The insertion of palindromic repeats in the evolution of proteins

Jean-Michel Claverie and Hiroyuki Ogata

Information Génétique et Structurale, CNRS-AVENTIS UMR 1889, Institut de Biologie Structurale et Microbiologie, IFR 88, 31 Chemin Joseph Aiguier, Marseille FR-13402, France

The current theory of protein evolution is that all contemporary proteins are derived from an ancestral subset. However, each new sequenced genome exhibits many genes with no detectable homologues in other species, leading to the paradoxical picture of a universal ancestor with more genes than any of its progeny. Standard explanations indicate that fast evolving genes might disappear into the 'twilight zone' of sequence similarity. Regardless of the size of the original ancestral subset, its origin and the potential mechanisms of its subsequent enlargement are rarely addressed. Sequencing of *Rickettsia conorii* genome recently led to the discovery of three families of repeat–mobile elements frequently inserted into the middle of protein coding genes. Although not yet identified in other species of bacteria, this discovery has provided the first clear evidence for the *de novo* creation of long protein segments (up to 50 amino acid residues) by repeat insertion. Based on previous results and theories on the coding potential of palindromic elements, we speculate that their insertion and mobility might have played a significant role in the early stages of protein evolution.

The complete sequencing and analysis of the *Rickettsia conorii* genome surprisingly revealed three new families of palindromic repeats capable of in-frame insertion into pre-existing open reading frames (ORFs): *Rickettsia* palindromic element (RPE)-1, RPE-2 and RPE-3 [1–3] (Fig. 1). The size of these repeats is 141 bp, 105 bp and 116 bp for RPE-1, RPE-2 and RPE-3, respectively. In-frame insertion was observed 23 times out of the 45 total occurrences of RPE-1. RPE-2 (5 out of 7) and RPE-3 (4 out of 4) also exhibited a high frequency of in-frame ORF insertions [3]. Homologous repeats were also identified within the ORFs of other *Rickettsia* species [1,3]. Many of the RPE-containing ORFs correspond to genes with important functions (i.e. DNA polymerase I). No intact forms of these genes are found in the *Rickettsia* genome. Although RPE insertions can be found anywhere along the gene sequences, they always appear at the surface of the 3D structure of the proteins in a way compatible with their original fold and function [1,3]. Experiments have shown that RPE-containing genes are transcribed [1] and translated normally, producing proteins with RPE-derived peptide inserts (35–50 amino acid residues) while maintaining their expected enzymatic activity (C. Abergel *et al.*, unpublished).

An RPE-1 sequence was also found to be inserted into a functional RNA gene (tmRNA) [3], making this type of repeat an ultimate molecular parasite, equally capable of propagating within intergenic regions, protein-coding genes, and RNA genes of bacterial genomes.

Initially, such an in-frame insertion phenomenon could be considered a one-in-a-billion-years freak evolutionary accident. However, the identification of three unrelated repeat families within the ORFs of *Rickettsia* – a bacterial genus closely related to the ancestor of mitochondria – indicates that in-frame repeat insertions indeed occurred recurrently, at least in this group of bacteria. Despite their specificity to *Rickettsia*, RPEs resemble other known intergenic palindromic repeats, such as 127-bp intergenic repeat unit (IRU) or 152-bp repeat sequence (RSA) [4,5], in terms of their structure, size and frequency of occurrence in the genome. However, a comprehensive database search of those repeats did not reveal a convincing case for internal ORF insertion [1]. So, why are these in-frame repeat insertion events only seen in *Rickettsia*?

A detailed analysis of the small genomes (~1 Mb) and gene contents of *R. conorii* and *R. prowazekii* did not identify any specific mechanism that could cause such a phenomenon to happen uniquely in *Rickettsia*. On the contrary, these genomes do not show high repeat-frequencies (*R. prowazekii* is in fact among the lowest [6]), and exhibit a very low level of genomic rearrangements [2].

We believe that in-frame repeat insertions are a general phenomenon resulting from the natural properties of ORFs and palindromes. Our opinion is that they occurred recurrently in the past in different bacteria but remained detectable only in the slow-evolving, sequestered obligate intracellular *Rickettsia* [3]. In this case, the insertion of palindromic repeats in pre-existing coding regions might have played a significant role in the overall evolution of protein domains.

The newly discovered coding RPEs highlight questions surrounding: (1) the coding potential of palindromic sequences, (2) the folding capacity of the resulting peptidic chain, and (3) the tolerance of proteins to relatively large peptidic insertions.

Corresponding author: Jean-Michel Claverie (jean-michel.claverie@igs.cnrs-mrs.fr).

http://tibs.trends.com 0968-0004/03/$ - see front matter doi:10.1016/S0968-0004(02)00036-1

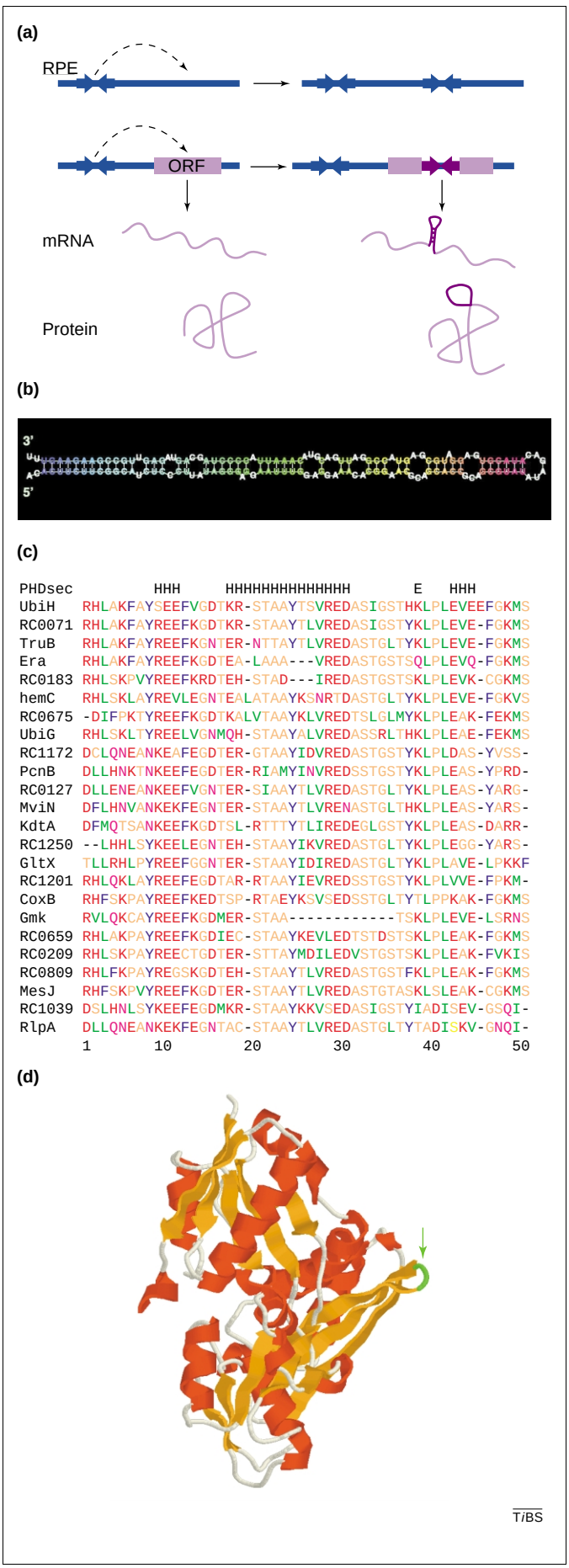

Fig. 1. General features of the *Rickettsia* palindromic elements (RPEs). (a) The mobile RPE randomly spreads over coding and non-coding regions of the *Rickettsia* genomes. When RPE is inserted in-frame within an existing open reading frame (ORF), the RPE-containing ORF is probably transcribed and translated normally. The RPE-derived peptides are predicted to be at the surface of the protein structure. (b) The predicted RNA secondary structure of the RPE-1 found in the glutamyl-tRNA synthetase gene of *Rickettsia conorii*. ©American Association for the Advancement of Science (2000). Reprinted, with permission, from [1]. (c) Alignment of the RPE-1-derived amino acid sequences identified in *R. conorii*. Amino acid residues are colored as follows: F, W and Y, blue; C, yellow; A, G, P, S and T, orange; I, L, M and V, green; D, E, H, K and R, red; N and Q, purple. The letters 'H' and 'E' in the first line represent predicted α-helical and extended conformations, respectively. This alignment can be found in the InterPro motif database (http://ftp.ebi.ac.uk/interpro/index.html; IPR 005728) (d) Crystal structure of the porphobilinogen deaminase from *Escherichia coli* (PDB code: 1PDA). The corresponding *R. conorii* protein is predicted to exhibit an extra peptide segment derived from RPE-1 at the location indicated by the green arrow.

Coding potential of a palindromic sequence

Within a palindromic sequence, the left half and the right half of the sequence from the same DNA strand are, by definition, complementary to each other. How could such a sequence therefore emerge by chance in the course of evolution? Although the detailed mechanisms are still unknown, the duplication of a DNA segment followed by its inverted insertion at one of the extremity of the original segment is the most probable scenario for the generation of a palindromic sequence (Fig. 2). There is little doubt that such a duplication–insertion process (a single event) is more probable than the accumulation of multiple point-mutations.

The DNA segment from which the palindrome first arises could be either a noncoding or a coding sequence (i.e. an ORF). Given the high ORF density in bacterial genome (~80%), a coding segment is more probable. Furthermore, if the duplication–inversion process is mediated by reverse transcription of mRNA [7], transcribed sequences (i.e. genes and thus ORFs) become an even more likely source of palindromes.

The high probability that a newly created palindrome originates from a coding sequence has important consequences for the statistical property of the palindromic sequence along its entire length. Several authors have examined the property of the complementary sequence of coding strands, and observed that the antisense reading frame, RF − 1, of existing genes tends to exhibit less stop codons and larger ORFs than is expected in a random sequence [8–12] (see Box 1 for the definition of different frames). In the *Escherichia coli* K-12 genome (49.2% A + T content), we found that the RF − 1, RF − 2 and RF − 3 frames exhibited stop codons at a frequency of 2.5%, 3.6% and 4.8%, respectively [the expectation is 3/64 (4.7%) for a random sequence with equal proportions of A, T, G and C]. In a frame with 2.5% of stop codons, the statistically expected ORF size is, on average, 39 amino acids, and the expectation for the largest ORF in a whole (1 Mb) bacterial genome is ~1500 ± 150 nucleotides [13]. In agreement with these statistical properties, half of the annotated *E. coli* genes exhibit an antisense ORF longer than 300 nucleotides (the standard threshold in genome annotation). In *R. conorii*, we found 204 genes exhibiting an antisense

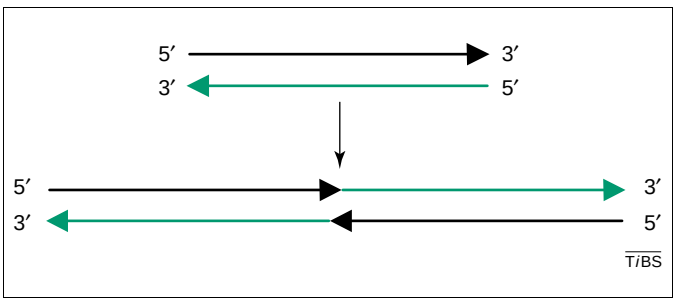

Fig. 2. Hypothetical origin of a palindromic sequence. Duplication followed by nearby insertion into an inverted orientation naturally leads to the formation of palindromic elements. This phenomenon can be mediated by DNA replication (fork) errors, mRNA reverse-translation, or both. The palindromic structure is known to promote the mobility of DNA segments.

ORF longer than 300 nucleotides, although this genome is AT-rich (67.6%) making it more probable that stop codons occur by chance (TAA, TAG and TGA are AT-rich). It is worth noting that the RPEs are not as AT-rich (57–60%) as is the rest of the *Rickettsia* genome [3].

As long antisense ORFs appear to be common in all bacterial genomes, palindromic sequences generated by the duplication–inversion of a pre-existing protein coding sequence (Fig. 2) will therefore often exhibit an ORF over their entire lengths. The spontaneous emergence of palindromic elements that code for peptides the size of a typical small protein domain (>50 amino acid residues) is thus likely to be a frequent mutational event.

Table 1. Amino acid compositions (%) of the conceptual translations on *Escherichia coli* K-12 genes in the six different reading frames[a]

aa[b]	+1	+2	+3	−1	−2	−3
A	9.5	7.9	9.6	7.7	9.9	9.5
C	1.2	4.9	4.6	2.8	3.8	3.5
D	5.2	0.6	4.2	4.2	1.3	3.6
E	5.8	0.8	2.7	2.9	2.1	2.3
F	3.9	3.1	3.2	7.5	3.9	3.1
G	7.4	3.5	10.1	6.5	4.6	5.3
H	2.3	1.9	3.8	5.3	1.0	4.8
I	6.0	5.5	2.4	6.8	6.8	2.0
K	4.4	5.9	2.4	3.4	3.4	2.3
L	10.7	9.8	6.4	7.4	7.6	8.0
M	2.5	2.8	0.4	1.3	2.1	0.7
N	4.0	3.4	2.9	5.0	3.5	2.4
P	4.5	6.7	5.4	3.4	7.7	9.9
Q	4.4	2.6	3.8	6.9	1.8	5.3
R	5.5	11.6	14.3	9.9	8.8	13.2
S	5.8	10.1	8.7	3.9	16.2	9.3
T	5.4	6.6	4.9	6.1	7.4	4.2
V	7.1	6.3	4.2	6.5	5.6	4.8
W	1.5	4.4	2.1	0.9	1.4	1.9
Y	2.9	1.7	3.9	1.6	1.0	3.8
χ^2	**0**	**41.3**	**38.0**	**20.4**	**40.8**	**37.7**

Abbreviations: aa, amino acids; +1, RF + 1; +2, RF + 2; +3, RF + 3; −1, RF − 1; −2, RF − 2; −3, RF − 3.

[a] The χ^2 value $\left(\Sigma\frac{(O-E)^2}{E}\right)$ was computed to quantify the goodness-of-fit between the normal amino acid composition (E = RF + 1) and the alternative reading frames [O = (RF + 2, RF + 3, RF − 1, RF − 2, RF − 3)].

[b] Stop codons are not counted.

Palindromic ORFs lead to well-behaved putative proteins

In addition to a higher probability of being more 'open' than the other antisense reading frames, RF − 1 corresponds to amino acid frequencies close to the composition of actual proteins [9]. This is shown in Table 1, where the χ^2 value was computed to measure the difference between the typical composition of actual proteins (RF + 1) and proteins derived from other frames. Using this criterion, the amino acid composition derived from RF − 1 is closest to that of normal proteins (RF + 1). This implies that, in addition to probably not containing stop codons, palindromic sequences generated in the RF + 1/RF − 1 arrangement statistically lead to peptidic products with more similarity to actual proteins in terms of overall physicochemical properties (e.g. solubility, pI and secondary structure propensity). These newly created protein products will thus be less prone to the formation of aggregates detrimental to the microbial cell, and have greater chance of surviving further evolutionary challenges.

Protein folding

Blalock's molecular recognition theory [14,15] claims that a peptide derived from the antisense RF − 1 exhibits a more than random binding affinity to the peptide derived from the sense RF + 1. Still controversial, this theory is based on a tendency for the 'antipeptides' encoded on the antisense strand (RF − 1) to exhibit hydropathy profiles that are somewhat complementary to the protein encoded by the sense ORF (RF + 1) [15]. Although the mechanisms of the molecular interaction between 'complementary' peptides are not known, the idea was repeatedly applied to the design of active peptides against receptors ([16,17] and references therein). Independent studies have also shown that the binary pattern of hydrophobic and hydrophilic amino acid residues can serve as a good predictor of the peptide properties, as demonstrated by the successful *ab initio* design of four-helix bundle folding peptides [18].

These arguments indicate a further advantage for the palindromic elements generated in the RF + 1/RF − 1 configuration. That is, a certain degree of internal complementarity within the putative protein might contribute to an improved propensity to fold into a self-contained domain-like structure. As in Dwyer's 'trexon' hypothesis, the two complementary peptidic halves would have a tendency for dimerization in a head-to-tail orientation [19]. A fraction of these new peptidic sequences might, thus, be immune from the proteolytic cellular mechanisms directed against misfolded proteins and, hence, be able to perform new functions.

Structure of the RPEs

The arguments outlined previously suggest that palindromic elements generated in the RF + 1/RF − 1 configuration have: (1) a high coding probability; (2) probably lead to a soluble peptide; and (3) might have a tendency to adopt a compactly folded, self-contained domain-like structure. This leads to the prediction that

Box 1. Sense and antisense frame configuration in *Rickettsia* palindromic elements

Definitions of the various reading frames relative to the sense-ORF RF + 1 are as follows:

RF + 1 1 2 3 **1 2 3** 1 2 3 (coding frame)
RF + 2 3 1 2 3 **1 2 3** 1 2
RF + 3 2 3 1 2 3 **1 2 3** 1
RF − 1 3 2 1 **3 2 1** 3 2 1
RF − 2 2 1 **3 2 1** 3 2 1 3
RF − 3 1 **3 2 1** 3 2 1 3 2

The nine possible base-pairing patterns between the left and right strands of a protein-coding perfect palindrome fall into one of the three cases (Table I). The three different frame configurations, RF + 1/RF − 1, RF + 1/RF − 2 and RF + 1/RF − 3, correspond to Case 1, Case 2 and Case 3, respectively.

Table II shows the observed frequencies of different base-pairing patterns in three different *Rickettsia* palindromic element repeats. The left part of each palindrome was optimally paired with the right (complementary) part. The resulting nucleotide pairs were then classified according to their position within the codons (Table I).

Table I. Base-pair categories in a coding palindrome[a]

	R 1st base of codon	R 2nd base of codon	R 3rd base of codon
L 1st base of codon	Case 3	Case 2	Case 1
L 2nd base of codon	Case 2	Case 1	Case 3
L 3rd base of codon	Case 1	Case 3	Case 2

[a]L and R represent the respective left and right strands of a protein-coding perfect palindrome.

Table II. Observed base pairing[a]

	R 1st base of codon	R 2nd base of codon	R 3rd base of codon
RPE–1: Case 1 (0.54), Case 2 (0.32), Case 3 (0.13)			
L 1st base of codon	0.043	0.126	**0.178**
L 2nd base of codon	0.094	**0.172**	0.044
L 3rd base of codon	0.191	0.047	0.105
RPE–2: Case 1 (0.80), Case 2 (0.10), Case 3 (0.09)			
L 1st base of codon	0.021	0.052	**0.271**
L 2nd base of codon	0.010	**0.281**	0.031
L 3rd base of codon	**0.250**	0.042	0.042
RPE–3: Case 1 (0.13), Case 2 (0.34), **Case 3 (0.53)**			
L 1st base of codon	**0.194**	0.102	0.046
L 2nd base of codon	0.102	0.028	**0.157**
L 3rd base of codon	0.056	**0.176**	0.139

[a]L and R represent the respective left and right strands of a protein-coding perfect palindrome. The total frequency of the base pairs supporting each case (Table I) is indicated in parenthesis. The best supported cases are in bold.

RF + 1/RF − 1 should be the dominant configuration for the identified RPEs.

Testing this prediction with the RPE sequences of today is not straightforward as they have accumulated numerous mutations since their birth >40 million years ago [2]. For a perfect palindrome, the optimal base-pairing of the theoretical single-stranded molecule (i.e. treating DNA as RNA) is a totally annealed hairpin. In the symmetrical RF + 1/RF − 1 configuration, the base pairs are formed such that bases 1, 2 and 3 of the codons in the first half of the palindrome are facing bases 3, 2 and 1 of the codons from the second half (Box 1). Owing to their evolution, the palindromes of today are imperfect, and their predicted RNA secondary structures exhibit bulges and loops of different sizes and locations. Yet, some of the statistical properties of the RPE sequences of today can be used to infer the original sense and/or antisense frame configuration. The optimal base-pairing pattern (i.e. the predicted RNA structure) was first computed for each RPE sequence [20]. All base-pairs were then classified according to their positions in their respective codons (Box 1). The number of base-pairs compatible with the three theoretical configurations (RF + 1/RF − 1, RF + 1/RF − 2 or RF + 1/RF − 3) were then computed (Box 1). For the two largest repeat families RPE-1 (45 members; 23 in ORFs) and RPE-2 (seven members; five in ORFs), the results strongly support the RF + 1/RF − 1 configuration. The result for the smaller RPE-3 family (four members; four in ORFs) best fits an RF + 1/RF − 3 model. Globally, this analysis is consistent with the predicted preference for the RF + 1/RF − 1 configuration.

Peptide insertion as a good evolutionary strategy

In contrast to other repeats, RPE insertions show no preference for noncoding sequences versus coding sequences. Within protein coding regions, the insertion sites of the RPE-derived peptides always appear to be at the surface of the protein structures [1,3]. In a typical bacterial genome, noncoding sequences and ORFs represent about 20% and 80% of the sequence, respectively. Considering that a quarter of a protein sequence corresponds to its surface residues [21], the target-sequence sizes become approximately equal for the coding and noncoding fraction of the genome. That the numbers of noncoding versus coding RPEs are approximately the same indicates that they are as well-tolerated at the surface of proteins as they are in non-coding regions. Although initially surprising, this observation is in fact compatible with our current understanding of protein structures and their mutation pattern. Globular proteins exhibit a compact hydrophobic core and relatively flexible surface loops. The latter are known to be much more tolerant to evolutionary changes than the protein core [22]. Experimental insertions of 7–17 residues into a loop of the chymotrypsin inhibitor-2 (64 amino acids) has little effect on protein stability and folding rate [23].

Interestingly, the addition of peptide segments of random sequence at flexible sites of a protein can even improve its function. Matsuura *et al.* [24] designed a population of catalase I from *Bacillus stearothermophilus* by the addition of random peptide tails to the C-terminal of the enzyme. When catalase mutants with much lower thermostability than the wild type were used, they found that the addition of random C-terminal tails could increase their stability above the wild-type level. In another set of experiments, Doi *et al.* [25] showed that insertions of random sequences (120–130 residues) at the surface loop of *E. coli* RNase H1, followed by a subsequent random mutagenesis, could lead to an increase in solubility and RNase activity of the protein. Thus, some natural proteins would not have optimal function and stability; the addition of extra sequences might provide a shortcut to better function and stability [25]. The insertion of RPE peptides in the *Rickettsia* proteins, initially proposed to be evolutionarily neutral, or slightly detrimental [1], might turn out to be beneficial to some of the target proteins. Experimental studies are currently being undertaken to better understand the functional consequence of RPE-peptide insertions.

The majority (93%) of the insertions and/or deletions identified in contemporary protein sequences are shorter than 10 residues [22]. However, this estimate is computed from the most reliable portions of sequence alignments (i.e. those containing only small insertions and deletions). The results are therefore probably biased towards small inserts that remain detectable over longer evolutionary divergence times. Indeed, structural domains have been found to be inserted in the middle of other known domains [26,27]. If such insertions have occurred in a recurrent fashion, the resulting arrangement of (partial) domains will not follow the simple linear arrangement of prototype domains that current domain detection programs expect and were designed for. Interestingly, the leading protein motif Pfam database recognizes recurrent domains in only 69% of SWISS-PROT protein sequences [28]. The identified Pfam domains span only 50% of these protein sequences [28]. According to our experience in annotating whole microbial genomes, it appears that, on average, again 50% of the protein sequences are not covered by any InterPro (the union of all leading domain databases, [29]) domain assignment. The simple model of proteins derived from ancestral sequences through classical mutational events is thus not supported by a significant fraction of their amino acid residues. The current paradigm interprets these apparently unique segments as being beyond the 'twilight zone' of homology detection. Our opinion is that at least part of these unique segments could originate from the complex sequence rearrangement induced by recurrent RPE-like insertions, actually creating new peptidic sequences. Figure 3 illustrates a possible 'Russian Doll' model of recurrent RPE-like insertion, by which the ancestral core of a protein could be successively expanded from the inside out, by the repetition of insertion events at its surface. In this model, new peptidic sequences are added by recurrent genomic DNA sampling through the use of mobile palindromes. Over time, this mechanism will mask the ancestral domain sequence, while classical mutational events (i.e. point mutations and small indels) will progressively erase the palindromic structures of previous insertions [3].

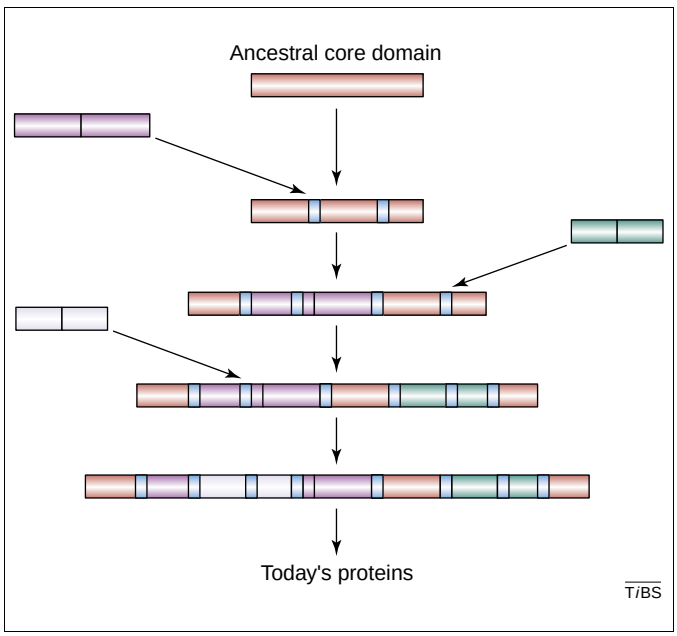

Fig. 3. 'Russian Doll' model of protein evolution. Starting from an ancestral core that can be common to many existing proteins, the successive insertions of palindromic coding segments in solvent-accessible regions of the molecule contribute to the creation of new peptide sequences, while progressively masking the original core domain structure and the palindromic nature of previous insertions. Blue boxes indicate insertion-tolerant segments at the surface of the protein molecule.

Creating new proteins from old repeats

The contribution of noncoding repeated elements to the evolution of proteins has been recurrently argued and remains controversial. It is clear that their mobility and selfish amplification enables them to play a major role in the plasticity of genomic sequences. Short tandem repeats of DNA oligomers, such as microsatellites, are abundant in both prokaryotic and eukaryotic genomes [30,31]. Their expansion mechanism is thought to involve slipped-strand mispairing, which might be the result of inadequate DNA mismatch repair [32]. Ohno *et al.* [33,34] proposed that primordial proteins were encoded by such oligomeric repeats (10 bp units), and that newly arisen coding sequences in modern organisms also derive from such repeats. The gene encoding antifreeze glycoprotein (AFGP) of an Antarctic fish provides clear evidence for such a case. A novel portion of the gene encoding AFGP (which has ice-binding function) is a tandem repeat of a unit, which itself is derived from a part of noncoding and coding sequence of an unrelated trypsinogen gene [35]. The role of much larger transposable elements in protein evolution has also been argued [36,37]. However, some initial reports of Alu-derived sequences in genes [38] were later recognized as artifacts [39,40], even prompting the inclusion of Alu warning entries in SWISS-PROT (P39188–P39195) [41]. A recent analysis of the human genome sequences again

found traces of transposable elements in 4% of human genes [42].

Concluding remarks

Until now, a clear case of a well-conserved large repeat family identified at high frequency in both the coding and non-coding fraction of a genome was missing. This is now provided by RPE-1 and, to a lesser extent, RPE-2 and RPE-3. These repeats exhibit a palindromic structure (required for mobility and amplification), a high entropy sequence (required for real protein creativity), a length compatible with stable self-contained folding (35–50 residues), and evidence for multiple insertions within unrelated proteins at many positions (N terminus, C terminus or middle). Finally, there is now evidence that the RPE-containing ORFs correspond to functional proteins.

Thus, despite their unique identification in *Rickettsia*, the newly discovered RPEs provide the required proof-of-principle that the *de novo* creation of protein segments by palindromic repeats is indeed possible, and has occurred in the past. We thus believe that this mechanism, together with classical mutational processes, should be taken into account in attempts to retrace the evolution of protein structures and sequences.

Note added in proof

For additional speculations about proteins arising from opposite strands of the same gene see Carter, C.W. and Duax, W.L. (2002) Did tRNA synthetase classes arise on opposite strands of the same gene? *Mol. Cell.* 10, 705–708.

Acknowledgements

We would like to thank Chantal Abergel for helpful discussions and for allowing access to her experimental work on *Rickettsia* palindromic element-containing proteins before publication. We also thank Karsten Suhre and David Pollock for their critical reading of this article.

References

1 Ogata, H. *et al.* (2000) Selfish DNA in protein-coding genes of *Rickettsia*. *Science* 290, 347–350
2 Ogata, H. *et al.* (2001) Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii*. *Science* 293, 2093–2098
3 Ogata, H. *et al.* (2002) Protein coding palindromes are a unique but recurrent feature in *Rickettsia*. *Genome Res.* 12, 808–816
4 Bachellier, S. *et al.* (1996) Repeated sequences. In *Escherichia coli and Salmonella* (2) (Neidhardt, F.C. *et al.*, eds), pp. 2012–2040, ASM Press
5 Rudd, K.E. (1999) Novel intergenic repeats of *Escherichia coli* K-12. *Res. Microbiol.* 150, 653–664
6 Frank, A.C. *et al.* (2002) Genome deterioration: loss of repeated sequences and accumulation of junk DNA. *Genetica* 115, 1–12
7 Inouye, S. and Inouye, M. (1995) Structure, function, and evolution of bacterial reverse transcriptase. *Virus Genes* 11, 81–94
8 Boles, E. and Zimmermann, F.K. (1994) Open reading frames in the antisense strands of genes coding for glycolytic enzymes in *Saccharomyces cerevisiae*. *Mol. Gen. Genet.* 243, 363–368
9 Yomo, T. and Urabe, I. (1994) A frame-specific symmetry of complementary strands of DNA suggests the existence of genes on the antisense strand. *J. Mol. Evol.* 38, 113–120
10 Forsdyke, D.R. (1995) Sense in antisense? *J. Mol. Evol.* 41, 582–586
11 Borodovsky, M. *et al.* (1994) Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res.* 22, 4756–4767
12 Boldogkoi, Z. *et al.* (1995) G and C accumulation at silent positions of codons produces additional ORFs. *Trends Genet.* 11, 125–126
13 Smith, T.F. *et al.* (1985) The statistical distribution of nucleic acid similarities. *Nucleic Acids Res.* 13, 645–656
14 Blalock, J.E. (1995) Genetic origins of protein shape and interaction rules. *Nat. Med.* 1, 876–878
15 Blalock, J.E. and Smith, E.M. (1984) Hydropathic anti-complementarity of amino acids based on the genetic code. *Biochem. Biophys. Res. Commun.* 121, 203–207
16 Bost, K.L. *et al.* (1985) Similarity between the corticotropin (ACTH) receptor and a peptide encoded by an RNA that is complementary to ACTH mRNA. *Proc. Natl. Acad. Sci. U. S. A.* 82, 1372–1375
17 Baranyi, L. *et al.* (1995) The antisense homology box: a new motif within proteins that encodes biologically active peptides. *Nat. Med.* 1, 894–901
18 Kamtekar, S. *et al.* (1993) Protein design by binary patterning of polar and nonpolar amino acids. *Science* 262, 1680–1685
19 Dwyer, D.S. (1998) Assembly of exons from unitary transposable genetic elements: implications for the evolution of protein–protein interactions. *J. Theor. Biol.* 194, 11–27
20 Hofacker, I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* 125, 167–188
21 Wootton, J.C. (1994) Sequences with 'unusual' amino acid compositions. *Curr. Opin. Struct. Biol.* 4, 413–421
22 Pascarella, S. and Argos, P. (1992) Analysis of insertions/deletions in protein structures. *J. Mol. Biol.* 224, 461–471
23 Ladurner, A.G. and Fersht, A.R. (1997) Glutamine, alanine or glycine repeats inserted into the loop of a protein have minimal effects on stability and folding rates. *J. Mol. Biol.* 273, 330–337
24 Matsuura, T. *et al.* (1999) Evolutionary molecular engineering by random elongation mutagenesis. *Nat. Biotechnol.* 17, 58–61
25 Doi, N. *et al.* (1997) Insertion of foreign random sequences of 120 amino acid residues into an active enzyme. *FEBS Lett.* 402, 177–180
26 Russell, R.B. (1994) Domain insertion. *Protein Eng.* 7, 1407–1410
27 Gibson, T.J. *et al.* (1994) PH domain: the first anniversary. *Trends Biochem. Sci.* 19, 349–353
28 Bateman, A. *et al.* (2002) The Pfam protein families database. *Nucleic Acids Res.* 30, 276–280
29 Apweiler, R. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* 29, 37–40
30 van Belkum, A. *et al.* (1998) Short-sequence DNA repeats in prokaryotic genomes. *Microbiol. Mol. Biol. Rev.* 62, 275–293
31 Metzgar, D. *et al.* (2002) Domain-level differences in microsatellite distribution and content result from different relative rates of insertion and deletion mutations. *Genome Res.* 12, 408–413
32 Strand, M. *et al.* (1993) Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* 365, 274–276
33 Ohno, S. and Epplen, J.T. (1983) The primitive code and repeats of base oligomers as the primordial protein-encoding sequence. *Proc. Natl. Acad. Sci. U. S. A.* 80, 3391–3395
34 Ohno, S. (1987) Evolution from primordial oligomeric repeats to modern coding sequences. *J. Mol. Evol.* 25, 325–329
35 Chen, L. *et al.* (1997) Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc. Natl. Acad. Sci. U. S. A.* 94, 3811–3816
36 Smit, A.F. (1999) Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* 9, 657–663
37 Miller, W.J. *et al.* (1997) Molecular domestication of mobile elements. *Genetica* 100, 261–270
38 Margalit, H. *et al.* (1994) A complete Alu element within the coding sequence of a central gene. *Cell* 78, 173–174
39 Tugendreich, S. *et al.* (1994) Alu sequences in RMSA-1 protein? *Nature* 370, 106
40 Yeo, J.P. *et al.* (1997) Erratum. *Nature* 388, 697
41 Claverie, J.M. and Makalowski, W. (1994) Alu alert. *Nature* 371, 752
42 Nekrutenko, A. and Li, W.H. (2001) Transposable elements are found in a large number of human protein-coding genes. *Trends Genet.* 17, 619–621