Curr Genet (2005) 47: 289–297
DOI 10.1007/s00294-005-0573-5

RESEARCH ARTICLE

Berislav Lisnić · Ivan-Krešimir Svetec
Hrvoje Šarić · Ivan Nikolić · Zoran Zgaga

Palindrome content of the yeast *Saccharomyces cerevisiae* genome

Received: 10 February 2005 / Accepted: 20 February 2005 / Published online: 18 March 2005

Abstract Palindromic sequences are important DNA motifs involved in the regulation of different cellular processes, but are also a potential source of genetic instability. In order to initiate a systematic study of palindromes at the whole genome level, we developed a computer program that can identify, locate and count palindromes in a given sequence in a strictly defined way. All palindromes, defined as identical inverted repeats without spacer DNA, can be analyzed and sorted according to their size, frequency, GC content or alphabetically. This program was then used to prepare a catalog of all palindromes present in the chromosomal DNA of the yeast *Saccharomyces cerevisiae*. For each palindrome size, the observed palindrome counts were significantly different from those in the randomly generated equivalents of the yeast genome. However, while the short palindromes (2–12 bp) were under-represented, the palindromes longer than 12 bp were over-represented, AT-rich and preferentially located in the intergenic regions. The 44-bp palindrome found between the genes *CDC53* and *LYS21* on chromosome IV was the longest palindrome identified and contained only two C-G base pairs. Avoidance of coding regions was also observed for palindromes of 4–12 bp, but was less pronounced. Dinucleotide analysis indicated a strong bias against palindromic dinucleotides that could explain the observed short palindrome avoidance. We discuss some possible mechanisms that may influence the evolutionary dynamics of palindromic sequences in the yeast genome.

Communicated by S. Hohmann

B. Lisnić · I.-K. Svetec · Z. Zgaga (✉)
Faculty of Food Technology and Biotechnology,
University of Zagreb, Pierottijeva 6,
10000 Zagreb, Croatia
E-mail: zgazo@pbf.hr
Tel.: +385-1-4836013
Fax: +385-1-4836016

H. Šarić · I. Nikolić
Sail Company Croatia Ltd., Ilica 412, 10000 Zagreb, Croatia

Keywords Palindrome · Inverted repeat · Dinucleotide · *Saccharomyces cerevisiae* · Sequence analysis

Introduction

Closely spaced inverted repeats (IRs), palindromes and quasipalindromes can be found in the DNA of natural plasmids, viral and bacterial genomes and eukaryotic chromosomes and organelles. In prokaryotes, they may serve as binding sites for regulatory proteins, while short perfect palindromes are known as recognition sites for type II restriction-modification systems (RMSs) that play a significant role in bacterial ecology and evolution (Gelfand and Koonin 1997; Rocha et al. 2001). Another important property of such motifs is their potential to form intra-strand hydrogen bonds within DNA molecules or in corresponding RNA transcripts. Therefore, they are contained in genes encoding functional RNA molecules, the structure of which depends on the formation of proper intra-strand bonding, and in different *cis*-acting genetic elements, like terminators, attenuators, plasmid and viral origins of replication. Protein binding and secondary structure formation are also modes of action for IRs and related motifs in eukaryotic cells. For example, palindromes with a spacer of one nucleotide were identified in yeast sequences regulating cellular response to the accumulation of unfolded proteins in the endoplasmic reticulum (Mori et al. 1998) and a heterodimeric complex was isolated that binds two palindromic sequences in the promoter region of the human *erbB-2* gene (Chen and Gill 1996). In mouse B lymphoma cells, palindromic and potential stem-loop motifs were identified as break-points during class switch recombination (Tashiro et al. 2001); and the formation of intra-strand secondary structures is essential in the process of immunoglobuline gene rearrangement known as V(D)J-joining (Cuomo et al. 1996).

However, in spite of their importance and functional versatility, longer palindromes and IRs were shown to be

very unstable in different organisms, from bacteria to mammalian cells. The recombinogenicity of such motifs is attributed to their potential to form secondary structures known as hairpins and cruciforms and the molecular models proposed to explain palindrome-induced genomic instability can be divided into two classes: first, based on template switching or slippage of the DNA polymerase during replication of DNA adopting secondary structures and, second, requiring an enzymatic activity that transforms cruciforms and hairpins to recombinogenic lesions, like double-strand breaks (DSBs; Leach 1994). Both types of models are supported by experimental data; and several human genetic disorders can be explained either by errors occurring during the replication of palindromic and quasipalindromic sequences (Bissler 1998; Gordenin and Resnick 1998) or by IR-induced illegitimate end-joining (Repping et al. 2002).

Given the importance of palindromic sequences in the regulation of different cellular processes on the one side and their influence on genetic stability on the other side, an analysis of the incidence of palindromes at the genomic level seems particularly interesting. LeBlanc et al. (2000) prepared a catalog of palindromes of 4–60 bp present on chromosomes III and X of the nematode *Caenorhabditis elegans*. Palindromes were classified as AT-rich, non-AT-rich or GC-rich and this analysis indicated that the long, AT-rich palindromes are much more frequent in the actual chromosomes than in the randomly generated sequences. The IRs separated by a single base pair were also included in their study and such "odd palindromes" were more frequent than perfect IRs without a spacer. Avoidance of short palindromic sequences was observed in the genomes of some bacteriophages (Sharp 1986), but also in the genomes of their bacterial hosts (Karlin et al. 1997; Gelfand and Koonin 1997; Rocha et al. 2001). The highest bias was frequently observed for the recognition sites of the type II restriction enzymes, supporting the view that RMSs influence genome evolution in prokaryotes. Consistent with this hypothesis, Fuglsang (2004) demonstrated that the succession of codons disfavoring palindrome formation is more pronounced for the bacterial species that have RMSs. In order to initiate a systematic study of palindromes present in different genomes, we decided first to develop a computer program that can score and count all palindromes present in a given sequence in a strictly defined way. This program was then used to analyze the palindrome content of the entire genome of the yeast *Saccharomyces cerevisiae* and the results of this analysis clearly indicate that the palindromic sequences conform to the specific rules of evolutionary dynamics.

Materials and methods

DNA sequences

The DNA sequence files for all yeast chromosomes (NC_001133.4, NC_001134.5, NC_001135.6, NC_001136.5, NC_001137.2, NC_001138.4, NC_001139.4, NC_001140.3, NC_001141.1, NC_001142.5, NC_001143.4, NC_001144.3, NC_001145.2, NC_001146.2, NC_001147.4, NC_001148.2) were downloaded from the NCBI (ftp://ftp.ncbi.nih.gov/genomes/Saccharomyces_cerevisiae/) on 3 April, 2004. The total length of downloaded yeast genomic DNA sequences (excluding mitochondrial DNA) was 12,070,766 bp.

The file containing the coding regions of the entire yeast genome (that is, ORF coding sequences only, without 5′UTR, 3′UTR, intron sequences, or bases not translated due to translational frameshifting) was downloaded on 12 July, 2004 from the SGD (ftp://genome-ftp.stanford. edu/pub/yeast/data_download/sequence/genomic_sequence/orf_dna/orf_coding.fasta.gz). This file includes all ORFs except dubious ORFs, and prior to examination of the palindrome and dinucleotide content in the coding regions, the sequences of mitochondrial DNA were removed from the file. The total length of the coding DNA sequences in this shortened file was 8,755,368 bp.

Ten random genomes (12,070,766 bp each) were generated with respect to the frequency of the four bases present in the yeast genome (A = 30.90%, C = 19.17%, G = 19.13%, T = 30.81%), so that the average proportions of the bases in the random genomes were: A = 30.90 ± 0.01%, C = 19.18 ± 0.01%, G = 19.13 ± 0.01% and T = 30.81 ± 0.01%.

Palindrome count

The Spinnaker program described in this work is available upon request.

Statistics

The numbers of dinucleotides and palindromes determined in the chromosomal DNA were called the observed numbers, while the numbers of dinucleotides and palindromes determined in the random genomes were called the expected numbers. To test whether the medians of the expected numbers of palindromes and dinucleotides differed significantly from the observed numbers, we employed the Wilcoxon signed rank test with a confidence interval of 99%, using MINITAB ver. 14.1.

Results

Palindrome scoring

Systematic study of palindromic sequences requires convenient software that will recognize and map all palindromes in a predetermined size-range at the genomic level. This is not a trivial problem, since individual

palindromes found within a given sequence can be scored in different ways. First, we may decide to score as individual palindromes only those palindromes that do not share common base pairs, but this approach can result in an underestimation of the actual number of palindromes, since some palindromes may remain undetected and different results can be obtained for the same sequence (compare Fig. 1a,b). Therefore, we decided to count also the palindromes that do share common base pairs and in this case we have the possibility either to score only the longest palindromes that may overlap, or to include also the shorter palindromes embedded within longer palindromes. There are two classes of such palindromes, those sharing the same center of symmetry (nested) and those having a different center of symmetry. This is shown in Fig. 1c, where all palindromes present in the test sequence are indicated, including both nested and non-nested palindromes entirely contained within longer palindromes. Embedded palindromes are omitted in Fig. 1d, while Fig. 1e presents a combination of these two modes of palindrome scoring where only non-nested embedded palindromes are presented.

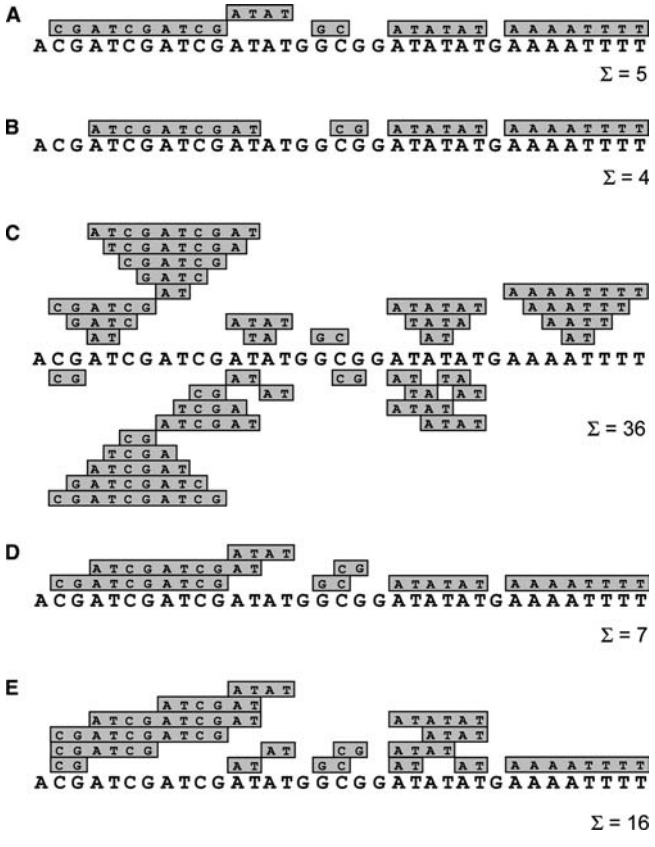

Fig. 1 Palindrome scoring. Different numbers of palindromes can be counted in the same sequence depending on the scoring criterion. **a**, **b** Non-overlapping palindromes do not share common base pairs. **c** Embedded palindromes include both non-overlapping and palindromes that share common base pairs as well as all shorter palindromes contained within a longer palindrome. **d** Overlapping palindromes include both non-overlapping and palindromes that share common base pairs but not shorter palindromes contained within a longer palindrome. **e** Embedded palindromes with nested palindromes excluded

Our computer program, named Spinnaker, was designed to score palindromes as indicated in Fig. 1c–e. Let us define first the frame as an array of nucleotides of even length. The location of the frame within a given DNA sequence is identified by position of its left boundary (L) and by its length. The maximum length of the frame (p) is propounded as a search parameter that corresponds to the length of the potentially largest palindrome. The position of the right boundary of the frame is given by $R = L + p - 1$. Additionally, we define P_1 as the right boundary of the last previously found palindrome. In the beginning, we set $L = 0$ and $P_1 = 0$. The algorithm for the search for overlapping palindromes can be described recursively as follows. Step 1: we set the frame size to p and move it one position to the right ($L := L + 1$). Step 2: within the frame, we check for complementarity in the Lth and Rth base pair, then the $(L + 1)$th and the $(R - 1)$th base pair and so on. If it is found that all base pairs are complementary, the frame is defined as a palindrome and its properties, including P_1, are recorded. However, if there is at least one non-complemantary base pair and if the current size of the frame is > 2 and $R > P_1$, we reduce the frame size by decreasing the position of its right boundary by two and then repeat step 2. Otherwise, we repeat step 1. The search for embedded palindromes (with or without unique symmetry axis) follows the same general procedure, ignoring the P_1 parameter and introducing a few additional internal structures needed for the affiliation of a palindrome to a given category.

The maximal size of a palindrome can be set to 500 nt and the results of sequence analysis can be sorted by size or frequency, alphabetically, or by GC content (percentage or absolute number of GC base pairs). Additionally, Spinnaker can graphically display the distribution of palindromes, where the analyzed sequence is presented as a horizontal line and palindromes of a given size as vertical lines. The results of a search can be saved as two separate files, one that includes the number and exact chromosomal locations of all palindromes and the other that includes the summarized results of a search. The program is user-friendly and works under Windows.

We compared Spinnaker with two programs designed for the identification of IRs: Reputer (Kurtz and Schleiermacher 1999) and Palindrome (Rice et al. 2000). Both programs recognize palindromic sequences, as shown in Fig. 1e, so that some of the palindromes contained within a longer palindrome remain undetected. For example, they detected only three 6-nt palindromes in the test sequence presented in Fig. 1, while Spinnaker recorded six such palindromes when embedded palindromes were counted. Reputer also recorded some non-palindromic sequences, so that 29 instead of 16 palindromic dinucleotides were found in our test se-

Table 1 Comparison of the numbers of palindromes in the yeast genome and in random genomes

Palindrome size	Overlapping palindromes		Embedded palindromes	
	Yeast genome	Random genomes[a]	Yeast genome	Random genomes[a]
2	*1,732,917*[b]	1,904,950.4	*2,770,675*[b]	3,183,592.3
4	*479,972*[b]	580,521.2	*705,275*[b]	839,462.8
6	*133,411*[b]	159,564.1	*194,874*[b]	221,347.3
8	*38,234*[b]	42,738.6	*57,086*[b]	58,424.4
10	*10,246*[b]	11,256.8	17,514[c]	15,382.8
12	*2,809*[b]	2,995.9	6,495[c]	4,095.1
14	939[c]	807	3,159[c]	1,096.3
16	282[c]	212	1,813[c]	289.1
18	96[c]	55.2	1,207[c]	77.1
20	59[c]	16.9	857[c]	21.9
22	65[c]	3.8	606[c]	5.0
24	24[c]	1.2	413[c]	1.2
26	29[c]	0	289[c]	0
28	17[c]	0	194[c]	0
30	9[c]	0	125[c]	0
32	9[c]	0	80[c]	0
34	8[c]	0	49[c]	0
36	4[c]	0	25[c]	0
38	3[c]	0	13[c]	0
40	3[c]	0	6[c]	0
42	2[c]	0	3[c]	0
44	1[c]	0	1[c]	0

[a]The average number from ten random genomes
[b]Under-represented palindromes ($P = 0.006$), also represented in *italics*
[c]Over-represented palindromes ($P = 0.006$)

quence; and the minimal palindrome size for Palindrome is 4 nt.

Palindromic sequences in the yeast genome

We initiated our study of yeast palindromic sequences by identifying all palindromes present in individual chromosomes (data not shown). We did not observe any particular pattern in the distribution of palindromic sequences along different chromosomes, although this feature was beyond the scope of this study and was not systematically analyzed. The complete catalog of palindromes present in all yeast chromosomes, determined both as overlapped and embedded palindrome counts, is presented in Table 1. As expected, a sharp size-dependent decrease in the numbers of palindromes was observed, but only for shorter palindromes. For example, the number of palindromes containing 22 bp was 65, while the number of 20-bp palindromes was 59. Similarly, there were more 26-bp than 24-bp palindromes identified. The palindrome content determined in the ten randomly generated genomes is also included in Table 1 and can be compared with the palindrome content of the actual genome. For each palindrome size, the values observed in the yeast genome were significantly different from those obtained in artificial genomes ($P = 0.006$). However, palindromes up to 12 bp were under-represented, while longer palindromes were over-represented. The longest palindromes found in ten randomly generated sequences contained 24 bp, while 85 palindromes longer than 24 bp were found in the yeast genome. Similar results were obtained when embedded palindromes were included in the palindrome count, but here even the 10-bp palindromes were over-represented. This is not unexpected, since in this case, all the 10-bp palindromes present in longer palindromes were also added to the sum. It is important to note that, even when embedded palindromes were counted, palindromes shorter than 10 bp were under-represented in comparison with randomly generated genomes.

Long palindromes are AT-rich and are placed in intergenic regions

Each palindrome identified by our program can be located on the physical map of the yeast genome (http://www.yeastgenome.org), as shown for the six longest palindromes found in the yeast chromosomal DNA (Fig. 2). They were all placed in intergenic regions and were AT-rich, so we decided to analyze these features more systematically. We calculated the A + T content for all palindromes found in the yeast genome and for the palindromes detected in ten randomly generated sequences. The A + T content of palindromes found in artificial sequences was around 70%, while slightly lower values were observed in genomic palindromes up to 12 nt (Fig. 3). However, a high increase in A + T content was observed with longer palindromes that were almost exclusively composed of A-T base pairs (Fig. 2). The remarkable exception was the presence of a GC-rich (52.4%) 42-nt palindrome found between the *RFX1* gene and YLR177W on chromosome XII (Fig. 2).

In order to find out the position of palindromic sequences with respect to annotated coding regions, we determined the proportion of palindromes that were placed in intergenic (non-coding) regions. The sequence

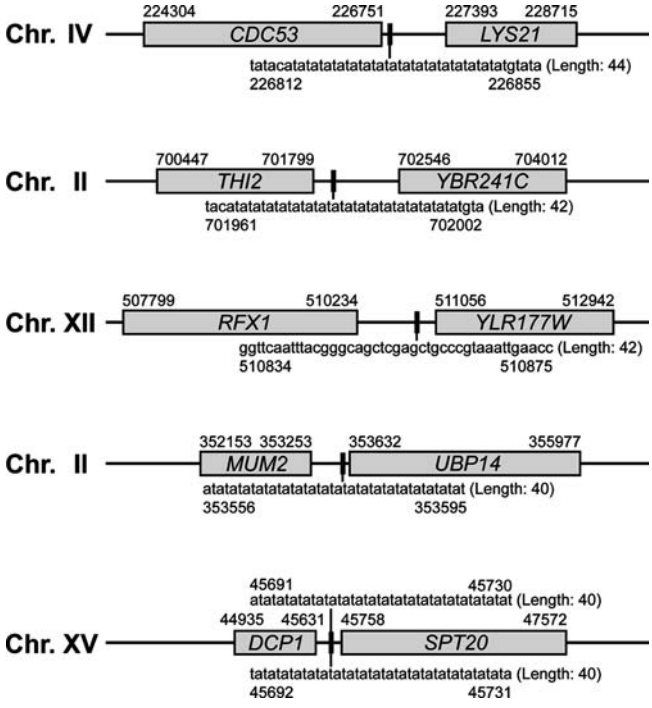

Fig. 2 Sequences and chromosomal locations of the six longest palindromes found in the yeast genome. *Numbers* indicate the start and end locations for ORFs and palindromes

containing only the yeast coding regions represented 72.73% of the total yeast genome (see Materials and methods), but even the short palindromes were preferentially (30–35%) placed in intergenic regions. Only the proportion of palindromic dinucleotides found in the intergenic regions (27.65%) was very close to that expected from random distribution between the coding and non-coding part of the genome. A preferential location in intergenic regions was particularly pronounced for palindromes longer than 10 bp, so that more than 97% of palindromes longer than 18 bp were identified outside of the coding regions (Fig. 3).

Palindromic dinucleotides in the yeast genome

Four dinucleotides, AT, TA, GC and CG, may be considered as the shortest palindromic sequences, but they are also found in the center of any longer palindrome. Our analysis indicated that the short palindromes, including palindromic dinucleotides, were less frequent in the yeast genome than in the random genomes (Table 1) and we decided to determine relative frequencies for all 16 dinucleotides (Fig. 4). Indeed, the two most under-represented dinucleotides were palindromic dinucleotides containing purine on the 3′ end, TA (0.77) and CG (0.80), while AT (0.94) was moderately under-represented and GC (1.02) was only slightly over-represented. For non-palindromic dinucleotides, very close values were observed for complementary dinucleotides and only the AC and GT couple was under-represented (0.89). The ratio between non-palindromic and palindromic dinucleotides composed either of A-T or G-C base pairs illustrates well the avoidance of palindromic dinucleotides in the yeast genome. TT + AA dinucleotides are 1.33 times more frequent than TA + AT dinucleotides, while CC + GG dinucleotides are 1.17 times more frequent than CG + GC dinucleotides (Table 2). The same analysis was also done for dinucleotides present in all palindromes longer than

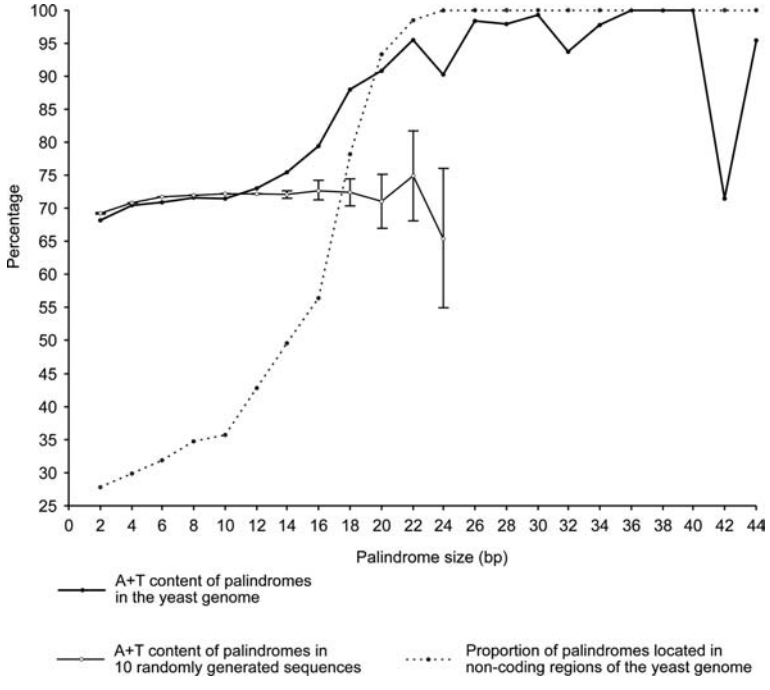

Fig. 3 Analysis of the A + T content and location of palindromes with respect to the non-coding regions in the yeast genome. Palindromes were scored as presented in Fig. 1d (overlapping palindromes). *Error bars* indicate one standard deviation

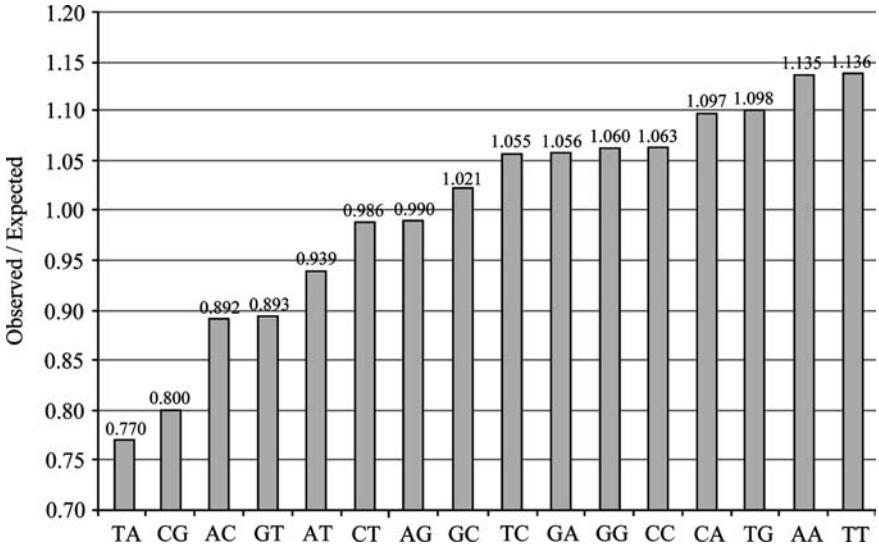

Fig. 4 Occurrence of dinucleotides in the yeast genome. For each dinucleotide, the ratio between the number determined in the yeast genome (*observed*) and the average number obtained in ten random genomes (*expected*) is presented

14 bp and the results obtained were quite different from those observed with the entire yeast genome, indicating that the long palindromes are specifically enriched in palindromic dinucleotides. However, the bias was much more pronounced for the AT and TA dinucleotides, which were 5.6 times more frequent than the TT + AA dinucleotides (Table 2).

The data presented in Fig. 4 also indicated a strong bias in the frequency of palindromic dinucleotides composed of identical base pairs, so that the AT/TA ratio was 1.22 and the GC/CG ratio was 1.28. This analysis was also done for the coding complement of the yeast genome and for palindromes longer than 14 bp (Table 2). The AT/TA ratio rose to 1.31 in coding DNA, while in long palindromes it was 1.036, indicating that the two dinucleotides are almost equally frequent. This can be explained by the presence of long runs of alternating A and T nucleotides frequently found in the long palindromes (Fig. 2; data not shown). In contrast, the GC/CG ratio in the coding complement was very close to that observed for the entire genome and in the long palindromes it even increased to 1.33.

Discussion

Palindromes, quasipalindromes and closely spaced IRs can be found in the genomes of all organisms and are involved in various cellular processes. In the present work, we focused on palindromes, defined as perfect IRs without spacer DNA. In order to make possible a systematic study of these important DNA motifs in different genomes, we first developed a computer program that can identify, locate and count palindromes in a given nucleotide sequence in a strictly defined way. This program was then used to prepare a catalog of all palindromic sequences present in the *S. cerevisiae* genome and in randomly generated equivalents of the yeast genome. Comparison between actual and random genomes revealed several important differences that are discussed below.

Table 2 Relative frequencies of palindromic dinucleotides in the whole genome, coding DNA and palindromes longer than 14 bp. *ND* Not determined

Ratio	Whole genome	Coding DNA	Palindromes > 14 bp
(AA + TT)/(AT + TA)	1.328	ND	0.178
(GG + CC)/(GC + CG)	1.166	ND	0.800
AT/TA	1.220	1.308	1.036
GC/CG	1.277	1.291	1.330

Identification of palindromic sequences

As discussed before, palindrome scoring can be done in different ways and we decided to consider both separated (non-overlapping) and overlapping palindromes, together with all smaller palindromes contained within longer palindromes, as individual palindromes. Such an approach ensures that no palindromic motif that may be hidden within a longer palindrome can be omitted; and this is very important for the detection of specific palindromic sites, like restriction enzyme sites or regulatory units. In addition to the embedded palindrome count, we also used a simplified way of palindrome scoring where partial overlapping of individual palindromes is allowed, but all smaller palindromes completely contained within longer palindromes are ignored. This way of palindrome scoring (overlapped palindromes) indicates individual palindromic loci within a given sequence and each palindrome is counted only once. In this way, a more realistic picture of the numbers and distribution of palindromic loci is obtained, especially when we consider the long runs of alternating complementary dinucleotides frequently present in a genome. For example, ten AT dinucleotides will be scored as a single 20 nt

palindrome, while in the same sequence 100 embedded palindromes can be found. For these reasons, we decided to use both embedded and overlapping palindrome counts in the analysis of palindrome content (Fig. 1c,d). As illustrated by our analysis of the yeast genome, this distinction is important. For example, both 10-nt and 12-nt palindromes are under-represented when counted as overlapped palindromes, but are over-represented when scored as embedded palindromes (Table 1). Palindromic dinucleotides were also analyzed for two reasons. First, there is no clear reason to include tetranucleotides and to exclude dinucleotides from the palindrome count and, second, each palindrome contains one of the four palindromic dinucleotides in the center and the analysis of their incidence could be useful for the general study of palindromic sequences. This was demonstrated by our analysis of the yeast genome, where the palindromic dinucleotides are particularly under-represented compared with non-palindromic dinucleotides with the same base pair composition. In the *C. elegans* genome, the IRs separated by one nucleotide are more frequent than the repeats without a spacer (LeBlanc et al. 2000). One possible explanation for this interesting observation could be that the palindromic dinucleotides are also avoided in the *C. elegans* genome, but this type of analysis has not been performed.

Palindromes in the yeast genome

The palindrome content of the yeast genome was sorted by size and compared with the palindrome content of the randomly generated equivalents of the genome. Further analysis indicated several interesting numerical trends that allowed us to make a clear distinction between "short" and "long" palindromes, with the breaking point at 10–12 bp. While positive selection could account for the relative abundance of long palindromes, under-representation of short palindromes seems more intriguing. The avoidance of short palindromes was already observed in viral and bacterial genomes and was attributed either to the activity of the restriction enzymes present in the cell, or introduced by horizontal gene transfer (Sharp 1986; Gelfand and Koonin 1997; Rocha et al. 2001). Obviously, such an interpretation could not explain the paucity of short palindromes in the yeast genome, where no restriction/modification system has been detected. Since palindromic dinucleotides were also under-represented in the yeast genome, we decided to perform a systematic analysis of the dinucleotide content of the yeast genome. This analysis indicated a different bias for each palindromic dinucleotide in the following order: TA > CG > AT > GC; and only the GC dinucleotide was slightly over-represented. The strong bias observed for palindromic dinucleotides can also explain the under-representation of other short palindromes, since each palindrome contains a palindromic dinucleotide as a center of symmetry. If, for some reason, such centers of symmetry are less frequently formed, the occurrence of all palindromic sequences may also be expected to be less frequent. Palindromic dinucleotides composed of A-T base pairs were particularly disfavored, consistent with our finding that the A + T content of short palindromes present in the yeast genome is decreased in comparison with the random sequence.

The observed differences in the occurrence of dinucleotides are not specific for the yeast genome. Early sequence analyses indicated strong biases in dinucleotide frequencies in prokaryotes, eukaryotes, mitochondrial and viral genomes (Nussinov 1984). Karlin et al. (1997) also found that TA and CG are among the most under-represented dinucleotides in different microbial genomes; and they extensively discussed the mechanisms that could create a genomic signature. For example, one possible explanation for the observed biases in the dinucleotide frequencies could reflect the yeast codon bias (Sharp and Cowe 1991), since a high proportion of the yeast genome consists of protein-coding sequences. However, our data indicated that 72.4% of all palindromic dinucleotides were located within coding regions that represent 72.7% of the entire genome. This result suggests that the high avoidance of palindromic dinucleotides in the yeast genome (Table 1) is not limited only to the coding DNA. We also observed that the dinucleotide GC-CG ratio was 1.28 for the entire genome and 1.29 for the protein-coding part of the genome, while in palindromes longer than 14 bp it was even increased to 1.33. These results indicate that the observed dinucleotide bias is not a consequence of the bias in codon usage but could rather reflect the intrinsic bias in spontaneous mutagenesis or a bias in replication/repair, as proposed by Karlin et al. (1997). For example, the fact that the TT dinucleotide is 1.47 times more frequent than the TA dinucleotide in the yeast genome could be due to the biased incorporation of non-complementary nucleotides during DNA replication and/or repair synthesis, to the bias in the mismatch repair process, or because the probability for TA dinucleotides to mutate is slightly higher than that for TT dinucleotides. Such biases could be beyond the level of detection in any biochemical assay, but could leave an imprint in the genomic DNA sequences during evolutionary time, like the genome-wide under-representation of short palindromes described here. Moreover, it is possible that a similar mutagenic bias, rather than the activity of restriction enzymes, resulted in the short palindrome/restriction site avoidance observed in bacterial genomes. A systematic study of their palindrome contents, like the one described here, may contribute to a better understanding of the role of these two processes in the evolution of bacterial genomes.

A high preference for non-coding regions was observed for long palindromes, but even the short (4–10 bp) palindromes showed size-dependent avoidance of the protein-coding regions. Therefore, although biased dinucleotide composition could be the main reason for the decreased frequency of short palindromes at the level

of the whole genome, there could be some other mechanism(s) regulating their incidence in the coding regions. Avoidance of the 6-bp palindromes was also observed in the genes of several bacterial species, including *Buchnera* sp. and *Wigglesworthia* sp., which apparently have no RMSs (Fuglsang 2004). It is tempting to speculate that the presence of palindromes in the genes is avoided because they could affect in some way the metabolism of mRNAs. It should be noted that, in addition to the formation of intramolecular hairpins, palindromic sequences can also promote associations between two identical mRNA molecules in opposite orientations. The RNA/RNA duplex created between two palindromic sequences may interfere with subsequent step(s) in protein synthesis. For example, longer palindromes present in mRNAs may produce an effect similar to that of the interfering RNA molecules (Novina and Sharp 2004), stimulating mRNA degradation.

Our data clearly indicate that the number of large palindromes present in the yeast genome is much higher than expected from the analysis of random sequences. Starting from 10–12 bp, palindromes tend to be not only more frequent than expected, but also more AT-rich and preferentially located in the non-coding regions. Palindromes longer than 18 bp are usually built mainly of A-T base pairs and are almost exclusively located in the intergenic regions, comprising only 27.7% of the entire genome. The longest palindrome detected in the yeast genome (12.1 Mb) has 44 bp and contains only two C-G base pairs. An even more pronounced bias for long, AT-rich palindromes was observed in *C. elegans* chromosomes III and X (LeBlanc et al. 2000). Four 60-bp palindromes were found on chromosome III (11 Mb) and 17 on chromosome X (16 Mb) and they were built exclusively of A-T base pairs. Over-representation of large palindromes is consistent with their presumed role in different cellular processes, like the regulation of gene expression or initiation of chromosomal replication; and the high A + T content may facilitate local DNA melting and adoption of secondary structures. Systematic deletions of the longest palindromes detected in the yeast genome, followed by phenotype analysis of the corresponding strains, could shed more light on their possible biological functions.

Different mechanisms may regulate palindrome incidence

The data presented here, together with the results of other studies, may suggest an integrated view on the evolutionary dynamics of palindromic sequences in yeast (Fig. 5). The occurrence of short palindromes in the entire genome could be disfavored due to a slight bias in mutagenic DNA replication and/or the increased probability for palindromic dinucleotides TA, CG and AT to mutate, while additional mechanism(s) could contribute to palindrome avoidance within coding regions. AT-rich palindromes that acquire the critical size of 10–12 nt may become unstable and increase their size, mainly due to the insertion of AT or TA dinucleotides by slippage during DNA replication (Kruglyak et al. 1988; Toth et al. 2000). Since insertions in the coding regions lead to gene inactivation, they are tolerated only in the intergenic regions. Long, AT-rich palindromes could acquire novel functions, but could also present a starting point for the generation of other palindromic sequences, imperfect palindromes and IRs, enlarging the repertoire of possible *cis*-acting genetic elements. The upper size of a palindrome may be regulated by the potential of palindromic sequences to form cruciform structures in vivo. As mentioned before, such structures can be processed to DSBs (Lobachev et al. 2002) and are known to induce different types of recombination events (Gordenin et al. 1993; Leach 1994). The longest palindrome found in the yeast genome contains 44 bp, but the critical size of a palindrome that could act as an initiator of recombination still needs to be experimentally determined.

Concluding remarks

The new research tool for palindrome analysis described here may contribute to a deeper insight into the evolution and functions of these important DNA motifs. Our analysis of the first complete catalog of palindromic sequences present in the genome of a cellular organism revealed several interesting findings. For example, we show that the under-representation of short palindromes is not limited to the bacterial genomes but can also be observed in yeast, while size-dependent palindrome avoidance in the coding regions seems particularly intriguing. We also point out the relevance of dinucleotide bias analysis for the study of palindromes and we believe our computer program will also be useful for more specific tasks, like the identification of potential restriction sites or regulatory elements.

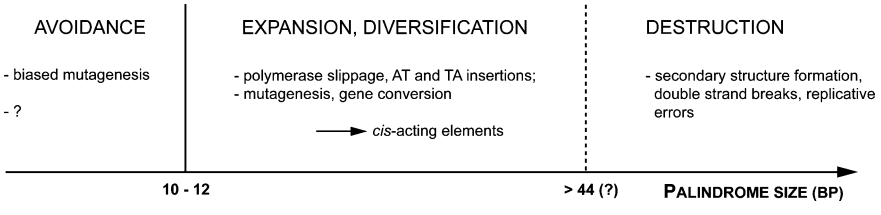

Fig. 5 Evolutionary dynamics of palindromic sequences in yeast. The *solid vertical line* indicates the size limit between short and long palindromes as determined in this work. The critical size leading to palindrome loss/destruction (*dashed vertical line*) needs to be experimentally determined

Acknowledgements We are grateful to Ana Vukelić for help in statistical analysis. This work was supported by grant 0058014 from the Croatian Ministry of Science, Education and Sports.

References

Bissler JJ (1998) DNA inverted repeats and human disease. Front Biosci 3:408–418

Chen Y, Gill GN (1996) A heteromeric nuclear protein complex binds two palindromic sequences in the proximal enhancer of the human *erbB-2* gene. J Biol Chem 271:5183–5188

Cuomo AC, Mundy CL, Oettinger MA (1996) DNA sequence and structure requirements for cleavage of V(D)J recombination signal sequences. Mol Cell Biol 16:5683–5690

Fuglsang A (2004) The relationship between palindrome avoidance and intergenic codon usage variations: a Monte Carlo study. Biochem Biophys Res Commun 316:755–762

Gelfand MS, Koonin EV (1997) Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. Nucleic Acids Res 25:2430–2439

Gordenin DA, Resnick MA (1998) Yeast ARMs (DNA at-risk motifs) can reveal sources of genome instability. Mutat Res 400:45–58

Gordenin DA, Lobachev KS, Degtyareva NP, Malkova AL, Perkins E, Resnick MA (1993) Inverted DNA repeats: a source of eukaryotic genomic instability. Mol Cell Biol 13:5315–5322

Karlin S, Mrazek J, Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179:1363–1370

Kruglyak S, Durret RT, Schug MD, Aquadro CE (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

Kurtz S, Schleiermacher C (1999) REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics 15:426–427

Leach DRF (1994) Long DNA palindromes, cruciform structures, genetic instability and secundary structure repair. Bioessays 16:893–898

LeBlanc MD, Aspeslagh G, Buggia NP, Dyer BD (2000) An annotated catalog of inverted repeats of *Caenorhabditis elegans* chromosomes III and X, with observations concerning odd/even biases and conserved motifs. Genome Res 10:1381–1392

Lobachev KS, Gordenin DA, Resnick MA (2002) The Mre11 complex is required for repair of hairpin-capped double-strand breaks and prevention of chromosome rearrangements. Cell 103:83–193

Mori K, Ogawa N, Kawahara T, Yanagi H, Yura T (1998) Palindrome with spacer of one nucleotide is characteristic of the *cis*-acting unfolded protein response element in *Saccharomyces cerevisiae*. J Biol Chem 273:9912–9929

Novina CD, Sharp PA (2004) The RNAi revolution. Nature 430:161–164

Nussinov R (1984) Doublet frequencies in evolutionary distinct groups. Nucleic Acids Res 12:1749–1763

Repping S, Skaletsky H, Lange J, Silber S, Veen F van der, Oates RD, Page DC, Rozen S (2002) Recombination between palindromes P5 and P1 on the human Y chromosome causes massive deletions and spermatogenic failure. Am J Hum Genet 71:906–922

Rice P, Longden I, Bleasby A (2000) EMBOSS—the european molecular biology open software suite. Trends Genet 15:276–278

Rocha EPC, Danchin A, Viari A (2001) Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. Genome Res 11:946–958

Sharp PM (1986) Molecular evolution of bacteriophages: evidence of selection against the recognition sites of host restriction enzymes. Mol Biol Evol 3:75–83

Sharp PM, Cowe E (1991) Synonymous codon usage in *Saccharomyces cerevisiae*. Yeast 7:657–678

Tashiro J, Kinoshita K, Honjo T (2001) Palindromic but not G-rich sequences are targets of class switch recombination. Int Immunol 13:495–505

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