nature genetics

Evolution of new protein topologies through multistep gene rearrangements

Sergio G Peisajovich, Liat Rockah & Dan S Tawfik

New protein folds have emerged throughout evolution, but it remains unclear how a protein fold can evolve while maintaining its function, particularly when fold changes require several sequential gene rearrangements. Here, we explored hypothetical evolutionary pathways linking different topological families of the DNA-methyltransferase superfamily. These pathways entail successive gene rearrangements through a series of intermediates, all of which should be sufficiently active to maintain the organism's fitness. By means of directed evolution, and starting from *Hae*III methyltransferase (M.*Hae*III), we selected all the required intermediates along these paths (a duplicated fused gene and duplicates partially truncated at their 5′ or 3′ coding regions) that maintained function *in vivo*. These intermediates led to new functional genes that resembled natural methyltransferases from three known classes or that belonged to a new class first seen in our evolution experiments and subsequently identified in natural genomes. Our findings show that new protein topologies can evolve gradually through multistep gene rearrangements and provide new insights regarding these processes.

Some of the clearest examples of protein folds that are presumed to be evolutionarily related are 'circular permutants'[1–4]. These can be visualized as the ligation of the N- and C-termini of a protein and the opening of the chain at another site to yield a new topology. Circular permutants are artificially created by ligating the 5′ and 3′ ends of a gene and opening it at another site[5]; in nature, however, such genomic rearrangements are unlikely. Naturally occurring circular permutations have therefore probably evolved through other types of gene rearrangements. A widely accepted model, known as 'permutation by duplication'[1,6–8] (**Fig. 1a**), postulates that gene duplication and in-frame fusion occur first, followed by partial degeneration of the 5′ coding region in the first copy of the duplicated gene (creating a new start codon) and of the 3′ coding region in the second copy (introducing a new stop codon), resulting in a new topology[1,6–8]. As these events are not likely to be concurrent, it must be assumed that there is a series of evolutionary intermediates that retain the activity of the original protein at a level sufficient to avoid a substantial reduction in the fitness of the organism. Unfortunately, the sequences of circularly permuted genes do not provide unequivocal evidence about their birth histories[9]. Thus, the plausibility of the permutation-by-duplication model, or indeed of other complex multistep rearrangements that may yield new folds[2–4,10], has not been established.

DNA methyltransferases are a prolific example of how different topologies might relate through circular permutation. They are composed of a small target-recognition domain (TRD) and a large catalytic domain with a Rossmann fold structure (**Fig. 1b**)[11]. The linear order of nine conserved sequence motifs in the catalytic domain[12,13] and the location of the TRD define at least seven families, or classes (**Fig. 1c**), that could be related by circular permutation. The occurrence of natural methyltransferases composed of two enzymes fused in tandem[14–16] inspired the permutation-by-duplication model[8]. This model has been validly questioned, however, in particular with regard to the stability and functionality of the truncated intermediates that are likely to have exposed hydrophobic surfaces[9]. Here, we investigate the mechanism by which gradual changes that occur through the intermediates predicted by this model ultimately lead to new protein topologies.

RESULTS

We started with M.*Hae*III, a 5-methylcytosine (m^5C)-class methyltransferase, and subjected it to the various steps hypothesized by the permutation-by-duplication model. A previously described *in vivo* selection[17] technique was applied to isolate all functional variants generated by this process. Briefly, gene libraries derived from the gene encoding M.*Hae*III were transformed into *Escherichia coli*. In the cultured cells, plasmids containing active M.*Hae*III-encoding genes became methylated at their GGCC sites and thereby gained resistance to the cognate restriction endonuclease *Hae*III. Transformation of the plasmid DNA that survived *Hae*III digestion into *E. coli* enriched the gene pool with genes encoding M.*Hae*III activity.

To assess the methyltransferase activity of the selected genes, protection of GGCC sites against *Hae*III digestion was measured after expression of the genes *in vivo* and *in vitro* (using cell-free extracts). To provide a realistic evolutionary scenario, selection and *in vivo* activity tests were carried out under very low expression levels.

Department of Biological Chemistry, Weizmann Institute of Science, Rehovot, 76100 Israel. Correspondence should be addressed to D.S.T. (dan.tawfik@weizmann.ac.il).

Received 8 September; accepted 21 November 2005; published online 15 January 2006; doi:10.1038/ng1717

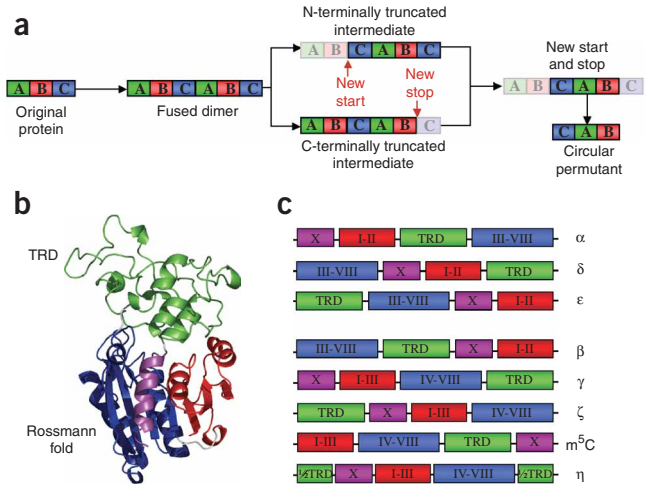

Figure 1 The permutation-by-duplication model is presumed to account for the evolution of various classes of DNA methyltransferases. (**a**) Permutation-by-duplication model. Gene duplication and in-frame fusion lead to a fused dimer. Partial degeneration of the 5′ and 3′ coding regions of the first and second copies of the duplicated genes, respectively, gradually removes redundant regions and eventually leads to circular permutants. (**b**) Three-dimensional structure of the m^5C class M.*Hae*III. The large catalytic domain (Rossmann fold) comprises a SAM-binding subdomain (residues 1–59, motifs I–III, red) connected by a loop to the catalytic subdomain (residues 63–178, motifs IV–VIII, blue) and followed by a TRD (green). Motif X closes the Rossmann fold with a C-terminal helix (magenta). (**c**) DNA methyltransferases are classified according to the linear order of their conserved sequence motifs (I–X) and the location of the TRD, thus defining at least seven classes: α, δ and ε are related by circular permutations, as are β, γ, ζ and m^5C (refs. 9,12,25). Class η was predicted by our laboratory evolution experiments and identified in natural genomes.

Using the same vector, strain and growth conditions used for the methyltransferase selection experiments, the expression levels measured with a reporter gene were ~230 protein molecules per cell.

The evolutionary intermediates are functional

The first step along the permutation-by-duplication pathway is gene duplication and in-frame fusion leading to a dimer arranged in tandem (fused dimer, **Fig. 1a**). We mimicked this step by connecting two copies of wild-type M.*Hae*III through a five-residue linker. The plasmid encoding this construct was fully protected against digestion by *Hae*III (**Fig. 2a**), indicating that the fused dimer was active *in vivo*. We also measured the activity of the fused dimer *in vitro* and found that it methylated 50% of the target sites ($t_{50\%}$) in ~0.3 min, compared to wild-type M.*Hae*III, which had a $t_{50\%}$ of ~0.2 min. Thus, dimerization does not involve substantial reductions in function, as suggested by the relatively high occurrence of fused methyltransferase dimers in natural genomes (refs. 14–16 and **Supplementary Fig. 1** online). As a control, we confirmed that removing a fragment, delimited by *Nde*I sites, from the coding regions of all selected intermediates abolished their methyltransferase activity (**Supplementary Fig. 3**).

The next step in the pathway involves partial degeneration of either the 5′ coding region of the first copy or the 3′ coding region of the second copy of the fused dimer. We mimicked this step using the incremental truncation for the creation of hybrid enzymes (ITCHY) methodology[18,19]. Briefly, a library of N-terminally truncated intermediates was generated by introducing start codons at random locations along the first copy of the genes encoding M.*Hae*III (**Supplementary Fig. 2** online). Analogously, a library of C-terminally truncated intermediates was generated by introducing stop codons at random locations along the second copy. Analysis of these libraries revealed a random distribution of the locations of the new start and stop codons. Both libraries were then selected *in vivo* for M.*Hae*III activity.

The C-terminally truncated library became almost fully resistant to *Hae*III digestion after three rounds of selection (**Fig. 2b**), whereas the N-terminally truncated library showed only partial protection (data not shown). Sequencing of isolated clones indicated that about half were revertants in which one of the two fused genes had been fully, or almost fully (up to ten redundant residues kept), eliminated by the truncation. In contrast, the frequency of revertants in the naive library was substantially lower (≤10 of 330 possible incremental truncations, or 0.03) than that of the *bona fide* truncated intermediates (320 of 330, or 0.97). Given that the ratio of revertants to intermediates was ~1:1 after selection, we deduced that the revertants were ~30 times more

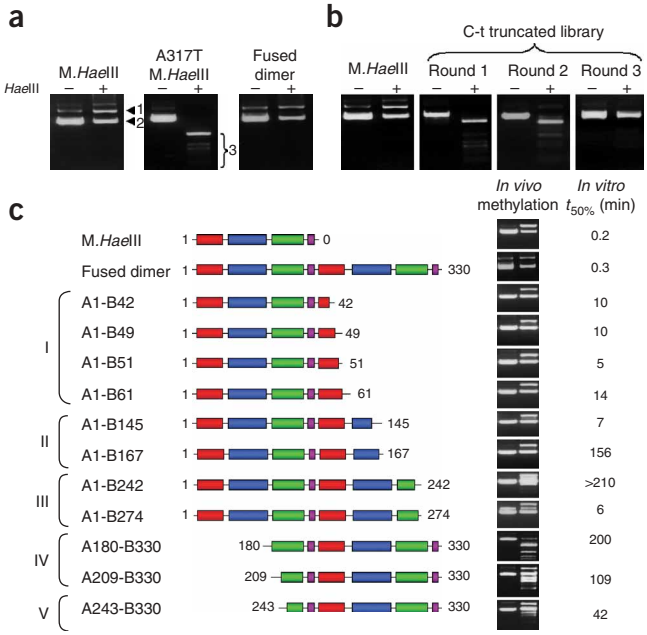

Figure 2 Reproducing the first two steps of the permutation-by-duplication model by directed evolution. (**a**) Fused dimer is active *in vivo*. Agarose gel electrophoresis of plasmids encoding M.*Hae*III, an inactive mutant (A317T M.*Hae*III) and the fused dimer were incubated with (left lane) or without (right lane) *Hae*III endonuclease. Plasmids encoding active methyltransferases were resistant to *Hae*III digestion. The upper band (1) corresponds to circular DNA; lower band (2) corresponds to supercoiled DNA. In contrast, plasmid encoding inactive A317T mutant was fully digested to give a range of short fragments (3). (**b**) Selection of C-terminally truncated evolutionary intermediates. Plasmids encoding M.*Hae*III (lanes 1 and 2) and C-terminally (C-t) truncated libraries at various rounds of selection were digested with *Hae*III (–) or left untreated (+). (**c**) Active N- and C-terminally truncated evolutionary intermediates isolated from the *in vivo* selections. Colored blocks refer to subdomains of M.*Hae*III described in **Figure 1b**. Clones were named as follows: 'A' number indicates residue at which the first copy of the two tandem-arranged M.*Hae*III genes starts; 'B' number indicates residue at which the second copy ends. Also shown for each clone are *in vivo* methyltransferase activity (left lane, untreated plasmid; right lane, plasmid incubated with *Hae*III) and *in vitro* methyltransferase activity (expressed as time required to methylate 50% of M.*Hae*III target sites).

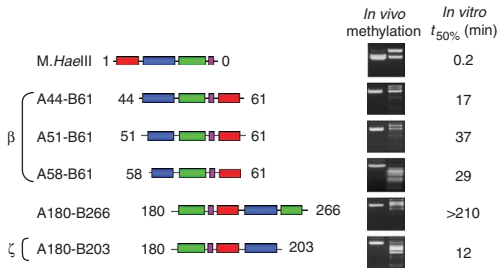

Figure 3 *In vivo* and *in vitro* activity of circular permutants generated by truncation of cluster I (clones ending at residue B61) and cluster IV (clones starting at residue A180) intermediates. Cluster I–derived permutants resemble the β class, and cluster IV–derived permutant belongs to the ζ class. Variants are named and color-coded as in **Figures 1b** and **2b**.

likely to be fixated than were the truncated intermediates. The remaining clones corresponded to eight different C-terminally truncated variants and three different N-terminally truncated variants out of a potential repertoire of ~320 different truncated intermediates present in each of the naive libraries.

Mapping the location of the new start and stop codons onto the three-dimensional structure of M.*Hae*III[20] showed that they were roughly clustered around regions connecting domains or subdomains (**Fig. 2c**). The C-terminally truncated intermediates were clustered in three groups: cluster I included variants with a redundant C-terminal *S*-adenosylmethionine (SAM)-binding subdomain (about half of the Rossmann fold), cluster II included variants with an almost completely redundant Rossmann fold and cluster III included variants with a redundant Rossmann fold and approximately half of the TRD. The activities of the N-terminally truncated selected clones were significantly lower, and their diversity was much narrower; they contained either a complete (cluster IV) or partially (cluster V) redundant N-terminal TRD.

Overall, most of the intermediates isolated corresponded to the putative intermediates that might have led to the natural methyltransferase classes related to m^5C by circular permutation (**Fig. 1c**). For example, intermediates from clusters I and II would yield β- and ζ-class enzymes, respectively, upon N-terminal truncation, and intermediates from cluster IV would yield ζ-class enzymes upon C-terminal truncation. We isolated no intermediates that would lead to γ-class enzymes but rather intermediates with divided TRD (clusters III and V). As thus far, no DNA methyltransferases have been identified with divided TRD, we named this potential new class η.

The truncated intermediates lead to circular permutants

We next investigated whether the selected N- and C-terminally truncated intermediates could indeed lead to circular permutants. We randomly truncated the selected N-terminally truncated intermediates at their C termini and randomly truncated the selected C-terminally truncated intermediates at their N termini, generating six different libraries of potential circular permutants. In particular, we created N-terminally truncated libraries based on the four cluster I variants and C-terminally truncated libraries based on the two cluster IV variants. The cluster I–derived variants selected after four rounds (**Supplementary Fig. 4** online) were β-class circular permutants starting at residues 44, 51 or 58 of the first copy of the fused M.*Hae*III genes and ending at residue 61 of the second copy (**Fig. 3**). Active variants derived from cluster IV intermediates initially yielded clones that had lost only few C-terminal residues. We therefore created new C-terminally truncated libraries while introducing a deleterious mutation (A317T) at the C termini of these intermediates. These libraries yielded two active variants (**Fig. 3**): A180-B203 resembled the ζ-class enzymes, whereas A180-B266 contained a longer duplicated region (about half of the TRD) and could still be regarded as an intermediate. As a control, we confirmed that removing a fragment, delimited by *Nde*I sites, from the coding regions of all selected circular permutants abolished their methyltransferase activity (**Supplementary Fig. 5**).

Point mutations can favor one topology over others

Although all of the truncated intermediates and circular permutants described above were sufficiently active to allow the survival of their encoding plasmids, many of them were substantially inferior to wild-type M.*Hae*III (**Fig. 3**) and often reverted to the wild-type topology. In nature, point mutations that compensate for the changes in topology are likely to occur concomitantly with, or after completion of, the gene rearrangements that lead to the permutations. We explored this scenario by randomly mutating the N-terminally truncated intermediate A180-B330, one of the least active intermediates isolated (**Fig. 2**), and the circular permutant A44-B61 and selecting for M.*Hae*III activity. In both cases, clones isolated after four rounds fully protected their encoding plasmids *in vivo* and showed markedly enhanced *in vitro* activities compared with their respective starting points (**Fig. 4**). The selected clones derived from intermediate A180-B330 carried diverse mutations, mostly in their redundant N-terminal TRD and motif X helix (fragment A; **Fig. 4a**). Only one clone carried an additional mutation, F181L, in the loop before the C-terminal TRD. The exact role of these mutations remains to be

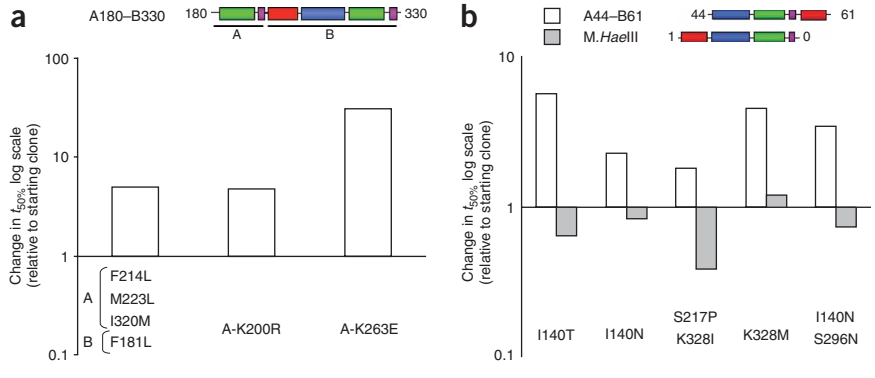

Figure 4 Point mutations can favor one topology over others. (**a**) Substitutions can improve the activity of a truncated intermediate. Intermediate A180-B330 was subjected to random mutagenesis and selection for higher methyltransferase activity. *In vitro* methyltransferase activity of the selected mutants is shown as $1/t_{50\%}$ relative to the activity of A180-B330. 'A' corresponds to mutations in N-terminal gene fragment A; 'B' refers to mutations in C-terminal gene fragment B. (**b**) Substitutions selected for an increase in the methyltransferase activity of the β-class permutant A44-B61 are largely neutral in wild-type M.*Hae*III. Permutant A44-B61 was subjected to random mutagenesis and selection for higher methyltransferase activity. *In vitro* methyltransferase activity of the selected mutants is shown as $1/t_{50\%}$ relative to the activity of A44-B61. *In vitro* methyltransferase activity of M.*Hae*III variants with wild-type topology carrying the same mutations is shown as $1/t_{50\%}$ relative to the activity of wild-type M.*Hae*III.

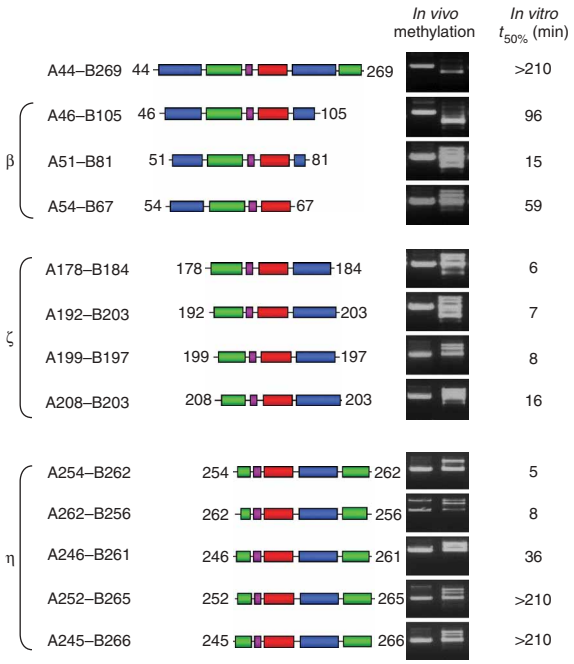

Figure 5 Directly selected circular permutants. The fused dimer of M.*Hae*III was truncated simultaneously at both the N and C termini and subjected to selection. The active variants are named and color-coded as in **Figures 1b** and **2b**. *In vivo* and *in vitro* activity of these variants is also shown.

determined, although some of them may encode a revertant by having a deleterious effect on the N-terminal TRD.

Selection of libraries derived from circular permutant A44-B61 yielded clones in which four residues were repeatedly replaced: Ile140 by the more polar threonine or asparagine; Ser217 by proline; Ser296 by asparagine; and Lys328, which was close to the linker between the original N- and C-termini, by the more hydrophobic methionine or isoleucine. Notably, the same mutations in wild-type M.*Hae*III were neutral or slightly detrimental (**Fig. 4b**). Thus, certain point mutations, such as I140N, could initially accumulate as 'neutral' drift with no apparent effects on the wild-type gene[21] and facilitate the divergence of a new topology at a later stage. Other mutations, such as S217P, may selectively direct the process by favoring the new topology while disfavoring the wild-type topology.

Protein modularity defines the allowed topologies

The N and C termini of the selected variants, which coincide with those of natural permutants and split methyltransferase variants (consisting of two separate polypeptide chains) found in nature or created in the laboratory[19], are clustered, defining four modules with both the Rossman fold and the TRD roughly divided in halves (**Figs. 2** and **3**). This suggests that structural and/or functional constraints restrict the location of the N and C termini of the truncated evolutionary intermediates. If these restrictions do not apply to the end products of the circular permutation process, there may be other classes or topologies of DNA methyltransferases not seen in nature or in our laboratory evolution experiments that proceeded through truncated intermediates. In other words, the fate of this process could be determined by the functionality of the intermediates and not of the end products. In fact, in artificial circular permutations created by directly linking the 5′ and 3′ ends of a gene and opening it at another location, N and C termini were largely scattered along the entire length of the polypeptide chain[5,22,23]. To explore this possibility, we selected a library of all possible circular permutants of M.*Hae*III without having gone through intermediates. A doubly incrementally truncated library, containing random start and stop codons in the first and second copies of the fused genes encoding M.*Hae*III, respectively, was therefore selected (**Supplementary Fig. 6** online).

The isolated clones from the doubly truncated library (**Fig. 5**) resembled the permutants derived from the C-terminally or N-terminally truncated intermediates (β or ζ classes, respectively) or the circular permutants of the putative η class (that is, intermediates belonging to cluster III and V (**Fig. 2c**) are indeed the putative intermediates that would lead to class η circular permutants). Thus, selection either through intermediates or directly for permutants yielded similar results. This suggests that, in the case of DNA methyltransferases, the same structural constraints apply to both the intermediates and the end products. It remains unknown precisely why permutants belonging to other topologies known among natural DNA methyltransferases were not obtained. In the case of the γ class, introducing new N and C termini in the region between conserved motifs IX and X in M.*Hae*III may be unfavorable, as they would be located very close to the active site (the side chain of a putative C terminus, Asn306, was only 4.5 Å from the most important active site residue, Cys71). It is likely that the γ class or other methyltransferase topologies not obtained here (such as DRM2; ref. 9) could be obtained through permutation by duplication of genes from classes other than m^5C.

The plausibility of the permutation by duplication model has been validly questioned regarding the foldability and functionality of truncated intermediates that are expected to have exposed hydrophobic surfaces[9]. However, these intermediates would fold properly if the modules that make up the methyltransferase fold were to behave, to some extent, as independent folding units with minimal exposed hydrophobic surfaces. This seemed to be the case in our experiments, as the N and C termini of our laboratory-evolved variants were clustered. To further support this hypothesis, we calculated the nonpolar accessible surface area (A_{NP}) of a series of hypothetical proteins generated by removing one residue at a time from the N or C termini of the three-dimensional structure of M.*Hae*III. We compared these calculated surface areas to the expected values based on the molecular weight (MW) of each hypothetical protein following the power law A_{NP} (Å^2) α MW$^{0.73}$, which generally holds true for all globular proteins[24]. The C-terminal truncations indeed defined two

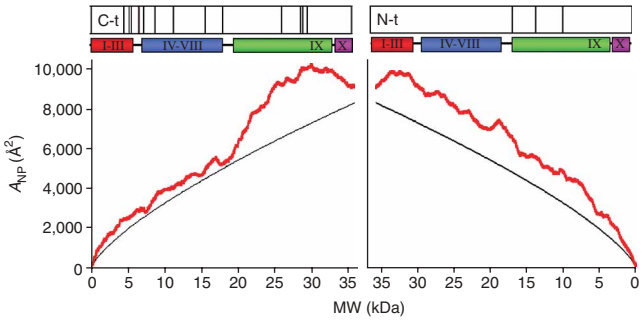

Figure 6 Dependence of A_{NP} on chain length of M.*Hae*III. A_{NP} values (red lines) were calculated for hypothetical M.*Hae*III variants of diminishing lengths (and, consequently, of diminishing MW) as residues were incrementally removed from the C terminus (left) or N terminus (right). Black lines indicate calculated values based on the power law $A_{NP} = f_{NP}6.3*MW^{0.73}$. Locations of the C (C-t) and N (N-t) termini of selected truncated intermediates are shown above.

Figure 7 Identification of η-class members in natural genomes. M2.*Bsa*I, M.*Eco*31IC and M.*Esp*3I are homologous to m^{5}C M.*Bsp*LU11. As observed in laboratory-evolved η-class variants (**Fig. 5**), however, their TRDs are divided between the N and C termini. (**a**) For clarity, alignment between the m^{5}C DNA methyltransferases M.*Hae*III (of known three-dimensional structure) and M.*Bsp*LU11 unambiguously defines the boundaries of the M.*Bsp*LU11 TRD. The region in the M.*Hae*III TRD where the N and C termini of η-class permutants are located is highlighted in yellow. (**b**) Multiple alignment of M2.*Bsa*I, M.*Eco*31IC, M.*Esp*3I and a η-class permuted version of M.*Bsp*LU11 (M.*Bsp*LU11*)

modules of growing size (**Fig. 6**) that roughly corresponded to the N-terminal SAM-binding subdomain (residues ~1–65) and the Rossmann fold (residues ~1–170), as well as the full-length protein (residues 1–330), with exposed hydrophobicities typical of well-folded proteins. We could not identify in this way the truncation that divided the TRD into two regions, although this probably reflected the fact that the TRD has a large DNA-interaction surface and therefore does not follow the power law even when intact. The N-terminal truncation did not predict modules with acceptable levels of exposed hydrophobicity, coinciding with the N-terminally truncated intermediates being scarce and having low activity. These results suggest that the modules that make up the methyltransferase fold might behave, to some extent, as independent folding units with minimal exposed hydrophobic surfaces.

Identification of natural η-class DNA methyltransferases

There exists a clear overlap between most of the intermediates and circular permutants evolved in the laboratory and the classes identified in nature, with the exception that no topology has been thus far identified among natural DNA methyltransferases in which the TRD is halved[25]. Our experiments indicate that this topology is functional, as observed in intermediates from clusters III and V and in permutants of the η class, some of which were among the most active selected circular permutants. This encouraged us to look for the η class in natural genomes using the laboratory-evolved η-class genes as bait for homology searches. In this way, we identified *Bacillus stearothermophilus* M2.*Bsa*I, *E. coli* M.*Eco*31IC and *Hafnia alvei* M.*Esp*3I as members of the η class (**Fig. 7**). These methyltransferases are homologous to the m^{5}C methyltransferase M.*Bsp*LU11 (which, in turn, is homologous to the m^{5}C M.*Hae*III), but their TRDs are divided between the N and the C termini, as observed in the laboratory-evolved η-class permutants. In particular, the first ~50 residues of the M.*Bsp*LU11 TRD, identified by their homology to the M.*Hae*III TRD (whose structure is known), are at the C termini of the η-class M2.*Bsa*I, M.*Eco*31IC and M.*Esp*3I. In contrast, the remaining ~100–130 residues of the TRD, depending on the particular enzyme, are located at the N termini.

DISCUSSION

Our results show that new protein topologies can evolve gradually through multistep gene rearrangements such as the permutation-by-duplication mechanism. We observed that intrinsic protein modularity defines the boundaries of allowed topological space, and point mutations direct the process toward a particular topology. The nature of the intermediates of this process is revealed here; these can be functional and thereby serve as evolutionary nodes, in this case connecting different classes of DNA methyltransferases. Our results also indicate that, although viable, the steps that follow duplication and fusion might be accompanied with a substantial decrease in activity. Indeed, under the selection pressure applied in our experiments, revertants were ~30 times more likely to be fixated than were truncated intermediates. During periods of relaxed selection pressure, however, these less-active intermediate forms might be

neutral. Additional point mutations and subsequent positive selection might yield more active forms, as shown here. Alternatively, changes in target specificity—namely the emergence of a new function—could drive the evolution of natural DNA methyltransferase permutants[26].

We also found that the inherent modularity of the methyltransferase fold dictates which intermediates will fold, possess function and eventually yield a new topology. An issue for future analysis is the possibility that, in some intermediates, duplicated modules could already be swapped at the protein level, even before permutation has been completed at the gene level. It is also possible that the existence of putative autonomous modules exposed upon C-terminal, but not N-terminal, truncation could result from the adaptation of the protein to cotranslational (vectorial) folding, in which N-terminal modules, such as the SAM-binding subdomain, fold first as they emerge from the ribosome. This modular adaptation, in turn, would limit the allowed topologies of the truncated intermediates. Thus, after gene duplication and fusion, the preferable route may involve the introduction of a stop codon first and a new translation initiation site at a second stage. In fact, stop codons are more likely to appear, considering that certain nucleotide substitutions and most indels would result in a premature stop codon.

Evolution often finds different solutions to a single problem, so it is possible that mechanisms other than permutation by duplication have also contributed to the diversity observed among DNA methyltransferases[9]. Regardless of the particular mechanism, however, the modularity of the methyltransferase fold is likely to be central to their feasibility. Moreover, as shown here, the topological space afforded by the fold of M.*Hae*III, and indeed by any other m^5C methyltransferase, seems to have been exhausted by natural evolution. Thus, none of our selected intermediates had N and C termini that did not correspond to known DNA methyltransferases, and the η class first identified in our laboratory evolution experiments turned out to be represented among natural DNA methyltransferases.

METHODS

Detailed experimental protocols are provided in **Supplementary Methods** online.

Creation of a fused dimer of M.*Hae*III. A tandem M.*Hae*III enzyme dimer was created by PCR amplification of the gene and joining of two identical copies through a GNASG linker. The dimer and all other M.*Hae*III variants described here were ligated to pIVEX 2.2 (Roche) and transformed into *E. coli* strain ER2267 (*EcoKr*$^-$*m*$^-$*McrA*$^-$*McrBC*$^-$*Mrr*$^-$) in which GGCC methylation (the methylated base is underlined) is not lethal.

Estimation of protein expression levels. The expression levels of M.*Hae*III variants were estimated by measuring the expression of the reporter gene encoding *Pseudomonas diminuta* phosphotriesterase, which was evolved in our laboratory[27] for optimum *E. coli* expression and was therefore likely to provide an upper limit for our estimation. The reporter gene was subcloned into the same vector in the same bacterial strain under the same growth conditions.

Creation of N- or C-terminally incrementally truncated libraries. The N- and C-terminally incrementally truncated libraries were created using the ITCHY methodology[19,28] with minor modifications (**Supplementary Fig. 2**). Both libraries had identical sequences between the ribosomal binding site and the start codon to minimize differences in expression resulting from unequal translation efficiencies. DNA sequencing of individual clones from the unselected libraries indicated that the start and stop codons were randomly introduced all along the tandem-arranged genes. The doubly incrementally truncated library was created by two consecutive ITCHYs with no selection between them. To minimize the presence of clones corresponding to the m^5C class of M.*Hae*III, the library was built on a tandem gene in which the second copy carried the inactivating A317T mutation.

Selection and characterization of active clones. Libraries cloned in pIVEX 2.2 were transformed, and the resulting bacteria were grown on agar plates at 30 °C (A44-B61–derived libraries were selected after growth at 37 °C). Plasmid DNA was extracted and digested with 20 units of *Hae*III for 6 h at 37 °C. The surviving plasmid DNA was purified and transformed back into *E. coli* ER2267, thus enriching the gene pool with genes encoding M.*Hae*III activity. Several rounds of selection were typically carried out, after which active clones were individually isolated and sequenced. *In vitro* compartmentalization selection in water-in-oil emulsions[29] is ideally suited for the selection of DNA methyltransferases, but it involves PCR amplification of the surviving genes after each round. Unfortunately, we found that the truncated intermediate genes, which possessed duplicated DNA regions, could not be PCR-amplified without recombination artifacts. This precluded the use of *in vitro* compartmentalization in this work.

In vivo M.*Hae*III activity levels were determined by the resistance of plasmid DNA isolated from these clones to digestion with *Hae*III. All active variants were subjected to the following controls. First, plasmid DNA was retransformed, and three randomly chosen individual colonies were retested to confirm their activity *in vivo*. Second, plasmids were digested with *Nde*I and religated, thus removing a fragment from the coding region of the fused gene encoding M.*Hae*III and abolishing their methyltransferase activity. The truncated plasmids were transformed back into *E. coli* ER2267, and the disappearance of M.*Hae*III activity was confirmed (**Supplementary Figs. 3, 5** and **7** online). The activity of the selected clones was also tested *in vitro*: plasmid DNA was transcribed and translated in S30 *E. coli* extract (Eco Pro T7; Novagen), and the resulting methyltransferase activity was assayed with a linear DNA substrate using a digoxigenin-biotin ELISA at 27 °C (ref. 29). The *in vitro* and *in vivo* activities correlated quite well, although some variants (particularly those with C termini beyond B263) behaved differently in living bacteria than in cell-free extracts. For example, the *in vitro* activity of variant A180-B266 was unexpectedly low ($t_{50\%} > 210$ min) despite its *in vivo* activity being comparable to that of other circular permutants.

Calculation of A_{NP}. A series of individual coordinate files was created from the coordinates of M.*Hae*III (1dct) by incrementally removing the atomic coordinates corresponding to one N- or C-terminal residue after another. The A_{NP} for this series of truncated proteins was calculated with the software GETAREA 1.1 (ref. 30). The theoretical dependence of A_{NP} on the MW was calculated using the formula $A_{NP} = f_{NP}6.3{*}MW^{0.73}$, where A_{NP} is in Å^2 and f_{NP} is the nonpolar fraction of the total area of the full-length protein (modified from ref. 24).

Random and site-directed mutagenesis. Libraries were created by error-prone PCR using DNA encoding clones A180-B330 or A44-B61 as a template and varying the Mn^{2+} concentrations and CA/TG bias, or using the base analogs 8-oxo-2′-deoxyadenosine and 2′-deoxy-*P*-nucleoside-5′-triphosphate, resulting in two to seven nucleotide substitutions per gene. The libraries were recloned in pIVEX 2.2, grown on agar plates at 37 °C and subjected to selection as described above. Mutations I140T, I140N, S217P, S296N, K328M or K328I were introduced in wild-type M.*Hae*III by PCR. The resultant genes were fully sequenced to confirm the corresponding mutation and to verify that no other, undesired mutations were generated.

Accession codes. SwissProt: M2.*Bsa*I, Q6SPF5; M.*Eco*31IC, Q8RNY6; M.*Esp*3I, Q8RNY3.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank the Minerva Foundation and the Israel Science Foundation for financial support, M. Babor for the software that generated the truncated PDB files and A. Levy for helpful comments on the manuscript. S.G.P. is the recipient of a Dewey D. Stone Postdoctoral Fellowship, L.R. is the recipient of a Feinberg Graduate School Fellowship and D.S.T. is the incumbent of the Elaine Blond Career Development Chair.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Ponting, C.P. & Russell, R.B. Swaposins: circular permutations within genes encoding saposin homologues. *Trends Biochem. Sci.* **20**, 179–180 (1995).
2. Grishin, N.V. Fold change in evolution of protein structures. *J. Struct. Biol.* **134**, 167–185 (2001).
3. Koonin, E.V., Wolf, Y.I. & Karev, G.P. The structure of the protein universe and genome evolution. *Nature* **420**, 218–223 (2002).
4. Aravind, L., Mazumder, R., Vasudevan, S. & Koonin, E.V. Trends in protein evolution inferred from sequence and structure analysis. *Curr. Opin. Struct. Biol.* **12**, 392–399 (2002).
5. Graf, R. & Schachman, H.K. Random circular permutation of genes and expressed polypeptide chains: application of the method to the catalytic chains of aspartate transcarbamoylase. *Proc. Natl. Acad. Sci. USA* **93**, 11591–11596 (1996).
6. Cunningham, B.A., Hemperly, J.J., Hopp, T.P. & Edelman, G.E. Favin versus concanavalin A: circularly permuted amino acid sequences. *Proc. Natl. Acad. Sci. USA* **76**, 3218–3222 (1979).
7. Luger, K., Hommel, U., Herold, M., Hofsteenge, J. & Kirschner, K. Correct folding of circularly permuted variants of a beta alpha barrel enzyme *in vivo*. *Science* **243**, 206–210 (1989).
8. Jeltsch, A. Circular permutations in the molecular evolution of DNA methyltransferases. *J. Mol. Evol.* **49**, 161–164 (1999).
9. Bujnicki, J.M. Sequence permutations in the molecular evolution of DNA methyltransferases. *BMC Evol. Biol.* **2**, 3 (2002).
10. Richardson, J.S. The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* **34**, 167–339 (1981).
11. Martin, J.L. & McMillan, F.M. SAM (dependent) I AM: the S-adenosylmethionine-dependent methyltransferase fold. *Curr. Opin. Struct. Biol.* **12**, 783–793 (2002).
12. Malone, T., Blumenthal, R.M. & Cheng, X. Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyltransferases, and suggests a catalytic mechanism for these enzymes. *J. Mol. Biol.* **253**, 618–632 (1995).
13. Klimasauskas, S. *et al.* Sequence motifs characteristic of DNA[cytosine-N4]methyltransferases: similarity to adenine and cytosine-C5 DNA-methylases. *Nucleic Acids Res.* **17**, 9823–9832 (1989).
14. Kita, K., Kotani, H., Sugisaki, H. & Takanami, M. The fokI restriction-modification system. I. Organization and nucleotide sequences of the restriction and modification genes. *J. Biol. Chem.* **264**, 5751–5756 (1989).
15. Sugisaki, H., Kita, K. & Takanami, M. The FokI restriction-modification system. II. Presence of two domains in FokI methylase responsible for modification of different DNA strands. *J. Biol. Chem.* **264**, 5757–5761 (1989).
16. Leismann, O., Roth, M., Friedrich, T., Wende, W. & Jeltsch, A. The Flavobacterium okeanokoites adenine-N6-specific DNA-methyltransferase M.FokI is a tandem enzyme of two independent domains with very different kinetic properties. *Eur. J. Biochem.* **251**, 899–906 (1998).
17. Szomolanyi, E., Kiss, A. & Venetianer, P. Cloning the modification methylase gene of Bacillus sphaericus R in *Escherichia coli*. *Gene* **10**, 219–225 (1980).
18. Ostermeier, M., Shim, J.H. & Benkovic, S.J. A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotechnol.* **17**, 1205–1209 (1999).
19. Choe, W., Chandrasegaran, S. & Ostermeier, M. Protein fragment complementation in M.HhaI DNA methyltransferase. *Biochem. Biophys. Res. Commun.* **334**, 1233–1240 (2005).
20. Reinisch, K.M., Chen, L., Verdine, G.L. & Lipscomb, W.N. The crystal structure of HaeIII methyltransferase convalently complexed to DNA: an extrahelical cytosine and rearranged base pairing. *Cell* **82**, 143–153 (1995).
21. Wagner, A. Robustness, evolvability, and neutrality. *FEBS Lett.* **579**, 1772–1778 (2005).
22. Hennecke, J., Sebbel, P. & Glockshuber, R. Random circular permutation of DsbA reveals segments that are essential for protein folding and stability. *J. Mol. Biol.* **286**, 1197–1215 (1999).
23. Iwakura, M., Nakamura, T., Yamane, C. & Maki, K. Systematic circular permutation of an entire protein reveals essential folding elements. *Nat. Struct. Biol.* **7**, 580–585 (2000).
24. Chothia, C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105**, 1–12 (1976).
25. Jeltsch, A. Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. *ChemBioChem* **3**, 274–293 (2002).
26. Kusano, K., Naito, T., Handa, N. & Kobayashi, I. Restriction-modification systems as genomic parasites in competition for specific sequences. *Proc. Natl. Acad. Sci. USA* **92**, 11095–11099 (1995).
27. Roodveldt, C. & Tawfik, D.S. Directed evolution of phosphotriesterase from *Pseudomonas diminuta* for heterologous expression in *Escherichia coli* results in stabilization of the metal-free state. *Protein Eng. Des. Sel.* (2005).
28. Ostermeier, M., Nixon, A.E., Shim, J.H. & Benkovic, S.J. Combinatorial protein engineering by incremental truncation. *Proc. Natl. Acad. Sci. USA* **96**, 3562–3567 (1999).
29. Tawfik, D.S. & Griffiths, A.D. Man-made cell-like compartments for molecular evolution. *Nat. Biotechnol.* **16**, 652–656 (1998).
30. Fraczkiewicz, R. & Braun, W. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comp. Chem.* **19**, 319–333 (1998).