Preprint of *FEBS Letters* 581(13): 2523-2526 (2007) [http://dx.doi.org/doi:10.1016/j.febslet.2007.04.073]

Divergent microsatellite evolution in the human and chimpanzee lineages

Zoltán Gáspári[a,1], Csaba Ortutay[b,1] and Gábor Tóth[c,*]

[a]Institute of Chemistry, Eötvös Loránd University, Pázmány Péter sétány 1/A, H-1117 Budapest, Hungary.

[b]Institute of Medical Technology, FI-33014 University of Tampere, Tampere, Finland.

[c]Bioinformatics Group, Agricultural Biotechnology Center, Szent-Györgyi Albert u. 4, H-2100 Gödöllő, Hungary.

[1]These authors have contributed equally to this work.

Abstract

Comparison of the complete human genome sequence to one of its closest relatives, the chimpanzee genome, provides a unique opportunity for exploring recent evolutionary events affecting the microsatellites in these species. A simple assumption on microsatellite distribution is that the total length of perfect repeats is constant compared to that of imperfect ones regardless of the repeat sequence. In this paper we show that this is valid for most of the chimpanzee genome but not for a number of human chromosomes. Our results suggest accelerated evolution of microsatellites in the human genome relative to the chimpanzee lineage.

Keywords: Microsatellites; Tandem repeats; Genome; Human; Chimpanzee; Comparative genomics; Molecular evolution

* Corresponding author: Gábor Tóth,
Agricultural Biotechnology Center, Gödöllő, Szent-Györgyi Albert u. 4., H-2100 Hungary
Phone: +36 28 526 224, Fax: +36 28 526 101, E-mail: tothg@abc.hu

1. Introduction

Microsatellites, or simple sequence repeats (SSRs), are more abundant in eukaryotic genomes than expected from statistical considerations [1]. Trinucleotide repeats constitute a subset of microsatellites that are used as genetic markers in studies of human evolution [2]. They are responsible for a number of genetic diseases [3,4] and are also considered to be important factors in eukaryotic genome evolution [5], since they contribute to genetic variability at multiple levels [6,7,8]. Their role in the evolution of intrinsically unstructured proteins has also been presented [9]. There is no stringent consensus in the literature on the definition of SSRs and on the use of the most appropriate method to measure the abundance or frequency of SSRs in genomic sequences [1]. Commonly, only perfect repeats are examined, since they can be identified in a much more objective manner than imperfect ones (see Scheme 1).

In this paper we introduce a new approach to gain insight into the evolution of trinucleotide repeats. We propose that the separate identification and analysis of perfect and imperfect repeats at identical loci within a genome is a valuable tool to assess SSR evolution. Instead of tracking the evolutionary history of individual repeats by using interspecies sequence alignments, we compare overall chromosomal repeat distributions, without the identification of orthologous SSRs in the human and chimpanzee genomes. We also determine whether results and conclusions obtained for perfect repeats can be directly extrapolated to imperfect ones.

2. Methods

A detailed description of the applied methods can be found in the Supplementary Material. Human (*Homo sapiens*) and chimpanzee (*Pan troglodytes*) complete genome sequences were downloaded from NCBI and the corresponding annotations were used. For the identification of imperfect and perfect SSRs (≥ 12 bp with unit lengths 1-6 bp), we used TRF [11] and an in-house written program, respectively. SSRs were categorized on the basis of their repeat units (class [12,13]), and location as exonic, intronic or intergenic repeats using the CDS fields in the sequence files. For direct comparison, we used only imperfect–perfect pairs at the same locus with identical repeat class. To compare distributions of total repeat

numbers / lengths per class, we applied the chi-square contingency analysis yielding a measure of similarity as in the PRIDE protein structure comparison algorithm [14].

3. Results and discussion

Identification of perfect and imperfect repeats

Our initial scan for perfect and imperfect repeats revealed that the results of our two search programs did not correlate well with each other in many cases. Further analysis was therefore limited to repeats that were identified simultaneously as perfect and imperfect SSRs with the same repeat unit that lie entirely in the same region (i.e. exonic, intronic or intergenic segments). It is reasonable to assume that the majority of imperfect repeats identified this way are derived from perfect ones rather than generated by point mutations from a non-repeated sequence by chance. Using this approach, the theoretical expectation that perfect repeats comprise a subset of the imperfect ones (Scheme 1) is also met.

Perfect and imperfect trinucleotide repeats in the human and chimpanzee genomes

Our detailed results are available at the SSRDB web site (http://bioinformatics.abc.hu/ssr/). Figure 1 shows the distribution of trinucleotide repeats in the coding regions of human chromosome 19 as an example where the length distributions of perfect and imperfect repeats differ considerably, even though repeats at identical loci were considered only.

In a general comparison of the human and chimpanzee genomes, the distributions of perfect and imperfect trinucleotide repeats differ to a lesser extent in the chimpanzee genome. This is apparent from the remarkably higher average probability values obtained for the chimpanzee genome, indicating that the distributions of perfect and imperfect repeats are highly similar, and the scarcity of values $P < 0.5$ relative to the human chromosomes (Table 1), mostly on the smaller ones (19–22 and Y, using the consensus numbering of chimpanzee chromosomes [15]).

Comparison of trinucleotide repeat distributions on human–chimpanzee orthologous chromosome pairs also underlines the dissimilarity in imperfect repeat distributions. *P* values

indicating the similarity of the normalized distributions across repeat classes are generally smaller for imperfect than for perfect repeats in all of the genomic regions examined (Table 2). This phenomenon is also more pronounced on smaller chromosomes.

These results indicate that there is considerable difference between the imperfect trinucleotide repeat distributions of the two genomes. The human genome displays a biased pattern of imperfect relative to perfect repeats, whereas there is practically no bias in chimpanzees (see also Table 1). Our detailed results for chromosomes 19, 21 and Y can be found in the Supplementary Material.

SSRs in the context of chromosome evolution

At the genomic scale, several differences have already been identified between *Homo sapiens* and *Pan troglodytes*. Divergent evolution in mono- and dinucleotide repeats in the two species has been presented using alignments of orthologous human–chimpanzee sequences [16]. Importantly, the distribution of interspersed repeats also exhibits characteristic differences between the two species [10,17,18].

At the chromosomal level, our survey shows that uneven distributions of perfect and imperfect SSRs are more apparent on smaller human chromosomes (16, 19, 20, 21, 22, Y; Table 1). Chromosomes 16 and 19 have high contents of interspersed repeats (~47% and ~55%, respectively, compared to the genomic average of ~45%) [19,20,21]. On the other hand, chromosomes 19 and 22 are noted for their high gene density [20,22]. The Y chromosome is particularly poor in repeats, but it was shown to be subject to frequent gene conversion events between its palindromic sequences [23,24]. However, these features cannot be unambiguously linked to the overall picture of SSR evolution. We emphasize that our results refer to whole chromosomes, which makes statistical considerations more robust, although it is clear that certain chromosomal segments have a distinct evolutionary history [24], and a number of molecular events can contribute to the observed pattern of trinucleotide repeats.

Our findings demonstrate that microsatellites of different repeat classes and in different regions of the human and chimpanzee genomes are subject to different evolutionary pressures to maintain or disrupt their regular pattern of repeats. Our observations are consistent with a scenario where human microsatellites underwent expansion after divergence from the human-

chimpanzee common ancestor and the presently observed high density of imperfect repeats is the result of disrupting mutations accumulated since then. Although this hypothesis remains to be proven, it is more likely than assuming that the majority of imperfect repeats around perfect ones were formed by point mutations in the human lineage. We also point out that identifying only perfect repeats, which is done in most studies, yields an incomplete picture of repeat abundance and distribution because such results cannot always be extended to imperfect repeats. Consequently, identifying both perfect and imperfect repeats is crucial for describing microsatellites and understanding their roles in evolutionary processes.

Acknowledgements

We thank Kathryn Rannikko and Anna K. Füzéry for proofreading of this manuscript. G.T. was supported by a János Bolyai postdoctoral fellowship from the Hungarian Academy of Sciences.

Supplementary Material

Full description of the applied methods, detailed analysis of our results on chromosomes 19, 21 and Y as well as the full author lists of references 10, 19, 20, 22 and 24 are supplied as Supplementary Material.

References

[1] Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* 5, 435-445.

[2] Harpending, H. and Rogers, A. (2000). Genetic perspectives on human origins and differentiation. *Annu. Rev. Genomics Hum. Genet.* 1, 361-385.

[3] Bates, G. and Lehrach, H. (1994) Trinucleotide repeat expansions and human genetic disease. *Bioessays* 16, 277-284.

[4] Richards, R.I. and Sutherland, G.R. (1992). Dynamic mutations: a new class of mutations causing human disease. *Cell* 70, 709-712.

[5] Hancock, J.M. (1995). The contribution of slippage-like processes to genome evolution. *J. Mol. Evol.* 41,1038-1047.

[6] Fondon, 3rd, J.W. and Garner, H.R. (2004). Molecular origins of rapid and continuous morphological evolution. *Proc. Natl. Acad. Sci. U S A* 101, 18058-18063.

[7] Kashi, Y. and King, D.G. (2006). Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.* 22, 253-259.

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

[9] Tompa, P. (2003). Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* 25, 847-855.

[10] Kuroki, Y., et al. (2006). Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nat. Genet.* 38, 158-167.

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

[12] Jurka, J. and Pethiyagoda, C. (1995). Simple repetitive DNA sequences from primates: compilation and analysis. *J. Mol. Evol.* 40, 120-126.

[13] Tóth, G., Gáspári, Z. and Jurka, J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* 10, 967-981.

[14] Carugo, O. and Pongor, S. (2002). Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison. *J. Mol. Biol.* 315, 887-898.

[15] McConkey, E.H. (2004). Orthologous numbering of great ape and human chromosomes is essential for comparative genomics. *Cytogenet. Genome Res.* 105, 157-158.

[16] Webster, M.T., Smith, N.G. and Ellegren, H. (2002). Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proc. Natl. Acad. Sci. USA* 99, 8748-8753.

[17] The Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437, 69-87.

[18] The International Chimpanzee Chromosome 22 Consortium (2004). DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* 429, 382-388.

[19] Martin, J., et al. (2004). The sequence and analysis of duplication-rich human chromosome 16. *Nature* 432, 988-994.

[20] Grimwood, J., et al. (2004). The DNA sequence and biology of human chromosome 19. *Nature* 428, 529-535.

[21] The International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

[22] Dunham, I., et al. (1999). The DNA sequence of human chromosome 22. *Nature* 402, 489-495.

[23] Rozen, S., Skaletsky, H., Marszalek, J.D., Minx, P.J., Cordum, H.S., Waterston, R.H., Wilson, R.K. and Page, D.C. (2003). Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423, 873-876.

[24] Skaletsky, H., et al. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423, 825-837.

Table 1

Probability of identity[a] of perfect and imperfect trinucleotide repeat distributions[b] on human and chimpanzee chromosomes

	Human chromosome																							
Region	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**X**	**Y**
All sequences	0.83	0.92	0.86	0.92	0.93	0.97	0.53	0.84	0.87	0.88	0.87	**0.48**	0.94	0.89	0.79	**0.27**	0.69	0.80	**0.06**	**0.39**	0.83	**0.31**	0.84	**0.12**
Coding regions	0.79	0.78	0.75	**0.07**	**0.35**	0.98	0.82	0.74	0.93	0.78	0.95	0.97	**0.49**	0.60	**0.33**	0.93	0.76	**0.04**	**0.10**	**0.31**	**0.42**	0.80	0.93	**0.25**
Introns	0.83	0.90	0.79	0.74	0.88	0.98	0.51	0.61	0.80	0.79	0.90	0.69	0.88	0.87	0.54	**0.31**	0.84	0.75	**0.09**	0.96	**0.27**	**0.14**	0.74	**0.00**
Intergenic regions	0.78	0.92	0.86	0.94	0.92	0.88	0.53	0.89	0.80	0.88	0.79	**0.27**	0.95	0.89	0.58	**0.15**	**0.25**	0.75	**0.03**	**0.07**	0.74	**0.09**	0.70	**0.23**
	Chimpanzee chromosome																							
	1	**2A**	**2B**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**X**	**Y**
All sequences	0.98	0.99	0.99	0.98	0.93	0.98	0.97	0.92	0.96	0.95	0.98	0.97	0.97	0.97	1.00	0.89	0.85	0.98	0.93	0.60	0.96	0.92	0.76	0.98	0.97
Coding regions	0.91	0.84	**0.31**	0.93	0.76	0.94	0.97	0.97	0.93	0.99	0.81	0.84	0.90	0.50	0.84	0.96	0.93	0.99	0.84	**0.20**	0.84	0.65	0.92	0.93	**0.04**
Introns	0.99	0.98	0.98	0.99	0.98	0.99	0.90	0.94	0.93	0.99	0.97	0.99	0.92	0.96	0.95	0.77	0.92	0.99	0.62	0.70	0.92	1.00	0.69	1.00	**0.00**
Intergenic regions	0.90	0.96	0.99	0.97	0.92	0.93	0.95	0.92	0.96	0.91	0.93	0.95	0.92	0.93	1.00	0.94	0.67	0.86	0.84	**0.37**	0.89	0.69	**0.38**	0.97	1.00

[a] Contingency probability values (P) were obtained by comparing the proportions of trinucleotide repeat classes (using the total lengths in each class) observed for perfect and imperfect repeats at identical loci in different regions of the chromosomes. Each contingency probability value is interpreted as a measure of similarity between the two distributions compared, thus numbers close to 1 indicate close similarity while lower values, primarily below 0.5, are interpreted as discrepancies between the distributions. Probabilities indicating highly different distributions ($P < 0.5$) are marked bold.

[b] Repeat abundance was measured as the cumulative length of the repeat class (in bp) divided by the cumulative length of the region (in Mbp).

Table 2

Region-specific comparison of perfect and imperfect repeats on homologous human–chimpanzee chromosome pairs[a]

	Human / chimpanzee chromosome																								
Perfect repeats	1	2A[b]	2B[b]	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
All sequences	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.99	1.00	0.97	0.96	1.00	**0.25**
Coding regions	1.00	0.86	0.76	0.64	0.61	0.83	0.88	0.99	0.95	0.91	0.88	0.89	0.58	**0.04**	0.97	0.93	0.76	0.85	0.51	0.90	0.81	**0.01**	**0.46**	**0.46**	**0.31**
Introns	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.98	0.84	0.88	0.89	1.00	0.99	0.92	1.00	**0.12**
Intergenic regions	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.96	0.97	0.99	0.99	1.00	1.00	0.82	0.87	1.00	**0.30**
Imperfect repeats	1	2A[b]	2B[b]	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y
All sequences	0.94	0.99	0.93	0.98	0.99	0.94	0.99	0.72	0.97	0.98	0.96	0.98	0.75	0.99	0.85	0.93	0.75	0.78	0.87	**0.32**	0.53	**0.29**	0.67	0.97	**0.01**
Coding regions	0.87	**0.28**	**0.19**	0.60	**0.05**	**0.17**	0.93	0.79	0.56	0.59	0.84	0.70	0.59	**0.05**	0.91	0.53	0.75	**0.43**	**0.00**	0.81	0.64	**0.45**	**0.31**	0.50	0.50
Introns	0.94	0.99	0.93	0.86	0.78	0.70	0.98	0.78	0.65	0.73	0.87	0.97	0.86	0.97	0.67	0.99	**0.44**	0.56	0.74	**0.30**	0.94	**0.12**	**0.08**	0.80	**0.00**
Intergenic regions	0.93	0.99	0.90	1.00	0.99	0.95	0.98	0.74	1.00	1.00	0.99	0.97	0.52	1.00	0.89	0.66	0.71	0.51	0.86	0.58	**0.25**	**0.38**	0.61	0.92	**0.07**

[a] For each pair and region, the probability of identity is calculated using the abundance data obtained as described in the footnote of Table 1.
[b] Human chromosome 2 is separately compared to the homologous chimpanzee chromosomes 2A and 2B.

Scheme and figure legends

Scheme 1. Different examples of perfect and imperfect SSRs. An imperfect SSR may, but does not necessarily, have mismatches, insertions and/or deletions relative to its perfect counterpart. Therefore, the relation between perfect and imperfect SSRs at a given locus may fall into one of three different categories: a perfect SSR makes up a segment of the corresponding imperfect one (a), the perfect SSR may be identical to the imperfect one (i.e. the SSR detected by the program TRF) (b), and multiple, isolated perfect SSRs occur in the imperfect one (c).

Figure 1. Distribution of the total lengths of trinucleotide repeats in the coding region of human chromosome 19. Observed data for perfect and imperfect repeats as well as expected data for imperfect repeats (calculated by uniformly multiplying the observed lengths of perfect repeats by 2.4, the average ratio) are shown. Comparison of perfect and imperfect repeats using the expected distribution would yield a probability value of 1.0 in contrast to 0.1 obtained from the observed data.

a)	gttctagt**acgacgacgacgacg**accacgactct	perfect repeat
	gt**acgaccacgacgacgacgacgaccacgac**tct	imperfect repeat
b)	gttctacc**acgacgacgacgacg**accccgactct	perfect repeat
	gttctacc**acgacgacgacgacg**accccgactct	“imperfect” repeat
c)	cacgtcg**acgacgacgacg**ccgtcg**acgacgacgacgacg**atgacg	perfect repeat
	c**acgtcgacgacgacgacgccgtcgacgacgacgacgacgatgacg**	imperfect repeat

Scheme 1.

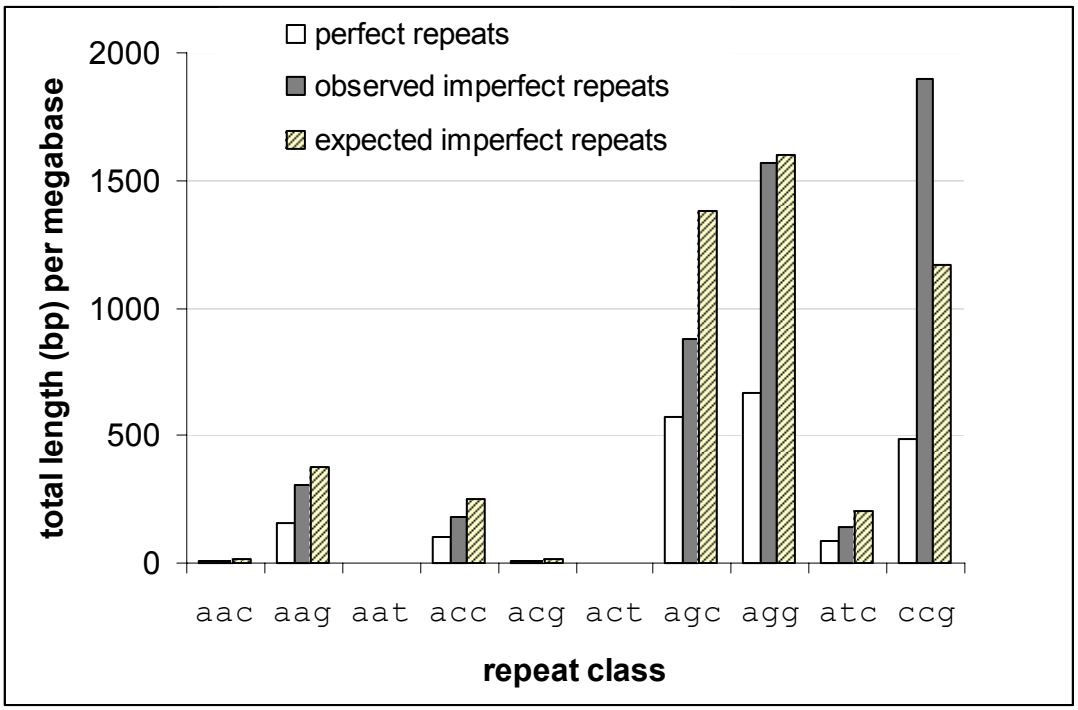
perfect repeats
observed imperfect repeats
expected imperfect repeats
total length (bp) per megabase
2000
1500
1000
500
0
aac
aag
aat
acc
acg
act
agc
agg
atc
ccg
repeat class

Figure 1.