Acta Biochimica Polonica

Vol. 53 No. 2/2006, 279–287

on-line at: www.actabp.pl

Regular paper

Tandemly repeated trinucleotides — comparative analysis

Monika Piwowar[1], Jan Meus[1], Piotr Piwowar[2], Zdzisław Wiśniowski[1], Justyna Stefaniak[3] and Irena Roterman[1]✉

[1]*Department of Bioinformatics and Telemedicine, Collegium Medicum, Jagiellonian University, Kraków, Poland;* [2]*Department of Measurement and Instrumentation, AGH-University of Science and Technology, Kraków, Poland;* [3]*Institute of Mathematics, Jagiellonian University, Kraków, Poland;*
✉*e-mail: myroterm@cyf-kr.edu.pl*

Received: 26 October, 2005; revised: 12 February, 2006; accepted: 09 May, 2006
available on-line: 29 May, 2006

Characteristics of 64 possible tandem trinucleotide repeats (TSSR) from *Homo sapiens (hs)*, *Mus musculus (mm)* and *Rattus norvegicus (rn)* genomes are presented. Comparative analysis of TSSR frequency depending on their repetitiveness and similarity of the TSSR length distributions is shown. Comparative analysis of TSSR sequence motifs and association between type of motif and its length (*n*) using ϱ-coefficient method (quantitatively measuring the association between variables in contingency tables) is presented. These analyses were carried out in the context of neurodegenerative diseases based on trinucleotide tandems. The length of these tandems and their relation to other TSSR is estimated. It was found that the higher repetitiveness (*n*) the lower frequency of trinucleotides tandems. Differences between genomes under consideration, especially in longer than n=9 TSSR were discussed. A significantly higher frequency off A- and T-rich tandems is observed in the human genome (as well as in human mRNA). This observation also applies to *mm* and *rn*, although lower abundant in proportion to human genomes was found. The origin of elongation (or shortening) of TSSR seems to be neither frequency nor length dependent. The results of TSSR analysis presented in this work suggest that neurodegenerative disease-related microsatellites do not differ *versus* the other except the lower frequency *versus* the other TSSR. CAG occurs with relatively high frequency in human mRNA, although there are other TSSR with higher frequency that do not cause comparable disease disorders. It suggests that the mechanism of TSSR instability is not the only origin of neurodegenerative diseases.

Keywords: trinucleotide tandem repeats, microsatellites, neurodegenerative disease

INTRODUCTION

Mammalian genomes scattered with simple sequence repeats (SSRs) (minisatellites and microsatellites) comprise about 3% of the human genome.

SSRs are liable to fluctuate, introducing deletion or insertion of one or more repeat units, probably because of the so-called slipped strand mispairing, which predisposes to pathogenic deletions and frameshifting insertions.

SSRs — microsatellites are the most abundant and have a more uniform distribution than minisatellites with the greatest single contribution originating from dinucleotide repeats, the ones most widely used in genetic analyses, which are now central to medicine, agriculture, evolutionary biology, and forensic science (Waterston *et al.*, 2002).

Trinucleotide microsatellites (TSSR) focus the attention of researchers nowadays because of their high polymorphism, which is very useful in genetic studies. They produce a very large number of alleles because of their high variability. This generates a very high degree of heterozygosity or polymorphic information content at each locus, making them some of the most informative genomic markers for genetic analyses (Bickeboller & Clerget-Drapoux, 1995). This variability influences their expandability as well as contractibility. Although expansions and

Abbrevations: *at*, *Arabidopsis thaliana*; *hs*, *Homo sapiens*; *mm*, *Mus musculus*; Msh2, mismatch repair genes; NIIs, neuronal intranuclear inclusions; *oi*, *Oceanobacillus iheyensis*; SSRs, simple sequence repeats; TSSR, trinucleotide simple sequence repeats; *rn*, *Ratus norvegicus*; *sc*, *Saccharomyces cerevisiae*.

contractions can occur, a bias towards expansion is mostly observed. It is characteristic that repeats below a certain length are stable in mitosis and meiosis, while above a certain threshold length the repeats become extremely unstable (Strachan & Read, 1999). The amplitude between expandability and contraction may differ significantly between particular sequence motifs (Jurka & Pethiyagoda, 1995).

Most SSRs (including TSSR) are known currently as useful in linkage studies of mouse and human, because of their polymorphism in populations. Such SSRs arising through replication errors might be largely equivalent between mouse and human, but impressive differences between these two species are observed (Beckman & Weber, 1992).

On the other hand, the ability of TSSR to change size is also responsible for some genetic diseases, the origin of which is mainly the elongation of trinucleotide repetitive sequences. This abnormality in TSSR distribution is registered frequently in neurodegenerative diseases (see Table 1), the background of which draws the attention of many researchers. The trinucleotide repeat disorders are a growing list of genetic neurodegenerative diseases characterized by the expansion of normally polymorphic repeated tripled nucleotides. They are dangerous because some kinds of tandems (located in genes) may be possibly responsible for modulation of protein–protein interaction. The length of tandem repeats is essential to the interaction between proteins. For example, protein containing polyglutamine tracts causes neurological disorders because of the different lengths of polyglutamine repeats associated with different affinities to transcription factors (Ashley & Warren, 1995; Margolis *et al.*, 1997). Expansion of tandem CAG-nucleotides causing neurodegenerative disorders results from to a toxic gain of function of mutant expanded proteins. Occurrence of NIIs is characteristic. Protein misfolding, interference with DNA transcription and RNA processing, activation of apoptosis and dysfunction of cytoplasmic elements have all been invoked in the toxic process (Everett & Wood, 2004). Interesting is the androgen receptor gene, mutation of which causes spinal and bulbar muscular atrophy. It is caused by expansion of the trinucleotide (CAG) repeat that codes for a polyglutamine tract in the transactivation domain of the receptor. Infertility is also associated with CAG expansion in the androgen receptor. It is known that infertile men are more likely to have longer than normal CAG repeats in the androgen-receptor gene than fertile men (Dowsing *et al.*, 1999). Lower numbers of CAG repeats in the androgen-receptor gene have been associated with higher incidence of prostate cancer (Gsur *et al.*, 2002; Strom *et al.*, 2004).

Two models have been proposed to account for variability in the number of repeat units. The first one is the initial co-mobilization of SSRs with dispersed repeats as a result of transposition. The length and composition of repeat units would be the result of unequal strand exchange followed by nucleotide divergence. In the second model the majority of length variants would result from mistakes (they are thought to arise by slippage) during DNA replication or during sister chromatid exchange (Toth *et al.*, 1987; Levinson & Gutman, 1987; Schlotterer & Tautz, 1992; Kruglyak *et al.*, 1998). Some data support a model in which expansion in the germ cells arises by gap repair and depends on a complex containing Msh2. Expansion occurs during gap-filling synthesis when DNA loops comprising the CAG trinucleotide repeats are incorporated into the DNA strand (Pearson *et al.*, 1997; Kovtun & McMurray, 2001).

TSSR instability may result in the creation of non-standard structures of DNA, particularly in the non-coding regions (Sinden *et al.*, 2002), which disturbs the natural functioning of genetic processes, or it can result in perturbation of gene expression, causing synthesis of defective proteins (when the TSSR occurs in the gene or in the close vicinity of the gene).

Despite the large number of publications on this subject, the mechanism of tri-nucleotide microsatellite expansion is not completely identified. That is why analysis of repetitive sequence instability is of such interest.

The characteristics of tandemly repeated homogenous trinucleotides in genomes of selected organisms are presented in this paper. The relation between sequence motifs of TSSR and their lengths was the main object of this research. This relation was compared in different organisms in the context of disease-related TSSR expandability (some of the repeats recognized as disease-related are listed in Table 1.

The distribution of TSSR in the organisms under consideration and a comparative analysis of the quantitative estimation of association between the sequence motif and its repetitiveness is presented in this paper. The distribution was approximated to a function, the parameters of which allowed a qualitative comparative analysis. A scale quantitatively measuring the association between two parameters (sequence motif and repetitiveness) was introduced based on the ϱ-coefficient. The method of ϱ-coefficient calculation allows ranking according to the strength of the mutual dependence. If the dependence is found, the results can show, which form of the pair of variables is mostly responsible for this association and which ones play a negligible role (Goodman & Kruskal, 1959; 1963; 1972). Among a few methods to calculate dependency in contingency tables (Goodman & Kruskal, 1954; 1959; 1963; 1972;

Table 1. Selected trinucleotides (and their gene locations), expansion of which causes neurodegenerative diseases

Disorder	Gene location	Motif
Huntington Disease; HD	4p16.3	CAG
Machado–Joseph (MJD)/ SCA III	14.q24.3	CAG
Atrophy, X-linked; (SBMA)	Xq11	CAG
SCA I - (ATX1)	6q23	CAG
SCA II - (ATX2)	2q24	CAG
SCA VI - (CACNA1A)	19q13	CAG
SCA VII - (AT7)	3p21.1-p12	CAG
DRPLA	12q13	CAG
Myotonin Dystrophy	19q13.2	CTG
Fragile X type A (FMR-1)	Xq27.3	CCG
Fragile X typ E	Xq28	CGG/CCG
Fragile X typ F	Xq28	CGG/CCG
Fragile X typ16A	?	CGG/CCG
Friedriech Ataxia (FRDA)	9q	CAA

Björnstad, 1979) the ϱ-coefficient analysis not only gives information about the presence of the association but also allows the form of the particular pairs of variables to be validated. This method allows assess dependency between both grouped and subdivided variables.

MATERIALS AND METHODS

Genomic DNA and cDNA (mRNA) data were taken to computational analyze of TSSR. The sequences of the following model organisms were studied in the context of the presence of poly-trinucleotides: *Homo sapiens*, *Ratus norvegicus*, *Mus musculus* (The data from 14.04.2003) were from National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/). The sequence of human mRNA was also taken from NCBI.

Computational analysis

TSSR distribution comparison. TSSR were detected by tracing the $(XXX)_n$ sequence repeats in genomes, where n=4 to 14 and XXX = sequence motif under consideration. The size of n in TSSR frequency calculation was selected as n=4 to 14 based on data of (Margolis *et al*., 1997) PERL scripts were used for counting.

The function (found according to MATLAB to be the best one)

$$y = k_1 e^{k_2 x^{k_3}} \tag{1}$$

was applied to approximate the mammalian (*hs, mm, rn*) TSSR distribution depending on the length. Three parameters (k_1, k_2, k_3) found for each distribution were compared allowing comparison of genomes under consideration.

The exponential function parameters (k_1, k_2, k_3) turned out to be simple and easy for interpretation to show how frequencies of trinucleotide sequences within different genomes are differentiated.

The meanings of the parameters are as follows:

k_1 — directional coefficient

k_2 — an increase in the values of the k_2 parameter causes a slight decrease of the function values together with an increase of the x-value (size of tandem). In our case, it proves the slight decrease of frequency of repeated sequences, from short to long tandems, or it proves a sudden increase of long tandems

k_3 — however, a high value of the k_3 parameter causes a high value of function for a low x-value. With the increase of the x-value (size of tandem) the value of the function drastically decreases. The higher the value of k_3 the shorter the length of TSSR.

The method allows the assessment of differences between TSSR within one genome and also permits the inter-genome comparative analysis of TSSR.

Dependence in contingency table measurements. The ϱ-coefficient was applied to measure the association between two qualitative variables (sequence size of tandem **A** = $\{A_i\}$ — columns; its sequence motif **B** = $\{B_j\}$ — rows). The ϱ-coefficient was defined as follows for a 2 × 2 table:

$$Z^2(A_i : B_j) = 1 - \left[P(B_j)\frac{1-P(A_i \mid B_j)}{P(\overline{A}_i)} \cdot \frac{1-P(A_i \mid B_j)}{P(A_i)} + P(\overline{B}_j)\frac{1-P(A_i \mid \overline{B}_j)}{P(\overline{A}_i)} \cdot \frac{1-P(\overline{A}_i \mid \overline{B}_j)}{P(A_i)} \right] \tag{2}$$

and was used to evaluate the mutual dependence between sequences consisting of a particular sequence motif and its tandem size (repetitiveness).

Briefly, the method is as follows:

Assume that the contingency table below represents the observed (empirical) probabilities for c different realizations of variable **A** (qualitative) and for r different realizations of variable **B** (qualitative). For the problem presented in this paper, assume that **A** represents sequence motives and **B** their repetitiveness (see Table 2).

To estimate whether p_{ij} expresses relatively high or low probability (high or low coupling of a

Table 2. Contingency table representing *c* different realizations of variable A (size of tandem) and for *r* different realizations of variable B (trinucleotide)

Size of tandem / Trinucleotide	$\mathbf{A_1}$	$\mathbf{A_2}$	$\mathbf{A_3}$	....	$\mathbf{A_i}$	....	$\mathbf{A_c}$
$\mathbf{B_1}$	p_{11}	p_{21}	p_{31}	...	p_{i1}	...	p_{c1}
$\mathbf{B_2}$	p_{12}	p_{22}	p_{32}	...	p_{i2}	...	p_{c2}
$\mathbf{B_3}$	p_{13}	p_{23}	p_{33}	...	p_{i3}	...	p_{c3}
.....	...	...	...	...	...	...	...
$\mathbf{B_j}$	p_{1j}	p_{2j}	p_{3j}	...	p_{ij}	...	p_{cj}
.....	...	...	...	...	...	...	...
$\mathbf{B_r}$	p_{1r}	...	p_{3i}	...	p_{ir}	...	p_{cr}

particular *i*-th sequence with a particular *j*-th structure, its value is compared with all possibilities for solutions of other **A** (excluding the *i*-th) and other **B** (excluding the *j*-th). Each pair of *i*-th and *j*-th realizations of **A** and **B** can be represented using a 2 × 2 contingency table (see Table 3).

The value of the ϱ-coefficient can be calculated for each *i*-th and *j*-th realization of **A** and **B**. The ranking order permits comparisons between particular pairs over the whole contingency table, allowing selection of those that play an important role in the general dependence of **A** and **B** (structure-to-sequence). High ϱ values distinguish pairs whose participation in general dependence is high. Others with lower ϱ values are not necessarily responsible for the dependence (relation) under consideration.

The method based on the ϱ-coefficient enables the assessment of the strength of the pair-wise mutual dependence. In other words, we can find the preferences for particular sequences to occur with a particular length. Thus the ϱ-coefficient based method allows for distinguishing the sequences with tendency to low and high dispersion all over the genome. The value of the ϱ-coefficient can be interpreted as a measure of the strength of association between a particular length and a particular kind of motif.

RESULTS

The analysis was focused on the search for similarities and differences of TSSR between the species studied, taking the length (size of tandems) and sequence motifs into account.

Comparative analysis of TSSR frequency depending on their repetitiveness

Polynucleotide repeats are overrepresented in the genomes of most eukaryotes. The amount of TSSR in *sc, at, oi* and *at* is significantly lower than in mammalians like *hs, mm* and *rn*. Among the bacterial genomes present in our analysis (not published),

Table 3. Contingency table (2 × 2) representing *i*-th and *j*-th realizations of A and B variables.

Variables	$\mathbf{A_{ij}}$	$\mathbf{A_{ik}}$ k=1,...,c and k=j excluded
$\mathbf{B_{ji}}$	$P(A_i \mid B_j)$	$P(\overline{A_i} \mid B_j)$
$\mathbf{B_{jn}}$ n=1,..r and n=i excluded	$P(A_i \mid \overline{B_j})$	$P(\overline{A_i} \mid \overline{B_j})$

the *oi* genome contains the highest number of TSSR. There are repetitive trinucleotide sequences only for n<= 5. The general tendency (on the basis of selected organisms) found for TSSR frequencies are that the higher the number of n the lower the number of trinucleotides in tandemly repeated fragments. The most frequent are the tandems of n=4 (especially in *hs, mm* and *rn* genomes; see Fig. 1A). The decrease of frequency treated as dependent on tandem size has a hyperbolic shape, although the functions differ.

Long tandems (n above 10) are more frequent in *mm* and *rn* than in *hs* (see Fig. 1B).

The number of TSSR in the *mm* genome was found to be higher than in the *rn* genome.

Similarity of the TSSR length distributions

The frequency distribution was approximated to the function presented in Methods. Only *hs, mm* and *rn* were incorporated into this analysis. Three parameters were obtained for each approximated function calculated for the distribution of tandems (n=4 to 14) (see Fig. 2). Mean standard errors of approximation in particular species were as shown below (see Table 4).

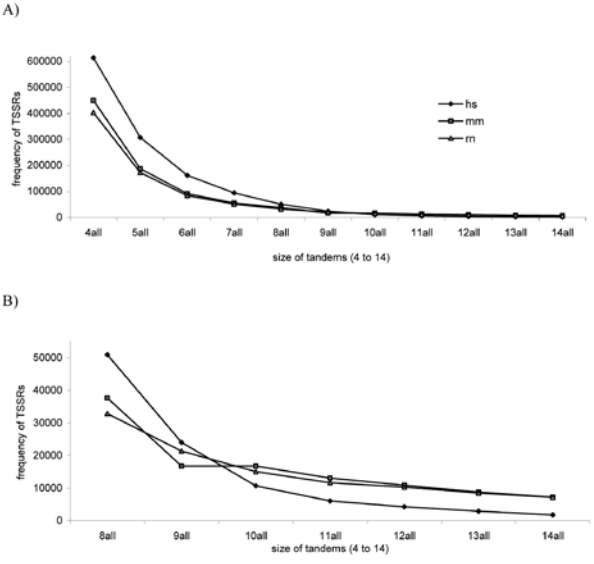

Figure 1. Frequency of TSSR of different lengths in selected genomes (*hs, rn, mm*).
A. frequency of TSSR with n=4 to 14; B. frequency of TSSR with n=8 to 14.

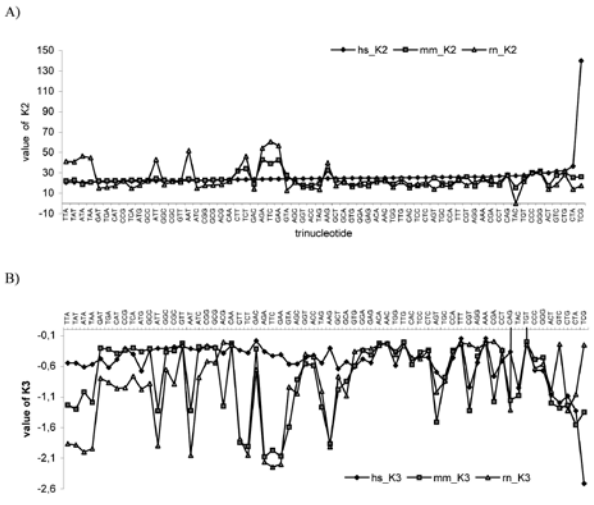

Figure 2. Values of parameters of the function approximated to the distribution of TSSR (n=4 to 14) in *hs*, *mm* and *rn* genome.
A. k_2 Parameter distribution function; B. k_3 parameter distribution function.

Parameter k_2 in *hs* is rather stable. Its value slightly increases together with the increase of the presence of C, G and T (see Fig. 2A). An exception was $(TCG)_n$, for which parameter k_2 is equal to 139.936, due to the relatively low frequency of TSSR and the presence of longer repeats with higher frequency and the absence of some short tandems.

Approximation of parameter k_3 allowed estimation of the "preference" of particular TSSR to be present in tandems of low size. A high value of k_3 suggests a tendency to lower *n* values (see Fig. 2B). Interpretation of the k_3 value for all trinucleotides in *hs* shows that tandems $(AAA)_n$ and $(TTT)_n$ appear mostly for low *n* (n=4). An increase of *n* for those trinucleotides causes a decrease of their frequency. The lowest k_3 was obtained for $(TCG)_n$. The lowest and the highest values of k_3 are shown in table (see Table 5).

Table 5. Lowest and highest values of parameter k_3 for *hs*, *mm*, *rn*

Highest k_3		Lowest k_3	
	hs		*hs*
AAA	–0.14609	TCG	–2.51055
TTT	–0.14861		*mm*
	mm	AGA	–2.07692
TGT	–0.20436	GAA	–2.0663
TTG	–0.20787		*rn*
	rn	TTC	–2.24412
TAC	***17.77355***	GAA	–2.19672
CGA	–0.19716	AGA	–2.16192
ACG	–0.20995	AAT	–2.05222
		TCT	–2.05147

Comparative analysis of TSSR sequence motifs

The analysis reveals that the TSSR of highest frequency are the trinucleotides (AAA) and (TTT) especially in the human genome (Fig. 3A). These sequences are also abundant in *mm* and *rn* genomes. Tandems (AAA) in the human genome appear to have a 0.343366 share of all TSSR under consideration (n=4 to 14). Their shares in the other analyzed organisms are *mm*=0.19435, *rn*=0.211723. TSSR $(AAA)_n$, and $(TTT)_n$ appeared significantly frequently in *at* and *sc* genomes.

The characteristics of $(AAA)_n$ and $(TTT)_n$ seem similar for all organisms, but other sequence motifs rich in „A" and „T" appeared to differ significantly between organisms. It is surprising that TSSR very abundant in the *mm* and *rn* genomes are not frequent in the *hs* genome (see Fig. 3B). These motifs are $(TAT)_n$, $(ATA)_n$, $(TTA)_n$, $(TAA)_n$, $(ATT)_n$, $(AAT)_n$ (combinations of nucleotides „A" and „T"), $(GGC)_n$, $(GCC)_n$, $(GCG)_n$, $(CGC)_n$, and also $(CGG)_n$, and $(CCG)_n$. On the other hand, $(TTG)_n$, $(CAA)_n$,

Table 4. Mean standard errors of approximation (Mse) of the function to the distribution of TSSR (n=4 to 14) in *hs*, *mm* and *rn*.

The distribution of $(TAC)_n$ in *rn* was the worst approximation, with 46% error.

Species	Mse (%)
hs	3.6
mm	3
rn	3.4
rn $(TAC)_n$	46

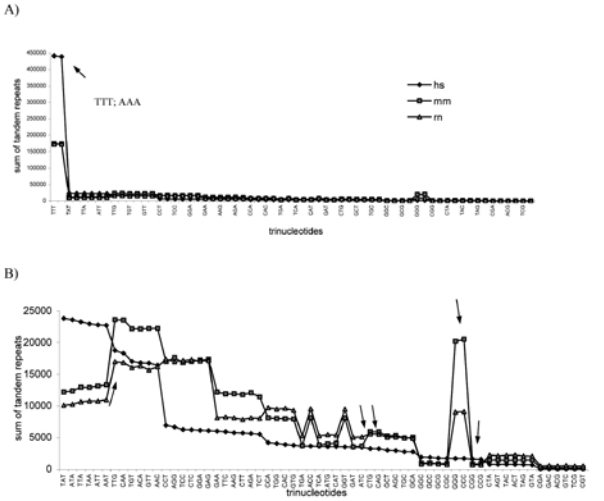

Figure 3. Frequencies of TSSR of a particular motif in *hs*, *mm* and *rn* genomes.
A. For all possible trinucleotides. It reveals the dominant role of $(AAA)_n$ and $(TTT)_n$, especially in the *hs* genome. These trinucleotides are also very frequent in other mammalian genomes (similar genome size). B. Without $(TTT)_n$ and $(AAA)_n$. Arrows show trinucleotides responsible for neurodegenerative diseases.

Table 6. Partial (%) participation of disease-related trinucleotides *versus* total TSSR (n=4 to 14) (*hs, mm, rn*)

Motif	*hs*	*mm*	*rn*
CAG(%)	0.25405797	0.66156	0.67388
CTG(%)	0.2553091	0.66312	0.68121
CGG(%)	0.12698989	0.08773	0.08641
CCG(%)	0.12386206	0.09778	0.08483
CAA(%)	1.43371894	2.63206	2.05828
GAA(%)	0.47175492	1.36062	0.99319

$(TGT)_n$, $(ACA)_n$, $(GTT)_n$, $(AAC)_n$, $(CTC)_n$, $(AGG)_n$, $(TCC)_n$, $(CTC)_n$, $(GGA)_n$, $(GAG)_n$, $(CTA)_n$, $(AGT)_n$, $(TAC)_n$, $(ACT)_n$, $(TAG)_n$, and $(GTA)_n$ are relatively abundant.

The abundance of TSSR $(CCC)_n$ and $(GGG)_n$, in the *mm* and *rn* genomes, is also remarkable while in *hs* these sequences occur with low frequency.

Disease-related trinucleotides are present in the analyzed genomes with relatively low frequency *versus* the frequency of other TSSR (see Table 6).

The proportions of disease-related tandems in coding sequence mRNA differ. Thus, tandems $(CAG)_n$ (whose expansion is responsible for most neurodegenerative diseases) are only 0.25% of the complete amount of TSSR; in mRNA their presence is expressed by 2.05% of TSSR (n=4 to 14) (see Fig. 4). Their presence in mRNA can be even as high as 5% in the case of exclusion of $(AAA)_n$ and $(TTT)_n$ (see Table 7).

The distribution of the remaining disease-related trinucleotides in mRNA is also shown in Fig. 4. It reveals that TSSR like $(CAA)_n$ are not many, while in the whole genome they have higher participation.

Association between type of motif (sequence motif) and its length (size of tandem)

The association between the size of tandems and the trinucleotide sequence was searched using ϱ-coefficient. The contingency table was created, with columns expressing the length of the tandem (A_i) (4 to 14) and rows expressing the sequence of the trinucleotide (B_j) (64 combinations of four different nucleotides in the trinucleotide sequence).

Table 7. Share of whole TSSR (%) (with and without $(AAA)_n$ and $(TTT)_n$ tandems)

Motif	All TSSR	Excluding $(AAA)_n$ and $(TTT)_n$
CAG (%)	2.05	4.7
CTG (%)	1.52	3.55
CGG (%)	2.19	5.12
CCG (%)	1.55	3.61
CAA (%)	0.28	0.66
GAA (%)	0.9	2.09

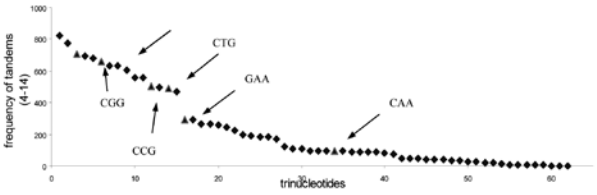

Figure 4. TSSR responsible for neurodegenerative diseases in human mRNA.
Distribution of TSSR (n=4 to 14) with disease-related ones highlighted.

The ϱ-coefficient was calculated for *hs, mm* and *rn*. ϱ-Coefficient calculation reveals that associations can be found for particular sequences and their lengths. Moreover, the procedure of ϱ-coefficient calculation for each cell of the contingency table can validate a particular dependence quantitatively.

The contingency table with ϱ-coefficients calculated for each cell is presented in Fig. 5 using a color scale (legend included). The highest ϱ-coefficient values are explained precisely.

Trinucleotides TCT and TCG appeared to be highly associated with the multiple of 4 in the *hs* genome (see Fig. 5a). It may be inferred that these trinucleotides represent a low tendency to polymorphism. The interpretation is as follows: If the TSSR of the sequence $(TCT)_n$ or $(TCG)_n$ is found in the human genome, one can predict that its size is n=4. There is a low probability that longer fragments can be found, and one may be sure that it does not represent a length of n=10. Although the calculated 0.075 value of the ϱ-coefficient is very low its value shall be interpreted relative to the contingency table under consideration. This value is highest in comparison with all others obtained for this contingency table, and reveals the association between the presented trinucleotides and the tendency to occur in a multiple of 4.

Trinucleotides CTT and GGA (see Fig. 5b, c) of *mm* and *rn* appeared to represent the highest association with the length of tandems n=4.

The disease-related trinucleotides do not reveal a high association with a particular length of TSSR that can be interpreted as sequences of high polymorphic character. No association was found in *mm* and *rn* between disease-related trinucleotides and their length suggesting the absence of evolution-dependent processes.

The high association of sequence-to-length analysis in *hs* mRNA found for $(TCG)_n$, and n=4 (see Fig. 6) supports findings from the complete *hs* genome. The low polymorphic character of this sequence seems more reliable.

Trinucleotides of high expandability, such as (CAG) in human mRNA, do not exhibit any association with a particular length.

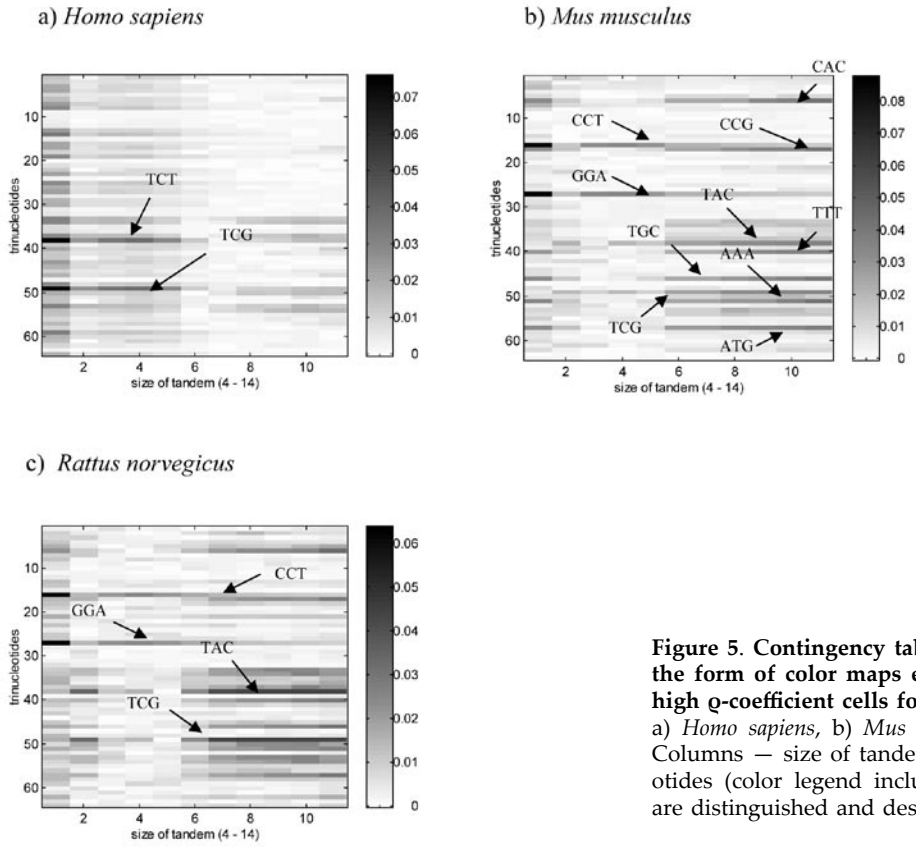

Figure 5. **Contingency tables for complete genomes in the form of color maps expressing the distribution of high ϱ-coefficient cells for:**
a) *Homo sapiens*, b) *Mus musculus*, c) *Rattus norvegicus*. Columns — size of tandems (4 to 14), rows — trinucleotides (color legend included). High ϱ-coefficient cells are distinguished and described.

DISCUSSION

The characteristics of TSSR allow localization of polymorphic sequences. It is also important to distinguish dominant and marginal fractions in the genome. Estimation of the association between the sequence of the repetitive unit and its length may allow analysis in the context of evolution, particularly when the genomes of different species are compared.

Numbers and proportions of TSSR are different in every species. The comparison of genomes of non-mammalian organisms (e.g. bacteria, fungi, plants, etc.) with genomes of mammals (e.g. human, rat, mice, etc.) reveals significant increase of tandem sequence accumulation. The higher the genome organization the higher the number of TSSR. On the other hand, for comparably developed organisms (for example mammals) both the number and the kind of repetitions are also differentiated, which indicates that those sequences can be involved in many biochemical processes that are typical for organisms with genomes in which they occur.

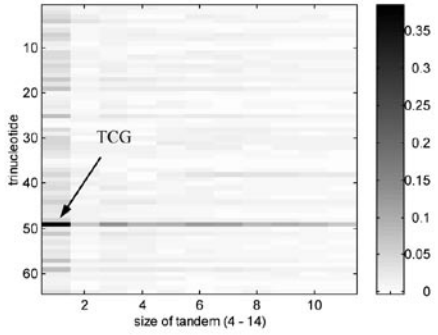

Figure 6. **Contingency tables for human mRNA in the form of color maps expressing the distribution of high ϱ-coefficient cells**.
Columns — size of tandems (4 to 14), rows — trinucleotides (color legend included). The high ϱ-coefficient cells are distinguished and described.

Some TSSR are important for proper protein interactions in transcriptional complexes (Ashley & Warren, 1995; Margolis *et al.*, 1997). One may assume that also some TSSR, when expressed, can influence many special biochemical interactions on the protein level. In spite of this they may be responsible for proper interactions between proteins occurring in biochemical processes in cells. On the other hand, tandem trinucleotide repeated sequences occurring in noncoding fragments probably can play an important role in the organization of genetic material.

It can be concluded that highly organized genomes have developed a system of repeated sequences for a better organization of genetic material and a better precision in a process of expressing information stored in genomes.

Particularly important is comparative analysis of neurodegenerative disease-related microsatellites and other repetitive sequences without a disease-related phenotype.

Analysis of TSSR provided information about their characteristics in genome. It was found that the higher the number of *n* the lower the number of trinucleotides in tandemly repeated fragments. The differences were found between *mm*, *rn* and *hs* especially in longer than n=9 TSSR. There are much fewer of them in human genome than in *mm* and *rn*. A significantly higher frequency off A- and T-rich tandems is observed in the human genome (as well as in human mRNA). This observation also applies to *mm* and *rn*. Trinucleotide microsatellites composed of A and T nucleotides are abundant in *hs* genome. It is caused by their retroposons origin particularly by Alu sequences (90% of A-rich SSRs in human are provided by or spawned from poly(A) tails of Alu and L1 elements) (Toth *et al.,* 2000). Very abundant and poly(A)-rich microsatellite classes, such as ATA, are frequently associated with an evolutionarily older subclass of Alu repeats (Yandava *et al.,* 1997).

Different frequencies of trinucleotide tandems appear for combinations of A, T, C, G; particularly high frequencies of C- and G-rich tandems are present in the *rn* and *mm* genome. These two genomes are also characterized by a proportionally high frequency of homogenous tandems like $(CCC)_n$ and $(GGG)_n$, whose frequency in the human genome is rather low.

The mechanism of microsatellite elongation (or shortening) is to errors of polymerase in the replication process. Repetition of trinucleotides results in a local increase of flexibility of DNA, which may lead to the creation of single-stranded hairpins, triplexes or quadruplexes (Hartenstine *et al.,* 2000; Sinden *et al.,* 2002). The problem of selectivity of the sequences which undergo this phenomena has been considered.

The results of TSSR analysis presented in this work suggest that neurodegenerative disease-related microsatellites do not differ from the others, at least from the point of view of the parameters studied here. They occur with low frequencies, comparable to other TSSR. CAG occurs with relatively high frequency in human mRNA, although there are other TSSR with higher frequency that do not cause comparable diseases.

The analysis of association between the unit sequence and size of tandems revealed a lack of mutual dependence for disease-related TSSR. The characteristic association of TSSR with their length allows to locate polymorphic sequences. One shall mention that the most non-polymorphic trinucleotide sequences are TCG (both in *hs* genome and mRNA) and TCT (*hs* genome) whereas many others are polymorphic.

The origin of elongation (or shortening) of TSSR seems to be neither their frequency nor their length. The elongation (or shortening) may come from neighbor sequences causing "slipped-strands" of polymerase. This supposition is in agreement with the observation that the elongation is correlated with the number of C+G flanking regions, particularly in the case of CAG and GCC in exons (Jurka & Pethiyagoda, 1995). Also important is the observation that the change of TSSR length influences known diseases of the nervous and muscle-neuronal systems. It suggests that tissue specificity may play an important role in disease etiology. Another problem is mosaicism (the instability of TSSR in mitosis), which may occur during life of patients with some neurodegenerative diseases (Gusella & MacDonald, 1996). It confirms that not only TSSR are responsible for changing their length.

In conclusion, it seems that the mechanism of TSSR instability is not the only origin of neurodegenerative diseases.

REFERENCES

Ashley CT, Warren ST (1995) Trinucleotide repeat expansion and human disease. *Annu Rev Genet* **29**: 703–728.

Beckman JS, Weber JL (1992) Survey of human and rat microsatellites. *Genomics* **12:** 627–631.

Bickeboller H, Clerget-Drapoux F (1995) Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. *Genet Epidemiol* **12**: 865–870.

Björnstad JF (1979) Inference theory in contingency tables. *Statistical Research Report Oslo* pp 1–26.

Dowsing AT, Yong EL, Clark M, McLachlan RI, Kretser DM, Trounson OA (1999) Linkage between male infertility and trinucleotide repeat expansion in the androgen-receptor gene. *Lancet* **354**: 640–643.

Everett CM, Wood NW (2004) Trinucleotide repeats and neurodegenerative disease. *Brain* **127**(Pt 11): 2385–2405.

Goetz CG, Pappert EJ (1999) *Textbook of Clinical Neurology*. W.B. Sounders Company, Philadelphia, London, Toronto, Montreal, Sydney, Tokyo.

Goodman LA, Kruskal WH (1954) Measures of association for cross classifications. *J Am Stat Assoc* **49**: 732–764.

Goodman LA, Kruskal WH (1959) Measures of association for cross classifications. II: Further discussion and references. *J Am Stat Assoc* **54**: 123–163.

Goodman LA, Kruskal WH (1963) Measures of association for cross classifications. III: Approximate sampling theory. *J Am Stat Assoc* **58**: 310–364.

Goodman LA, Kruskal WH (1972) Measures of association for cross classifications. IV. Simplification of asymptotic variances. *J Am Stat Assoc* **67**: 415–421.

Grewal RP (1999) Neurodegeneration in Xeroderma Pigmentosum: a trinucleotide repeat mutation analysis. *J Neurol Sci* **163**: 183–186

Gusella JF, MacDonald ME (1996) Trinucleotide instability: a repeating theme in human inherited disorders. *Annu Rev Med* **47**: 201–9.

Gsur A, Preyer M, Haidinger G, Zidek T, Madersbacher S, Schatzl G, Marberger M, Vutuc C, Micksche M (2002) Polymorphic CAG repeats in the androgen receptor gene, prostate-specific antigen polymorphism and prostate cancer risk. *Carcinogenesis* **23**: 1647–1651.

Hartenstine MJ, Goodman MF, Petruska J (2000) Base stacking and even/odd behaviour of harpin loops in DNA polymerase. *J Biol Chem* **275**: 18382–18390.

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

Kovtun IV, McMurray CT (2001) Trinucleotide expansion in haploid germ cells by gap repair. *Nat Genetics* **27**: 407–411.

Kruglyak S, Durret RT, Schug MD, Aquadro CF (1998) Equilibrium distribution of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci USA* **95**: 10774–10778.

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* **4**: 203–221.

Margolis RL, Abraham MR, Gatchell S, Li SH, Kidwai AS, Breschel TS, Stine OC, Callahan C, McInnis MG, Ross CA (1997) cDNAs with long CAG trinucleotide repeats from human brain. *Hum Genet* **100**: 114–122.

Pearson CE, Ewel A, Acharya S, Fishel RA, Sinden RR (1997) Human MSH2 binds to trinucleotide repeat DNA structures associated with neurodegenerative diseases. *Hum Mol Genet* **6**: 1117–1123.

Schlotterer C, Tautz D (1992) Slippage synthesis of simple sequence DNA. *Nucleic Acids Res* **20:** 211–215.

Sinden RR, Potaman VN, Oussatcheva EA, Pearson CE, Lyubchenko YL, Shlyakhtenko LS (2002) Triplet repeat DNA structures and human genetic disease: dynamic mutations from dynamic DNA. *J Biosci* **27:** 53–65.

Strom SS, Gu Y, Zhang H, Troncoso P, Babaian RJ, Pettaway CA, Shete S, Spitz MR, Logothetis CJ (2004) Androgen receptor polymorphisms and risk of biochemical failure among prostatectomy patients. *Prostate* **60**: 343–351.

Strachan T, Read AP (1999) *Human Molecular Genetics 2* 2nd edn. BIOS Scientific Publishers Ltd, Oxford, UK.

Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* **10**: 967–981.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**: 520–562.

Yandava CN, Gastier JM, Pulido JC, Brody T, Sheffield V, Murray J, Buetow K, Duyk GM (1997) Characterization of *Alu* repeats that are associated with trinucleotide and tetranucleotide repeat microsatellites. *Genome Res* **7**: 716–724.