DNA repetitive sequences-types, distribution and function: A review

Satyawada Rama RAO*,1, Seema TRIVEDI2, Deepika EMMANUEL2, Keisham MERITA1 and Marlykynti HYNNIEWTA1

1Cytogenetics and Molecular Biology Laboratory, Department of Biotechnology and Bioinformatics, North-Eastern Hill University, Permanent Campus, Mawkyrnoh, Umning, Shillong- 793022, Meghalaya (INDIA)
2Department of Zoology, Jai Narain Vyas University, Jodhpur- 342005, Rajasthan (INDIA)
(* author for correspondence; sr Rao@yahoo.com)

Received: 21 September 2009; Accepted: 14 May 2010

Abstract

The development and use of molecular markers for the detection and exploitation of DNA polymorphism is one of the most significant developments in the field of molecular genetics. DNA based molecular markers have acted as versatile tools and have found their own position in various fields like taxonomy, physiology, embryology, genetic engineering etc. A major step forward in genetic identification is the discovery that about 30-90% of the genome is constituted by regions of repetitive DNA which are highly polymorphic in nature. Microsatellites are multilocus probes creating complex banding patterns and are usually non-species specific occurring ubiquitously. They form an ideal marker system and are dominant fingerprinting markers and co-dominant STMS (sequence tagged microsatellites) markers. Microsatellites markers have been used successfully to determine the degree of relatedness among individuals or groups of accessions to clarify the genetic structure or partitioning of variation among individuals, accessions, populations and species. Repetitive sequences have been widely used for examining genome and species relationships by in situ and by Southern hybridization.

Keywords: Satellites, microsatellites, minisatellites, retroposons and proretroviral transposons

DNA repetitive sequences-types, distribution and function: A review


Anahtar Sözcükler: Satelit, mikrosatellitler, minisatellitler, retropozonlar ve proretroviral transpozonlar
Introduction

The analysis of genetic diversity and relatedness between or within different species, populations and individuals is a central task for many disciplines of biological science. Classical strategies of evaluating genetic variability are comparative anatomy, morphology, embryology and physiology. These are complemented by analysis of chemical constituents like plant secondary compounds or with specific characterization of macromolecules and allozymes. In recent years, focus has been shifted to the development of molecular markers based on DNA or protein polymorphism. The importance of these studies lies in exploitation of uniqueness of DNA sequences that facilitate research in diverse disciplines such as taxonomy, phylogenetics, ecology, genetics and plant breeding.

Establishing an individual’s identity is one of the uses of DNA sequence information that highlight uniqueness of a particular sample. The methodology focuses on ways to reduce complexity of DNA into simple patterns that are representative of the sample. This type of analysis is called fingerprinting, profiling, genotyping or identity testing. Jeffreys et al. (1985) introduced this term to describe a method for the simultaneous detection of variable DNA loci by hybridization of specific multilocus probes with electrophoretically separated restriction fragments. DNA fingerprinting is useful for forensic identification, determination of family relationship, linkage mapping, antenatal diagnosis, localization of disease loci, determination of genetic variation, molecular archaeology and epidemiology (Watkins, 1988; Donis-Keller et al., 1987; Landegren et al., 1988; Paabo, 1989; Golenberg et al., 1990). Molecular markers have been used for identification of individuals, clones, close relatives, paternity testing or in studies of reproductive behavior and mating success.

Repetitive sequences as molecular markers

A repeat is recurrence of a pattern whereby DNA exhibits recurrence of many features. The number of occurrences of a pattern is called copy number. The number of copies in a particular tandem repeat region is termed region copy number. The term genome copy number refers to number of copies of tandem or interspersed repeats in genome.

The repetitive DNA family(ies) may be widely distributed in a taxonomic family or a genus, or may be specific for a species or chromosome. Repeats may occur in specific locations in a genome, e.g. in telomeric regions or scattered throughout the genome. They may acquire large scale variation in the sequence and copy number over evolutionary time-scale. The repetitive elements are under different evolutionary constraints as compared to the genes. Hybrid polyploids are excellent models for studying evolution of repetitive sequences (Kubis et al., 1998). These variations are the basis of utilization of repetitive sequences for taxonomic and phylogenetic studies (Smith and Flavell, 1974).

There are many classifications of repetitive DNA based on characteristics measured by different techniques but consolidation of these systems defines five broad classes: satellites, microsatellites and minisatellites, retroposons and retroviral transposons. The classification scheme makes a distinction between repetitive regions exhibiting tandem repetition and interspersed repetition but is not precise since each class retains the characteristics of both. Some of these repeats are described as follows:

Moderately repetitive DNA includes reiterations of genes like tRNA, tRNA, hemoglobin etc. that retain similar or nearly similar sequences due to duplication. Some of these duplications result in pseudogenes and may have many copies in the genome. Some repetitive DNA sequences are transposable elements since they cannot enhance the success of the cell (or organism) they reside in, behave selfishly and also accumulate to the levels restricted only by the resources available to them. The selfish DNA hypothesis of Doolittle and Sapienza, (1980) assumes that repetitive DNA can behave in a selfish manner because it is not functional. Indeed, there is some evidence that its presence can result in losses of fitness of the host cell due to mutations caused by transposable elements. However, some moderately-repetitive DNA has functions for example, in directing chromosome movement in eukaryotes (Vogt, 1990). Variations in selfish DNA have the potential for evolutionary changes, especially when it changes without having any deleterious effects on the organism (Flavell et al., 1977). Susumo Ohno (1970) asserted that "natural selection merely modified while redundancy created". Duplication of genes can thus be internal source of novelty in
the genome. If repetitive DNA is transposable, it may create novel genes. Repetitive DNA is therefore the "Research & Development" laboratory of genome, creating both redundancy and novel sequences that may prove valuable for genome. However, these repetitive sequences are generally not used for DNA fingerprinting.

**Tandem and interspersed repeats**

Tandem repetitions are consecutive head-to-tail, direct, repetition of a pattern due to local duplication. Interspersed repetitions are recurrence of patterns that may or may not be proximal, formed by either non-local duplication or multiple introductions of the same or similar extraneous DNA segments. These repeats are dispersed throughout the genome and have no restriction on the relative positions of identical occurrences occurring in tandem locations. Research indicates that interspersed repeats are inserts since they resemble either processed RNAs i.e. retroposons, or viruses i.e. proretroviral transposons. In addition, a suspected target sequence for insertion occurs at both ends of these repeats as expected for a circular DNA crossover insertion. Furthermore, some repeats actively move within the genome, such as jumping genes in maize.

DNA repeat patterns also classify as direct, indirect, complement, reverse complement or palindrome. A direct or forward repeat is the recurrence of a pattern on the same strand in the same nucleotide order; e.g. ACCG recurs as ACCG. An indirect, inverse or reverse repeat recurs on the same strand but the order of the nucleotides is reverse, e.g. the indirect recurrence of ACCG is GCCA. Complement repeats are repeats where the nucleotides are complemented according to Watson Crick pairing, e.g. the complement of ACCG is TGGC. A reverse complement repeat recurs on the same strand but, the nucleotides are complemented and the order of the nucleotides is reversed; e.g. the reverse complement of ACCG is CGGT. In DNA, most repetitions occur as forward or reverse complement repeats and rarely as reverse or complement repeats (Grumbach and Tah, 1994). Palindrome is a combination of two consecutive occurrences in opposite orientations and read the same when read from left to right or vice-versa.

Repetitive DNA sequences, divided into high-repeat satellite DNA which replicates thousands or millions of times and "moderate-repeat" minisatellite and microsatellite DNA which replicates tens to perhaps a thousand times, account for varying proportions of the genome of multicellular eukaryotes. An example of representative data from eukaryotes has been given in Table 1.

Prokaryotes contain little or no repetitive sequences. Noncoding repetitive DNA varies from one group of organisms to another; individual to individual and therefore used as DNA fingerprinting tool.

**Table 1. Proportion of repetitive sequences of genomic DNA in different eukaryotes.**

<table>
<thead>
<tr>
<th></th>
<th>Drosophila</th>
<th>Xenopus</th>
<th>Mouse</th>
<th>Tobacco</th>
</tr>
</thead>
<tbody>
<tr>
<td>High repeat</td>
<td>13%</td>
<td>3%</td>
<td>10%</td>
<td>5%</td>
</tr>
<tr>
<td>Moderate repeat</td>
<td>13%</td>
<td>43%</td>
<td>20%</td>
<td>65%</td>
</tr>
<tr>
<td>Non repetitive</td>
<td>74%</td>
<td>54%</td>
<td>70%</td>
<td>30%</td>
</tr>
</tbody>
</table>

**Tandem repeats and Satellite DNA**

As repeats were discovered in different locations exhibiting different copy numbers, new terms arose such as satellite, minisatellite and microsatellite. Some researchers refer to all types of satellites as tandem repeats and describe a specific tandem repeat region according to its location within the genome, its periodicity, pattern structure and copy number. These repeats were first identified on a cesium chloride buoyant density gradient as peaks separate from the primary DNA peak. The separate or satellite peaks were composed of array of highly conserved tandem repeats localized to heterochromatic regions of chromosomes like centromeres (Schueler et al., 2001). The structure of a tandem repeat region has well-conserved
pattern but varies in size from less than 20 bp to several thousand bp.

**Structural and functional roles**

Tandem repeats play significant structural and functional roles. They occur in abundance in structural areas such as telomeres, centromeres and histone binding regions. They play a regulatory role near genes and perhaps even within genes.

**Transcription**

The precise role of tandem repeats in transcription regulation is not known. Since nucleosomes can repress or enhance transcription initiation and elongation (Hartzog and Winston, 1997; Kornberg and Lorch, 1999) repeats may influence transcription by affecting nucleosome positioning and stability. Tighter bonds between the histone complex and repeats restrict access for RNA polymerase and regulatory proteins (Dai & Rothman-Denes, 1999). This may happen by changing the degree and direction of DNA supercoiling or forming alternative DNA structures such as cruciforms and hairpins (Shlyakntenko et al., 1998; Ohyama, 2001). Tandem repeats having an alternating purine (R=A or G) pyrimidine (Y=C or U/T) pattern forms Z-DNA (Yang et al., 1996) and repeats with a RRY or a YRY pattern form triplex DNA structures (Grabcyzk and Usdin, 2000). The degree of repression is directly proportional to repeat length.

**Centromeric and subtelomeric satellite DNA families**

The tandem satellite DNA sequences exhibit characteristic chromosomal locations, usually at subtelomeric (or intercalary repetitive sequences) and centromeric regions (Heslop-Harrison et al., 2003; Jiang et al., 2003). Satellite DNA families may arise de novo due to molecular mechanisms like unequal crossing over, rolling circle amplification, replication slippage and mutation. Satellite DNA have variable repeat unit length (sometimes equivalent to micro or minisatellite length), often forming arrays spanning up to 100 Mb (Charlesworth et al., 1994; Kubis et al., 1998; Schmidt and Heslop-Harrison, 1998; Vergnaud and Denœud, 2000). However, satellite repeat monomer lengths of 140 – 180 bp and 300 – 360 bp, corresponding to the length of the mono and dinucleosomes are most the common (Hemleben, 1990; Traut, 1991; Macas et al., 2002).

Centromeric tandem repeats ranging from 150-200 bp in length (Henikoff et al., 2001) are essential components of a functional centromere. A functional centromere has been defined as the DNA sequence which interacts with the kinetochore where the interaction between centromere-kinetochore appears to be mediated by DNA-protein recognition process (Jiang et al., 2003). The core sufficient for centromeric function is an alpha satellite about 3 Mbp long having a 171 bp pattern recurring in a tandem fashion. (Schueler et al., 2001; Zhong et al., 2002).

A highly repetitive 180 bps centromeric satellite DNA family constituting between 2-5% of the Arabidopsis thaliana genome is the key component of its centromere/kinetochore complex (Nagaki et al., 2003a,b). These repeats are occasionally interrupted by the Athila retrotransposons, although the latter are mainly clustered in pericentromeric regions (Heslop-Harrison et al., 1999; Nagaki et al., 2003a,b). Similarly, centromeric DNA in several plants species including rice, maize, wheat, Beta species and Zingeria biebersteiniana mainly contain satellite sequence repeats and retro-transposons (Gindullis et al., 2001; Kishii et al., 2001; Kumekawa et al., 2001; Saunders and Houben, 2001; Cheng et al., 2002; Nagaki et al., 2003a,b). A high monomer divergence is observed within several centromeric repetitive DNA families thereby indicating presence of chromosome specific variant sequences (Harrison and Helsop-Harrison, 1995; Nagaki et al., 1998; Helsop-Harrison et al., 2003). For example, chromosome specific 180 bp satellite repeat variants in Arabidopsis thaliana may be explained by the possibility that either the repeat sequences on each chromosome have been homogenizes independently or specific variants of the satellite sequence have been amplified on each chromosome (Heslop-Harrison et al., 1999).

The subtelomeric regions also contain repetitive sequences (review in Pryde et al., 1997). Not all species have the same structure but all have structures containing tandem repeats, interspersed repeats or both (Pryde et al., 1997). Degenerate TTAGGG repeats enable alignment other subtelomeric regions allowing sequence exchange between subtelomerases (Flint et al., 1997).
Minisatellite and Microsatellite DNA

Hypervariable regions, also known as variable number of tandem repeats (VNTRs) classified as minisatellites and microsatellites are regions that contain a variable copy number. These repeats are found throughout the genome (Vogt, 1990) but rarely within genes. Most regions contain short to moderate region copy number (Jeffreys, 1985). DNA fingerprinting capitalizes on the differences between alleles at specific VNTR loci. Various human diseases are attributed to high copy numbers associated with some VNTR locus.

Minisatellites are characterized by moderate length patterns, usually less than 50 bp (Jeffreys, 1985) with an array of 0.5 - 30kb. Two types of variability are observed, viz., one displays copy number variation with each replication event whereas the other displays distinct alleles within a population such that different alleles contain different copy numbers.

Microsatellites, also known as simple sequence repeats (SSRs) or simple tandem repeats (STRs) have a short well-conserved pattern length of 2 to 6 bp and region copy number of 10 to 40 pattern copies. Microsatellites have been found in non-centromeric regions, many of them being located either near or within genes.

Automatic identification and characterization of tandem repeats is crucial as genome projects generate an ever-increasing quantity of sequence data. Tandem repeats increase the complexity of genome sequence analysis algorithms. For instance, the process of generating full chromosome sequences often utilizes the sequence assembly procedure; a procedure that stitches short, similar fragments together to reconstruct a larger sequence. The consecutive recurrence of a pattern associated with tandem repeats confuses this process. Some commercially available algorithms avoid assembling tandem repeat regions. Others often assemble moderate-sized tandem repeat regions improperly. At present, algorithms are being developed for handling tandem repeat regions.

The mechanism responsible for minisatellite and simple sequence polymorphisms

Minisatellites and simple sequences are often characterized by high mutation rates (up to 5%), which may involve either internal heterogeneity of repeats or their number. Mutation rates also show positive correlation with the total size of the array of repeats. In accordance with these observations, high molecular weight bands within a multilocus fingerprint are often more variable than bands occurring in the low molecular weight range. The molecular basis of both minisatellites and simple sequence variability is still debatable. Possible mechanism include replication slippage, transposition, recombinational events and/or unequal exchange between sister chromatids or between homologous chromosomes and gene conversion (reviewed by Jarman and Wells, 1989; Jeffreys et al., 1990; Richards and Sutherland 1992; Wolff et al., 1991.)

The slippage hypothesis implicates mispairing of slipped-strand during the replication process. Strand slippage may happen due to shift in origin of replication especially during lagging strand synthesis. Strand slippage and mismatch appear to be nucleotide specific. Differential activities of mismatch pair of (CAG)_n repeats occur but not of (CTG)_n repeats. Certain factors like the length of the repeats and replication direction play a role in destabilizing (CAG)_n (CTG)_n repeat. Such positioning effects results in loop formation due to strand slippage and results in expansion or reduction of repeat during replication.

Several lines of evidence have lent support to the recombination hypothesis:
- A variety of minisatellite core sequences share homology of the bacterial recombination signal chi.
- Minisatellite - like sequences have been found at sites of meiotic crossing over.
- Both minisatellite and macrosatellites behave as recombinational hot spots in transfected mammalian cells.

Wolff et al., (1991) observed no exchange of flanking markers in a newly created minisatellite allele, thus ruling out unequal exchange between homologous chromosomes as a mutational mechanism. In human minisatellite locus, MS32 (reviewed by Jeffreys et al., 1985), 5’ end of the array has a strong mutation bias, suggesting existence of a mutational hot spot. Some mutant alleles contain segments from both parental alleles, providing evidence for interallelic exchange. It is suggested that the major mutational process involves nonreciprocal transfer of repeats from a donor allele to the 5’ end of a recipient allele. Therefore, recombinational processes as well as replication slippage may contribute to the creation of minisatellite and simple sequence variability. However, other (yet unidentified) mechanisms may
also be involved, especially in case of the explosive amplification of microsatellite like trinucleotide repeats associated with human genetic diseases and polymorphism. Structural analysis of mutated vs. parental alleles may help to gain more information about the mutational mechanisms. In this respect, transgenic systems will be informative, since successive deletion of the flanking DNA will allow precise location of mutational hot spots.

**Retroposons**

Retroposons resemble processed RNAs and transpose passively via RNA intermediate (Weiner, 1986). Each element is composed of an A-rich tail at the 3' end and short target site duplications (direct repeats of 5-21 bp) flanking the repeat (Rabin, 1985). Two main subclasses dominate this class:

**Short Interspersed Elements (SINEs)**

These are distributed throughout the non-centromeric regions of genome (over 100,000 copies per genome) (Weiner, 1986). A SINE contains one or more RNA polymerase III, promoter sites and an A-rich region. One subfamily is composed of a head-to-tail catenation of two promoter site, A-rich region pairs (Weiner, 1986). Both subfamilies are flanked by short direct repeats of 5 to 21 bp. Primate specific Alu sequence (5 to 9 kbp) is a SINE with two promoter sites and a dimer. The uniqueness of Alu sequences provides a wonderful tool for separating primate DNA from that of other species. SINEs present challenges to sequence assembly due to their high genome copy number (300,000 to 500,000 copies) (Rogers, 1985).

**Long Interspersed Elements (LINEs)**

LINEs are composed open reading frames (ORFs) followed by a 3' A-rich region having 20,000 to 50,000 copies per genome (Hutchison et al., 1989; Weiner, 1986). Direct repeats of 6-15 bp flank the element. L1 family (primary LINE family) is 6 to 7 kbp long. The consensus structure of the family is well defined but not well conserved because L1 element can deviate significantly from the structure such that entire structural components are deleted or duplicated (Weiner, 1986).

**Proretroviral transposons**

Proretroviral transposons are mobile elements that transpose via RNA intermediate (Varmus and Brown, 1989). Their structure and content resembles integrated viruses and often contain genes encoding viral products, e.g. protease, reverse transcriptase and integrase (Boeuf and Corces, 1989). The LTRs contain transcriptional signals for initiating and terminating transcripts, a promoter, an enhancer and a polyadenylation signal (Temin, 1985; Schmid et al., 1990). Inverse repeats exist at the ends of each LTR and always begin with the bases, TG, and end with CA (Temin, 1985). The two LTRs and the genes are flanked by 4 to 6 bp direct repeats.

**Other recurring genetic features**

DNA contains many recurring features that do not classify as tandem or interspersed repeat. A gene cluster is a group of proximal genes having similar sequence and often, similar structure but, different function. There may be requirement for multiple copies of functional genes tRNA or rRNA genes. Copies of promoters and other regulatory regions associated with many genes also do not classify as repetitive DNA.

**Telomeres**

Telomeric DNA is G-rich consisting of the 3’ overhang and adjacent tandem repeat with wide variation in length across species (reviewed in Blackburn, 1991; Hemann and Greider, 1999). For example, length of telomere TTAGGG repeats in humans is 5 to 15 kbp but in mouse (Mus musculus) it is ~50 kbp. Yeast, Saccharomyces cerevisiae, has irregular pattern of TG1-3 and repeat length of ~300 bp. A recent model suggests that this region does a d-loop-t-loop by having the 3’ overhang invade the tandem repeat (Griffith et al., 1999). This invasion forms a triplex DNA structure, d-loop, and encloses a large segment of duplex DNA in a terminal loop or t-loop. Telomere length and size of loops is species specific (Shore, 2001). Universal presence of this structure across species is not clear though there may be telomeres that are unable to form a t-loop (Griffith et al., 1999).
DNA repetitive sequences

Nucleosomes

Periodicity of di-nucleotides (TATA-tetrads) or tandem repeat with a 10 bp pattern of 5’ TATAA(A/C)CG(T/C)C 3’ band DNA and form association with histone proteins (Widlund et al., 1997). However, tandem repeats may increase or decrease nucleosome stability. For example, a tandem repeat having a CAG (=CTG) pattern located close to a nucleosome increases its stability (Wang et al., 1994; Wang and Griffith, 1995; Godde and Wolffé, 1996). On the other hand, tandem repeat CGG (=CCG) has no impact unless it is methylated. Methylated CGG (=CCG) with a limited copy number increase the nucleosome stability while those with large copy numbers decrease nucleosome stability (Godde et al., 1996; Wang and Griffith, 1996).

Tandem repeats in genes

Tandem repeat hypervariability enables identification of genes e.g. antifreeze gene and several degenerative diseases. Repeats may help in stability of transcripts or proteins but repeat expansions and instability (particularly of trinucleotide repeats) lead to neurological disorders and cancer (Ashley and Warren, 1995; Mitas, 1997). Long stretch of CAG repeats translated into polyglutamine tracts result in a gain-of-function, possibly a toxin (Perutz et al., 1994; Baldi et al., 1999). CGG, AGG and TGG repeats form quadruplex and GAA repeats form triplex structures that can block or reduce transcription and DNA replication (Sinden, 1999). CGG repeats also destabilize nucleosomes (Sinden, 1999) due to CpG hypermethylation leading to promoter repression and lack of gene expression (Nelson 1995, Baldi et al., 1999). On the other hand, CTG repeats stabilize nucleosomes and block replication forks in E. coli (Sinden, 1999).

Evolution

Repeats have a role in genome evolution and possibly in C-value paradox. Variation in nuclear DNA amount in higher plants species exemplifies this. The variation (>2500 fold) in 1C DNA content in angiosperms ranges from 0.05 picograms in Cardamine amara to 127.4 picograms in Fritillaria assyriaca (Bennett, 1985). Part of such variation is due to the numerical changes in chromosomes but in many, there is substantial variation resulting from amplification or deletion of DNA sequences. Chromosomes of many monocot and dicot species contain fast reassociating highly repetitive fraction, slow reassociating middle repetitive fraction and single copy sequences (Britten and Kohne, 1968; Smith and Flavell, 1974; Flavell et al., 1977; Katsiotis et al., 2000). These sequences may be dispersed repetitive sequences including transposeable elements or tandem repeats. The retroelement class forms sometimes upto 50% component of plant genomes (Guidet et al., 1991; Heuros et al., 1993; Kubis et al., 1998; Bennetzen, 2000; Katsiotis et al., 2000; Linares et al., 2000; Ananiev et al., 2002).

References


DNA repetitive sequences


Rabin M. In "Discovering Repetitions in Strings," Combinatorial Algorithms on Words.


Saunders VA and Houben A. The pericentromeric heterochromatin of the grass Zingeria biebersteiniana (2n = 4) is composed of Zbcen1-type tandem repeats that are intermingled with accumulated dispersedly organized sequences. Genome. 44: 955-961, 2001.


