BIOINFORMATICS

Vol. 18 no. 1 2002
Pages 28–35

PlantSat: a specialized database for plant satellite repeats

Jiří Macas, Tibor Mészáros and Marcela Nouzová*

Institute of Plant Molecular Biology, Laboratory of Molecular Cytogenetics, Branišovská 31, České Budějovice, CZ-37005, Czech Republic

Received on May 20, 2001; revised on July 20, 2001; accepted on August 15, 2001

ABSTRACT

Motivation: Tandemly organized repetitive sequences (satellite DNA) are widespread in complex eukaryotic genomes. In plants, satellite repeats often represent a substantial part of nuclear DNA but only a little is known about the molecular mechanisms of their amplification and their possible role(s) in genome evolution and function. Unfortunately, addressing these questions via characterization of general sequence properties of known satellite repeats has been hindered by a difficulty in obtaining a complete and unbiased set of sequence data for this analysis. This is mainly due to the presence of multiple entries of homologous sequences and of single entries that contain more than one repeated unit (monomer) in the public databases.

Results: We have established a computer database specialized for plant satellite repeats (PlantSat) that integrates sequence data available from various resources with supplementary information including repeat consensus sequences, abundances, and chromosomal localizations. The sequences are stored as individual repeat monomers grouped into families, which simplifies their computer analysis and makes it more accurate. Using this feature, we have performed a basic sequence analysis of the whole set of plant satellite repeats with respect to their monomer length and nucleotide composition. The analysis revealed several preferred length ranges of the monomers (~165 bp and its multiples) and an over-representation of the AA/TT dinucleotide in the repeats. We have also detected an enrichment of satellite DNA sequences for the motif CAAAA that is supposed to be involved in breakage–reunion of repeated sequences.

Availability: The PlantSat database is accessible via a web interface (http://w3lamc.umbr.cas.cz/PlantSat) and can be searched for keywords, sequence motifs, and sequence homologies, or it can be used as a source of organized sequence data for further analyses.

Contact: macas@umbr.cas.cz

*To whom correspondence should be addressed.

INTRODUCTION

Highly abundant, tandemly arranged DNA repeats referred to as satellite DNA (satDNA) are widespread in complex eukaryotic genomes. In contrast to micro- and minisatellites, their monomers are tens to thousands of nucleotides long and often form continuous arrays spanning up to 100 Mbp (Charlesworth *et al.*, 1994; Schmidt and Heslop-Harrison, 1998; Kubis *et al.*, 1998). In higher plants, individual families of satDNA can comprise up to 20% of the nuclear genome (Ingham *et al.*, 1993), corresponding to 10^6–10^7 copies per haploid genome (Kato *et al.*, 1984; Ingham *et al.*, 1993; Irifune *et al.*, 1995; Macas *et al.*, 2000). The lengths of the repeated units and their nucleotide sequences vary significantly between satDNA families, as does the degree of their amplification even in evolutionary related species (Deumling, 1981; De Kochko *et al.*, 1991; Schmidt and Heslop-Harrison, 1993; Nouzová *et al.*, 1999; Macas *et al.*, 2000). Although several models have been proposed to explain amplification and maintenance of satellite DNA in eukaryotes (Smith, 1976; Walsh, 1987; Charlesworth *et al.*, 1994; Stephan and Cho, 1994), the precise molecular mechanisms are still unknown. Similar uncertainity concerns a possible role of satDNA in plant genomes since only a fraction of tandem repeats was found to have a specific function (Kubis *et al.*, 1998; Schmidt and Heslop-Harrison, 1998), none of these being typical satellite DNA. However, despite their nucleotide sequence divergences many satDNA sequences share common features like intrinsic curvature and specific chromatin folding structure (Vogt, 1992). Whether these features are required for a specific function or merely arise as a side effect of mechanisms involved in satDNA amplification and maintenance in the genome is still a matter of investigation.

Recently, novel sequences are being reported at an increasing rate which provides material for addressing the questions of satDNA evolution and function using approaches based on its computer analysis. However, such analysis is hampered by a difficulty to retrieve a complete and unbiased set of plant satDNA sequences from databanks such as GenBank and EMBL which

contain all types of sequences (Wheeler *et al.*, 2001; Stoesser *et al.*, 2001). This is mainly due to inconsistent annotations that make it difficult to distinguish satellite DNAs from other classes of tandemly repeated sequences, all of them being marked as tandem repeats or even merely as repetitive DNA. Yet another problem is the presence of multiple entries of homologous sequences and of single entries that contain more than one repeated unit (monomers) in the public databanks. This makes it impossible, for example, to analyze the distribution of monomer lengths by simply using the lengths of the retrieved databank entries. Therefore, we established a database of plant satellite DNA that is organized such that it reflects specific features of this type of repetitive sequences. In addition to offering a possibility to retrieve and analyze nucleotide sequences of individual repeat families, a web interface to the database offers text- and sequence-based searches including a BLAST homology search, and provides easy access to additional information regarding individual repeats.

SYSTEM AND METHODS

Data acquisition

Searching the GenBank database was done using the Entrez retrieval system (Wheeler *et al.*, 2001) and the following set of keywords used separately or in suitable combinations: sat*, satellite*, tandem, rep*, repet*, repeat*. The search was limited to seed plants (*Spermatophyta*) and the entries containing microsatellite and minisatellite sequences were excluded. We also excluded rDNA genes and subrepeats present in intergenic spacers (IGS) of rDNA genes; however, IGS-related satellite sequences known to be amplified outside the rDNA loci in several plant species were included. As the major databases (GenBank/EMBL/DDBJ) mirror their data on a daily basis (Stoesser *et al.*, 2001) we presume that the GenBank search was representative and covered most of the currently available sequences. The same keywords were used to search the Web of Science citation database (http://wos.cuni.cz) and Medline (using the Entrez browser) for papers describing plant satellite DNA sequences. Additionally, some older papers were retrieved from references cited in other publications or from the database maintained in our laboratory. If the described sequences were not available from GenBank they were entered into PlantSat manually; such entries were marked with the suffix '_noGB' added to their monomer names.

The PlantSat database

The database was implemented on a PC running under a SuSE Linux operating system. It is composed of text files organized into subdirectories representing individual repeat families. Each subdirectory contains two basic files storing information about the family and its monomer sequences, respectively. If available, additional information, like sequence logos or images, is stored in separate files. The data from these files are made available through web pages which are dynamically generated using PHP 3.0 scripts running under the Apache web server (http://w3lamc.umbr.cas.cz/PlantSat). The web interface was optimized for viewing using Netscape Navigator under both Linux and MS Windows operating systems; however, it has also been successfully tested in MS Internet Explorer and the Linux version of Lynx (a terminal-based browser).

Sequence analyses

Programs for sequence analyses were written in C and run on the server hosting the PlantSat database. The output data were visualized after importing into a StarOffice 5.2 (Sun Microsystems) spreadsheet. All analyses were performed on monomer sequences and subsequently averaged for individual repeat families. Dinucleotide relative abundances were determined using the method of Burge *et al.* (1992). The source codes of the programs as well as PHP scripts are available upon request.

IMPLEMENTATION

Data acquisition and processing

Most sequence data were acquired from GenBank following its searching using a broad range of appropriate keywords. The search terms were selected such that they retrieved any tandemly repeated sequences, and the entries that did not represent satDNAs were then discarded. Similar searches were performed in indexes of scientific publications in order to find sequences that were published without deposition into the sequence databanks.

In the next step, we sorted the sequences into families based on two main criteria: (i) mutual sequence homologies, and (ii) the sizes of basic repeated units. Sequence assignments into families followed original sequence annotations and published experimental data, provided these two criteria were met (several rearrangements were made involving highly homologous repeats with the same monomer lengths that were grouped together). In the case of homologous sequences that are known to be amplified in some genomes as distinct variants differing in their monomer lengths (as determined using Southern blot analysis), different families were assigned. In an exceptional case of *Alstroemeria* repeats the families were assembled using sequence homologies only, since the basic repeated units could not be defined due to their complex character and the lack of experimental data. Although these sequences are included in the database they were not used for the sequence analyses described in this paper.

Table 1. Representation of plant families in the PlantSat database

Division	Class	Family	Number of repeat families
Coniferophyta (conifers)	Coniferopsida	Pinaceae	1
Magnoliophyta (flowering plants)	Eudicotyledons	Solanaceae	20
		Fabaceae	18
		Brassicaceae	14
		Chenopodiaceae	12
		Cucurbitaceae	6
		Asteraceae	5
		Polygonaceae	3
		Oleaceae	2
		Rutaceae	2
		Actinidiaceae	1
		Caryophyllaceae	1
		Malvaceae	1
		Ranunculaceae	1
		Rosaceae	1
		Salicaceae	1
		Scrophulariaceae	1
	Liliopsida	Poaceae	50
		Alliaceae	4
		Alstroemeriaceae	4
		Hyacinthaceae	3
		Iridaceae	1
		Total	152

The table lists numbers of repeats isolated from individual taxa.

Since most satDNAs are family, genus or species-specific, the repeat names were derived from the name of the respective taxon followed with the length of the repeat monomer (e.g. Allium_370), or preferably with a commonly used repeat name in case it exists (e.g. Poaceae_Afa). For each family, the sequence data were stored as a set of monomer sequences in FASTA format (if necessary, the original sequences were split into monomers). Additional information regarding repeat abundance and chromosomal localization in individual species was also collected and deposited in the database.

The database is supposed to be updated on a regular basis by the authors, however, any data provided by other investigators will be greatly appreciated. At the time of this manuscript preparation, the database contains 152 satDNA families represented by between one and 69 sequenced monomers (849 monomers in total). As expected, most sequences were isolated from extensively investigated taxa such as Poaceae (50), Solanaceae (20), and Fabaceae (18). However, there are a total 22 plant families that are represented in the database by at least one entry (Table 1).

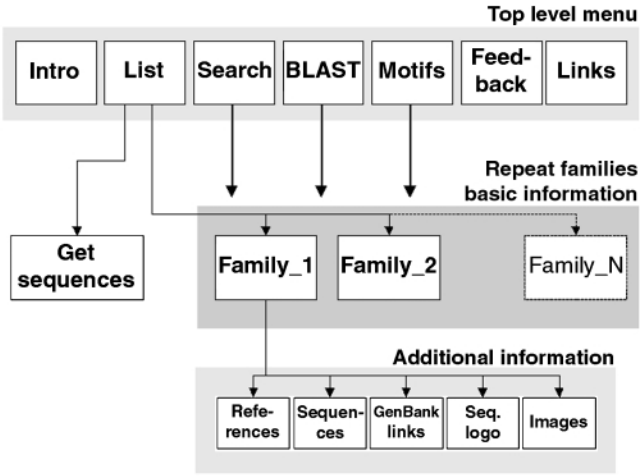

Fig. 1. Simplified overview of the PlantSat web interface.

A web interface to the PlantSat database

To allow for an easy and efficient extraction of information, we designed a web interface to the PlantSat database. As depicted in Figure 1, the interface reflects the arrangement of the data according to individual repeat families and provides several tools for accessing them. An index to all families in the database can be obtained using the *List* option. It offers links to 'homepages' of individual repeats containing basic information about the respective family and links to pages with additional information (Figure 2). These include references, monomer sequences, pointers to original GenBank files, sequence logos (Schneider and Stephens, 1990) and images of repeat localizations on chromosomes *in situ*. The latter two are currently available only for the repeats investigated in our laboratory (for examples see Vicia_VicTR_A and Vicia_VicTR_B) but we expect to include images provided by other researches in the near future.

The *List* page can also be used to display or download monomer sequences from a single, or several selected, families or to directly download the complete set of sequences in a file that can serve as an input for external programs. This plain text file contains FASTA-formatted monomer sequences sorted into families that are separated with special tags ('FAMILY : NAME'). This makes further processing of the data easy and computer programs can be designed that recognize and analyze the monomer sequences within individual families before making inter-family comparisons.

Another way to find data of interest is by using the *Search* page, which generates links to repeat families containing the search term in their description or additional information files. It can be used to find repeats present in a

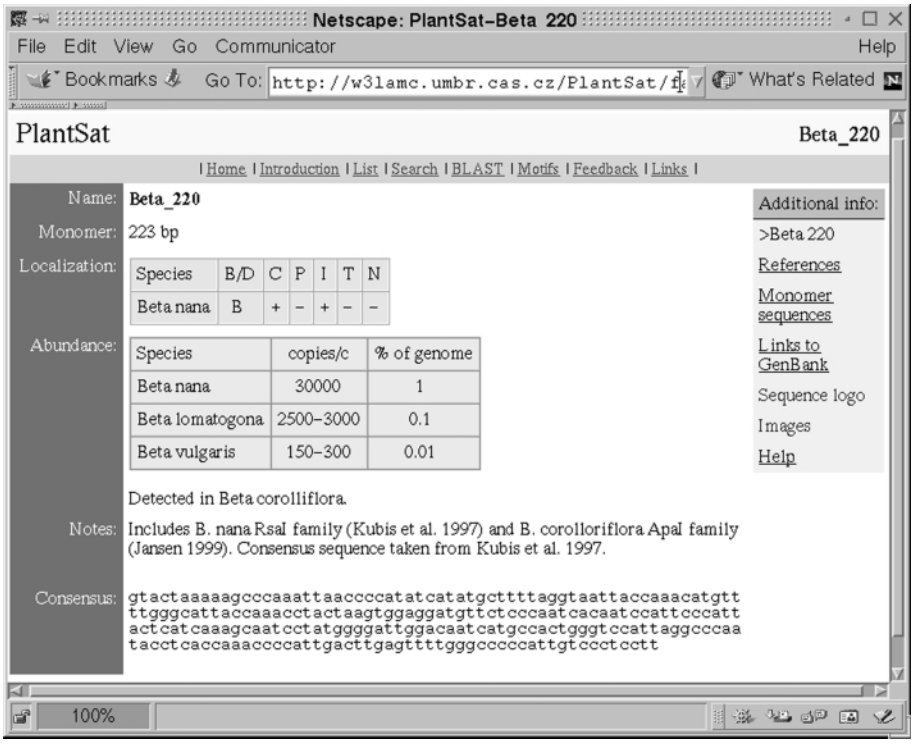

Fig. 2. An example of the repeat family main page displaying basic information about the family and hyperlinks to additional data. The information is arranged into the following fields: monomer—average length of basic repeated unit. Localization—chromosomal localization of the repeat as detected using *in situ* hybridization or PRINS. The 'B', 'D' or 'A' in the column following the species name stands for the appearance of the signal in distinct bands (B), as dispersed labeling (D), or as a combination of both (A). The '+' or '−' signs represent the presence or absence of the signal in individual chromosomal regions (C = centromeric, P = paracentromeric, I = intercalary, T = (sub-) telomeric, N = NOR (secondary constriction)). Abundance—is given in copy numbers of the repeat monomers per haploid genome (1C) or as percentage of the total genome size, depending on data available from the literature. If the abundance has not been determined but the repeat was detected by means of Southern or dot-blot hybridizations, the species name is listed under the table. Notes—any relevant information that does not fit into other categories. Consensus—the consensus sequence derived from available monomers. In some cases, the consensus copied from published paper(s) is given (for example, if the published consensus is based on a larger number of monomers that were not published or deposited into sequence databanks).

given plant species or in a higher taxon as genus or family, or to search for an author's name. Additionally, there are two special search routines available, allowing the retrieval of sequences according to GenBank accession numbers and chromosomal localization, respectively. The former is intended for a quick search for monomer sequences derived from a particular GenBank entry or for checking if the sequence is present in PlantSat. The latter provides a list of families that were detected in selected chromosomal regions. It should be noted, however, that chromosomal localization has been determined for only a fraction of known satDNAs, and that the repeats are often located in several different regions of the chromosome.

Nucleotide sequence-based searches include *BLAST* (Altschul *et al.*, 1990), which provides a means for detecting homologies between user-entered sequences and PlantSat entries. It uses a locally implemented stand-alone blastall program (Altschul *et al.*, 1997) and performs searches against either monomer or consensus sequences. The results are displayed in a form of score lists and alignments supplemented with hypertext links to the corresponding repeat families. As the analysis of satDNAs often includes detection of specific sequence patterns, we also provide a tool for detecting them in PlantSat database entries. *Motif search* allows searching for relatively complicated patterns that may include ambiguously defined bases and regular expression-like statements. It can be used, for example, to find out if the motif that is conserved in a sequence of interest is present in a wider range of satDNA families and thus possibly might have a functional or structural significance.

As in any database of this kind, keeping the data free of errors and omissions is in part based on feedback from its users and authors of the source data. For this purpose

we provide a simple *Feedback* form that can be used to enter comments or corrections. All user-added data will be properly acknowledged in the section 'References' of the repeat family additional information.

Sequence analyses

We used the sequences downloaded from the PlantSat database to perform several analyses using computer programs developed for this purpose. Taking advantage of sequence family assignments, the calculations were first done for each family to obtain average values that were subsequently used for comparisons between individual repeat families. Thus, in the following analyses each family is represented by only a single data point.

First, we analyzed monomer length distribution and nucleotide composition of plant satellite sequences. The repeats ranged from 33 bp to almost 4 kbp, however, the distribution of monomer lengths between these extremes was not uniform. The majority (91%) of the repeats had monomers shorter then 600 bp and were concentrated into several size ranges, the most prominent ones being between 135–195 and 315–375 bp, respectively (Figure 3a). The highest peak was centered around 165 bp and included 57 (38.5%) repeats. Although the AT/GC content differed significantly among the repeats, ranging from 22 to 75% A + T, this feature did not correlate with the monomer lengths (Figure 3b). The proportion of A + T of most satDNA sequences was above 50% (58% in average).

Dinucleotide composition of plant satellite DNA was analyzed using relative abundance (odds ratio) representations (Burge *et al.*, 1992). In principle, these calculations assess dinucleotide bias as ratios of expected and observed frequencies and are independent of nucleotide composition and strand orientation of analyzed sequences. The calculations for all ten possible dinucleotides were performed separately for each family (data not shown) and then averaged for all families to get values representative for plant satellite sequences (Table 2). These data show that dinucleotide AA/TT is significantly over-represented and TA is significantly under-represented in plant satellite repeats. This is also reflected on a family level, as AA/TT is over-represented in 51% of the repeats and there is no family exhibiting its suppression. The bias is even more evident for TA, as it is suppressed in 74% of the repeats and only a single family (Vicia_faba_TIII15; 0.7%) is enriched for this motif. Although the average values for other dinucleotides do not show such high deviations, three motifs (CC/GG, CG, and GC) are biased in more than 40% of the families. It is interesting that in addition to the 44.3% of repeats depleted for CG there is a large fraction (25.5%) that is enriched for this sequence (Table 2).

To demonstrate the *Motif search* algorithm implemented as a part of the PlantSat web interface, we performed

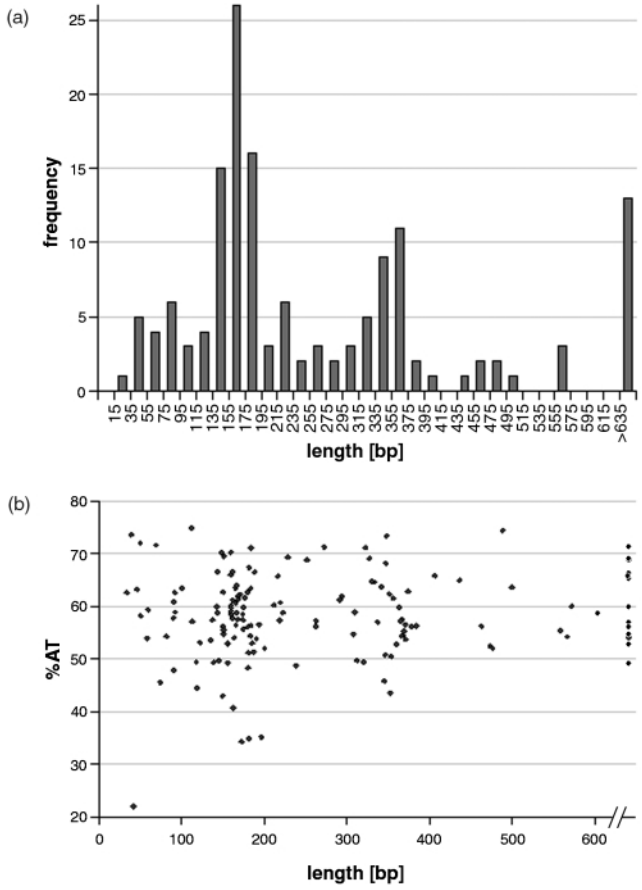

Fig. 3. The size distribution of repeat monomers (a), and the proportion of A + T versus monomer length (b) of plant satDNAs. Each datapoint represents the average value for one repeat family.

a search for the pentanucleotide CAAAA, which is supposed to be involved in a breakage–reunion mechanism of repeated sequences (Appels *et al.*, 1986; Katsiotis *et al.*, 1998). This sequence has been previously found in several repeat families (Katsiotis *et al.*, 1998; Macas *et al.*, 2000), however, the extent of its occurrence in satDNAs was unknown. The search revealed that the motif is present in the consensus sequences of 120 out of 152 analyzed repeat families (78.9%). When all monomer sequences were analyzed, the motif was detected in 132 families (86.8%). In order to test if such a high frequency of appearance of the CAAAA motif is not simply due to its short length and A-rich sequence, we also searched for other possible permutations of this sequence. However, none of these variants were detected as frequently as CAAAA (Table 3); on average they occurred in 57.4% of repeat families (or in 69.2% for analyzed monomers). Similar results were obtained for the motifs in which the C at the first position was replaced by T or G (Table 3).

Table 2. Dinucleotide relative abundances (ρ^*_{XY}) of plant satDNA sequences

	AA/TT	CC/GG	CA/TG	GA/TC	AC/GT	AT	CG	GC	AG/CT	TA
Average ρ^*_{XY}	1.24	1.18	1.06	1.03	0.92	0.92	0.88	0.88	0.87	0.71
	Distribution (%)									
ρ^*_{XY} range	AA/TT	CC/GG	CA/TG	GA/TC	AC/GT	AT	CG	GC	AG/CT	TA
<0.79	0.0	6.0	9.4	10.1	18.8	18.8	39.6	36.2	30.9	65.1
0.79–0.82	0.0	1.3	4.0	3.4	6.0	8.7	4.7	4.7	6.7	8.7
0.83–1.19	49.0	46.3	57.0	65.8	70.5	63.8	30.2	47.0	55.0	25.5
1.20–1.22	2.7	6.7	4.7	4.7	0.7	4.0	1.3	1.3	0.7	0.0
>1.22	48.3	39.6	24.8	16.1	4.0	4.7	24.2	10.7	6.7	0.7

Average values for ten possible dinucleotides calculated from all repeat families are given in the upper part of the table. A deviation of ρ^*_{XY} value from 1 reflects marginal (1.20–1.22) or extreme (>1.22) over-representation, or marginal (0.79–0.82) or extreme (<0.79) under-representation of a given dinucleotide (Karlin and Burge, 1995). The proportion of the families having their ρ^*_{XY} values in one of these ranges is expressed as a percentage of their total number (149) and is given in the bottom part of the table.

DISCUSSION

Recent accumulation of sequencing data from many different organisms causes an increasing demand for new tools allowing easy access and meaningful analysis of this information. One of the logic outcomes of this demand is the development of specialized databases oriented to a particular class of sequences or organisms (Abdrakhmanov *et al.*, 2000; Bell *et al.*, 2001; Garcia-Martinez *et al.*, 2001; Shimko *et al.*, 2001). Since these databases are designed to reflect specific features of the sequence type or organism of interest, they can provide more efficient tools for data retrieval and analysis than general-purpose databanks such as EMBL/GenBank/DDBJ. In the case of the PlantSat database, we used single monomers as the basic entries that are grouped into families defined by sequence homologies and monomer lengths. This classification of satellite repeats into families is widely accepted among plant genome researchers and in this paper we demonstrate that it is also useful for accurate analysis of the whole set of plant satDNA sequences.

One of the obstacles in assembling data for the PlantSat database was caused by the definition of satellite DNA, being most often described as consisting of highly abundant, tandemly arranged repeats organized in large contiguous blocks (Charlesworth *et al.*, 1994; Kubis *et al.*, 1998). However, many GenBank entries refer to unpublished data and thus lack the information needed for determination if they fully conform with this definition. Therefore, we decided to use more relaxed criteria for selecting data for PlantSat and to include all tandem repeats that do not belong to microsatellites and minisatellites (see section **System and methods**). However, since the information about copy numbers and other characteristics of individual repeat families is preserved in PlantSat, it can still be used to identify the repeats that are known to be amplified to certain copy numbers or to form distinct bands on mitotic chromosomes. It should also be noted that our primary rule in assembling the database was to use the existing names and definitions of basic repeated units (monomers) of satDNA families as much as possible. Therefore, except for including taxon names to all repeat family names we did not attempt to establish a uniform nomenclature of satellite repeats (see section **Implementation**). However, the simple rules we used for arranging data in this database might serve as a starting point for a discussion about conventions for assigning names to satDNA sequences as well as for discrimination of individual satellite repeat families. We hope that PlantSat will become a platform for these discussions and we will maintain a corresponding interface for them on its web page.

The sequence analyses presented in this paper provide

Table 3. Detection of the motif CAAAA and its permutations in plant satDNA repeats

Motif	Consensus		Monomer	
	Number of families	(%)	Number of families	(%)
CAAAA	120	78.9	132	86.8
ACAAA	86	56.6	107	70.4
AACAA	78	51.3	105	69.1
AAACA	85	55.9	97	63.8
AAAAC	100	65.8	112	73.7
TAAAA	97	63.8	114	75.0
GAAAA	100	65.8	118	77.6

The number of repeat families containing the respective motif is given and expressed as a percentage of all families in the database. The analysis was performed separately on consensus and monomer sequences; in the latter case the families were considered to contain the motif if it occurred in at least one monomer sequence.

the first comprehensive and unbiased data about plant satellite DNAs. The observed distribution of monomer lengths revealed a preference for size ranges around 165 bp and its multiples. This confirms previously published observations that the basic repeated units of satDNAs often correspond to the length of DNA wrapped around a nucleosome particle (Kubis *et al.*, 1998; Schmidt and Heslop-Harrison, 1998). Although nucleosome phasing has been demonstrated on several plant satellite repeats (Gazdová *et al.*, 1995; Matyášek *et al.*, 1997; Vershinin and Heslop-Harrison, 1998), its relation to the size preference of satDNA monomers is yet to be investigated.

The dinucleotide relative abundances revealed some interesting features of satDNA sequences. Compared to the data published for plant genomes (Karlin and Burge, 1995; Karlin *et al.*, 1998) the over-representation of AA/TT appears to be specific for satellite repeats and probably reflects a frequent occurrence of adenine runs in their sequences (data not shown). It is known that the adenine runs cause intrinsic bending of DNA molecules (Koo *et al.*, 1986; Dlakic and Harrington, 1996) and thus may provide specific structural properties required for the amplification/maintenance of satDNA in the genome, or to be a consequence of such processes. This would accord with the finding of a frequent occurrence of the CAAAA motif that is presumably involved in recombination events between the repeats (Appels *et al.*, 1986; Katsiotis *et al.*, 1998).

In contrast to AA/TT, the biased representation of TA in satellite repeats is in agreement with a general suppression of this dinucleotide in plant and other eukaryotic genomes (Karlin and Burge, 1995; Karlin *et al.*, 1998). Similar suppression could also be expected for CG, as this sequence is a frequent target for cytosine methylation which may cause its elimination due to conversion of 5-methylcytosine to thymine (Karlin and Burge, 1995). Surprisingly, this suppression occurs only in a part of satDNA families, while there is also a considerable number of families where this motif is over-represented, and the average relative abundance of CG is the same as for GC (Table 2). It is also interesting that although the CG under-representation was observed in dicot but not in monocot plants (Karlin and Burge, 1995; Karlin *et al.*, 1998), the distribution of satDNAs showing CG under- or over-representation in these two groups is roughly the same (data not shown). Thus, this probably reflects specific features of the repeats rather than overall genome composition. This phenomenon, together with its possible correlation to relative dinucleotide frequencies in individual repeat families will be a subject of further study.

ACKNOWLEDGEMENTS

We thank Robert Wolf for technical help with configuring the PlantSat web server, Susanne M.Rafelski for assistance in preparation of the manuscript, and members of our laboratory for database testing and many useful discussions. This work was supported by grants GA CR 521/96/K117 and AVOZ 5051902.

REFERENCES

Abdrakhmanov,I., Lodygin,D., Geroth,P., Arakawa,H., Law,A., Plachý,J., Korn,B. and Buerstedde,J.M. (2000) A large database of chicken bursal ESTs as a resource for the analysis of vertebrate gene function. *Genome Res.*, **10**, 2062–2069.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Appels,R., Moran,L.B. and Gustafson,J.P. (1986) Rye heterochromatin. I. Studies on clusters of the major repeating sequence and the identification of a new dispersed repetitive sequence element. *Can. J. Genet. Cytol.*, **28**, 645–657.

Bell,C.J., Dixon,R.A., Farmer,A.D., Flores,R., Inman,J., Gonzales,R.A., Harrison,M.J., Paiva,N.L., Scott,A.D., Weller,J.W. and May,G.D. (2001) The *Medicago* genome initiative: a model legume database. *Nucleic Acids Res.*, **29**, 114–117.

Burge,C., Campbell,A.M. and Karlin,S. (1992) Over-representation and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl Acad. Sci. USA*, **89**, 1358–1362.

Charlesworth,B., Sniegowski,P. and Stephan,W. (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, **371**, 215–220.

De Kochko,A., Kiefer,M.C., Cordesse,F., Reddy,A.S. and Delseny,M. (1991) Distribution and organization of a tandemly repeated 352-bp sequence in the Oryzae family. *Theor. Appl. Genet.*, **82**, 57–64.

Deumling,B. (1981) Sequence arrangement of a highly methylated satellite DNA of a plant, *Scilla*: a tandemly repeated inverted repeat. *Proc. Natl Acad. Sci. USA*, **78**, 338–342.

Dlakic,M. and Harrington,R.E. (1996) The effects of sequence context on DNA curvature. *Proc. Natl Acad. Sci. USA*, **93**, 3847–3852.

Garcia-Martinez,J., Bescos,I., Rodriguez-Sala,J.J. and Rodriguez-Valera,F. (2001) RISSC: a novel database for ribosomal 16S–23S RNA genes spacer regions. *Nucleic Acids Res.*, **29**, 178–180.

Gazdová,B., Široký,J., Fajkus,J., Brzobohatý,B., Kenton,A., Parokonny,A., Heslop-Harrison,J.S., Palme,K. and Bezděk,M. (1995) Characterization of a new family of tobacco highly repetitive DNA, GRS, specific for the *Nicotiana tomentosiformis* genomic component. *Chromosome Res.*, **3**, 245–254.

Ingham,L.D., Hanna,W.W., Baier,J.W. and Hannah,L.C. (1993) Origin of the main class of repetitive DNA within selected *Pennisetum* species. *Mol. Gen. Genet.*, **238**, 350–356.

Irifune,K., Hirai,K., Zheng,J., Tanaka,R. and Morikawa,H. (1995) Nucleotide-sequence of a highly repeated DNA sequence and its chromosomal localization in *Allium fistulosum*. *Theor. Appl. Genet.*, **90**, 312–316.

Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes—a genomic signature. *Trends Genet.*, **11**, 283–290.

Karlin,S., Campbell,A.M. and Mrázek,J. (1998) Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.*, **32**, 185–225.

Kato,A., Yakura,K. and Tanifuji,S. (1984) Sequence analysis of *Vicia faba* repeated DNA, the FokI repeat element. *Nucleic Acids Res.*, **12**, 6415–6426.

Katsiotis,A., Hagidimitriou,M., Douka,A. and Hatzopoulos,P. (1998) Genomic organization, sequence interrelationship, and physical localization using in situ hybridization of two tandemly repeated DNA sequences in the genus *Olea*. *Genome*, **41**, 527–534.

Koo,H.S., Wu,H.M. and Crothers,D.M. (1986) DNA bending at adenine–thymine tracts. *Nature*, **320**, 501–506.

Kubis,S., Schmidt,T. and Heslop-Harrison,J.S. (1998) Repetitive DNA elements as a major component of plant genomes. *Ann. Bot.*, **82**, 45–55.

Macas,J., Požárková,D., Navrátilová,A., Nouzová,M. and Neumann,P. (2000) Two new families of tandem repeats isolated from genus *Vicia* using genomic self-priming PCR. *Mol. Gen. Genet.*, **263**, 741–751.

Matyášek,R., Gazdová,B., Fajkus,J. and Bezděk,M. (1997) NTRS, a new family of highly repetitive DNAs specific for the T1 chromosome of tobacco. *Chromosoma*, **106**, 369–379.

Nouzová,M., Kubaláková,M., Doleželová,M., Koblížková,A., Neumann,P., Doležel,J. and Macas,J. (1999) Cloning and characterization of new repetitive sequences in field bean (*Vicia faba* L.). *Ann. Bot.*, **83**, 535–541.

Schmidt,T. and Heslop-Harrison,J.S. (1993) Variability and evolution of highly repeated DNA sequences in the genus *Beta*. *Genome*, **36**, 1074–1079.

Schmidt,T. and Heslop-Harrison,J.S. (1998) Genomes, genes and junk: the large-scale organization of plant chromosomes. *Trends Plant Sci.*, **3**, 195–199.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos—a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Shimko,N., Liu,L., Lang,B.F. and Burger,G. (2001) GOBASE: the organelle genome database. *Nucleic Acids Res.*, **29**, 128–132.

Smith,G.P. (1976) Evolution of repeated DNA sequences by unequal crossover. *Science*, **191**, 528–535.

Stephan,W. and Cho,S. (1994) Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. *Genetics*, **136**, 333–341.

Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Lombard,V., Lopez,R., Parkinson,H., Redaschi,N., Sterk,P., Stoehr,P. and Tuli,M.A. (2001) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **29**, 17–21.

Vershinin,A.V. and Heslop-Harrison,J.S. (1998) Comparative analysis of the nucleosomal structure of rye, wheat and their relatives. *Plant Mol.Biol.*, **36**, 149–161.

Vogt,P. (1992) Code domains in tandem repetitive DNA sequence structures. *Chromosoma*, **101**, 585–589.

Walsh,J.B. (1987) Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. *Genetics*, **115**, 553–567.

Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. and Rapp,B.A. (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **29**, 11–16.