Duplication, divergence and formation of novel protein topologies

Christine Vogel\(^1\)* and Veronica Morea\(^2\)

Summary

The rearrangement or permutation of protein substru-
structures is an important mode of divergence. Recent work\(^{(1)}\) explored one possible underlying mechanism called permutation-by-duplication, which produces special forms of motif rearrangements called circular permutations. Permutation-by-duplication, involving gene duplication, fusion and truncation, can produce fully functional intermediate proteins\(^{(1)}\) and thus represents a feasible mechanism of protein evolution. In spite of this, circular permutations are relatively rare and we discuss possible reasons for their existence. BioEssays 28:973–978, 2006. © 2006 Wiley Periodicals, Inc.

Permutations are one form of protein divergence

Since the first protein structures and sequences became available more than 50 years ago, it became clear that gene duplication and divergence are major processes in the formation of the protein repertoire. Duplicated proteins often diverge via the introduction of point mutations into their sequences, but can also evolve through the rearrangement of sequence stretches longer than one amino acid. Such rearrangements in protein structures are called shufflings, recombinations or permutations. Permutations are identified in homologous proteins, which differ in the sequential order of their sequence elements.

A special form of permutation is represented by circular permutations (CPs), characterized by the presence of the same protein subsequences in the same linear order but different positions of the N and C termini (Fig. 1). Among the protein families containing CP members, bacterial SAM-dependent DNA-methyltransferases (DNA-MTases) are a well-studied example. The members of this family identified to date differ in the sequential organization of five permutation units, namely: the SAM-binding domain, catalytic domain, C-terminal helix and each of the two halves of the target recognition domain (Fig. 1). Proteins showing the same sequential order of these units are assigned to one of seven classes (\(a, \delta, e, m^{5}C, \beta, \gamma, \zeta, \eta\)) that belong to two groups of circular permutants (Fig. 2). Recent in vitro experiments by Peisajovich et al.\(^{(1)}\) explored the viability of permutation-by-duplication, one of the mechanisms proposed to be responsible for producing CP.

Molecular mechanisms producing permutations

While CPs are easy to visualise in a three-dimensional protein structure (Fig. 1), their formation at a genetic level is somewhat less straightforward to imagine, and several mechanisms have been proposed to explain their origins (Fig. 3).

(1) Permutations can arise after production of the protein, without changing the actual gene sequence. During such post-translational permutation, a new peptide bond is formed between the original N and C termini, and the chain is opened at a different location in the protein. One such example is flavin/concanavalin A, the first circular permutation reported.\(^{(2)}\)

(2) Permutations can also be produced by duplication and reinsertion of all or part of a gene sequence, followed by deletion of redundant segments\(^{(3,4)}\) (Fig. 3, left). These partial duplications and deletions can lead to both circular and non-circular permutations. The shuffled gene parts may, in eukaryotes, correspond to exons.\(^{(4,5)}\) A succession of partial shuffling events has also been hypothesized for the DNA-MTases,\(^{(6)}\) based on structural similarity between the SAM binding and the catalytic domain. These
two domains might have arisen from the duplication of the small-molecule binding centre of an ancestral AdoMet-binding protein, followed by an evolved ability of either the N- or C-terminal end to bind adenine and finally an insertion of the more variable target-recognition domain. A third potential mechanism responsible for circular permutations is that of permutation-by-duplication. During this process, the whole gene is duplicated and the two resulting copies fuse to form a tandem repeat. Subsequently, the two termini are truncated removing redundant subsequences (Fig. 3, right). Indeed, tandemly fused genes, which are the first intermediates proposed to occur, have been identified, e.g. amongst saposins/swaposins and also DNA-Mtases.

Figure 1. Circular permutations of DNA-methyltransferases. In DNA-methyltransferases (DNA-MTases), permutation units, i.e. protein substructures that are shuffled during circular permutations, are coloured as in Peisajovich et al.: red, SAM-binding domain; blue, catalytic domain; green, target recognition domain (TRD); and magenta, C-terminal helix. The ribbon diagram represents the X-ray three-dimensional structure of M. Haelll DNA-MTase of the m5C class (PDB ID: 1dct, left) and of a hypothetical circular permuntant protein of ζ class (right). The hypothetical permuntant has been modelled by connecting the C and N termini of M. Haelll with a three-residue linker (white) and by deleting one residue between the catalytic domain and the TRD of M. Haelll to open new N- and C-terminal ends (yellow). N-ter, N terminus; C-ter, C-terminus.

Figure 2. Cartoon representation of the arrangement of the five permutation units, represented by boxes, in the seven known classes of DNA-MTases. Here the TRD is coloured in different shades of green as it is split into two parts by circular permutations. We observe two distinct groups of variants, i.e. classes x, δ, and classes m5C, β, γ, ζ, η. Proteins within one group are circular permuntants of one another. Across the two groups, evolution can be imagined as a combination of partial duplication and deletion events (see Fig. 3).
Despite these hypotheses, the exact molecular mechanisms responsible for CPs and their relative contributions are still unclear.

**In vitro permutation-by-duplication**

Recently, Peisajovich et al.\(^{(1)}\) used an in vitro directed evolution approach to explore how viable the permutation-by-duplication mechanism is for production of DNA-MTase CP (Fig. 4).

(1) Starting with the gene of *Hae* III DNA-methyltransferase \((M\text{-}Hae\text{III})\) of the class \(m^{3}C\), the authors generated a genetic construct containing a tandem fusion of this gene, connected through a five-residue linker. The function of the protein product, i.e. its ability to methylate GGCC sites for protection against digestion by the cognate restriction endonuclease *Hae* III, was comparable to that of the starting gene both in vivo and in vitro.

(2) Start codons were introduced at random positions in the first gene copy, generating three N-terminally truncated proteins. Similarly, they introduced stop codons in the second gene copy, producing eight proteins truncated at the C-terminus. Despite N- or C-terminal truncation, the intermediates demonstrated partial or full resistance against *Hae* III digestion, respectively.

(3) Subsequently the N- and C-terminally truncated intermediates were randomly truncated at their C- and N-termini, respectively, generating sequences having the same sequence elements as the original *M. Hae* III gene, but in a circularly permuted order. Some of the CP intermediates were subjected to a few rounds of random point mutations and selection for *M. Hae* III activity, which, interestingly, resulted in the production of CP variants with increased activity.

(4) Similar results in terms of truncation positions and functional activity were obtained when producing CPs by simultaneously introducing random start codons in the first copy and stop codons in the second copy of the fused gene generated in step 1, followed by selection for methylase activity.

The final CP variants obtained correspond to the \(\beta\), \(\zeta\), and (previously unidentified) \(\eta\) classes, but no variant corresponding to the \(\gamma\) class was isolated.\(^{(1)}\) As suggested by Peisajovich et al, this might be due to the specific starting fold and/or to the specific sequence with that fold used in the experiment. However, in principle, if additional truncations could be ‘forced’ into some of the intermediates, they would produce CPs of the \(\gamma\) class as well.

The activity of both the intermediate and the final species produced in experiments mimicking permutation-by-duplication\(^{(1)}\) strongly argues for the mechanism’s viability, and disproves previous concerns about exposure of hydrophobic surfaces by redundant motifs, hampering correct folding.\(^{(3)}\) Further support is provided by naturally occurring DNA-MTases that still contain redundant copies of some of the motifs, and by tandem DNA-MTases.\(^{(9)}\) Indeed, one of the circular permutants of the newly identified \(\eta\)-class\(^{(1)}\) (*M. Esp3I*) appears to consist of at least two consecutive DNA-MTase domains of classes \(\gamma\) and \(\eta\), which differ in motif arrangements.
Notably, all the functional intermediates isolated by Peisajovich et al. contained at least one complete set of motifs.(1)

Circular permutations in nature

While we know of several example protein families with circular permutations,(10,11) and, in spite of experimental work that demonstrates that more protein structures can tolerate CPs than those selected during evolution,(12) CPs seem to be rare events in nature. CPs are observed in only ~5% of proteins of known structure;(13) inversions of relatively large sequence stretches (domains) are also rare, occurring with a frequency of <5%.(14,15) Thus, one may ask why CPs exist in nature at all, why they may be tolerated (or advantageous) in some proteins and not in others, and what possible functions they have.

Peisajovich et al., report that sequence termini of their related functional variants, which correspond to naturally occurring and previously engineered CPs, cluster in specific regions.(1) Therefore, the authors suggest that specific structural and/or functional constraints must restrict the location of these termini along the protein chain.(1) Based on the activity of their intermediates and calculations of non-polar solvent accessible surface area of derivatives of the M. HaeIII three-dimensional structure, they also suggest that the DNA-MTase fold might be made of independent folding units, each having a relatively low level of hydrophobic surface exposed to solvent. Regions where new chain termini might be introduced would therefore be allowed only at the boundaries between those ‘modules’,(1) consistent with previous reports.(16)

Interestingly, in the M. HaeIII variants obtained by Peisajovich et al.,(1) truncations occur generally but not exclusively outside secondary structure elements (data not shown). The same has been seen in other CPs, suggesting that permutation units do not necessarily coincide with boundaries between secondary structure elements.(10,17) Furthermore, although the close spatial proximity between the N- and C-termini of many globular domains(18) is considered supportive of CPs,(8) exceptions to this also exist.(13)

Thus, at present, it remains unclear whether there are structural reasons why CPs occur in some proteins but not in others, and how the shuffled subsequences are defined. Obviously, CPs, insertions and deletions will be more readily accommodated in protein structures if they do not affect the integrity of the functional core. Indeed, when mapping the N- and C-termini of the final and intermediate DNA-MTase circular permutants obtained by Peisajovich et al.(1) onto the known M. HaeIII structure,(19) most of the N- and C-termini appear to lie on the periphery (data not shown), where the
insertion or deletion of even relatively long sequence stretches hardly affects the active site.

The selection of CP variants may also result from their changes in function compared to the original gene. Due to the conservation of the relative spatial position of the shuffled subsequences, the overall function of CP variants is generally preserved. However, CPs can increase functional diversity by introducing changes in substrate specificity or even in enzyme activity. For example, for some domains whose homologues occur in different multi-domain proteins, a change in the position of the N- and C-termini relative to the functional site within the domain may be necessary for proper integration of the domain with respect to the functional sites of the other domains that are part of different proteins; one such example are PDZ domains (A. Murzin, personal communication).

Finally, the methyltransferase fold is well conserved among known members of the SAM-dependent methyltransferase superfamily (A. Murzin, personal communication) while CP variants occur only in DNA-MTases and not in MTases acting on other substrates. This observation suggests that CP variants in DNA-MTases may exist for reasons that do not depend on their particular structural arrangement. Indeed, in the case of bacterial DNA-MTases a very plausible reason for the existence of CPs is provided by their biological function and, in particular, by the strong selective pressure applied by the cognate restriction endonuclease enzymes. In bacteria, DNA-MTases and restriction endonucleases are coupled together into single genetic units thought to be ‘selfish elements’ that can be laterally transferred between different bacteria. As DNA-MTases protect restriction sites of the host from cleavage by the cognate endonuclease, mutations resulting in complete loss of methylase activity would lead to cell death by endonuclease activity, whereas partially active DNA-MTase mutants can survive and potentially proceed to increase fitness by introducing additional mutations.

Conclusion
While permutations of protein subsequences are a common mode of divergence, circular rearrangements, i.e. circular permutations, are rare. However, as it is often the oddities and curiosities that can tell us the most about general principles, circular permutations have sparked researchers’ curiosity for a long time. The permutants of DNA-MTases are one such oddity, and several different reasons may explain their existence.

In their paper, Peisajovich et al deliver an important proof of the plausibility of the permutation-by-duplication mechanism, proposed to generate CPs in DNA-MTases. The authors demonstrate that almost all natural classes of permuted DNA-MTases are obtained during the directed-evolution experiments and that all intermediate and final variants are functional. Thus, duplication and fusion of whole genes, followed by partial deletion of redundant parts at either chain terminus can produce circular permutations.

Based on these recent findings, future research may take several paths. For example, it remains to be seen what the relative contribution of permutation-by-duplication is in generating permutations, and to what extent other mechanisms, such as partial duplication and reshuffling or post-translational modifications, have a role in the generation of circular and non-circular rearrangements. Furthermore, the permutation-by-duplication mechanism could be generalized to produce non-circular permutations (in addition to circular ones) by the deletion of internal subsequences rather than only N- or C-termini. The outcome of directed evolution experiments may also be helpful when trying to improve our ability to predict what constitutes a recombinatory unit. Finally, future experiments may introduce CPs into proteins to provide them with favourable properties, or construct completely novel topologies; some exciting studies in this direction have been published recently.

Apart from their interest from an evolutionary point of view, circular permutations have also been extensively studied in protein engineering to investigate the rules governing protein stability, structure and folding; protein folding seems to be particularly sensitive to this kind of rearrangement. We expect CPs and experimental studies investigating them, like those designed by Peisajovich et al., to continue to play an important role in these investigations.

Acknowledgments
We are grateful to Arthur Lesk and Alexey Murzin for most valuable comments and suggestions, and we thank Julian Gough, Emma Hill, and Stella Veretnik for critical reading of the manuscript. We would also like to acknowledge the work by all authors that could not be cited due to space restrictions of this article.

References