Available online at www.sciencedirect.com
ScienceDirect
Mitochondrion 6 (2006) 218–224
Mitochondrion
www.elsevier.com/locate/mito

Short communication

RepeatAround: A software tool for finding and visualizing repeats in circular genomes and its application to a human mtDNA database

Ana Goios [a,b], José Meirinhos [a], Ricardo Rocha [c,d], Ricardo Lopes [c,d], António Amorim [a,b], Luísa Pereira [a,*]

[a] *IPATIMUP (Instituto de Patologia e Imunologia Molecular da Universidade do Porto), R. Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal*
[b] *Faculdade de Ciências da Universidade do Porto, Porto, Portugal*
[c] *LIACC (Laboratório de Inteligência Artificial e Ciências de Computadores), Portugal*
[d] *Departamento de Ciências de Computadores, Faculdade de Ciências da Universidade do Porto, Portugal*

Received 22 March 2006; received in revised form 9 May 2006; accepted 7 June 2006
Available online 14 June 2006

Abstract

RepeatAround is a Windows based software tool designed to find "direct repeats", "inverted repeats", "mirror repeats" and "complementary repeats", from 3 to 64 bp length, in circular genomes. It processes input files directly extracted from GenBank database, providing visualisation of the repeats location in the genomic structure, so that for instance, in most mtDNAs the user can check if the repeats are located in coding or non-coding region (and in the first case in which gene), and how far apart the repeat pair(s) are. Besides the visual tool, it provides other outputs in a spreadsheet containing information on the number and location of the repeats, facilitating graphic analyses. Several genomes can be inputed simultaneously, for phylogenetic comparison purposes. Other capabilities of the software are the generation of random circular genomes, for statistical evaluation of comparison between observed repeats distributions with their shuffled counterparts, as well as the search for specific motifs, allowing an easy confirmation of repeats flanking a newly detected rearrangement. As an example of the programme's applications we analysed the Direct Repeats distribution in a large human mtDNA database. Results showed that Direct Repeats, even the larger ones, are evenly distributed among the human mtDNA haplogroups, enabling us to state that, based only on the repetitive motifs, no haplogroup is particularly more or less prone to mtDNA macrodeletions.

Keywords: Direct repeats; Inverted repeats; Mirror repeats; Complementary repeats; Circular genomes

1. Introduction

Genomes are interspersed by repeated sequence motifs, which can be classified in four types: direct (e.g. AGTTC/AGTTC), inverted (AGTTC/GAACT), mirror (AGTTC/CTTGA) and complementary (AGTTC/TCAA G). These repeated motifs are potential places for the occurrence of gross genome rearrangements, such as deletions and duplications, leading to a variety of malfunctions and diseases. This is the etiology of many human pathologies, both in mitochondrial DNA (Brandon et al., 2005; Samuels et al., 2004) and in the nuclear genome, most of them somatic, and sometimes leading to cancer (Chuzhanova et al., 2003). Recently, Samuels (2004) showed that life span of mammalian species can be constrained by the size of the longest direct repeats present in their mitochondrial genome.

The high number of available softwares for repeat finding [e.g. *REPuter* (Kurtz and Schleiermacher, 1999); GCG-package (http://www.accelrys.com); DnkSet_Demo (http://www.dnkset.com)] also testifies the importance of these repeats for genomes. Nevertheless, some of these packages characteristics do limit its application for repeat finding in mtDNA genomes: (1) many do not work in

* Corresponding author. Tel.: +351 225570700; fax: +351 225570799.
E-mail address: lpereira@ipatimup.pt (L. Pereira).

1567-7249/$ - see front matter
doi:10.1016/j.mito.2006.06.001

Windows platform (e.g. GCG-package), extensively limiting its use by geneticists, (2) do not take into account that sub-motifs of larger repeats overestimate genome redundancy (DnkSet_Demo), and (3) are devised for linear DNA (all of them), while many genomes, such as most mtDNAs are circular.

Here we present a software for repeat search in circular genomes, that does work in Windows platform, and that can be freely downloaded from http://www.ipatimup.pt/app/default.htm. To further encourage its use by geneticists, the input files can be downloaded directly from GenBank® database (http://www.ncbi.nlm.nih.gov/Genbank). The genetic information contained in the input file (genes and its location) is used to construct an image of the genome and the location of the repeated motifs found. But, if the user wants to analyse a sequence not yet submitted to GenBank, an option for inputting a text file is also provided.

2. Implementation

RepeatAround uses suffix trees for analyzing the sequence and finding the repeats or particular motifs. A suffix tree is a data-structure that allows many string (sequences of characters) problems to be solved very efficiently.

A suffix tree can be defined as a data-structure that exposes the internal structure of a string (Gusfield, 1997). As the name implies, it is a tree representation of all the possible string suffixes within a text. Therefore each leaf and node of the tree is associated with a suffix in the text, and the path from the root to a node or leaf represents a substring of the text common to all suffixes of the respective node/leaf. Then if one of this substrings has several different suffixes, this substring is clearly repeated in the text. So, it might be one of the repeats we are looking for. Often, these repeats occur as a substring of another larger repeat, and so, these should not be accounted for, as they would overestimate the genome redundancy.

We define a pair of a repeated string as two separate occurrences of this string whose both left and right characters are different. For example, for a text *x**s**y**w**s**y**w**s**z*, where ***s*** is a string and *x*, *y*, *w*, *z* are characters, the direct repeat pairs are ***s**y* (the first and second occurrences of ***s***), *w**s*** (the second and third occurrences of ***s***), and ***s*** at the first and third occurrences. The ***s*** pair is considered just once, the only instance where it does not occur as a substring of a longer repeated string.

The computation of the *inverted*, *mirror* and *complementary* repeats is similar. First the entire sequence is inserted in the suffix tree, but altered to match the desired repeats. For example, the sequence is inverted and the bases complemented to find the *inverted* repeats, or just inverted to find the *mirror* repeats. Then the algorithm for locating the repeated pairs is basically the same as for *direct* repeats.

When working with DNA sequences, the implementation of the suffix trees is simpler, as there are only four different symbols in the sequence. Also, because this software only analyses circular genomes, there is no need to add a different symbol for the end of the sequence.

3. Program displayings

The minimum and maximum lengths of the repeats can be defined by the user, between 3 and 64 bp (by default, the program finds repeats between 8 and 30 bp). The sequences analysed can either be generated randomly (the user sets the proportion of each of the four DNA bases), loaded from a text file in GenBank® format, or even a personal .txt file containing a sequence not yet submitted to the database.

When a genome is loaded the program computes all the *direct*, *inverted*, *mirror* and *complementary* repeats (between the minimum and maximum lengths defined) and shows them in a tree control, on the left side of the window, sorted by type and length of the repeats (Fig. 1). Each line in the tree has a number inside parenthesis showing the number of pairs found for a given repeat type, repeat length or specific repeat. By selecting a specific repeat, all the occurrences are highlighted in the sequence (green for the *direct* repeats, and red for *inverted*, *mirror*, and *complementary* repeats).

The entire set of a single type of repeat pair, or the number of repeated pairs (by length) is presented in a Microsoft™ Excel spreadsheet [the user must have Microsoft™ Excel 9.0 (2000) or latter installed in their own computer for this functionality to work]. This makes it easier to analyse the results produced by the program, and operate on them subsequently.

The GenBank® files, besides the actual DNA sequence, also contain information about genes and gene products, as well as regions of biological significance reported in the sequence. With this information, RepeatAround constructs a diagram of the circular sequence showing these regions and the points where a selected repeat occurs (again, marked green for *direct* repeats and red for the others; Fig. 1).

In order to save time in phylogenetic surveys (as conducted in Samuels, 2004), it is possible to simultaneously compare several genomes, by using the option "Compare Genomes". This tool produces a summary table of the number of direct repeats present in each one of the genomes being compared.

Finally, it is also possible to search for a specific subsequence (such as a small motif or a primer). This last feature allows an easy investigation on whether the limits flanking a certain newly detected rearrangement (which are continuously being discovered; e.g. Goios et al., 2005) are classifiable as repeats, and if so, its genetic location in the genome.

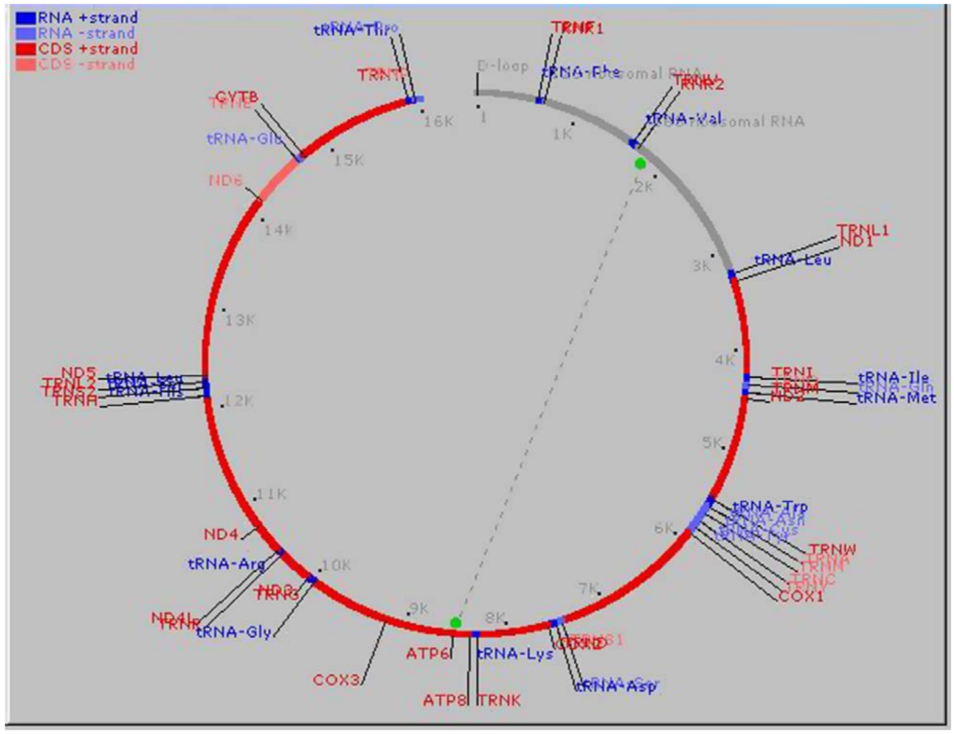

Fig. 1. Diagram output from RepeatAround when selecting the 12 bp DR (motif AAAAATTATAAC) detected in the human mitochondrial genome deposited in the GenBank with Accession Number NC_001807.

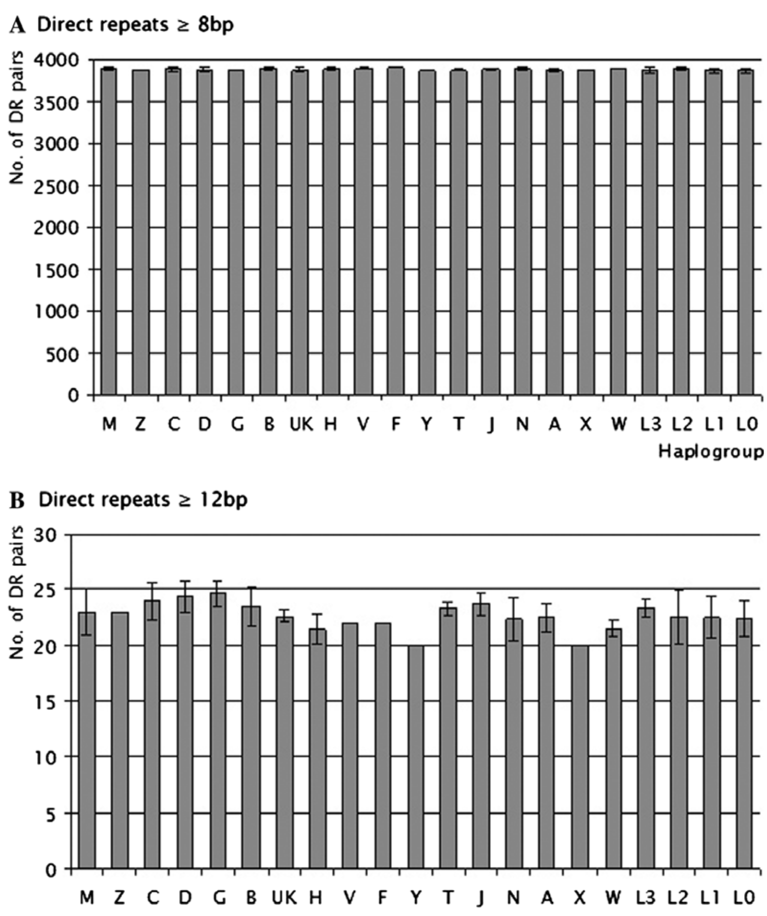

Fig. 2. Average number of direct repeat pairs in sequences of the different human haplogroups. Standard deviation is shown by the whisker around the top of each bar.

Table 1
The longer DR in the 102 human mtDNA sequences from diverse haplogroups (Hap)

		Frequent Direct Repeats									Frequent Direct Repeats						
Acc. No.	Hap.	15bp CAAACTCAA ACTACG	13bp ACCTCCCTC ACCA	13bp TCTATCACC CTAT	13bp CTCAACACC CACT	13bp CCCATACCC CGAA	13bp GTACATAGC ACAT	Sporadic Direct Repeats	Acc. No.	Hap.	15bp CAAACTCA AACTACG	13bp ACCTCCCT CACCA	13bp TCTATCAC CCTAT	13bp CTCAACAC CCACT	13bp CCCATACC CCGAA	13bp GTACATAG CACAT	Sporadic Direct Repeats
AF346971	A								AF347009	L0							
AY195786	A								AY195777	L0							
AY195771	A								AY195780	L0							
AY195760	A								AF346987	L1							14bp CCTAGACCAAACCT
AF346993	B								AF346986	L1							
AY195770	B								AF346992	L1							14bp CCTAGACCAAACCT
AY195749	B							13 bp CAAACCCCCCCC	AF346968	L1							14bp CCTAGACCAAACCT
AF347001	B								AF346969	L1							14bp CCTAGACCAAACCT
AF347007	B								AF346996	L1							
AF347011	B								AF346997	L1							14bp CCTAGACCAAACCT
AY195759	C							13bp GCTTCATTCACTG	AY195789	L1							14bp CCTAGACCAAACCT
AY195772	C								AY195783	L1							
AY195763	C								AF346995	L2							15bp AACTCATACCCCCAT
AY195753	C								AF346976	L2							
AF346991	C								AF346977	L2							13bp CCCTACTACTATC
AF346979	C								AY195788	L2							
AF346970	C								AY195776	L2							
AF347012	C								AY195766	L2							15bp AACTCATACCCCCAT
AF347013	C								AY195785	L2							15bp AACTCATACCCCCAT
AF346984	D								AF346994	L3							
AF346989	D								AF346980	L3							
AF346990	D								AF346967	L3							
AY195790	D								AY195784	L3							
AY195748	D								AY195782	L3							14bp TCATCCTAGCCCTA

(continued on next page)

AF346973	D							13bp TCCCCTTGTAAAT	AF347014	L3							
AF347010	D								AF347015	L3							
AY195791	F								AF346965	M							13bp ACCAACAAACTTA
AF346972	F								AF347003	M							
AY195762	G							13bp AACCCCCCTCCCC	AY195756	N							
AY195755	G							13bp AACCCCCCTCCCC	AY195769	N							
AF346966	G							13bp CAAAACAAATGAT	AF346964	N							
AY195758	H								AF346963	N							
AY195751	H								AF347002	R							
AF346975	H								AF347005	R							13bp CAAACCCCCCCCC
AY195747	H								AF347004	R							
NC_001807	H								AF346982	T							
AY195757	H								AY195745	T							
AY195746	H								AY195767	T							
AY195752	H								AF346988	UK							
AY195775	H								AY195764	UK							16bp CCCCATGTCTAACAAC
AF346974	H								AY195765	UK							13bp ATTAACCACTCAC
AF346981	H								AF346978	V							
AF346983	J							13bp ATTAACCACTCAC	AY195781	V							
AY195778	J								AY195750	V							
AY195754	J							13bp ATTAACCACTCAC	AF347006	V							
AY195774	J							13bp ATTAACCACTCAC	AY195779	W							
AF347000	L								AY195768	W							
AF346985	L0								AY195787	X							
AF346998	L0								AY195773	X							
AF346999	L0								AY195792	Y							
AF347008	L0								AY195761	Z							

For the "Frequent Direct Repeats" the gray cells mean presence of DR and the white cells absence.

4. Evaluating the distribution of direct repeats (DR) in a human mtDNA database

We used RepeatAround to analyse 102 human mtDNA sequences (including the Cambridge Reference Sequence – CRS – Andrews et al., 1999) belonging to different haplogroups (Mishmar et al., 2003; Ingman et al. 2000) and one *Pan troglodytes* sequence (GenBank Accession No.: NC_001643), and to generate and survey 20 randomly shuffled sequences with the same base composition as the human and chimpanzee references. Our aims were to: (1) compare distributions of DR between human haplogroups; (2) evaluate if particular haplogroup defining polymorphisms do disrupt the longer DR or do create new ones; (3) measure if DR are evenly distributed or not in the mtDNA molecule.

The results of our analysis of the human reference and chimpanzee sequences compared to the respective random sequences agree with those from Samuels (2004). The total number of observed pairs of DRs equal to or longer than 8 bp was about 1.3 times higher than expected, and in all 20 randomisations, there were fewer DRs $\geqslant$ 8 bp than in the human sequences. This excess of DRs was observed for every size class analysed. And the same pattern was observed in all individuals, with the total number of DRs $\geqslant$ 8 bp varying only between 3813 and 3908 (mean 3873, s.d. 20.4). The number of repeated motifs is strongly conserved in the distinct mtDNA sequences. This pattern is still present if we analyse the sequences by haplogroup, i.e. there are no differences in DR distributions between haplogroups (Fig. 2A), even considering only the longer motifs $\geqslant$12 bp (Fig. 2B).

Focusing on the actual motifs $\geqslant$12 bp present in the 102 human sequences, we observed that most of them are common to all individuals: besides the 15 bp repeat that is common to almost all except one of the human sequences analysed, there are also 4 DRs 13 bp long that are strongly conserved, one of which corresponds to the "Common Deletion" breakpoints (Table 1). Some new DRs seem to be caused by sporadic polymorphisms since they are not associated to any specific haplogroup, appearing in just a few individuals. But two cases call our attention: the 14 bp repeat appearing in six of the nine L1 samples, which is generated by the transversion T14000A (in the CRS numbering), characteristic of the subhaplogroup L1c; and the 13 bp DR GTACATAGCACAT, where the hotspot position 16311 is responsible for the frequent disruption of the motif.

It is also important to stress that, when DRs are sources of rearrangements, the places where the two motifs of a pair occur is not irrelevant, since deletions encompassing different regions of the molecule will have diverse consequences. Deleted molecules involving DR pairs on opposite sides of the origin of replication of the light chain (O_L: np 5721–5798), will generate a molecule unable to replicate, since the O_L is removed. Therefore these kinds of deletions are unlikely to reach levels allowing their detection, or leading to pathological states.

In order to visualise better whether DRs are uniformly distributed throughout the molecule, or whether they encompass one of the origins of replication, we plotted the pairs of DRs on a chart (Fig. 3), where it can be seen that repeat pairs are quite dispersed through the mtDNA molecule. By marking the origins of replication (O_H coincides with the *Y* axis), we observe that the DR pairs are not significantly grouped with respect to O_L, and consequently they have not been selected on the basis of their presumable pathogenicity.

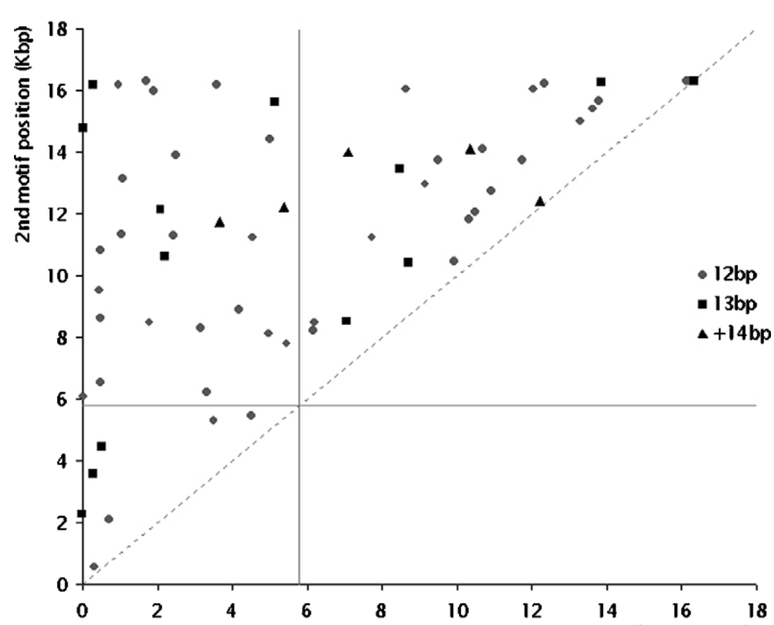

Fig. 3. Distribution of the direct repeat pairs equal to or longer than 12 bp on the mtDNA molecule. The vertical and horizontal lines show the origin of replication of the light chain (O_L, full lines).

5. Final remarks

RepeatAround is a user-friendly software, working in a Windows platform, with some useful tools for both practical and theoretical applications. The programme was successfully used to analyse raw data from a large database of human mtDNA sequences. This allowed the confirmation of the following important points: (1) all human mtDNA haplogroups show similar distributions of DRs; (2) the DRs are present at a higher number than expected at random; (3) DRs are evenly distributed along the molecule; and (4) considering only the DRs, no human mtDNA haplogroup is particularly more or less prone to mtDNA macrodeletions.

Acknowledgements

A.G. has a PhD grant (SFRH/BD/16518/2004) from Fundação para a Ciência e a Tecnologia. IPATIMUP is partially supported by Programa Operacional Ciência, Tecnologia e Inovação (POCTI), Quadro Comunitário de Apoio III.

References

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., Howell, N., 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat. Genet. 23, 147.

Brandon, M.C., Lott, M.T., Nguyen, K.C., Spolim, S., Navathe, S.B., Baldi, P., Wallace, D.C., 2005. MITOMAP: a human mitochondrial genome database–2004 update. Nucleic Acids Res. 33 (Database Issue), D611–D613.

Chuzhanova, N., Abeysinghe, S.S., Krawczak, M., Cooper, D.N., 2003. Translocation and gross deletion breakpoints in human inherited disease and cancer II: potential involvement of repetitive sequence elements in secondary structure formation between DNA ends. Hum. Mutat. 22, 245–251.

Goios, A., Nogueira, C., Pereira, C., Vilarinho, L., Amorim, A., Pereira, L., 2005. mtDNA single macrodeletions associated with myopathies: absence of haplogroup-related increased risk. J. Inherit. Metab. Dis. 28, 769–778.

Gusfield, D., 1997. Algorithms on Strings, Trees and Sequences. Cambridge University Press, New York.

Ingman, M., Kaessmann, H., Paabo, S., Gyllensten, U., 2000. Mitochondrial genome variation and the origin of modern humans. Nature 408, 708–713.

Kurtz, S., Schleiermacher, C., 1999. REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15, 426–427.

Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A.G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M.D., Sukernik, R.I., Olckers, A., Wallace, D.C., 2003. Natural selection shaped regional mtDNA variation in humans. Proc. Natl. Acad. Sci. USA 100, 171–176.

Samuels, D.C., 2004. Mitochondrial DNA repeats constrain the life span of mammals. Trends Genet. 20, 226–229.

Samuels, D.C., Schon, E.A., Chinnery, P.F., 2004. Two direct repeats cause most human mtDNA deletions. Trends Genet. 20, 393–398.