Palindromic repetitive DNA elements with coding potential in *Methanocaldococcus jannaschii*

Mikita Suyama[a], Warren C. Lathe III[a], Peer Bork[a,b,*]

[a] *EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany*
[b] *Max Delbrück Center for Molecular Medicine, Berlin-Buch, Germany*

Received 24 June 2005; revised 2 August 2005; accepted 18 August 2005

Available online 12 September 2005

Edited by Robert B. Russell

Abstract **We have identified 141 novel palindromic repetitive elements in the genome of euryarchaeon *Methanocaldococcus jannaschii*. The total length of these elements is 14.3 kb, which corresponds to 0.9% of the total genomic sequence and 6.3% of all extragenic regions. The elements can be divided into three groups (MJRE1–3) based on the sequence similarity. The low sequence identity within each of the groups suggests rather old origin of these elements in *M. jannaschii*. Three MJRE2 elements were located within the protein coding regions without disrupting the coding potential of the host genes, indicating that insertion of repeats might be a widespread mechanism to enhance sequence diversity in coding regions.**

Keywords: Archaea; Coding potential; Palindromic repetitive element; Stem–loop

1. Introduction

Repetitive elements are abundant in eukaryotic genomes. For example, more than half of the human genome is composed of them [1]; some of them even contribute to coding regions [2–4]. Bacterial genomes contain much less repeats resulting in small sized genomes and high gene density. Mostly transposons and insertion sequences as well as short repetitive elements that have the ability to form stem–loop structures are found in bacterial genomes [5]. These include PUs (palindromic units), also called REPs (repetitive extragenic palindromes), in *Escherichia coli* and *Salmonella typhimurium* [6,7], enterobacterial repetitive intergenic consensus (ERIC), also called intergenic repeat units (IRU), in *E. coli* [8,9], RPE (*Rickettsia* palindromic element) in *Rickettsia* [10,11], WPE (*Wolbachia* palindromic element) in *Wolbachia* [12], and RUP (repeat unit of pneumococcus) in *Streptococcus pneumoniae* [13]. Some of these repetitive elements have been found within coding regions [10–12], indicating that such a mechanism might contribute to the diversity of protein sequences.

Very little is known about repetitive elements in archaea, the third domain of life [14], which is subdivided into four groups: crenarchaea, euryarchaea, korarchaea, and nanoarchaea. Among these, crenarchaea and euryarchaea are the two main groups and several completed genomic sequences are available. Recently, short repetitive elements were found in the genome of crenarchaea *Sulfolobus solfataricus* [15]. These elements, *Sulfolobus* MITE (miniature inverted-repeat transposable element) or SM, are non-autonomous mobile elements with a total copy number of 143, corresponding to about 0.6% of the genome [15].

Except for low copy number repeats, such short and highly repetitive elements have not been reported so far in euryarchaea [16]. Here, we searched for repetitive elements in the genome of the euryarchaeon *Methanocaldococcus jannaschii* and found that there are three groups of palindromic repetitive elements that seem unique to this archaeal species and that have the ability to integrate into coding regions.

2. Materials and methods

The genome sequence of *M. jannaschii* [17] was taken from the GenBank database (Accession Number: L77117). All intergenic regions were collected based on the annotations in the database. In the first step, BLAST [18] was applied to all-against-all comparison among the intergenic regions. Significantly conserved intergenic regions were then aligned by using the SEAVIEW alignment editor [19]. Based on the alignment, HMM search [20] against the whole genome sequence of *M. jannaschii* was carried out to get more distantly related sequence elements. Newly identified sequences were incorporated into the alignment, and the genome sequence was searched again with the HMM constructed from the revised alignment. This cycle was iterated until convergence of the alignment. Other completed genome sequences were also searched with the HMM to check the existence of the repetitive elements. Possible stem–loop structures for these repetitive elements were predicted by the RNAfold program in the ViennaRNA package [21]. Secondary structure prediction was carried out on the PredictProtein server [22]. Domain analysis was carried out on the SMART server [23].

3. Results and discussion

3.1. Three groups of repetitive elements

We found in total 141 repetitive elements in the genome of *M. jannaschii* (for the list and the alignments of all the elements see Supplementary Table S1 and Fig. S1). The total length of

*Corresponding author. Fax: +49 6221 387 517.
E-mail address: bork@embl-heidelberg.de (P. Bork).

Abbreviations: MJRE, *Methanocaldococcus jannaschii* repetitive element; IS, insertion sequence; MITE, miniature inverted-repeat transposable element; SM, *Sulfolobus* MITE

0014-5793/$30.00
doi:10.1016/j.febslet.2005.08.051

the elements is 14.3 kb, which corresponds to 0.9% of the total genomic sequence and 6.3% of all extragenic regions. These elements can be divided into three groups based on the sequence similarities, and designated as *M. jannaschii* repetitive element 1–3 (MJRE1–3) (Table 1). One of the most prominent features, besides their abundance in the genome, is palindromic patterns that indicate the ability to form stable stem–loop structures. This is also supported by the presence of some compensatory mutations (Fig. 1A–C).

There are 59 copies of MJRE1 in the genome of *M. jannaschii*. The length of the elements ranges from 93 to 125 bp (108 bp on average), and the sequence identities within the group varies from 39% to 87% (on average 62%) (Fig. 1A). One of the interesting features in the alignment of MJRE1 is the existence of a pair of columns occupied not with complementary nucleotides but with the same nucleotides ("#" in the stem–loop line in Fig. 1A). The pair is flanked by residues that form a potential stem structure ("4" in the stem–loop line in Fig. 1A), and would introduce a conserved internal loop in the stem–loop structure, although its function is not clear.

Table 1
Basic statistics of MJREs

Group	Copy number	Average length (bases)	Average sequence identity[a] (%)	Coding[b]
MJRE1	59	108 ± 7	62 ± 8	0
MJRE2	69	96 ± 3	54 ± 9	3
MJRE3	13	101 ± 4	65 ± 8	0

[a]Average sequence identity is calculated from all possible pairs within a group.
[b]Number of the repetitive elements within coding regions.

There are 69 copies of MJRE2 in the genome. The length of the elements ranges from 87 to 106 bp (96 bp on average), and the sequence identities within the group varies from 32% to 83% (54% on average) (Fig. 1B).

MJRE3 occurs in 13 copies in the genome. The length of the elements ranges from 92 to 109 bp (101 bp on average), and the sequence identities within the group varies from 48% to 93% (65% on average) (Fig. 1C).

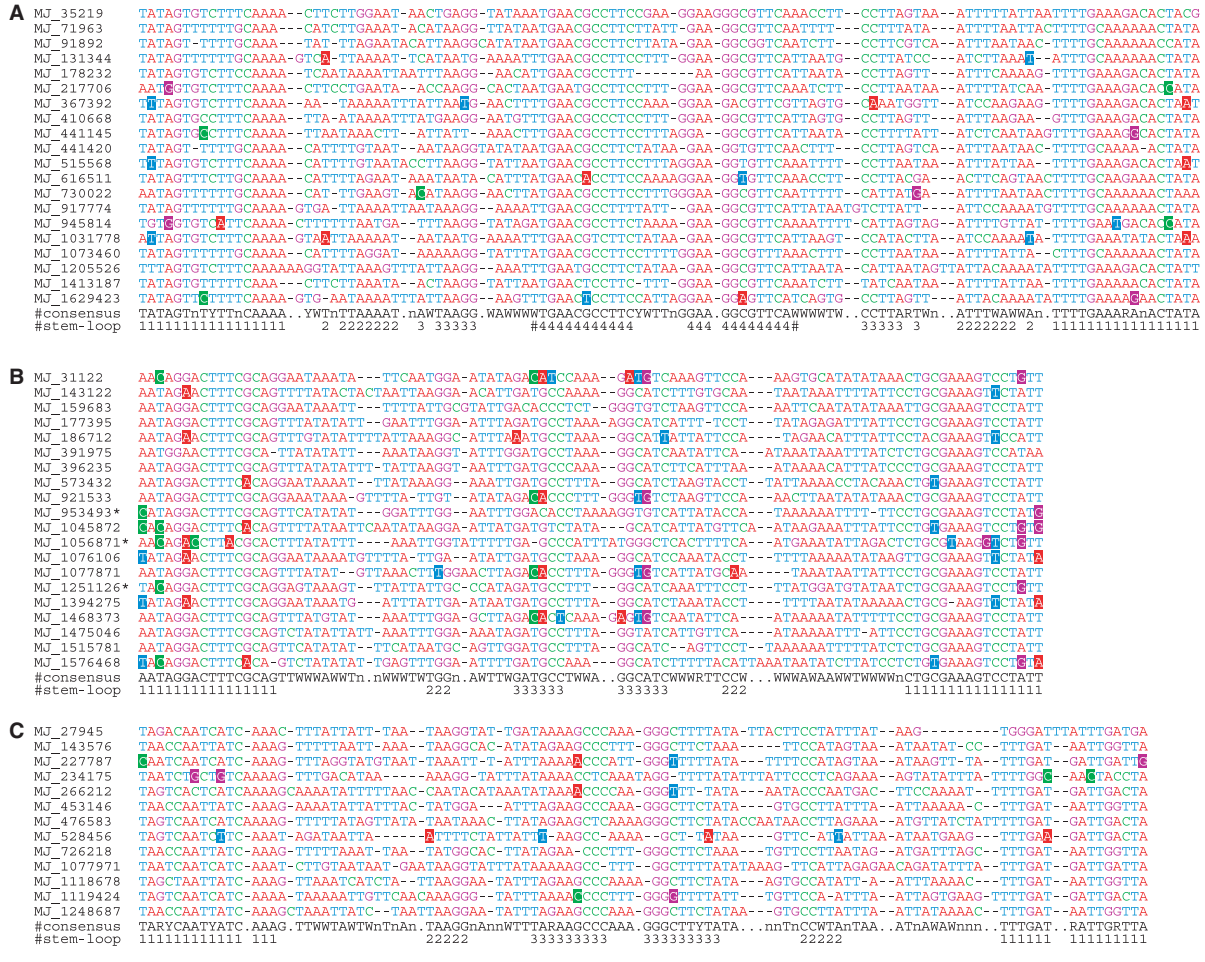

Fig. 1. Multiple sequence alignments of MJREs. (A) MJRE1, (B) MJRE2, and (C) MJRE3. The first column is the sequence identifier containing the positions in the genome. Compensatory changes are indicated by colored backgrounds. The alignment positions conserved in more than 60% of the sequences are indicated in the consensus line: R (A or G); Y (C or T); W (A or T). The residues which form pairs in the possible stem–loop structure are indicated under the consensus sequence by numbers, for example, 1111 at the left of the stem–loop line forms stem with the 1111 at the right. In MJRE1 (A), the pair of columns occupied with the same nucleotides, which might break the base pairing of the stem region, are indicated by "#". In MJRE2 (B), the identifiers of the three elements found within the coding regions are marked by "*". Only highly conserved 20 sequences are shown in MJRE1 (A) and MJRE2 (B) (for the alignments of all the elements see Supplementary Fig. S1).

Although the elements are divided into three groups, these three groups might have a common origin as they share some conserved features and show some limited sequence similarity (e.g., they appear in twilightzone of HMM search of distinct repeats).

It is suggested for SM elements in *S. solfataricus* that the elements are likely to be mobilized by transposases encoded by insertion sequence (IS) elements, because the inverted repeats in SM elements are similar to the terminal inverted repeats of IS elements [15]. To analyze the possible mechanism of proliferation of MJREs, we searched for the inverted repeats similar to MJREs in all IS elements in *M. jannaschii* [16] and in the IS database (http://www-is.biotoul.fr/is.html), and found that there is no IS elements with inverted repeats similar to MJREs, indicating that the mechanism of their emergence might be different from those of SM elements. The alternative explanation is that the IS elements responsible for the proliferation of MJREs might be lost in the course of evolution. Only with the currently available data, it is difficult to conclude about the insertion mechanism of MJREs.

The sequence identity within each group of MJREs is very low, indicating relatively old origin of the elements in the genome of *M. jannaschii*. Even though there is a high degree of sequence divergence in MJREs, there are distinct regions with conservation both at the nucleotide level and putative secondary structural level. This suggests that MJREs are evolving not neutrally but under some constraints for a certain function, although the function of MJREs is not clear (see Section 3.3).

3.2. Coding potential of the repetitive elements

When analyzing the location of the repeats in the genome, we identified three of the MJRE2 group within coding regions (MJ1022 hypothetical protein, MJ1116 asparagine synthetase, and MJ1303 polyferredoxin). In all three cases, the elements are inserted in-frame so that they do not introduce any truncation in the coding sequence such as in- frame stop codons or frame shifts (Fig. 2). When comparing the genes with their orthologs in another methanogen, *M. maripaludis* [24], the alignments clearly show that these are *M. jannaschii* specific insertions (Fig. 2 and Supplementary Fig. S2).

MJ1022 encodes a hypothetical protein with a clearcut ortholog (MMP1595) in *M. maripaludis*. Based on the secondary structure prediction [22], the repetitive element is inserted in the loop region between two α-helices (Fig. 2A). MJ1116 encodes asparagine synthetase, for which a three-dimensional model can be build based on its ortholog in *E. coli* [25]. In the three-dimensional structure, the position of the MJRE is located in the loop region between two α-helices (Fig. 2B). The loop is on the surface of the protein structure, indicating that the insertion does not affect the core of the protein (Supplementary Fig. S3). The third case was found in polyferredoxin protein (MJ1303), which contains 14 ferredoxin domains [23,26]. The element corresponds to the linker region that connects ninth and tenth domains (Fig. 2C and Supplementary Fig. S4), i.e., again does not seem to interfere with the fold.

Fig. 2. Three MJREs within the coding regions and their alignments with the orthologous genes in *M. maripaludis*. (A) MJRE in a hypothetical protein (MJ1022), (B) MJRE in asparagine synthetase (MJ1116) and (C) MJRE in polyferredoxin (MJ1303). The first row shows (A) the predicted secondary structures [22], (B) known secondary structures in the crystal structure of the ortholog in *E. coli* [25], and (C) polyferredoxin domain identified by SMART analysis [23], respectively. α-Helix and β-strand are represented by a green cylinder and a blue arrow, respectively. In each panel, the protein sequence alignment between the orthologous genes in *M. maripaludis* (the second row) and in *M. jannaschii* (the fourth row) is shown. The number in front of the sequence indicates position of the first residue in the protein. Identical residues are indicated in the third row. The corresponding DNA sequence of *M. jannaschii* is shown under the alignment. Note that the DNA sequence is splitted into three rows based on codon positions, and each codon corresponds to the three nucleotides which are aligned vertically. The DNA sequence region corresponding to MJRE is shown in red upper-case letters. Identifiers of MJREs are indicated below the DNA sequences. Note that all the three genes are on the complementary strand of the genome sequence in the database, so the DNA sequences in this figure are the complementary sequences of those in Fig. 1.

All the repetitive elements are inserted in different coding frames and the resulting translates have little similarity in protein sequence. This is different from the cases found in the bacterium *Rickettsia* and *Wolbachia*, where the repeats always inserted in the same frame and translated in similar protein sequences [10–12]. In MJ1022 and MJ1303, the length of the elements is not a multiple of three bases, indicating that the coding sequences of the host genes have been disrupted and then they regained the intact reading frames after the integration of the elements. Another possible explanation is that the integration happened without any disruption of the reading frames, but they look disrupted because of some insertions or deletions after integration. The proteins with MJREs in the coding sequences are likely to be functional since there is no disruption in the current form of reading frames and structures. Moreover, they have clear one-to-one orthologs with similar degrees of sequence identity to the other orthologs in *M. maripaludis*.

It has been shown that some repetitive sequences exist in coding regions of eukaryotic genes. One of such cases is Alu elements found in human transcripts [2,3]. A similar repeat with coding potential has been found in the bacterial genus *Rickettsia* and *Wolbachia* [10–12]. Here, we report the first example of repetitive sequence that is integrated into coding regions of archaeal genes. This suggests that incorporation of repetitive elements into coding regions is a widespread phenomenon, and might contribute to diversity in coding sequences [27]. Indeed, the cases we found use different coding frames introducing protein sequences with little similarities even though they originated from the same family of repetitive elements.

3.3. Random distribution of MJREs in the genome

To test for associated functions of MJREs in extragenic regions, we analyzed the directions of the neighboring genes and their annotated function.

The intergenic regions are grouped into three based on the direction of the neighboring genes: intergenic regions between unidirectional genes (→→), the regions between convergent genes (→←), and the regions between divergent genes (←→). For each group of the intergenic regions, the numbers of the intergenic regions and MJREs were counted (Supplementary Table S2). The χ^2 test was applied to these numbers to check whether there is significant preference of occurrence of MJREs in a certain configuration of intergenic regions, i.e., unidirectional or convergent or divergent. The null hypothesis is that there is no preference for the occurrence of the repetitive elements in a specific configuration of intergenic regions. In all the cases the χ^2 values are very small ($p = 0.51$ for MJRE1, $p = 0.28$ for MJRE2, $p = 0.57$ for MJRE3, and $p = 0.27$ for all MJREs) and even with high probability of rejecting a valid null hypothesis, such as $p = 0.1$ ($\chi^2_{df=2} = 4.605$), the null hypothesis is not rejected, indicating that there is no preferred configuration for MJREs and that the elements are located randomly in the intergenic regions in terms of configuration of the neighboring genes.

The functions of the neighboring genes were analyzed manually based on the annotation in the database. No association with a certain function was observed in the genes flanked by MJREs (for the location of MJREs in the genome see Supplementary Fig. S5). Although this indicates that these elements might not be involved in the regulation of transcription or translation, the elements might play a role in transcription attenuation or mRNA stability, which has been shown for some other microbial repeats [28,29].

To see whether there is a certain preference for location, we plotted the positions of MJREs around the genome (Fig. 3). From the plots, it seems all the three groups of MJREs are evenly distributed around the genome. We tested the statistical significance of the uniform distribution of each group of MJREs by Kuiper's test [30]. The null hypothesis here is that MJREs are uniformly distributed around the genome. The Kuiper statistics V for MJRE1, 2, and 3 are 1.05, 1.07, and 1.20, respectively. All of these values of V are smaller than the upper 5% point of Kuiper statistics V, in which the value $V = 1.75$, indicating that the null hypothesis is not rejected. Thus, we conclude that there is little evidence for a departure from uniform distribution.

Taken together, we were unable to find indications of a functional role of the repeats although their secondary structures are likely contain functional sites.

3.4. Uniqueness of the repetitive elements to M. jannaschii

We searched MJRE homologs in completed genomes (21 archaeal and 203 bacterial genomes) by HMM search using the HMM constructed from each of the MJRE alignments as a query. Although it is known that some of the genomes contain repetitive sequences with similar characteristics such as the length and the potential to form stem–loop structures, there are no highly significant hit (HMMsearch E-value < 0.00001) similar to the inverted repeats of MJREs in the surveyed genomes, even in the two plasmids of *M. jannaschii* (accession numbers: L77118 and L77119) and in another methanogen, *Methanococcus maripaludis* [24]. This suggests that MJREs are unique to *M. jannaschii*.

Recently, comparative genomic approach revealed that SM elements can facilitate genome rearrangement in *Sulfolobus* [31]. Moreover, from a comparison of 22 γ-proteobacteria

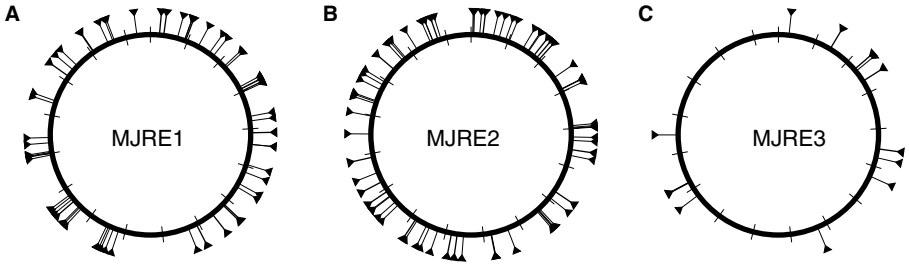

Fig. 3. Distribution of MJREs in the genome of *M. jannaschii*. (A) MJRE1; (B) MJRE2; and (C) MJRE3. Ticks are placed in every 100 kb.

genomes, it has been concluded that there is a strong correlation between repeat density and gene order conservation, indicating that the repetitive elements are involved in genome rearrangements [32]. To test the possible function that MJREs are responsible for genome rearrangements, we analyzed the conservation of gene order of orthologs between *M. jannaschii* and *M. maripaludis*, and measured the correspondence between location of MJREs and gene order breakpoints [33]. All the intergenic regions are divided into two categories, i.e., those located in gene order breakpoints (1278 intergenic regions) and outside breakpoints (480 intergenic regions). For these two categories, the numbers of intergenic regions which contain MJREs are 100 (in breakpoints; 7.8%) and 30 (outside breakpoints; 6.3%), respectively. Then we applied χ^2 test and obtained very high p value ($p = 0.26$), indicating that locations of MJREs are not significantly correlated with gene order breakpoints.

4. Conclusions

This is, to our knowledge, the first example of existence of palindromic repetitive elements with high copy number in euryarchaeon. The low sequence identity within each of the groups suggests relatively old origin of MJREs in the genome. Although MJREs have a potential to form a stem–loop structure, the function is still unclear as it is not for most of the other repetitive elements in prokaryotes. One interesting finding is that three MJREs exist within the coding regions without disrupting the coding potential of the host genes. Such a phenomenon was already reported in eukaryotes and bacteria [2–4,10–12], and MJREs are the first example of repetitive elements with coding potential in archaea, suggesting that this is a widespread mechanism in all three kingdoms of life to enhance diversity in protein sequences.

Acknowledgments: We thank anonymous reviewers for valuable comments. This work was supported by EU Grant (LSHG-CT-2003-503265) to P.B.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2005.08.051.

References

[1] International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409, 860–921.
[2] Makalowski, W., Mitchell, G.A. and Labuda, D. (1994) Alu sequences in the coding regions of mRNA: a source of protein variability. Trends Genet. 10, 188–193.
[3] Li, W.H., Gu, Z., Wang, H. and Nekrutenko, A. (2001) Evolutionary analyses of the human genome. Nature 409, 847–849.
[4] Zdobnov, E.M., Campillos, M., Harrington, E.D., Torrents, D. and Bork, P. (2005) Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. Nucleic Acids Res. 33, 946–954.
[5] Bachellier, S., Clément, J.M. and Hofnung, M. (1999) Short palindromic repetitive DNA elements in enterobacteria: a survey. Res. Microbiol. 150, 627–639.
[6] Higgins, C.F., Ames, G.F., Barnes, W.M., Clément, J.M. and Hofnung, M. (1982) A novel intercistronic regulatory element of prokaryotic operons. Nature 298, 760–762.
[7] Stern, M.J., Ames, G.F., Smith, N.H., Robinson, E.C. and Higgins, C.F. (1984) Repetitive extragenic palindromic sequences: a major component of the bacterial genome. Cell 37, 1015–1026.
[8] Sharples, G.J. and Lloyd, R.G. (1990) A novel repeated DNA sequence located in the intergenic regions of bacterial chromosomes. Nucleic Acids Res. 18, 6503–6508.
[9] Hulton, C.S., Higgins, C.F. and Sharp, P.M. (1991) ERIC sequences: a novel family of repetitive elements in the genomes of *Escherichia coli*, *Salmonella typhimurium* and other enterobacteria. Mol. Microbiol. 5, 825–834.
[10] Ogata, H., Audic, S., Barbe, V., Artiguenave, F., Fournier, P.E., Raoult, D. and Claverie, J.M. (2000) Selfish DNA in protein-coding genes of *Rickettsia*. Science 290, 347–350.
[11] Ogata, H., Audic, C., Abergel, C., Fournier, P.E. and Claverie, J.M. (2002) Protein coding palindromes are a unique but recurrent feature in *Rickettsia*. Genome Res. 12, 808–816.
[12] Ogata, H., Suhre, K. and Claverie, J.M. (2005) Discovery of protein-coding palindromic repeats in *Wolbachia*. Trends Microbiol. 13, 253–255.
[13] Oggioni, M.R. and Claverys, J.P. (1999) Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae*. Microbiol. 145, 2647–2653.
[14] Woese, C.R. and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA 74, 5088–5090.
[15] Redder, P., She, Q. and Garrett, R.A. (2001) Non-autonomous mobile elements in the crenarchaeon *Sulfolobus solfataricus*. J. Mol. Biol. 306, 1–6.
[16] Brügger, K., Redder, P., She, Q., Confalonieri, F., Zivanovic, Y. and Garrett, R.A. (2002) Mobile elements in archaeal genomes. FEMS Microbiol. Lett. 206, 131–141.
[17] Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A. and Gocayne, J.D., et al. (1996) Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii*. Science 273, 1058–1073.
[18] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., 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.
[19] Galtier, N., Gouy, M. and Gautier, C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. 12, 543–548.
[20] Eddy, S.R. (1998) Profile hidden Markov models. Bioinformatics 14, 755–763.
[21] Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M. and Schuster, P. (1994) Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie 125, 167–188.
[22] Rost, B., Yachdav, G. and Liu, J. (2004) The PredictProtein server. Nucleic Acids Res. 32, W321–W326.
[23] Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P. and Bork, P. (2004) SMART 4.0: towards genomic data integration. Nucleic Acids Res. 32, D142–D144.
[24] Hendrickson, E.L., Kaul, R., Zhou, Y., Bovee, D., Chapman, P., Chung, J., Conway de Macario, E., Dodsworth, J.A., Gillett, W. and Graham, D.E., et al. (2004) Complete genome sequence of the genetically tractable hydrogenotrophic methanogen *Methanococcus maripaludis*. J. Bacteriol. 186, 6956–6969.
[25] Larsen, T.M., Boehlein, S.K., Schuster, S.M., Richards, N.G.J., Thoden, J.B., Holden, H.M. and Rayment, I. (1999) Three-dimensional structure of *Escherichia coli* asparagine synthetase B:

a short journey from substrate to product. Biochemistry 38, 16146–16157.
[26] Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Maxon, S. and Sonnhammer, E.L., et al. (2004) The Pfam protein families database. Nucleic Acids Res. 32, D138–D141.
[27] Claverie, J.M. and Ogata, H. (2003) The insertion of palindromic repeats in the evolution of proteins. Trends Biochem. Sci. 28, 75–80.
[28] Espeli, O., Moulin, L. and Boccard, F. (2001) Transcription attenuation associated with bacterial repetitive extragenic BIME elements. J. Mol. Biol. 314, 375–386.
[29] Aranda-Olmedo, I., Tobes, R., Manzanera, M., Ramos, J.L. and Marqués, S. (2002) Species-specific repetitive extragenic palindromic (REP) sequences in Pseudomonas putida. Nucleic Acids Res. 30, 1826–1833.
[30] Fisher, N.I. (1993) Statistical Analysis of Circular Data, Cambridge University Press, Cambridge, UK.
[31] Brügger, K., Torarinsson, E., Redder, P., Chen, L. and Garrett, R.A. (2004) Shuffling of *Sulfolobus* genomes by autonomous and non-autonomous mobile elements. Biochem. Soc. Trans. 32, 179–183.
[32] Rocha, E.P.C. (2003) DNA repeats lead to the accelerated loss of gene order in bacteria. Trends Genet. 19, 600–603.
[33] Suyama, M. and Bork, P. (2001) Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet. 17, 10–13.