FEBS Letters 389 (1996) 92–95 FEBS 17137

Minireview

Genomic simple repetitive DNAs are targets for differential binding of nuclear proteins

Jörg T. Epplen*, Andreas Kyas, Winfried Mäueler

Molecular Human Genetics, MA5, Ruhr-University, 44780 Bochum, Germany

Received 29 April 1996

Abstract The biological meaning of abundant simple repetitive DNA sequences in eukaryote genomes is obscure. Therefore, $(GAA)_n$, $(GT)_n$, and composite $(GT)_n(GA)_m$ blocks were characterized for protein binding in the repeat and flanking sequences of cloned genomic DNA fragments. In gel mobility shift and competition assays the binding of nuclear proteins to the repeats was specific (including some flanking single copy sequences). DNase footprinting revealed the target sequences within and adjacent to the repeats. Chemical modifications (OsO_4, DEPC) demonstrated non-B DNA structures in the polypurine blocks. The binding of nuclear proteins in and around simple repeat sequences refute biological insignificance of all of these ubiquitously interspersed elements.

Key words: Composite simple repeat; DNA secondary structure; Microsatellite; Polypurine; H-DNA configuration; Z-DNA configuration

1. Introduction

Simple repetitive DNA stretches manifest themselves as conspicuously regular and monotonous paragons throughout the seemingly orderless genome. Simple repeats are defined as basic motifs of 1–6 bases in length which are perfectly and tandemly reiterated 5–100 times [1] or more. Several different basic motifs of these repeat elements are frequently represented and ubiquitously interspersed in eukaryote genomes [2], and they have mostly been regarded as 'junk' [3] without sequence-dependent meaning [4]. Therefore, such 'slippery DNA runs' are at times welcome targets of eloquent declamations [5], the decoding of which requires sufficient command of metaphilosophical allegories. Di-, tri- and tetranucleotide repeats are often highly polymorphic in their lengths and thus are optimal genetic markers [6] mostly investigated in the form of so-called microsatellites after amplification by the polymerase chain reaction (PCR) [1,7,8]. Assaying for repeat lengths of microsatellites is easily (semi-) automatable, and thus these simple repeats have superseded all other types of molecular genetic markers [6]. The high extent of polymorphism in simple repetitive DNA appears to reflect accelerated evolution in comparison to veritable single-copy sequences [9]. Hence, simple repeats are said to evolve quite rapidly. Yet, upon closer inspection, the speed of simple repeat evolution depends on the sequence content, in addition to length and also probably on the surrounding genomic environment [4,10,11]. The mutation rates of simple repeats often differ by several orders of magnitude for unknown reasons (op. cit.). These high mutation rates should be kept in mind, especially in forensic or diagnostic situations involving identity and relationship analyses. The characteristic polymorphic properties of this class of sequences have been mastered for use as efficient tools in modern genome research. But what are the basic biological functions of this class of DNA elements? Here we assemble arguments that in spite of the supposedly generally high mutation rate these repeats may be functionally significant.

In the past few years simple repeats have gained attention due to their ability to cause disease, but the cellular mechanisms are still mysterious in many cases. In database searches, longer stretches of simple mono-, di-, tetra- and pentanucleotide repeats have not been found to be represented in translated regions of mature mRNAs [2,4]. Nevertheless, certain trinucleotide motifs are situated in untranslated portions of mRNA [8] and they code for monomorphic amino acid blocks [12]. Based on so-called 'dynamic mutations' [13], some perfect trinucleotide units can haphazardly be increased in length over sharply definable thresholds in man. Expanded trinucleotides like $(CCG)_n$, $(CAG)_n$ and $(CTG)_n$ cause fatal, dominantly inherited diseases (for short overviews see [14,15]). Long $(CTG)_n$ and $(CCG)_n$ repeats are known to inhibit regulated gene expression, and the $(CAG)_n$ repeats are translated into polyglutamine stretches that might interfere with normal cellular proteins resulting in premature cell death and ensuing disease (like neurodegeneration) symptoms. Recent examples of such deleterious interactions include a few proteins harboring elongated polyglutamine stretches interacting with 'Huntingtin associated protein' as well as the enzyme GAPDH [16,17]. A substantially elongated $(GAA)_n$ repeat has been identified as the common mutation in the recessively inherited Friedreich ataxia [18] in humans. Expression studies on additional trinucleotide motifs like $(CAC)_n$ and frameshifts thereof are the subject of intense research efforts, in both normal and disease situations [19]. Thus, increasingly more motifs of genomic simple repeats are attracting considerable interest because of their propensity to mutate (elongate) and disrupt gene function. Besides the aforementioned trinucleotide entities, dinucleotide repeats as well as longer motifs (including minisatellites [20]) may have some bearing on the regulation of gene expression ([21]) via transcription factor binding ([22] see also below). More than 4000 transcription factor binding motifs are enlisted in EMBL/Genebank databases but no proteins are yet known that bind to double-stranded, simple repetitive DNA, e.g. to the $(GAA)_n$, $(GT)_n$ or $(GT)_n(GA)_m$ blocks covered below (latest data bank check on April 22nd, 1996; but see [23]).

*Corresponding author. Fax: (49) (234) 700 4196.
E-mail: joerg.t.epplen@rz.ruhr-uni-bochum.de

PII S0014-5793(96)00526-1

In the wider context of biological significance, the question arose if simple repeats are essentially naked DNA stretches in the chromatin of eukaryotic chromosomes. A quick survey using chemically synthesized 30-mer oligonucleotides, each containing one of 14 different di- and trinucleotide motifs, suggested that practically all of them were able to bind different nuclear proteins [24]. In general, perfectly complementary oligonucleotides with *non*-redundant sequence content are standard targets for nucleic acid/protein interaction studies. A few methodological shortcomings of the aforementioned investigations involving simple repeats [24] include the lack of convincing affinity studies and of appropriate controls for partial single strandedness of the reannealed 30-mer targets. Therefore, the general relevance of these oligonucleotide data for repeat elements in large eukaryote genomes can hardly be evaluated at present. In addition, any influences of flanking sequences would not have been registered in these investigations involving 'pure' simple repeats. We detail here our understanding of nuclear protein binding to simple repetitive sequences by shortly summarizing 5 years of experimentation in this field [25–30].

2. Nuclear proteins bind to genome-derived simple repetitive sequence elements

A number of methodological difficulties and artifactual protein binding to simple repetitive DNA have been delineated [27,28]. Consequently, the interpretation of nuclear protein binding to simple repeats is not trivial. We concentrated our efforts on three classes of basic motifs: di- and trinucleotide polypurines and alternating purine/pyrimidine stretches as well as a certain composite form thereof. Fig. 1 is a representative gel retardation experiment depicting that the di- and trinucleotide motifs covered here bind HeLa nuclear proteins. Clearly, the binding characteristics for different motifs are different. Different targets bind different nuclear proteins [25–30] (Table 1). The binding affinities are not only based on the motifs, but they also depend on a minimal length of the simple repeat block (see Fig. 1 [24,27]).

One of the repeats which we studied for binding nuclear proteins was $(GAA)_n$. The motif $(GAA)_n$ was first cloned in the context of a most intriguing single-copy locus in the chicken that defies Mendelian inheritance laws [31]. Non-Mendelian transmittance was also observed independently for a $(GAA)_n$ locus in the rabbit by oligonucleotide fingerprinting (Epplen et al., unpublished data). Concomitantly, the distribution and organization of $(GAA)_n$ blocks was characterized in the human genome: These repeats appear considerably less frequent than the most abundant $(GT)_n$ dinucleotide stretches (see below). In addition, the motif $(GAA)_n$ is hardly ever present in mature mRNA as was deduced on the basis of Northern blot hybridization experiments [32]. However, these polypurine trinucleotides may exert profoundly negative effects on the transcription/RNA processing of certain genes whenever elongated substantially as demonstrated recently in the first intron of the *frataxin* gene [18]. More subtle expression effects have been claimed 'under physiological conditions' for highly polymorphic $(GAA)_n$ repeats in the 5′ untranslated portion of the mRNA encoding e.g. the HLA-F transplantation antigen [33]. The protein binding to a double-stranded $(GAA)_{24}$ block (Fig. 1, right-hand side) is characterized by high affinity binding comparable to that measured in experiments involving dinucleotide repeats (see below and [29]). DNA footprinting results show that only the $(GAA)_n$ repeat of the polypurine strand is preferentially protected by the bound protein against DNase I digestion, whereas the complement is not. In addition, chemical modification reactions (OsO_4 and DEPC treatment) revealed that the $(GAA)_{24}$ repeat exhibits a complex non-B DNA conformation (for further details see Mäueler et al., submitted).

The second repeat element studied for nuclear protein binding was $(GT)_n$. $(GT)_n$ elements are the most abundant and polymorphic in humans as well as other species. We concentrated on the protein binding properties of polymorphic $(GT)_n$ blocks in certain variable elements of antigen receptor genes of T lymphocytes [28,30]. As a prerequisite for specific protein binding to $(GT)_n$ repeats there must be a minimum length. Nuclear proteins do not bind to $(GT)_n$ motifs of 6 dinucleotide units whereas 13-mers are bound [29]. This conclusion is supported by the observation that double-stranded $(GT)_8/(AC)_8$ oligonucleotides cannot compete for binding [29]. Footprint analyses revealed that, in $(GT)_n$ blocks, similarly to the mixed $(GT)_n(GA)_m$ (see below and [29]), repeat-adjacent, non-tandem sequences are obligatorily protected from DNase I digestion (Mäueler et al., in preparation).

Like the polypurine trinucleotide $(GAA)_n$, perfect $(GA)_m$ repeat elements are less frequent compared to over-abundant $(GT)_n$ blocks [11]. The $(GA)_m$ stretches are also known to form special secondary structures including triplex formation [34]. These elements were studied for nuclear protein binding flanked 5′ by a perfect $(GT)_n$ dinucleotide repeat. One such composite element has been traced in all vertebrate MHC class II *DRB* genes at the very same position 50 bases downstream from the exon 2/intron boundary (for a review see [35]). Hence, this composite simple repeat block is remarkably preserved in evolution – absolutely in contrast to the expectation of rapid evolution and selective neutrality for such elements. In *DRB* pseudogenes this perfect tandem organization of the simple repeat block appears interrupted and degenerating arguing for functional significance of this repeat element [4]. Arguments have been made that this composite repetitive block has a role in directing genomic exchanges (cross-over events) during evolution [35]. By comparing nuclear protein binding to genome-derived (cloned) targets in addition to single-stranded and double-stranded, sequence-matched synthetic

Table 1
Results of competition experiments between various genome-derived target sequences containing simple repeat blocks as primary components[a]

	Labelled targets	
	$(GT)_{22}(GA)_{15}$	$(GAA)_{24}$
Competitors		
$(GT)_{22}(GA)_{15}$[b]	+++	–
$(GT)_{25}(GA)_{10}CA(GA)_3CA(GA)_6$[b]	+++	n.d.
$(GT)_6$	–	–
$(GT)_{13}$[b]	–	–
$(GT)_{25}$[b]	–	–
$(GAA)_{24}$[b]	–	+++

[a]For all methodological details and further explanations see [25–29].
[b]Each of the simple repeat targets was an efficient competitor for protein binding whenever itself was the labelled target. +++, competition (K_D range $0.8–1.5\times10^{-8}$ M/l); –, no competition demonstrable; n.d., not done.

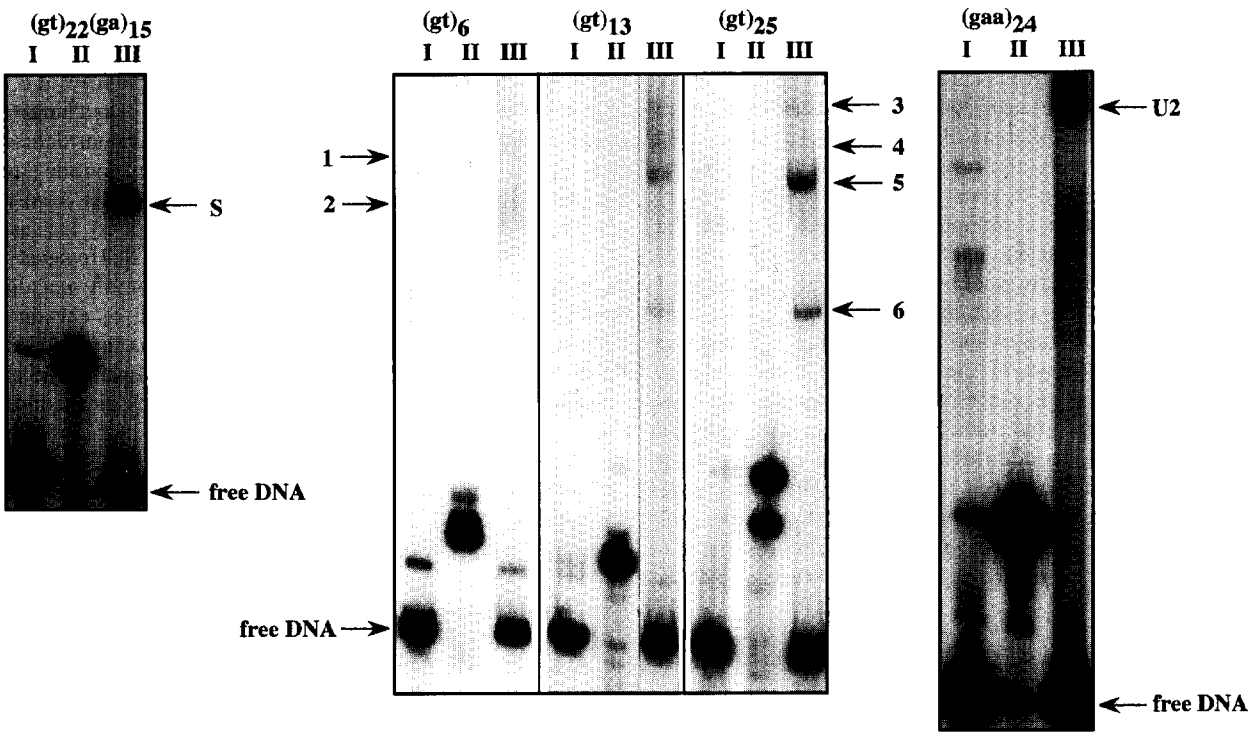

Fig. 1. Gel shift analyses of various nuclear proteins from HeLa cells bind differentially to genome-derived target sequences containing simple repeat blocks as major components. The simple repeat composition is depicted on the top, the origins of the targets are detailed in the original references [25–29]. Arrows in the upper half point to shifted bands designated in the same manner as in the original publications (op. cit.). Note that the differing intensities of the shifted bands reflect different protein binding efficiencies. I, double-stranded, labelled target; II, single-stranded, labelled target after denaturation (and immediate chilling); III, double-stranded, labelled target plus nuclear protein extract.

oligonucleotides, it was shown that binding differences included repeat flanking sequences. These results agree with interpretations of DNase I footprints [29] in that several bases outside the repeat region are protected from DNase I digestion suggesting that the binding protein (complex) also covers adjacent, non-repetitive territory.

3. Implications

A literature review on protein binding to single-stranded nucleic acids containing simple repeat stretches reveals a plethora of isolated, so-far non-cohesive data (see e.g. [36,37]). A critical discussion appears warranted since several of the simple sequence elements are able to form secondary structures with partly unpaired DNA strands and then bind proteins [38]. The relevance of such data is twofold, regarding both the genomic organization as well as primary or mature RNA transcripts. Altered secondary structure in the DNA has implications for the regulation of gene expression. Examples of this are the influence on gene expression of tandem repetitive sequences (GT rich minisatellite) 5′ to the insulin gene [20,21] and a $(GT)_n$ microsatellite in the Nramp gene (cited according to the oral presentation of Dr. Jenny Blackwell, Manchester, March 15, 1996). In combination with differential stability of microsatellites due to slippage mutations [39], a whole range of minor and major effects of repeat lengths could be expected theoretically with respect to gene activity. Given the extensively high numbers of simple repeats e.g. in the human genome ($\sim 10^6$), it is immediately obvious that each of the different elements cannot bind its own or even several specific proteins. Since the total gene number is less than 10^5 [40] only combinatorial options of protein complexes consisting of several independently encoded entities offered truly specific interactions. In addition, the potential for variability is based on different repeat lengths and the genomic environment. Interestingly, extracts from nuclei of various cell types (B lymphocytes, T cells, HeLa cells) differ only slightly, but especially with respect to relative band intensities in gel shift experiments involving, e.g. the composite $(GT)_n(GA)_m$ repeat [2].

4. Conclusions

Protein interactions with genomic simple repeat blocks are interesting in their own right. Yet the relevance of simple repetitive DNAs in human genetic conditions makes their thorough characterization mandatory. The causal pathogenesis of trinucleotide diseases as well as additional novel genomic alteration/interaction phenomena are still to be elucidated. Different secondary structures of the simple repeats could be used physiologically as landmarks in the wide genomic 'desert' or in the nuclear architecture. Differential protein binding depends on the genomic environments and on the lengths of the perfect simple repeat blocks. On the basis of allelic length variations, differential protein binding has enormous potential for influences on the regulation of gene expression and/or genomic rearrangements. Thus simple repeats have gained interest as markers as well as as targets for subtle effects on the expression of genes via differential protein binding with respect to the genetic background of multifactorial diseases. Such regulatory effects are demonstrable in particularly well-characterized conditions, but the complexities of

such interactions do not allow firm and generalizable statements as to their biological relevance as of today. Nevertheless, nuclear protein binding to genome-derived, simple repetitive DNA sequences proves sufficiently complex to warrant more surprises.

Acknowledgements: We thank our departmental members for discussions, especially Dietrich Stephan for correcting the English.

References

[1] Tautz, D. (1993) in: DNA Fingerprinting: State of the Science (Pena, S.D.J., Chakraborty, R., Epplen, J.T. and Jeffreys, A.J. eds.) pp. 21–28, Birkhäuser, Basel.
[2] Epplen, C., Melmer, G., Siedlaczck, I., Schwaiger, F.-W., Mäueler, W. and Epplen J.T. (1993) in: DNA Fingerprinting: State of the Science (Pena, S.D.J., Chakraborty, R., Epplen, J.T. and Jeffreys, A.J. eds.) pp. 29–45, Birkhäuser, Basel.
[3] Ohno, S. (1972) in: Evolution of Genetic Systems (Smith, H.H. ed.) pp. 366–370, Gordon and Breach, New York
[4] Epplen, J.T., Santos, E.M. and Epplen, C. (1996) Trends Genet., submitted.
[5] Dover, G. (1995) Nat. Genet. 10, 254–256.
[6] Epplen, J.T., Buitkamp, J., Bocker, T. and Epplen, C. (1995) Gene 159, 49–55.
[7] Weber, J.L. and May, P.E. (1989) Am. J. Hum. Genet. 44, 388–396.
[8] Litt, M. and Luty, J.A. (1989) Am. J. Hum. Genet. 44, 397–401.
[9] Weber, J.L. and Wong, C. (1993) Hum. Mol. Genet. 8, 1123–1128.
[10] Stallings, R.L., Ford, A.F., Nelson, D., Torney, D.C., Hildebrand, C.E. and Moyzis, R.K. (1991) Genomics 10, 807–815.
[11] Stallings, R.L. (1995) Genomics 25, 107–113.
[12] Ashley, C.T. and Warren, S.T. (1995) Annu. Rev. Genet. 29, 703–728.
[13] Richards, R.I. and Sutherland, G.R. (1992) Cell 70, 709–712.
[14] Roses, A.D. (1996) Nat. Med. 2, 267–269.
[15] Warren, S.T. (1996) Science 271, 1374–1375.
[16] Li, X.-J., Li, S.-H., Sharp, A.H., Nucifora, F.C., Jr, Schilling G., Lanahan, A., Worley, P., Snyder, S.H. and Ross, C.A. (1995) Nature 378, 398–402.
[17] Burke, J.R., Enghild, J.J., Martin, M.E., Jou, Y.-S., Myers, R.M., Roses, A.D., Vance, J.M. and Strittmatter, W.J. (1996) Nat. Med. 2, 347–350.
[18] Campuzano,V., Montermini, L., Moltò, M.D., Pianese, L., Cossée, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., Monticelli, A., Zara, F., Canizares, J., Koutnikowa, H., Bidichandani, S.I., Gellera, C., Brice A., Trouillas, P., De Michele, G., Filla, A., De Frutos, R., Palau, F., Patel, P.I., Di Donato, S., Mandel, J.-L., Cocozza, S., Koenig, M. and Pandolfo, M. (1996) Science 271, 1423–1427.
[19] Epplen, C., Epplen, J.T. (1994) Hum. Genet. 93, 35–41.
[20] Benett, S.T., Lucassen, A.M., Gough, S.C.L., Powell, E.E., Undlien, D.E., Pritchard, L.E., Merriman, M.E., Kawaguchi, J., Dronsfield, M.J., Pociot, F., Nerup, J., Bouzekri, N., Cambon-Thompson, A., Ronningen, K.S., Barnett, A.H., Bain, S.C. and Todd, J.A. (1995) Nat. Genet. 9, 284–292.
[21] Hamada, H., Seidman, M., Howard, B.H. and Gorman, C.M. (1984) Mol. Cell. Biol. 4, 2622–2630.
[22] Kennedy, G.C., German, M.S. and Rutter, W.J. (1995) Nat. Genet. 9, 293–298.
[23] Gilmour, D.S., Thomas, G.H. and Elgin, S.C.R. (1989) Science 245, 1487–1490.
[24] Richards, R.I., Holman, K., Yu, S. and Sutherland, G.R. (1993) Hum. Mol. Genet. 2, 1429–1435.
[25] Mäueler, W., Muller, M., Köhne, A.C. and Epplen, J.T. (1992) Electrophoresis 13, 7–10.
[26] Mäueler, W., Frank, G., Siedlaczck, I., Epplen, J.T. and Melmer, G. (1992) Electrophoresis 13, 641–643.
[27] Mäueler, W., Kyas, A., Bröcker, F. and Epplen, J.T. (1994) Electrophoresis 15, 403–410.
[28] Mäueler, W. and Epplen, J.T. (1995) Trends Genet. 11, 170.
[29] Mäueler, W., Frank, G., Muller, M. and Epplen, J.T. (1994) J. Cell. Biochem. 56, 74–85.
[30] Epplen, J.T., Buitkamp, J., Epplen, C., Mäueler, W. and Rieß, O. (1995) Electrophoresis 16, 683–690.
[31] Epplen, J.T., Ammer, H., Kammerbauer, C., Schwaiger, W., Schmid, M. and Nanda, I. (1991) Adv. Mol. Genet. 4, 301–310.
[32] Siedlaczck, I., Epplen, C., Rieß, O. and Epplen, J.T. (1993) Electrophoresis 14, 973–977.
[33] Geraghty, D.E., Wei, X., Orr, H.T. and Koller, B.H. (1990) J. Exp. Med. 171, 1–18.
[34] Noonberg, S.B., Francois, J.-C., Garestier, T. and Hélène, C. (1995) Nucleic Acids Res. 23, 1956–1963.
[35] Schwaiger, F.-W. and Epplen, J.T. (1995) Immunol. Rev. 143, 199–224.
[36] Muraiso, T., Nomoto, S., Yamazaki, H., Mishima, Y. and Kominami, R. (1992) Nucleic Acids Res. 20, 6631–6635.
[37] Aharoni, A., Baran, N. and Manor, H. (1993) Nucleic Acids Res. 21, 5221–5228.
[38] Hollingsworth, M.A., Closken, C., Harris, A., McDonald, C.D., Pahwa, G.S. Maher, L.J. III (1994) Nucleic Acids Res. 22, 1138–1146.
[39] Levinson, G. and Gutman, G.A. (1987) Mol. Biol. Evol. 4, 203–221.
[40] Bird, A.P. (1995) Trends Genet. 11, 94–100.