations, insertions, and deletions in their structure.

We introduced and exemplified a new computational algorithm for the identification of repeats in genomic sequences, convenient for identification of both alpha satellites and HORs. This method, named key-string algorithm (KSA) (21), was used to scan the recent GenBank complete data for the human clone AC017075.8 (22) consisting of 193277 bp from the centromeric region of chromosome 7. This method combines computation and visual inspection of computed results, providing a high degree of robustness in the identification of the alpha satellite and HOR patterns or coexisting subdivisions of genomic sequence.

The KSA method is equally sensitive in the identification of repetitions of subsequences with small number of nucleotides as well as very large subsequences consisting of thousands of nucleotides, like HORs, or even of tens of thousands of nucleotides.

In the centromeric region of the human chromosome 7, the 2.7-kbp HOR region consisting of 16 alpha monomers was first determined by restriction enzymes (23-29). The principle of the computational KSA is analogical to the principle of restriction enzyme method: the computational KSA procedure can "cut" genomic sequence into alpha monomers or into HORs, depending on the proper choice of the key string, which could be considered analogous to a computationally simulated "enzyme". The KSA method is organized in 5 steps, as follows:

1. Selection of a short string of nucleotides (three to six), referred to as the key string.

2. Computational segmentation of a given genomic sequence into subsequences, referred to as key-string subsequences, each starting with the key string.

3. Formation of an array of lengths of key-string subsequences (length-array) in the order of appearance.

4. If present, periodic HORs repetitions can easily be identified by naked-eye inspection of the length-array formed in the third step.

5. Deviations from periodic repetitions in the length-array are used for an easy and robust direct identification of insertions, deletions, and point mutations, with respect to an ideal HOR structure.

Identification of Higher-order Repeats in AC017075.8 by Using CCGTTT Key String

To illustrate quick identification of HORs in the genomic sequence AC017075.8, let us select the CCGTTT key string. On average, this key string appears about once per one kbp in the genomic sequence AC017075.8, ie, the frequency of its appearance is 190/193277.

The array of lengths of the CCGTTT key-string subsequences is shown in Table 1. In the front domain (about 30 kbp at the beginning of AC017075.8) and in the back domain (about 15 kbp at the end of AC017075.8), there is obviously no periodicity in the length array. In the central domain (from the position of about 31 kbp to about 178 kbp in AC017075.8), we can see the periodicity, with a periodic cell of three subsequences of 171 bp, 2393 bp, and 170 bp in length, respectively, starting at the position 31081 bp. This periodic pattern corresponds to the HOR of the following length:

171 bp + 2393 bp + 170 bp = 2734 bp.

This periodic repetition of the key-string subsequence lengths can be graphically presented in the form of a map, referred to as KSA length-map (Fig. 1).

Figure 1. Graphical presentation of the CCGTTT key-string subsequence lengths for the higher-order repeat (HOR) segment of genomic sequence AC017075.8 in the interval from 45092 bp to 50730 bp.

Deviations from the periodicity in the central domain can be easily explained.

For example, the CCGTTT subsequence starting at the position 33986 bp has the length of 2905 bp, instead of the periodic value 2393 bp, whereas the following expected 170-bp periodic subsequence is missing. This can be easily explained by observing that the 2905-bp subsequence can be segmented as follows:

2905 bp = 2393 bp + 2×171 bp + 170 bp.

In other words, the periodic pattern is modified by insertion of a 2×171-bp subsequence between the 2393-bp and 170-bp subsequences in the ideal periodic pattern.

171 bp, 2393 bp, and 170 bp.

We conclude that, at the position 33986, instead of the ideal periodic array, there is a 2905-bp subsequence obtained by fusion of subsequences

2393 bp, 171 bp, 171 bp, and 170 bp.

The ideal periodic segment differs from the modified 2905-bp subsequence by an insertion of the two 171-bp subsequences and their fusion.

Another example of deviation from periodicity, ie, from the ideal HOR, is the length of subsequence starting at the position 174613 bp. Its length is 2394

Table 1. Array of lengths of CCGTTT key-string subsequences in AC017075.8, with start position of the subsequence and the length of the subsequence

CCGTTT key-string subsequences	
position (bp)	length (bp)
14604	342
14946	13841
28787	683
29470	759
30229	852
31081	171
31252	2393
33645	170
33815	171
33986	2905
36891	171
37062	2393
39455	170
39625	171
39796	2393
42189	170
42359	171
42530	2392
44922	170
45092	171
45263	2393
47656	170
47826	171
47997	2393
50390	170
50560	171
50731	2393
53124	170
53294	171
53465	2393
55858	170
56028	171
56199	60
56259	2333
58592	170
58762	171
58933	2393
.....	....
174442	171
174613	2394
177007	170
177177	171
177348	3761
181109	341
181450	682
.....	....
191088	342
191430	340
191770	170
191940	342
192282	170
192452	171
192623	342
192965	171
193136	142

bp instead of 2393 bp. This indicates a point addition in the ideal periodic pattern.

Deviations from periodicity in the length-array also appear in subsequences starting at the position 56199 bp. The lengths of these subsequences are 60 bp and 2333 bp, instead of the 2393 bp, as of subsequence in the ideal periodic pattern. In this case, an additional CCGTTT key string was formed by a point mutation, segmenting the 2393-bp subsequence into two subsequences as follows:

2393 bp = 60 bp + 2333 bp.

By inspecting the array of lengths of key-string subsequences, we can immediately see that the 2734-bp HORs (16mer) are present in the central domain of the sequence AC017075.8, whereas the front- and back-domains are HORs-free.

Identification of Higher-order Repeats and Exact Determination of Insertions, Deletions, and Point Mutations by Using Key-string GTA

KSA approach reveals deviations from the ideal periodic pattern in the length-array of the central domain of AC017075.8. These deviations can be recognized as insertions, deletions, or point mutations, with respect to the ideal periodic pattern, ie, HOR consensus (21). For more precise identifications and positioning of these deviations, it is more convenient to use a key string with higher frequency of appearance. As an illustration, let us consider the GTA key string, with the frequency of appearance 1119/193277. This key string appears once per 0.2 kbp on average.

At the start of the genomic sequence AC017075.8, the GTA subsequence lengths form an irregular-length array (in bp) as follows: 429, 103, 51, 721, 171, 34, 49, 292, 38, 419, 947, 248, 634, 731, 34, 72, and so on. The first KSA subsequence of 429 bp starts at the position of 11 bp. The absence of periodic regularity reflects the absence of HORs.

The first subsequence in the KSA length-array showing periodicity is an 8-bp subsequence GTAA GGTT starting at the position 31336 bp. It is followed by a 154-bp subsequence GTACTTTGTT AGTTG AATGC ACCCATCAAA..., starting at 31344 bp position, which is followed by a 28-bp subsequence GTATCAGAAG GAAATTTCAA CTATGTGA starting at 31498 bp position.

The KSA length-array map in the central domain of AC017075.8 shows periodicity (Fig. 2). The periodic cell corresponding to the GTA key string is (in bp) 8, 154, 28, 21, 552, 280, 152, 26, 164, 264, 250, 143, 49, 96, 133, and 414. The sum of these lengths of the periodic cell is 2734 bp. This represents a consensus KSA length-array of the 2734-bp HOR when the GTA key string is used.

The KSA length-maps in Figures 1 and 2 represent two different segmentations of the 2734-bp HOR, corresponding to two different key strings, CCGTTT and GTA, respectively. The GTA key string has a sizably higher frequency than CCGTTT, and therefore leads to the segmentation of the genomic sequence AC017075.8 into shorter subsequences. This allows a more precise location of insertions, deletions, and point mutations with respect to consensus HOR (Fig. 3).

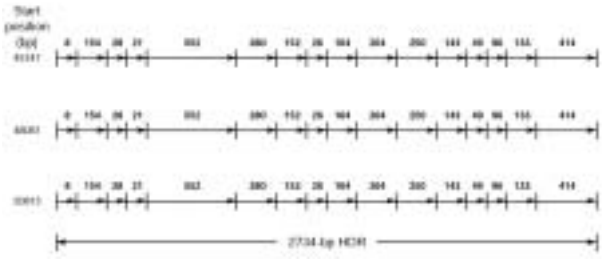

Figure 2. Graphical presentation of the GTA key-string subsequence lengths for the higher-order repeat (HOR) segment of AC017075.8 in the interval from 45347 bp to 53548 bp. For the sake of clarity, subsequences are not drawn to scale.

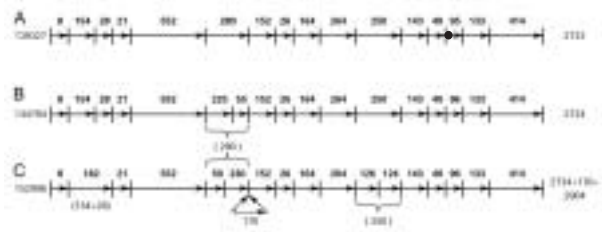

Figure 3. Graphical presentation of the GTA key-string subsequence lengths for some higher-order repeat (HOR) copies of AC017075.8 with deviations from the ideal 2734-bp HOR pattern. **A.** HOR segment starting at the position 139327 bp. All GTA-lengths are the same as for the ideal HOR in Fig. 2, except for the subsequence starting at 141418 bp, which has a length of 95 bp instead of 96 bp, ie, this length differs by one point deletion with respect to the ideal HOR sequence. This is indicated by a closed circle in the map. **B.** HOR segment starting at the position 144794 bp. In the subsequence starting at the position 145782 an additional GTA string is present due to point mutation with respect to the ideal GTA subsequence pattern, leading to a segmentation of the 280-bp subsequence into 225 bp + 55 bp subsequences. **C.** HOR segment starting at the position 152996 bp. We find by visual inspection three deviations from the ideal KSA length-array. The 154-bp and 28-bp subsequences are fused into one subsequence with a length of 154 bp + 28 bp = 182 bp. The 250-bp subsequence length is segmented into 126 bp + 124 bp subsequences due to point mutation in the 250-bp subsequence of the ideal KSA length-array. Another case is a segmentation of the 280-bp subsequence into 230 bp + 50 bp subsequences. A 170-bp insertion is at the position 154039 bp, between the 230-bp and 152-bp subsequences of the ideal length-array.

Identification of Alpha Satellites and Higher-order Repeats with TTTC Key String

The key strings we have used so far led to a straightforward identification of HORs and revealed deviations from their regular pattern. Let us now use the KSA method for the segmentation of AC017075.8 into alpha monomers, ie, into approximately 171-bp subsequences. It is possible to segment a sequence into alpha monomers in various ways by assigning different initial positions of alpha monomers. In the KSA method, this corresponds to the choice of different key strings. The key strings play the role of "computer enzymes" that cut the sequence at different po-

sitions. In our previous study (21), we presented the segmentation of the AC017075.8 genomic sequence into alpha monomers in two different ways: using the AAAC and the GTTTCT key string. In the first case, each HOR was segmented into nine 171-bp, five 170-bp, one 172-bp, and one 173-bp alpha monomers in particular order. In the second case, the same HOR pattern was segmented into ten 171-bp, four 170-bp, and two 172-bp alpha monomers (21).

In this study, we investigated other four-bp key strings, searching for a most convenient key string that segments the genomic sequence into approximately 171-bp monomers (171-bp, 170-bp, and 172-bp). Our study led to the identification of six convenient four-bp key strings (Table 2).

Table 2. Frequency of appearance of approximately 171-bp alpha monomers in AC017075.8 for six choices of the key string, in order of decreasing frequency of appearance of the 171-bp alpha monomer

	Frequency of appearance				
Key string	171 bp	170 bp	172 bp	169 bp	173 bp
TTTC	647	245	113	15	2
GTTT	634	232	103	11	1
CTTC	577	177	108	10	0
GATA	574	178	103	4	2
TTCT	535	183	72	15	1
TCTG	522	182	60	13	0

The largest number of the 171-bp subsequences within the AC017075.8 sequence appeared when the TTTC key string was used. Segmentation of AC017075.8 with the TTTC key string in the central domain is characterized by the periodic cell consisting of ten 171-bp, four 170-bp, and two 172-bp monomers organized into a 16mer, as follows:

170, 171, 171, 170, 171, 170, 171, 171, 171, 172, 171, 171, 171, 172, 171, 170.

The corresponding KSA subsequence array in the central domain is illustrated in Figure 4.

Figure 4. Graphical presentation of the TTTC key-string subsequence lengths for a higher-order repeat (HOR) segment of AC017075.8 in the interval from 81103 bp to 89305 bp.

When we compare the segmentation of HORs from the central domain into alpha monomers using different key groups, the KSA subsequence maps coincide for the TTTC and CTTC key groups (Fig. 5). The last segmentation displayed corresponds to the AAAC key string, which we used in our previous study (21).

On the other hand, in the front and back domains of AC017075.8, the length-arrays showed no periodicity, ie, HORs were not present. However, TTTC key

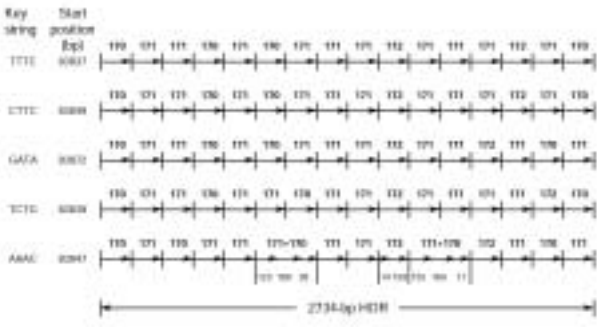

Figure 5. Comparison of segmentation of the 2734-bp HOR are presented for different key strings: TTTC, CTTC, GATA, TCTG, and AAAC.

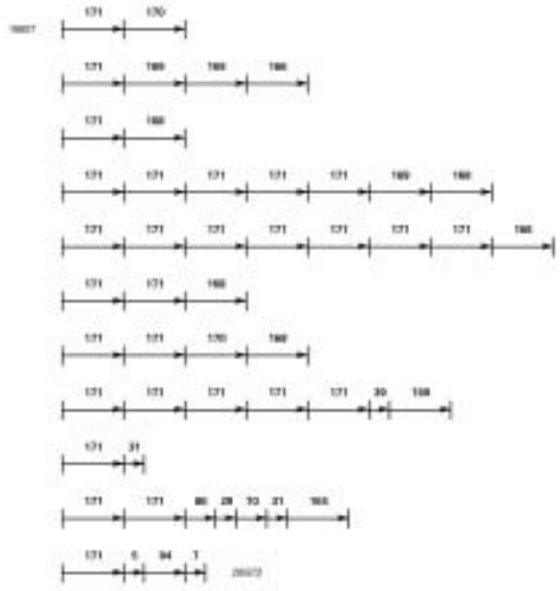

Figure 6. Illustration of segmentation of the front domain of AC017075.8 by using the TTTC key string.

string segmented the AC017075.8 sequence in non-HOR regions into alpha monomers (Fig. 6).

In the central domain, we identified larger insertions or deletions with respect to the HOR pattern (21). There is, for example, a 341-bp gap after the position 111893. The missing segment corresponds to the sequence at the 109160-109500 positions in the neighboring regular HOR copy. We found six distorted HORs in the central-domain that accounted for insertions and deletions (21). For example, in the HOR-copy No. 33, there is an insertion of the 856-bp sequence after the position of monomer m13; this insertion consists of an additional inserted 5-mer, m09+m10 +m11+m12+m13. This distorted HOR copy contains 21 instead of 16 monomers. We found that all insertions and deletions appearing in the AC017075.8 sequence contained only blocks which consist of monomers.

TTTTTTAAAAA – Higher-order Repeat Key String

We searched further for a characteristic HOR key string for segmentation of genomic sequence directly

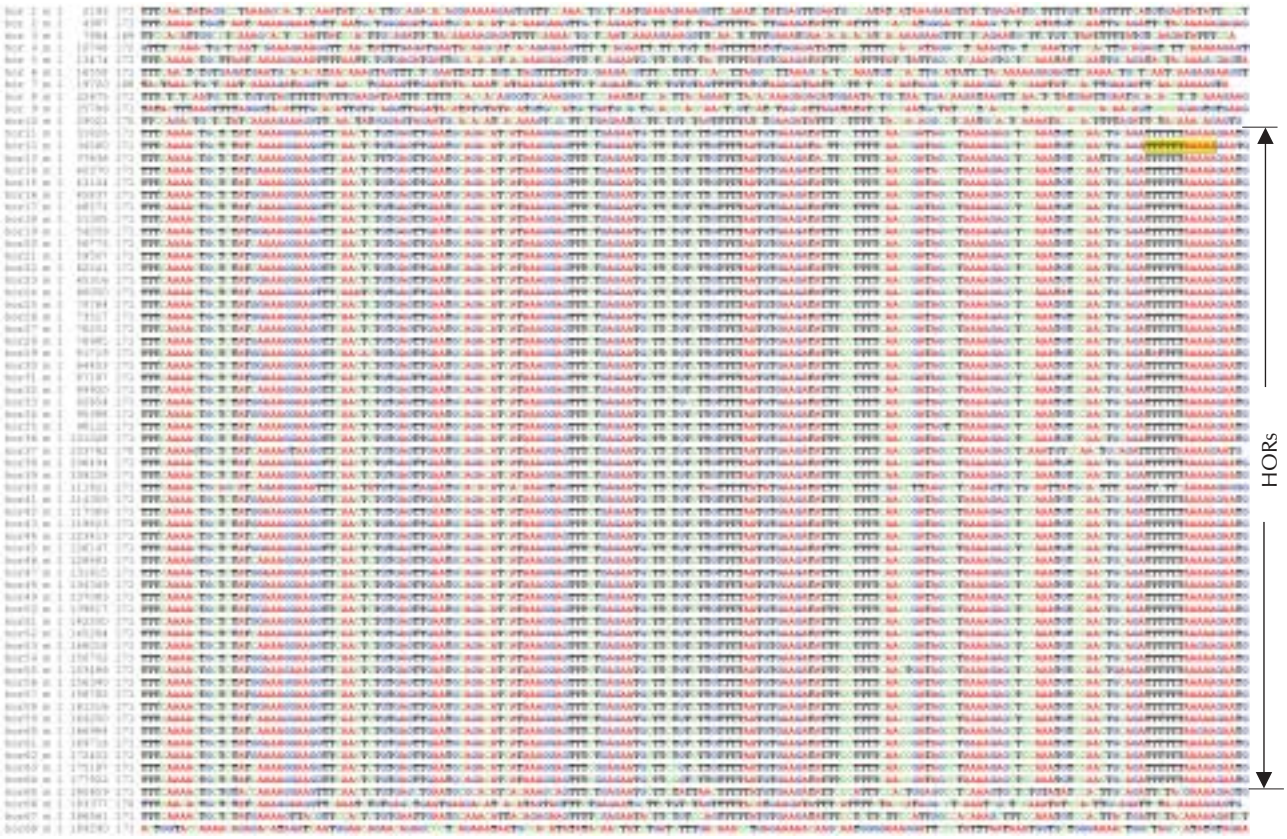

Figure 7. Segmentation of genomic sequence AC017075.8 by using the key string TTTC. Only the first 171-bp segment of each higher-order repeat (HOR) is displayed, one below the other. The "beautiful" string TTTTTTAAAAA is marked. The remaining 15 segments of approximately 171-bp, not shown in the figure, do not contain the «beautiful» string. Difference between the corresponding alpha satellites in the central domain is about 0.7%. This small difference is the key for exact identification of mutations.* Bases in the HOR-regions show almost perfect regularity, which is disturbed only by mutations. In the front and back domains, such crystal-like structure is not present.

into HORs. Such a key string is the 11-bp string TTTTTTAAAAA, referred to as the "beautiful" string. The beautiful string appears exclusively in the central domain of the clone, only once in each HOR copy and always at the same position (Fig. 7). This key string exhibits the highest degree of robustness, ie, the number of mutations with respect to genomic sequence is the smallest. Segmentation of the clone AC017075.8 with this key string reveals 55 HORs with 880 alpha-satellites. The first key string starts at the position 31981 bp and the last at 178077 bp. Thus, a more precise positioning of the three domains identified with the TTTTTTAAAAA key-string analysis are the front domain (31 kbp), the central domain (147 kbp), and the back domain (15 kbp).

At the positions 81876, 148182, and 167148, there is one point mutation in the corresponding key string, giving the respective modified forms: TATT TTAAAAA, TTTTTTAAGAA, and TTTTTTAAGAA. There are two point mutations in two strings at the positions 111779 and 111782 (giving TTCTTCAAAAA), and at the positions 120835 and 120836 (giving TCCTTTAAAAA) (Fig. 7). In the front-domain and in the back-domain there are 199 alpha monomers that are not organized into HORs.

**Web-extra:* Full segmentation of all 55 HORs is available at *www.cmj.hr*.

Sixteen Variants of Alpha Monomers within Higher-order Repeat (HOR) and Consensus HOR in AC017075.8

Sixteen alpha monomers in each HOR are denoted as follows: m01, m02...., m16. Divergence in sequence alignment of any two monomers within the same HOR copy is about 20% on average. This divergence was illustrated by displaying the alignment of two neighboring 171 bp sequences (21).

Divergence of monomers within the same HOR copy can be easily observed by displaying different variants of alpha monomer one below the other in the form of HOR-vector:

$$\begin{bmatrix} m01 \\ m02 \\ \cdots\cdots \\ m16 \end{bmatrix} \downarrow$$

As an example, we have displayed the variants of alpha monomer from HORs in this way (Fig. 8).

On the other hand, the sequence alignment between different 2734-bp HOR copies exhibits much weaker divergence. Divergence of different HOR copies with respect to each other is less than 1%.

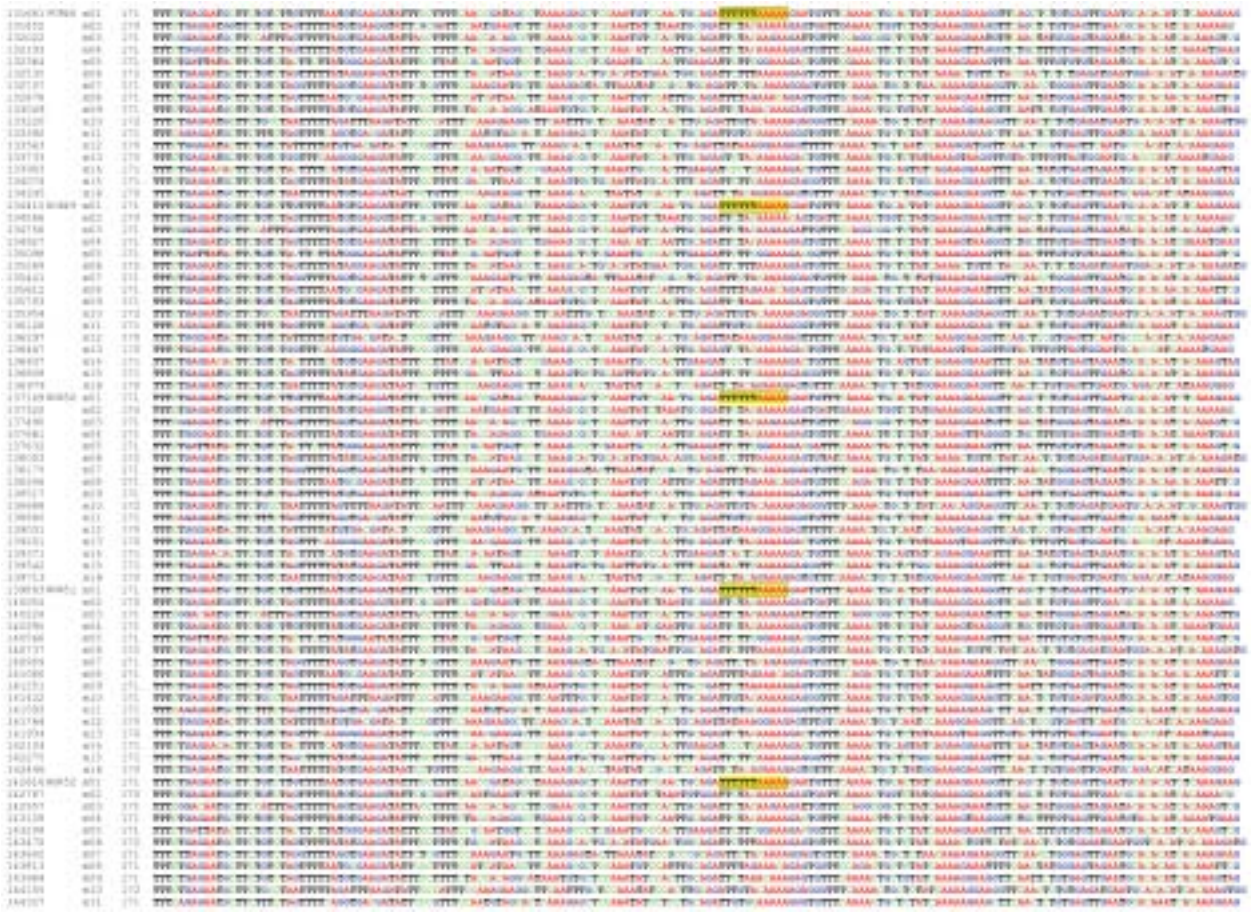

Figure 8. Segmentation of AC017075.8 into alpha monomers displayed one below the other according to schematic presentation of HOR-vector. The key group is TTTC. The «beautiful» string is marked yellow. Five higher-order repeats (HORs) are presented (h48-h52). Divergence between alpha satellites within the same HOR is about 20%.The pattern of bases does not show a crystal-like regularity of the type presented in the Figure 7 and point mutations are not visible like in Figure 7.

This pattern is consistent with previous findings of sizeable differences in nucleotide structure between the 171-bp monomer variants within each HOR copy (20-40%), while the differences between different HOR copies ere much smaller, less than a few percent (23-29).

Using the GTTTCT key string, with only one appearance in each alpha monomer, we have computationally constructed a consensus sequence in the central-domain of human chromosome 7 (21). Divergence of different HOR copies with respect to HOR consensus ranges from 0.3% (HOR 26) to 2% (HOR 01), with an average divergence of 0.7%. Thus, the concept of consensus HOR may serve as a convenient guideline and reference for computational analyses of mutations in satellite DNAs in eukaryotes.

Higher-order Repeat-matrix and Visual Display of Mutations

Let us define the map of 55 HORs using the KSA method in the case of AC017075.8. These HORs can be displayed in the form of the 55×2734 matrix (55 rows and 2734 columns). For example, using a chosen key group, we can display the first HOR in the first row, the second HOR in the second row, and so on until the 55th HOR in the 55th row. Such a HOR-matrix has an ordered alpha-monomer structure, with monomers m01 in the first 171 columns, m02 in the second 171 columns, and so on. Thus, 55 monomers m01 from 55 HORs are displayed one below the other, 55 monomers m02 from 55 HORs are displayed one below the other, and so on, as indicated schematically below in the HOR-matrix:

$$\begin{bmatrix} HOR01 \\ HOR02 \\ ..... \\ HOR55 \end{bmatrix} \rightarrow \begin{bmatrix} m01\, m02\, m03 \ldots m16 \\ m01\, m02\, m03 \ldots m16 \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ m01\, m02\, m03 \ldots m16 \end{bmatrix}$$

An example of two vertical segments from the HOR-matrix for AC017075.8 with TTTC key string is displayed in Figure 7 (corresponding to the m01-columns of the HOR-matrix) and in Figure 9 (corresponding to the m13-columns of the HOR-matrix).

The HOR-matrix presents a visually regular form of compact crystal-like structure. Within such a structure, each point mutation can be readily observed: in each column of a particular base or a group of bases, the changed base can be easily identified. Coloring each type of bases with the same color, the mutations can be straightforwardly recognized visually. For larger insertions and deletions, a disturbance of order-

Figure 9. The m13-columns of HOR-matrix for genomic sequence AC017075.8 with the key string TTTC. Centromere protein CENP-B box (in orange, read from rhs to lhs) as well as each of the CAAA and TTTC strings at three fixed positions are marked (color). Difference between alpha satellites m13 is about 0.7%. This small difference is the key for exact identification of mutations.

ing of bases is observed. More careful consideration would allow us to identify and follow all deviations from the regularity in the sequence.

Overview of Other Available Algorithms for Identification of Repetitive Pattern and Comparison with Key-string Algorithm for Higher-order Repeats

Many methods and programs have been developed for computational analyses of genomic sequences, all trying to achieve a compromise between efficiency and sensitivity requirements, which are different in different applications. However, despite many efforts and algorithms available, the identification of repetitive pattern within a given genomic sequence remains a challenge for computational genomics.

There are several algorithms available for identifying tandem repeats with a simple pattern structure, each having specific limitations (30-42). Repeat identification is complicated because of imperfect conservation due to substitutions, insertions, and deletions as well as complex structure of HORs.

Several algorithms for identifying recurrences in DNA sequences are based on suffix-tree approach (30-34) and compression-based approach (35-37). A priori period- and pattern-based algorithms use straightforward methods for locating tandem repeat regions. Algorithms were developed to find homologies between consensus sequences and a given genomic sequence (38,39). A similar approach is based on creating a sequence database of a genomic sequence and performing a homology search of the sequence itself, using tools like FASTA (43) and BLAST (44).

The problem of alignment has expanded and became substantially more complicated. Whereas the original alignment problem asked for a comparison of two short sequences, the current sequences being analyzed often demand robust alignment algorithms that satisfy many criteria and requirements. Alignment programs permit adjustment of numerous parameters and thresholds. However, it is a tricky business to test whether a particular combination of values is doing a good job (45). Moreover, different classes of parameters and thresholds might be tested in different ways. The maxim *"It is an order of magnitude easier to design two good programs than to tell which one is better"* seems appropriate (45).

Many difficult tasks have to be accomplished before the problem of aligning genome sequences is adequately solved (46). Also, a sophisticated yet efficient alignment program that combines the best of both local and global alignment algorithms still needs to be developed.

RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity sequences (47). The program provides position and type of repeats. RepeatMasker is not reliable for accurate annotation of simple repeat sequences for three reasons. The interspersed repeats are masked before the simple repeats, hiding many simple repeats contained within an interspersed repeat before these can be detected. Only dimeric, trimeric, tetrameric and some pentameric simple repeats are scanned for, and often regions that are indicated to be a simple repeat are highly divergent.

Several algorithms find repeats by using the results from the above tool; for example, Repeat Tandem Toolkit (RPT) identifies families of repeats (48).

MaskerAids masks repeats from sequence data (49). Unknown period- and pattern-based algorithms locate tandem repeat regions without prior knowledge of the period or the pattern. One basic algorithm in this framework tests ever longer constructed patterns by locating substrings and linking them into trains (50). However, such an algorithm is designed only for smaller periods (30-40 nucleotide range).

Another algorithm, Tandem Repeats Finder, looks for k-length substrings a common distance apart and uses statistical analysis to set thresholds for selecting distances that indicate a potential tandem repeat region (51-53). In this approach, the window length determines the sensitivity of the algorithm.

Hauth (54,55) used a similar sliding-length window technique, but the selection of potential tandem repeat periodicities differed from the Benson's (51,52) approach in that a threshold-based approach looked for clusters of identical distances and identification of regions having complex pattern structures that correlate to multiple periodicities within a tandem repeat region. Instead of a statistical model, a filter coupled with techniques for the identification of data sequence differences was used.

It was argued that complications due to imperfect conservation and complexity of HORs cannot be addressed by a single technique (54). Thus, an algorithm was proposed as a collection of algorithmic techniques, each designed to address a fundamental issue observed in genomic sequences (55).

When the motifs are known *a priori*, the dictionary approach provides a more efficient solution to the problem. This approach uses a dictionary of known motifs and scans the input sequence, searching for dictionary entries. A more recent program of that type is Tandem Repeat Occurrence Locator (TROLL) (56), which uses the dictionary approach to find tandem repeats of preselected motifs.

Let us now compare RepeatMasker and Tandem Repeat Finder with the KSA method when applied to the genomic sequence AC017075.8, which is characterized by HORs.

Although the RepeatMasker does a good job in masking simple repeats to avoid spurious matches in database searches, it is not written to find and indicate all possibly polymorphic simple repeat sequences.

When applied to the sequence AC017075.8, Repeat Masker shows that 97.7% of the sequence belongs to alpha satellites, but it cannot identify HORs.

The most famous program that lists tandem repeats from a genomic sequence is Tandem Repeat Finder (52), which finds repeats where motif is repeated with a certain degree of variation. The summary table, obtained by program analysis, contains information about each repeat, including its location, size, number of copies, and nucleotide content. Tandem Repeat Finder has been the most general and comprehensive algorithm for identification of tandem repeats available so far. However, only repeats with pattern size ranging from 1 to 500 bases can be detected. Thus, Tandem Repeat Finder program cannot be used to identify HORs in the sequence AC017 075.8. For example, this is shown for the segment from the position 75396 to 86857, which is built from HORs, as shown by KSA. Applying the Tandem Repeat Finder (52) with parameters 2 7 7 80 10 50 500, we can see that the program determined only the repeats of the lengths 171, 341, and 14 in that segment (Table 3).

Table 3. Result of identifying repeats using Tandem Repeats Finder in the interval from the position 75396 to 86857

Indices	Period	Copy number
75396-80378	171	29.2
76412-81403	341	14.6
76548-81546	171	29.3
77490-82489	171	29.3
82746-82777	14	2.3
78130-83112	171	29.2
78672-83671	171	29.3
79119-84118	341	14.6
79282-84280	171	29.3
80179-85163	341	14.6
80208-85207	171	29.3
80864-85846	171	29.2
81590-86589	171	29.3
81880-86857	341	14.6

Higher-order Repeats and DNA Folding

By analyzing each of the 16 alpha monomers in the central domain of AC017075.8 with HOR-matrix, we can see that all monomers are characterized by an increased A + T content (60-65%). However, with respect to the A + T content, their structure is segmented into three segments. The middle segment of length 68-69 bp differs from the two lateral segments of lengths 56-57 bp and 46-47 bp according to the base composition. The middle segment is dominated by thymine (40%) and the lateral segments by adenine (40%). This structure of alpha monomers is in connection with nucleosomes – arrangements of DNA and centromere proteins. Centromere protein A (CENP-A) is a centromere-specific histone H_3 homologue, a central component in kinetochore formation and centromere function (1,57,58). Centromere protein B (CENP-B) exists in every other satellite repeat (59). There is also centromere protein C (CENP-C). These three CENPs form regular spherical structures of folded eukaryotic centromere chromatin (1). Namely, antibodies to kinetochore proteins, such as CENP-A, localize only to a portion of alpha satellite DNA (1).

On the other hand, CENP-B specifically binds a 17-bp DNA sequence in alpha satellite named the CENP-B box, which positions nucleosomes adjacently (59, 60).

We propose that the middle segment of alpha monomer belongs to a linker, which binds nucleosomes, whereas the lateral segments are twice folded around CENP-A, creating a spherical structure of nucleosomes. Thus, it follows that the alpha monomer obtained by KSA computer analysis using TTTC or AAAC key strings binds neighboring nucleosomes in the following way: one fold from the first nucleosome + linker + one fold from the second nucleosome (Fig. 10). This is supported by the fact that we found the CENP-B box in the m13 of alpha satellites in HORs of AC017075.8. It was positioned at the end of linker bordering with the first fold around a nucleosome (Fig. 9).

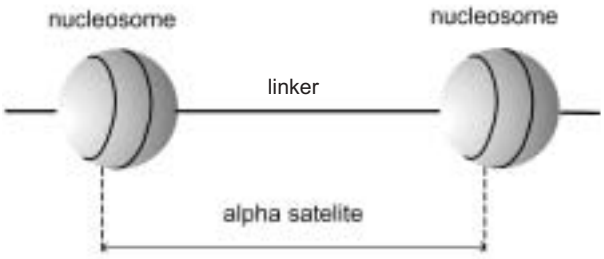

Figure 10. Schematic presentation of DNA folding around centromere proteins (CENPs).

Although the 16 alpha monomers within each HOR differ from each other by about 20%, a more careful analysis shows that despite significant overall divergence, there are two strings of bases present in all alpha monomers at the same position: CAAA and TTTC. While the CAAA strings in our HOR-matrix scheme are located almost exclusively in DNA segments folded around CENPs and between the two folds in the same nucleosome, the TTTC groups are present in the linker between the two folds just in front of CAAA. We can see this structure with the beginning of alpha monomer identified by TTTC or AAAC key string (Fig. 9).

The CAAA string is the most frequent four-nucleotide string in AC017075.8, which appears at the frequency of 5258, followed by AAAC (frequency 4144) and GAAA (frequency 1055). The four-nucleotide strings with the lowest frequency of appearance in AC017075.8 are, for example, CCGG (frequency 1), CGGG (frequency 2), and GCGT (frequency 8), and so on. Adenine (up to 5-bp strings) and thymine (up to 6-bp strings) cluster in the central HOR domain, while cytosine and guanine do not appear in strings larger than 4 bp.

We found the appearance of three to six segments of almost conserved CAAA motif in each of the 16 monomers within 55 HORs in AC017075.8 (Figs. 7- 9). In four of 16 monomers, some of the CAAA motifs were replaced by TAAA, GAAA, or GAGA. The monomer containing TTTTTTAAAAA (the "beautiful string") was the only monomer where the string TAAA replaced twice the CAAA string in robust columns (Fig. 7). It is particularly significant that in all monomers within HORs, the CAAA motif appeared at three analog positions, forming three robust chains. The other three positions of the CAAA string were localized inside most of the alpha monomers at analog positions, but not in all. According to our computational scheme of the central domain with HORs, three fixed positions of CAAA motif were placed at the beginning and the end of folding around nucleosomes, and between the two folds.

The next four-nucleotide string at three fixed identical positions in each of the 16 monomers within HORs was TTTC. Two positions were in the linker, and the third between the two folds around nucleosomes, immediately before the CAAA string (Fig. 9). The TTTC motif appeared 39 times in each HOR, and the variants TTC and TTTTC four times, and TTTTTC once. Cytosine was always present at the end of the motif TTTC or its variants.

In the middle segment of each of the 16 monomers within HORs, there was one additional larger and dominant thymine string consisting of 2-5 thymines (10×TTTTT, 4×TTTT, 1×TTT, 1×TT) in individual monomers. A nucleotide before and after the thymine string in each monomer was variable, and most often (11/16) guanine came before and adenine after the thymine string.

A similar structure can be found outside of the central-domain with HORs, but with less regular pattern.

What can be deduced from the robust identical strings at fixed positions in each of the 16 monomers within HORs? If we accept that the alpha-monomer is a basic unit of folding around nucleosomes (two folds around CENPs plus a linker), then regardless of nucleosome sequences in the folding direction and the number of monomers in one plane, fixed strings will always appear exactly one below the other. Thus, regardless of how big the central domain with HORs in the clone is, these fixed positions of the same string will present an axis in folding, contributing to the strength of centromeres. This could be compared to iron rods in reinforced concrete. The same motifs at fixed positions within monomers form a crystal-like structure inside the HORs that additionally lead to compactness of the centromere/kinetochore construction in the region of the kinetochore microtubules attachments.

New Look at Centromere Folding as Effect of Higher-order Repeats (HORs) and Super-HOR in AC017075.8

On the basis of our KSA computational method in the analysis of alpha-satellites and HOR structure, we discuss the problem of constructing the centromere folding, starting from the established features: (a) the centromeric region is built from blocks of alpha-satellites, (b) each alpha satellite consists of approximately 171 bp, which are folded twice around CENP-A, (c) one plane of rotation involves six nucleosomes, and (d) 16 alpha monomers form a HOR with the 2734-bp consensus.

Each of the 16 monomers within HORs has a specific structure, which is unique for each position within HORs. Thus, it is not appropriate to consider the common monomer consensus for all alpha monomers in the HOR. Instead, a consensus monomer should be determined separately for each of the 16 monomer positions. Consequently, there are 16 consensus alpha monomers associated with 55 HOR sequences.Divergence among these consensus alpha monomers within the same HOR copy is about 20%, whereas the divergence between the two HORs is 0.7% on average. We would suggest that the asymmetry, arising from 16 different monomer variants within HORs, leads to folding.

The CAAA and TTTC motifs, each at three fixed positions within the same monomer, form a crystal-like structure that additionally contributes to compactness of the centromere/kinetochore construction (Fig. 11).

Taking into account that six nucleosomes with CENP-A appear in the same plane (29) (Fig. 12), which is analogous to folding around histones in other parts of chromosomes, we studied the folding of HORs in the centromere region. According to the mathematical pattern associated with the "perfect" number six, the starting point of folding is irrelevant (Fig. 13). Cyclic arrays of three HORs display a new cycle of folding. Horizontal arrays in our schematic drawing generate three pairs of columns of the same group of 16 alpha monomers. Organized in this way, they simultaneously form three vertical HOR patterns in the scheme. We can see the specific positions of each of the 16 alpha monomers, whose cycles are regularly repeated. Such a doubly intertwined HOR structure is referred to as a super-HOR (3×2734 bp = 8202 bp). In fact, the pattern involving six nucleosomes is the only combination that can generate a super-HOR consisting of three successive HORs. This basic structure always results in spiral construction (Fig. 13). In this way, folding is generated by regular sequences of base pairs, alpha monomers and HORs.

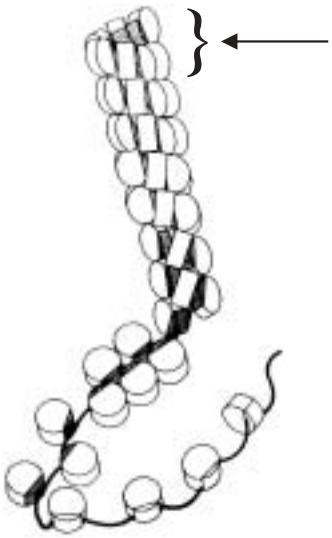

Figure 12. Schematic presentation of six nucleosomes in one plane (arrow).

A similar consideration can be expanded to other chromosomes regardless of the number of monomers within the HOR.

Such a type of folding reconstruction involves several levels of organization: double folding around each nucleosome, folding around nucleosomes according to the scheme based on the number six, and spiral folding generated by HORs in the horizontal and vertical arrays. We would suggest that such a regular mosaic organization could enable a unique construction of the centromeric region and a random attachment to the kinetochore microtubules.

Key-string Algorithm Analysis of Other Clones in the Centromere of Chromosome 7

Using the KSA analysis of other available clones in the centromeric region of chromosome 7, we identified arrays of alpha satellites in additional three clones. Employing the AAAA key string, we found al-

Figure 11. Key-string algorithm for the TTTC key string applied to a section of 16mer higher-order repeat (HOR) section of genomic sequence AC017075.8 in human chromosome 7. **A.** Genomic sequence corresponds to the odd-numbered alpha satellite monomers (1, 7, 13, 3, 9, 15, 5, 11) (odd-numbered vertical column in Fig. 13). **B.** Genomic sequence corresponds to the even-numbered alpha satellite monomers (6, 12, 2, 8, 14, 4, 10, 16) (even-numbered vertical column in Fig. 13). The "beautiful" string is marked yellow, centromere protein CENP-B box orange, TTTC string pink, and CAAA string green.

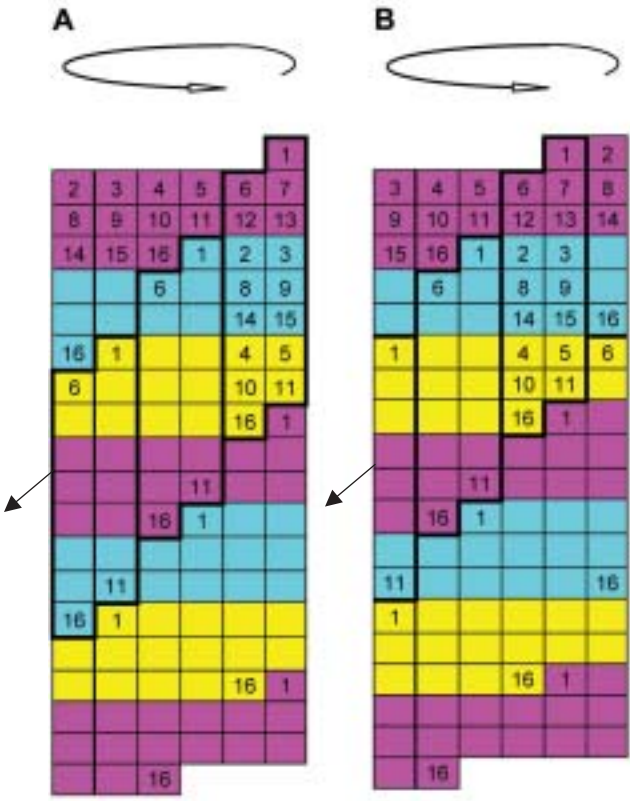

Figure 13. Schematic planar presentation of basic folding structure resulting from aligning alpha monomers and HORs. There are six monomers in each plane of rotation. **A.** Starting with monomer 1 within HOR. **B.** Starting with monomers 1 and 2. Three successive HORs (red, blue, and yellow) form one $2734 \times 3 = 8202$-bp super-HOR. In both presentations the analog HOR structure evolves horizontally (each characterized by one of three colors) and vertically (each encircled). This leads to a third, inclined direction in the scheme that corresponds to a spiral folding.

pha satellites in the clones AC092383 (137446 bp) and AC0104789 (142048 bp), and by using the TTTT key string we found them in the clone AC01906.3 (187282 bp). In the last 10,000 base pairs of the clone AC01906.3, we identified HOR dimers (171 bp, 169 bp). The other sequences of alpha satellites, interrupted at some positions by irregular shorter segments, did not exhibit HOR organization.

Novel Color-HOR Diagram for Identification of HORs

Here we introduce a novel graphical algorithm for a fast scan and identification of HORs, referred to as a Color-HOR diagram. This method of a straightforward visual identification of HORs is based on the KSA.

Using the KSA method, we formed a length-array, ie, an array of lengths of key-string subsequences in a given genomic sequence. We wanted to determine computationally the frequencies of subsequence lengths for a chosen key string and, based on this, to construct the Color-HOR diagram. Here follows the illustration of the method for AC017175.8, consisting of several steps.

First step. We choose a particular key string, for example GCC. Positions of the GCC key string within AC017075.8 are illustrated in Fig. 14. We determine computationally a set of lengths (Δ) of the corresponding subsequences.

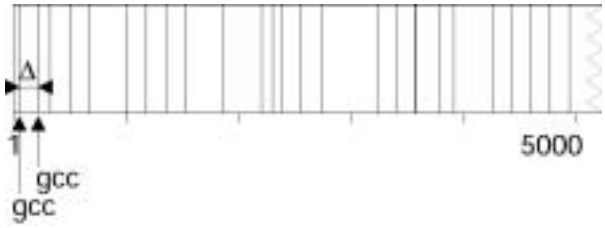

Figure 14. Illustration of positions of the key string GCC in the front section at positions 1-5000 bp of genomic sequence AC017075.8. Horizontal band displays genomic sequence; each vertical line displays position of the key string GCC along genomic sequence.

Second step. For each subsequence length Δ, the corresponding frequency *N* is computed. We display the length-frequency distribution, *N* versus Δ (Fig. 15). Each peak corresponds to a frequency *N* of a certain length Δ. From Figure 15, we identify a restricted set of lengths (five to ten) appearing with largest frequences (Fig. 15 and the first column in Table 4).

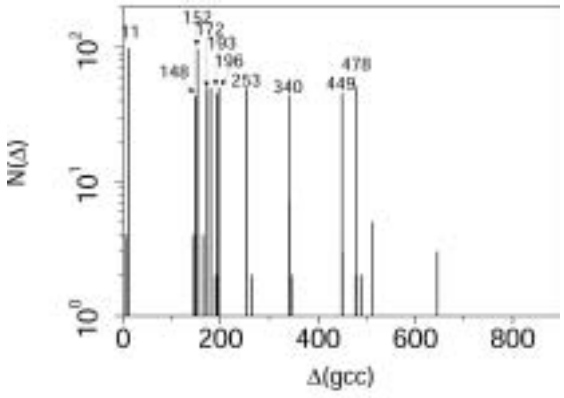

Figure 15. Graphical presentation of the length-frequency distribution (frequency *N* versus subsequence length Δ) for genomic sequence AC017075.8 with the key string GCC.

Third step. To each of most frequent lengths identified in the second step, we arbitrarily assign a color (Table 4). If the frequencies of lengths for a given key string are too scattered, so that a set of 5-10

Table 4. Ten most frequent subsequence lenghts in genomic sequence AC017075.8 for the key string GCC and their assigned colors

Length Δ(bp)	Color
478	black
449	red
340	green
253	blue
196	yellow
193	brown
172	grey
152	violet
148	cyan
11	magenta

lengths with pronounced frequencies does not appear, the same procedure is repeated for some other key string, until a convenient key string is found that generates a set of 5-10 lengths with pronounced frequencies.

Fourth step. Along the horizontal band that displays the genomic sequence (AC017075.8 in our case study), we present positions of all key strings appearing in the genomic sequence by vertical lines. This corresponds with extension of the band in Figure 14 to complete length of the genomic sequence AC017075.8 and its compression. In this way, the band is segmented into a series of intervals. Each interval in this map displays a KSA subsequence graphically and is colored according to the prescription chosen in Table 4: each interval of the length $\Delta = 152$ bp – violet, each interval of the length $\Delta = 11$ bp – magenta, each interval of the length $\Delta = 196$ bp – yellow, and so on. The intervals with lengths of low frequencies (not included in Table 4) are left blank.

In this way, we obtain a Color-HOR diagram for a given genomic sequence, AC017075.8 (Fig. 16).

The interval of alpha monomers not organized into HORs remains mostly blank in the Color-HOR diagram, because the lengths there are always very scattered. The domain with HORs displays a colored motif, which is periodically repeated (with some irregularities depending on insertions and deletions) along the map presenting the genomic sequence. We can also observe 53 repetitions of a basic color-motif, some of them distorted (Fig. 16). This repetitive color-motif can be seen in the central domain of the map, in accordance with the result of direct KSA segmentation. Thus, an easy identification of repetitive Color-HOR motif provides a simple visual way to determine the HOR region within a given genomic sequence.

Novel KSA Divergence Method

Key-string Algorithm Divergence Array, Divergence Matrix, and Divergence Diagram

Another KSA-based graphical method is the KSA-divergence method, which allows a simple visual identification of HORs in a given genomic sequence. The method illustrated here for the case study of AC017075.8 is organized in the following manner.

First step. A given genomic sequence (AC017 075.8) is segmented into KSA subsequences with a key-string algorithm that segments the sequence into alpha monomers, e.g., AAAC (21).

Second step. Key-string subsequences from the first step are transformed in order of appearance into modified KSA subsequences as follows:

- If the length of a subsequence is between 168 and 173 bp, the subsequence is left unchanged.
- If the length of a subsequence is larger than 173, the first 171 base pairs in the subsequence are retained, while the rest of the base pairs are deleted.
- If the length of a subsequence is less than 168 base pairs, it is deleted in the case that the following subsequence, in the order of appearance, is longer than 167 bp. For example, if a 79-bp subsequence is followed by a 170-bp subsequence, then the whole 79-bp subsequence is deleted.
- If the length of a subsequence is less than 168 base pairs and if the next subsequence is shorter than

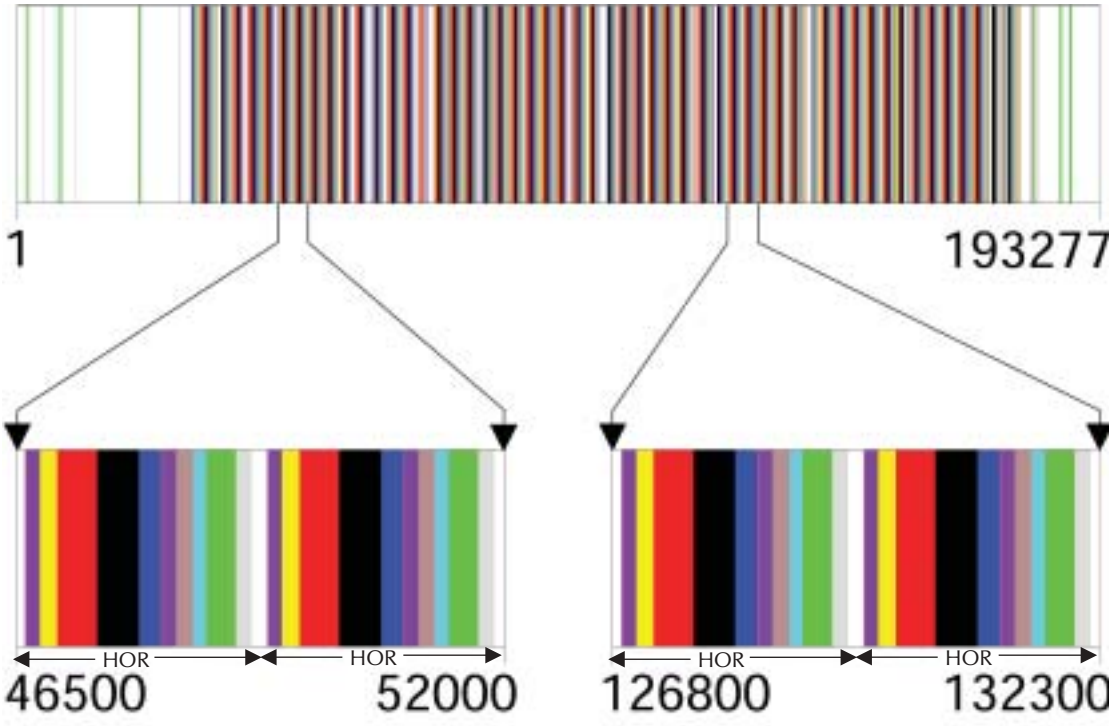

Figure 16. Color-HOR diagram for AC017075.8 with the key string GCC. The two smaller intervals of the Color-HOR diagram, containing two color-motifs each, are magnified for clarity. The front domain at the beginning and back domain at the end of the upper band, containing alpha monomers withouth HORs, are blank.

168 bp, both subsequences are fused into a single subsequence. After that, only the first 171 base pairs of the fused sequence are retained, while the rest are deleted. For example, if a 79-bp subsequence is followed by a 150-bp subsequence, they are fused into one subsequence of 79 bp + 150 bp = 229 bp (79 base pairs of the 79-bp subsequence followed by 150 base pairs of the 150-bp subsequence). In this fused 229-bp subsequence, only the first 171 base pairs are retained, while the rest are deleted.

In this way, a given genomic sequence (AC0170 75.8 in our case study) is transformed into an array of approximately 171-bp subsequences, referred to as KSA divergence-array. For AC017075.8, the KSA divergence-array based on the AAAC key string has the size of 1061 subsequences of approximately 171 bp.

Let us denote the divergence-array as m_1, m_2, m_3,... m_N, (N=1061 in our case study). Here, m_k denotes the kth subsequence of approximately 171 bp.

Third step. Now we compute the divergence between any two subsequences in divergence array. These results can be displayed by the following matrix, referred to as divergence matrix:

$$\begin{bmatrix} 0 & d_{12} & d_{13} & ... & d_{1N} \\ d_{21} & 0 & d_{23} & ... & d_{2N} \\ . & . & . & ... & . \\ d_{N1} & d_{N2} & d_{N3} & ... & 0 \end{bmatrix}$$

For example, d_{12} denotes divergence between the subsequences m_1 and m_2. In general, d_{ik} is divergence between the subsequences m_i and m_k.

Fourth step. To graphically display divergences d_{ik}, N divergence diagrams are used. In the first diagram (Fig. 17a) we display divergence of the subsequence m_1 with respect to all the other subsequences (m_2, m_3,... m_N). Here we refer to m_1 as the referent subsequence. The horizontal axis displays the enumerator k of a subsequence in the array (m_2, m_3,... m_N) and the vertical axis displays divergence with respect to the referent subsequence (m_1). For example, the closed circle showing divergence at the position k = 400 in Figure 17A corresponds to divergence between the referent subsequence m_1 and the subsequence m_{400}. Inspection of this diagram shows that divergence between the referent subsequence m_1 and all the other subsequences is in the range of about 15-75%.

The genomic sequence AC017075.8 can be segmented into three domains (Fig. 17A) in accordance with previous considerations by using other methods of analysis, as follows:

1. front domain, characterized by an irregular pattern, for $k \leq 150$, ie, corresponding to the front domain (≈30 kbp) of the genomic sequence;

2. central domain, characterized by a periodic pattern, for k between ≈150 and ≈1018, ie, for the middle domain of ≈150 kbp; and

3. back domain, characterized by an irregular pattern, for $k \geq 1018$, ie, for the back domain of ≈15 kbp.

The central domain contains periodically repeated HORs of 2734 bp (16mers), investigated by KSA segmentation in our previous study (21).

The divergence diagram for the referent subsequence m_2 (Fig. 17B) exhibits a qualitatively similar pattern to that in Fig. 17A.

Qualitatively similar behavior, with periodic central domain at 15% divergence, is found in the array of divergence diagrams for referent subsequences m_3, m_4, up to divergence diagram for the 152nd subsequence m_{152} (Fig. 17C). Some divergence diagrams have a reduced range of divergences, but they also show a periodic pattern in the central domain, with divergence larger than 15% (for some referent subsequence, the lower limit of divergence is sizably higher). This means that for any referent subsequence from the front domain, the divergences with respect to all other subsequences are larger than 15%.

The basic change of pattern appears first for the referent subsequence m_{153} (Fig. 17D). While the graph in the front and back domains is similar to previous cases, with all divergences greater than 15%, in the central domain every 16th subsequence shows dramatic drop in divergence to less than 1%. This reveals the presence of HORs. A similar pattern persists for referent subsequences m_{154}, m_{155}, up to the referent subsequence m_{1018}.

In the remaining diagrams, for referent subsequences to the very end of the back domain of AC017075.8, all divergences are again greater than 15%.

For example, for the referent subsequence m_{1017} (Fig. 17E), the minimal divergence in the central domain is smaller than 1%, whereas for m_{1019} (Fig. 17F) the minimal divergence is larger than 15%.

Due to the way (fusing and truncation) used in forming the divergence array (second step of the procedure), only the gross structure of HORs can be analyzed. The finer details can be analyzed by KSA segmentation, as described for the case study for genomic sequence AC017075.8.

Thus, divergences of any referent subsequence from the central domain with respect to all the other subsequences in the central domain of AC017075.8 exhibit pronounced periodicity, with very small divergence (1%) with respect to each 16th position in the central domain. A straightforward visual inspection of divergence diagrams provides an easy identification of that pattern, ie, of the 16mer HOR domain.

The KSA divergence-recipe for identification of HOR domain can be defined as follows:

First we arbitrarily choose a referent subsequence and computationally construct the corresponding divergence diagram (for example, Fig. 17A). In this diagram, we visually identify a domain with periodic pattern. Then we arbitrarily select as a new referent sequence the one from this periodic domain and computationally construct the corresponding divergence diagram (for example, Fig. 17D). This diagram

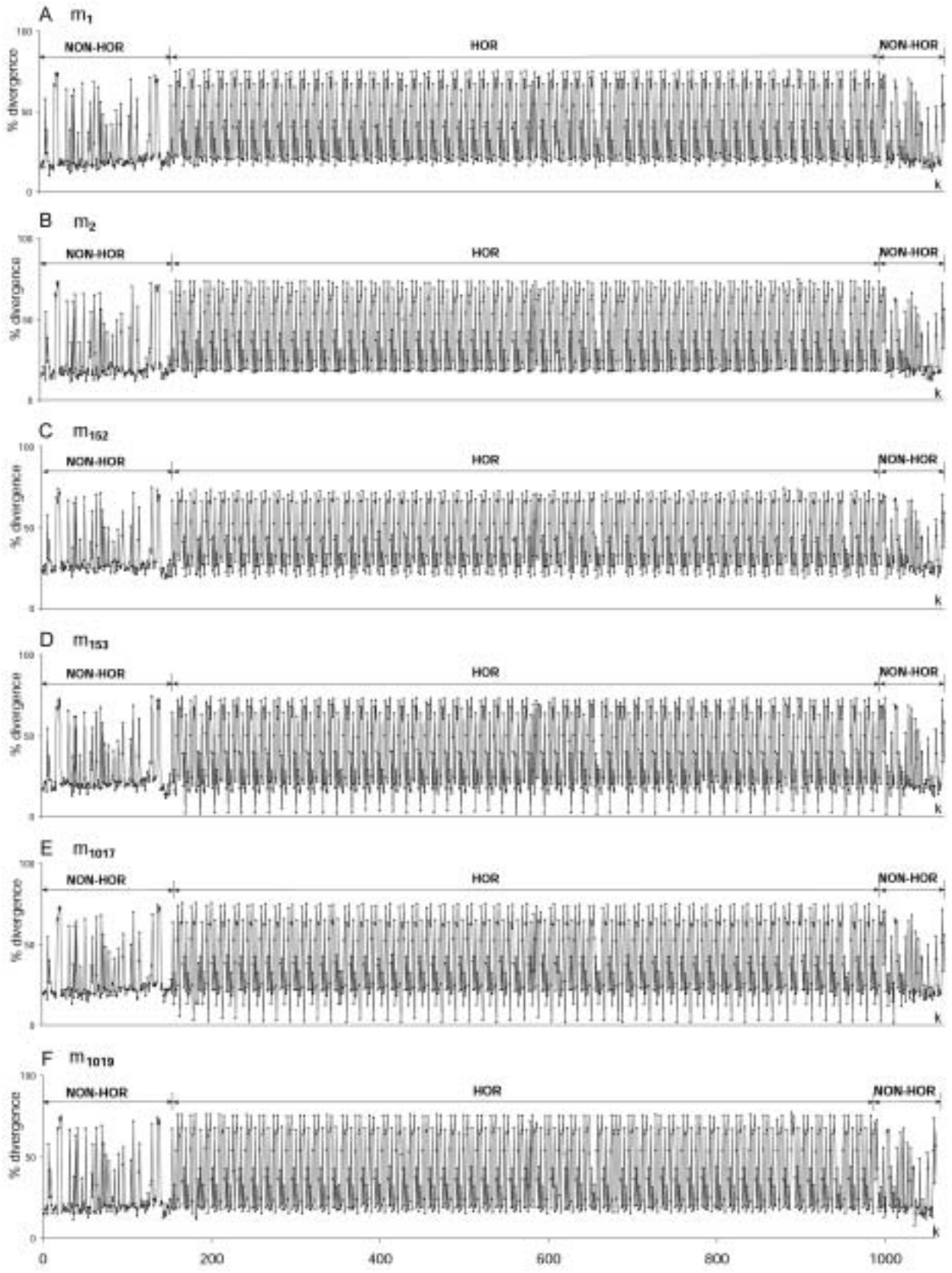

Figure 17. KSA minimum divergence-diagrams for genetic sequence AC017175.8 with key string AAAC. Referent subsequences: **A.** m_1, **B.** m_2, **C.** m_{152}, **D.** m_{153}, **E.** m_{1017}, and **F.** m_{1019}. **D** and **E**: HORs-divergence drops to less than 1%.

exhibits strongly pronounced periodic minima (at divergence less than 1 %) at every *n*th (16th in our case study) position in the periodic domain, identifying the 16mer HORs.

KSA Minimum-divergence Diagram

Each divergence diagram for each referent monomer corresponds to a row in the divergence matrix. Now we assign to each of *N* referent subsequences a minimum value of its divergence with respect to all other subsequences. We display graphically these minimum values of divergence along the divergence array, depending on the enumerator *k* of referent subsequences. For example, if the minimum value of all divergences in the divergence diagram for the *k*th referent subsequence is *d*, then the corresponding point in the KSA minimum-divergence diagram is a point with the horizontal coordinate *k* and vertical coordinate *d*.

Finally, to reduce local fluctuations in the above diagram, we perform an additional averaging in this diagram, as follows:

Let us consider a minimum divergence *d* that corresponds to the *k*th referent subsequence m_k. We form an interval of referent subsequences for 10 neighboring integers smaller than *k* (ie, for *k*-10, *k*-9, ... *k*-1) and for 10 neighboring integers larger than *k* (ie, for $k+1$, $k+2$, ... $k+10$) (Fig. 18). If there is at least one divergence less than 3% in the intervals on both sides of *k*, then we assign 0 instead of *d* to the *k*th referent subsequence. Otherwise, the computed value *d* is left unchanged. A diagram obtained in this way is referred to as the KSA minimum-divergence diagram (Fig. 19). Thus, in the KSA minimum-divergence diagram, the domain of HORs is directly identified as the interval with a straight horizontal line along the horizontal axis.

Figure 18. Interval for averaging procedure at the position *k* bp in a given genomic sequence.

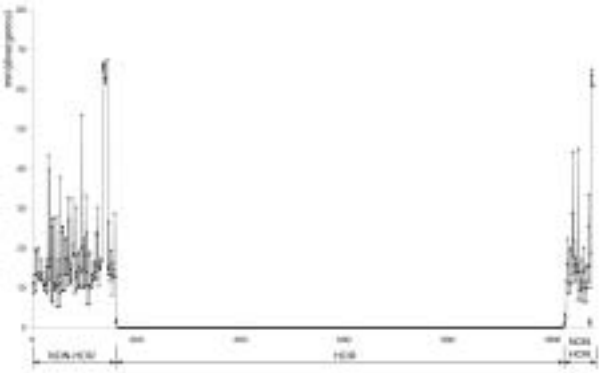

Figure 19. Key-string algorithm minimum-divergence diagram for AC017075.8 with key string AAAC.

171-bp Subsequence Convergence-diagram for Identification of Higher-order Repeats

Each monomer copy within a given genomic sequence diverges from other monomer copies in that sequence by 20-40%, except for selected alpha monomers within HORs. Each alpha monomer in one HOR copy differs from the corresponding alpha monomer in the other HOR copy by less than 1%: the first alpha monomers in all HOR copies are convergent to each other, the second monomers in all HOR copies are convergent to each other... (except for those having large insertions or deletions). Exploiting this feature, we have devised the following novel graphical method to identify HORs in a given genomic sequence.

First step. An array of 171-bp subsequences is formed: subsequence No. 1 – from position 1 to 171 in genomic sequence; subsequence No. 2 – from position 2 to 172 in genomic sequence; subsequence No. 3 – from position 3 to 173 in genomic sequence; and thus all the way to the end of the given genomic sequence. In the case study of AC017075.8, the last 171-bp subsequence is No. 193106 – from the position 193106 to 193277. In our case study, we formed a set of 193106 subsequences: (1-171), (2-172), (3-173),... (193106-193277), with, for example, (1-171) denoting the subsequence from the position 1 bp to the position 171 bp.

Second step. We compute divergence of the No. 1 subsequence with all 171-bp subsequences from the set. Then we assign to the No. 1 subsequence the enumerators of those subsequences that diverge from the No.1 subsequence less than 3%. In our case study, the No.1 subsequence has a divergence smaller than 3% only with itself, ie, to the No. 1 we assign only the No. 1. Thus, to the subsequence No. 1, we assign the pair of numbers $(k,l) = (1,1)$. A similar situation appears for all 171-bp subsequences within the front (ie, non-HOR) domain of AC01 7075.8: No. 2 is assigned to No. 2, No. 3 to No. 3, and so on. Thus, to No. 2 subsequence, we assign the pair of numbers $(k,l) = (2,2)$, to No. 3 subsequence we assign $(k,l) = (3,3)$, and so on. First, with the onset of the HOR-domain, the situation changes; for example, to the subsequence No. 34000, we assign subsequences No. 34000, $34000 + 2734 = 36734$, $34000 + 2\times2734 = 39468$, ... In this way, we have the corresponding pairs of numbers $(k,l) = (34000, 34000)$, (34000, 36734), (34000, 39468), and so on. Thus, in the HOR domain we assign several pairs of numbers $(k,l) = (k,k)$, $(k,k+2734)$, $(k,k+2\times2734)$, and so on, to each ordering number *k*.

Third step. Every pair of numbers from the second step are displayed graphically: the first number of the pair is taken as a coordinate on the horizontal axis and the second as a coordinate on the vertical axis (Fig. 20). Outside of the HOR domain, we obtain the points on a single monotonically increasing line, while in the HOR domain we obtain a series of parallel lines due to convergence between every 16th monomer. Some irregularities in the pattern of parallel lines reflect larger insertions and deletions in the HOR sequence.

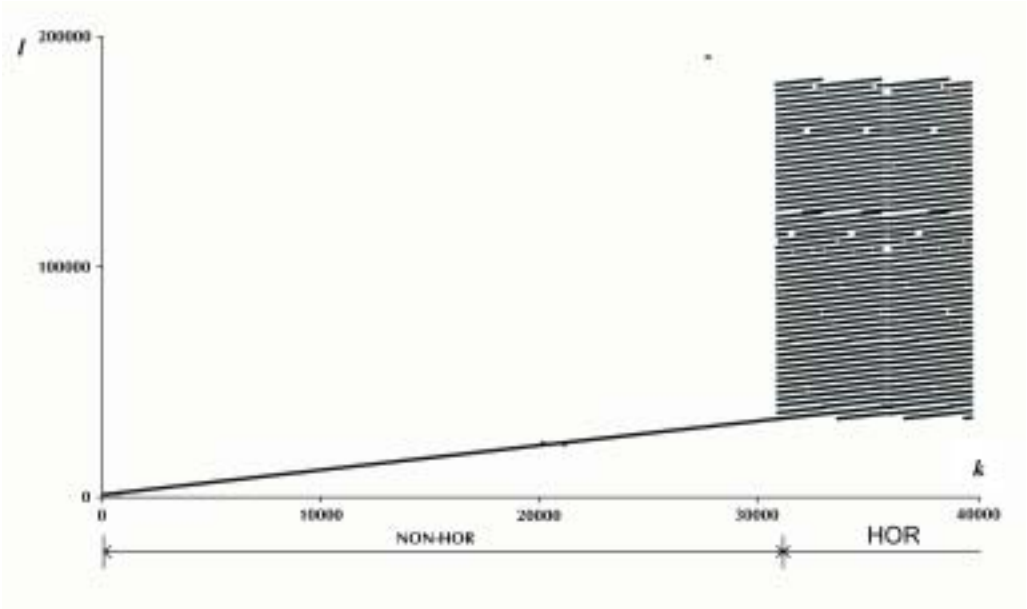

Figure 20. 171-bp subsequence convergence-diagram for AC017075.8. Horizontal axis: position *k* of the beginning of the 171-bp subsequence in AC017075.8. Vertical axis: position *l* assigned to *k* according to procedure described in text. Parallel lines are signature of the higher-order repeat domain.

Such convergence diagram provides a straightforward identification of HOR domain, as demonstrated in the case study for AC017075.8.

Total Frequency Distribution of the Key-string Subsequence Lengths

Let us now consider the set of all strings consisting of *r* bases. For example, for $r = 2$ there are 4^2, or 16, strings (AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, and TT). In general, for any *r* there are 4^r strings of *r* base pairs. In the case of equal probability of each *r*-bp string within a genomic sequence, the average distance between the two equal strings in a random distribution model is 4^r base pairs. For example, the average distance between the two equal 2-bp strings under assumption of equal probability is $4^2 = 16$ bp. On the basis of computer simulations, we can argue that a convenient referent frequency distribution for a given size of genomic sequence can be conveniently described by considering the set of all *r*-bp key strings, if the number 4^r is comparable to the HOR size.

For a convenient size *r* of the key strings, we compute in a given genomic sequence all KSA-subsequence lengths for all 4^r key strings with *r* basepairs. By superposing all subsequences, the total length frequencies are computed.

In the case study of the genomic sequence AC017075.8, a convenient referent size of the set of key strings for analysis of total length distribution is $r = 5$. Then, the total number of all *r*-bp key strings in the case of equal probability distribution is $4^5 = 1024$. Under the assumption of equal probability distribution, the expected average total frequency of appearance of two identical 5-bp strings is $193277/1024 = 190$. For the realistic genomic sequence AC017075.8, we will consider those total 5-bp subsequence lengths with frequency larger than five average frequencies, ie, larger than about 1000. In that case, one can expect strongly enhanced frequencies associated with alpha monomers and HORs.

Let us now investigate segmentations of the AC017075 using all 5-bp key strings. For each key string with five base pairs, we determine the corresponding KSA subsequence lengths. In each array of lengths, we determine computationally the frequency of appearance of each length. Then we perform summation of frequency of appearance of each value of length in subsequence arrays for all possible strings of five base pairs. Thus we obtain the total frequency distribution of lengths in all key-string subsequences. The most frequent length, with frequency of appearance greater than 1000, is 171 bp (frequency 29329), followed by 170 bp (10773), and 172 bp (5068) (Table 5). These three approximately 171-bp frequencies correspond to the basic alpha monomer.

Several other computed high-frequency lengths are those that can be segmented into *n*mers as follows:

341 (= 171 +170),

342 (= 171 +171),

512 (=170 + 2×171),

343 (= 171 + 172),

685 (= 3×171 + 172),

514 (= 2×171 + 172),

856 (= 4×171 + 172),

853 (= 3×171 + 2×170)...

The length 2734 bp, at high frequency of appearance (7323), corresponds to the 16mer HOR (10× 171 + 4×170 + 2×172).

On the other hand, inspection of shorter KSA subsequences, which are the segments of alpha mo-

Table 5. Total frequency distribution of key string subsequence lengths for 5-bp key strings in AC017075.8

Frequency of appearance	Length (bp)
29329	171
10773	170
7902	341
7323	2734
5640	342
5068	172
2322	512
2188	14
1866	343
1549	27
1446	29
1433	685
1096	514
1071	12
1070	46
1066	144
1066	856
1055	853

nomer, provide an indication for internal composition of approximately 171-bp alpha monomer. Short segments of high frequency presented in Table 5 are 14-bp, 27-bp, 29-bp, 12-bp, and 46-bp. They can be combined into three sequences of approximately 57 bp:

27 bp + 29 bp = 56 bp,

14 bp + 14 bp + 29 bp = 57 bp, and

12 bp + 46 bp = 58 bp,

which can in turn serve as building blocks for construction of approximately 171-bp alpha monomer:

171 bp = 56 bp + 57 bp + 58 bp,

170 bp = 56 bp + 56 bp + 58 bp, and

172 bp = 57 bp + 57 bp + 58 bp.

Thus, computed total frequencies of KSA subsequences reveal the presence of alpha monomer and HORs in a straightforward way, as well as a possible internal structure of alpha monomer.

Fast Fourier Transform Applied to Repetitive Sequences in Centromeric DNA

Fast Fourier analysis, which can identify repeats of certain segments of the same length in nucleotide sequences, was applied to search for hidden periodicities in DNA sequences and in studying long-range correlations. Previous investigations using Fourier analyses have been directed to exons and introns, whereas no specific study of alpha satellite DNA has been reported (61-65). We applied Fourier analyses to alpha satellite DNA sequences and investigated the associated periodicities and correlations. We first mapped the genomic sequence into a sequence of numbers by prescription A→4, T→3, C→2, G→1 (66). Our goal was to reveal periodicities in the short-, medium-, and long-range order, related both to monomer and higher-order repeats in human alpha satellite DNA. We wanted to compare Fourier analysis and KSA computational method in determination of periodicities by including alpha satellites and HOR in the genomic sequence AC017075.8.

As seen from the power spectrum in Figure 21a for the central (HOR) domain the prominent peak at the frequency $f_{16} = 0.005852$ bp^{-1} corresponds to the approximately 171-bp length (more precisely $1/f_{16} = 170.88$ bp), which is the alpha satellite monomer. The lowest frequency f_1 (spectral density $S_f = 2.71$) corresponds to the 2734-bp HOR having the role of fundamental frequency. We obtain:

$$f_1 = \frac{1}{16} f_{16}$$

A prominent difference of the front- and back-domain with respect to the central-domain in AC017075.8 is in the low-frequency region: there are no prominent peaks below the frequency corresponding to the 171-bp monomer. The peak at the frequency f_1 that would correspond to the 16-mer HOR is absent in the front- and back-domains (Fig. 21b), as well as those peaks that would correspond to its multiples nf_1 ($n = 1, 2, 3,\ldots 15$) below the frequency of the 171-bp monomer. This shows that the higher-order repeat is absent in the front- and back-domains, which is in accordance with results of the KSA method.

On the other hand, the equidistant pattern associated with prominent peaks above the frequency f_{16} (corresponding to the 171-bp monomer) in the front-

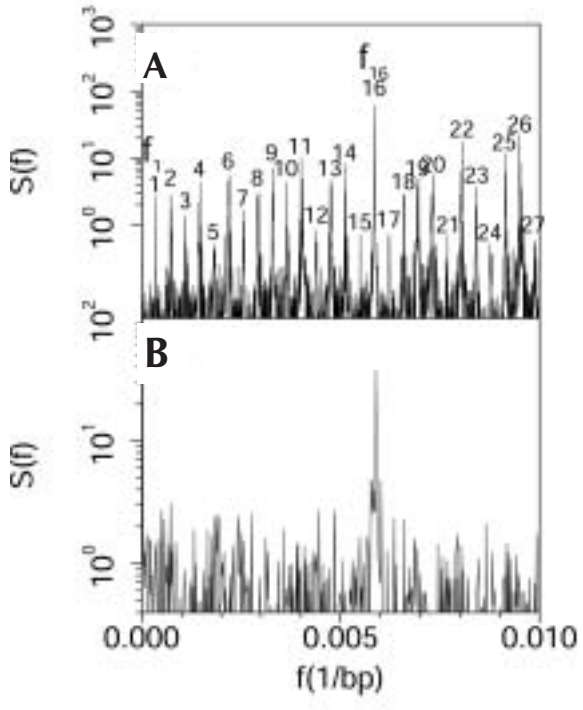

Figure 21. A. Power spectrum of the central domain of AC017075.8 obtained by Fast Fourier transform by using quartic mapping (66). The peak f1 corresponds to the 2734-higher-order repeat and the peak f16 to the approximately 171-bp monomer. **B.** Power spectrum of the front- and back-domain of AC017075.8 obtained by Fast Fourier transform computation using quartic mapping (66). The only pronounced peak corresponds to the 171-bp alpha satellite.

and back- domains is qualitatively similar to the central-domain (66).

On the basis of Fourier spectrum we have subdivided the 171-bp monomer unit into three approximately 56-bp subrepeats that were further subdivided into 12-bp, 14-bp, 17-bp, and 13-bp basic subrepeats. Taking into account segmentations:

13 bp + 14 bp = 27 bp,

12 bp + 17 bp = 29 bp,

12 bp + 17 bp + 17 bp = 46 bp,

this is consistent with the former KSA-based considerations.

All mutations, insertions, and deletions imposed on the ideal HOR structure (consensus HOR) have only a minor impact on the multiple-frequency pattern. Fourier transform provides a global method for identification HORs and internal monomer structures, insensitive to moderate deviations of periodicity. An exact analysis determining in details all mutations, deletions and insertions can be performed using the KSA method.

Computational Methods – List of Key-string Algorithm Analysis Software

We developed DNA sequence analysis software in programming language C, which performs the KSA analysis of a given genomic sequence for simple and higher-order repeats of any length. The use of software combines computations of small to moderate extent and straightforward naked-eye inspection. The software can be downloaded from our website *www.hazu.hr/KSA/,* and contains the following programs introduced, exemplified, and discussed in this paper:

www.hazu.hr/KSA/segmentation,
www.hazu.hr/KSA/HORmatrix,
www.hazu.hr/KSA/colorHOR
www.hazu.hr/KSA/divergence,
www.hazu.hr/KSA/mindivergence,
www.hazu.hr/KSA/subseqconvergence,
www.hazu.hr/KSA/frequency,
www.hazu.hr/KSA/HORFourier.

Conclusion

We introduced a novel algorithm for the analysis of non-coding genomic sequences, aimed primarily at identification and exact analysis of long repeats in anonymous sequences without any prior knowledge of a period or pattern. This is a simple and powerful method to identify very long repeats, for example, the alpha satellite HORs or long interspersed repeats of the period of tens of thousands of base pairs. Our basic idea was an effort to construct a computational method that corresponds to a "computationally simulated enzyme".

Thus, our KSA method is a HOR-oriented program, which makes it much simpler to identify HORs than the alpha monomer itself. Namely, almost any choice of the key string leads to a straightforward identification of HORs and insertions, deletions, and point mutations in its structure, while only some selected key strings reveal the underlying alpha monomer structure.

The KSA method exemplified in this case study of complete genomic sequence AC017075.8 from the GenBank database gave a full identification of 55 2734-bp HORs and determination of consensus and respective insertions, deletions, and point mutations. Furthermore, we showed the advantage of KSA with respect to the standard repeats-finding programs, RepeatMasker and Tandem Repeat Finder, as these standard programs were not able to identify HORs. However, the KSA is not a fully automated program: only the combination of modest-scale computation and visual inspection of computed results, which allows a high degree of flexibility in identification of the pattern and deviations, providing an efficient intertwining of computer calculation and human pattern recognition. Yet the method is simple and convenient to be used by "pedestrians" with only a moderate computational skills. We are in the process of the KSA analysis of centromeric regions of all human chromosomes. In addition, we preliminary identified several families of interspersed, intertwined very long repetitions (up to hundred thousands bp long) in the chromosome Y.

Acknowledgments

We thank S. Durajlija Žinić, Đ. Ugarković, B. Nagy, S. Popović, and I. Weygand Đurašević for discussions.

References

1 Sullivan BA, Blower MD, Karpen GH. Determining centromere identity: cyclical stories and forking paths. Nat Rev Genet 2001;2:584-96.

2 Pluta AF, Mackay AM, Ainsztein AM, Goldberg IG, Earnshaw WC. The centromere: hub of chromosomal activities. Science 1995;270:1591-4.

3 Clarke L, Carbon J. The structure and function of yeast centromeres. Annu Rev Genet 1985;19:29-55.

4 Clarke L. Centromeres of budding and fission yeast. Trends Genet 1990;6:150-4.

5 Murphy TA, Karpen GH. Localization of centromere function in a Drosophila minichromosome. Cell 1995; 82:599-609.

6 Sun X, Wahlstrom J, Karpen G. Molecular structure of a functional Drosophila centromere. Cell 1997;91: 1007-19.

7 Miklos GL. Localized highly repetitive DNA sequences in vertebrate and invertebrate genomes. In: MacIntyre JR, editor. Molecular evolutionary genetics. New York: Plenum Publishing Corp; 1985. p. 241-313.

8 Warburton PE, Haaf T, Gosden J, Lawson D, Willard HF. Characterization of a chromosome-specific chimpanzee alpha satellite subset-evolutionary relationship to subsets on human chromosomes. Genomics 1996; 33:220-8.

9 Manuelidis L. Complex and simple sequences in human repeated DNAs. Chromosoma 1978;66:1-21.

10 Manuelidis L. Chromosomal localization of complex and simple repeated human DNAs. Chromosoma 1978;66:23-32.

11 Mitchell AR, Gosden JR, Miller DA. A cloned sequence, p82H, of alphoid repeated DNA found at the centro-

meres of all human chromosomes. Chromosoma 1985; 92:369-77.

12 Yan J, Zhang XX, Fetni R, Drouin R. Trisomy 8 and monosomy 7 detected in bone marrow using primed in situ labeling, fluorescence in situ hybridization, and conventional cytogenetic analyses. A study of 54 cases with hematological disorders. Cancer Genet Cytogenet 2001;125:30-40.

13 Waye JS, England SB, Willard HF. Genomic organization of alpha satellite DNA on human chromosome 7: evidence for two distinct alphoid domains on a single chromosome. Mol Cell Biol 1987;7:349-56.

14 Amare Kadam PS, Varghese C, Bharde SH, Narasimhamoorthy NK, Desai S, Advani SH, et al. Proliferating cell nuclear antigen and epidermal growth factor receptor (EGFr) status in renal cell carcinoma patients with polysomy of chromosome 7. Cancer Genet Cytogenet 2001;125:139-46.

15 Wevrick R, Willard HF. Physical map of the centromeric region of human chromosome 7: relationship between two distinct alpha satellite arrays. Nucleic Acids Res 1991;19:2295-301.

16 Cooper KF, Fisher RB, Tyler-Smith C. Structure of the pericentric long arm region of the human Y chromosome. J Mol Biol 1992;228:421-32.

17 Jackson MS, Slijepčević P, Ponder BA. The organization of repetitive sequences in the pericentromeric region of human chromosome 10. Nucleic Acids Res 1993;21: 5865-74.

18 Trowell HE, Nagy A, Vissel B, Choo KH. Long-range analyses of the centromeric regions of human chromosomes 13, 14 and 21: identification of a narrow domain containing two key centromeric DNA elements. Hum Mol Genet 1993;2:1639-49.

19 Heiskanen M, Kallioniemi O, Palotie A. Fiber-FISH: experiences and a refined protocol. Genet Anal 1996;12: 179-84.

20 Heiskanen M, Hellsten E, Kallioniemi OP, Makela TP, Alitalo K, Peltonen L, et al. Visual mapping by fiber-FISH. Genomics 1995;30:31-6.

21 Rosandić M, Paar V, Basar I. Key-string segmentation algorithm and higher-order repeat 16mer (54 copies) in human alpha satellite DNA in chromosome 7. J Theor Biol 2003;221:29-37.

22 GenBank. Accession No. AC017075.8. Bethesda (MD): National Institute of Health; 2002.

23 Willard HF, Waye JS. Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. J Mol Evol 1987;25:207-14.

24 Wevrick R, Willard VP, Willard HF. Structure of DNA near long tandem arrays of alpha satellite DNA at the centromere of human chromosome 7. Genomics 1992;14:912-23.

25 Waye JS, Willard HF. Nucleotide sequence heterogeneity of alpha satellite repetitive DNA: a survey of alphoid sequences from different human chromosomes. Nucleic Acids Res 1987;15:7549-69.

26 Willard HF. Centromeres of mammalian chromosomes. Trends Genet 1990;6:410-6.

27 Vogt P. Potential genetic functions of tandem repeated DNA sequence blocks in the human genome are based on highly conserved "chromatin folding code". Hum Genet 1990;84:301-36.

28 Lee C, Wevrick R, Fisher RB, Ferguson-Smith MA, Lin CC. Human centromeric DNAs. Hum Genet 1997;100; 291-304.

29 Choo KH, Vissel B, Nagy A, Earle E, Kalitsis P. A survey of the genomic distribution of alpha satellite on all the human chromosomes, and derivation of a new consensus sequence. Nucleic Acids Res 1991;19:1179-82.

30 Kurtz S. Reducing the space requirement of suffix trees. Software Practice and Experience 1999;29:1149-71.

31 Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. REPuter: the manifold applications of repeat analysis on genomic scale. Nucleic Acids Res 2001;29:4633-42.

32 Kolpakov R, Kucherov G. Finding approximate repetitions under Hamming distance. In: auf der Heide FM. Proceedings of the 9th Annual European Symposium on Algorithms (ESA 2001); 2001 Aug 28-31; Aarhus, Denmark. Lecture Notes in Computer Science 2001; 2161:170-81.

33 Giraud M, Kucherov G. Maximal repetitions and applications to DNA sequences. Journées Ouvertes: Biologie, Informatique et Mathématiques; Montpellier, May 3-5, 2000; p.165-72.

34 Stoye J, Gusfield D. Simple and flexible detection of contiguous repeats using a suffix tree. Theor Comput Sci 2002;270:843-56.

35 Delgrange O, Dauchet M, Rivals E. Location of repetitive regions in sequences by optimizing a compression method. In: Altman R, editor. Proceedings of the 4th Pacific Symposium on Biocomputing; 1999 Jan 4-9; Hawai, USA. PBS 99 On-line Proceedings. Available at: *http://www-smi.stanford.edu/projects/helix/psb99/*. Accessed: May 30, 2003.

36 Matsumoto T, Sadakane K, Imai H. Biological sequence compression algorithms. Genome Inform Ser Workshop Genome Inform 2000;11:43-52.

37 Chen X, Kwong S, Li M. A compression algorithm for DNA sequences and its applications in genome comparison. Genome Inform Ser Workshop Genome Inform 1999;10:51-61.

38 Jurka J, Klonowski P, Dagman V, Pelton P. CENSOR – a program for identification and elimination of repetitive elements from DNA sequences. Comput Chem 1996; 20:119-21.

39 Jurka J. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 2000;16: 418-20.

40 Landau GM, Schmidt JP, Sokol D. An algorithm for approximate tandem repeats. J Comput Biol 2001;8:1-18.

41 Volfovsky N, Haas BJ, Salzberg SL. A clustering method for repeat analysis in DNA sequences. Genome Biol 2001;2:research0027.

42 Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC. Cross-species sequence comparisons: a review of methods and available resources. Genome Res 2003; 13:1-12.

43 Pearson WJ, Lipman DR. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988; 85:2444-8.

44 Altschul SF, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. J Mol Biol 1990;215: 403-10.

45 Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, et al. Human-mouse alignments with BLASTZ. Genome Res 2003;13:103-7.

46 Bray N, Dubchak I, Pachter L. AVID: a global alignment program. Genome Res 2003;13:97-102.

47 Smit AF, Green P. RepeatMasker documentation. Available from: *http://ftp.genome.washington.edu/RM/RepeatMasker.html*. Accessed: June 15, 2003.

48 Agarwal P, States DJ. The repeat pattern toolkit (RPT): analyzing the structure and evolution of the C. elegans genome. Proc Int Conf Intell Syst Mol Biol 1994;2:1-9.

49 Bedell JA, Korf I, Gish W. MaskerAid: a performance enhancement to RepeatMasker. Bioinformatics 2000; 16:1040-1.

50 Sagot MF, Myers EW. Identifying satellites and repetitions in biological sequences. J Comput Biol 1998;5: 539-53.

51 Benson G, Su X. On the distribution of the k-tuple matches for sequence homology: a constant time exact calculation of the variance. J Comput Biol 1998;5: 87-100.

52 Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999;27:573-80.

53 Klaerr-Blanchard M, Chiapello H, Coward E. Detecting localized repeats in genomic sequences: a new strategy and its application to Bacillus subtilis and Arabidopsis thaliana sequences. Comput Chem 2000;24:57-70.

54 Hauth AM. Identification of tandem repeats: simple and complex pattern structures in DNA sequences [dissertation]. Madison (WI): University of Wisconsin; 2002.

55 Hauth AM, Joseph DA. Beyond tandem repeats: complex pattern structures and distant regions of similarity. Bioinformatics 2002;18 Suppl 1:31-7.

56 Castelo AT, Martins W, Gao GR. TROLL – tandem repeat occurrence locator. Bioinformatics 2002;18:634-6.

57 Vafa O, Sullivan KF. Chromatin containing CENP-A and alpha-satellite DNA is a major component of the inner kinetochore plate. Curr Biol 1997;7:897-900.

58 Blower MD, Sullivan BA, Karpen GH. Conserved organization of centromeric chromatin in flies and humans. Dev Cell 2002;2:319-30.

59 Yoda K, Okazaki T. Site-specific base deletions in human alpha-satellite monomer DNAs are associated with regularly distributed CENP-B boxes. Chromosome Res 1997;5:207-11.

60 Tanaka Y, Nureki O, Kurumizaka H, Fukai S, Kawaguchi S, Ikuta M, et al. Crystal structure of the CENP-B protein-DNA complex: the DNA-binding domains CENP-B induce kinks in the CENP-B box DNA. EMBO J 2001;20:6612-8.

61 Voss RF. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. Phys Rev Lett 1992;68:3805-8.

62 Tiwari S, Ramachandran S, Bhattacharya A, Bhattacharya S, Ramaswamy R. Prediction of probable genes by Fourier analysis of genomic sequences. Comput Appl Biosci 1997;13:263-70.

63 Guharay S, Hunt BR, York JA, White OR. Correlations in DNA sequences across the three domains of life. Physica D 2000;146:388-96.

64 Jackson JH, George R, Herring PA. Vectors of Shannon information from Fourier signals characterizing base periodicity in genes and genomes. Biochem Biophys Res Commun 2000;268:289-92.

65 Nagai N, Kuwata K, Hayashi T, Kuwata H, Era S. Evolution of the periodicity and the self-similarity in DNA sequence: a Fourier transform analysis. Jpn J Physiol 2001;51:159-68.

66 Paar V, Pavin N, Rosandić M, Basar I, Luketin I. Spectral densities and frequencies in power spectrum of higher order repeat alpha satellite in human DNA molecule. Croat Chem Acta. In press 2004.

Received: June 10, 2003
Accepted: July 2, 2003

Correspondence to:
Vladimir Paar
Zagreb University Faculty of Science
Bijenička 32
10000 Zagreb, Croatia
paar@hazu.hr