Evolution of Circular Permutations in Multidomain Proteins

January Weiner 3rd and Erich Bornberg-Bauer

Division of Bioinformatics, School of Biological Sciences, University of Münster, Schlossplatz 4, Münster, Germany

Modular rearrangements play an important role in protein evolution. Functional modules, often tantamount to structural domains or smaller fragments, are in many cases well conserved but reoccur in a different order and across many protein families. The underlying genetic mechanisms are gene duplication, fusion, and loss of sequence fragments. As a consequence, the sequential order of domains can be inverted, leading to what is known as circularly permutated proteins. Using a recently developed algorithm, we have identified a large number of such rearrangements and analyzed their evolutionary history. We searched for examples which have arisen by one of the three postulated mechanisms: independent fusion/fission, "duplication/deletion," and plasmid-mediated "cut and paste." We conclude that all three mechanisms can be observed, with the independent fusion/fission being the most frequent. This can be partly attributed to highly mobile domains. Duplication/deletion has been found in modular proteins such as peptide synthases.

Introduction

Proteins are known to evolve not only at the level of amino acid sequence but also in a modular way, by rearrangements of larger sequence fragments. Domains are structural units and frequently determine the function of proteins (see, e.g., the Pfam database; Bateman et al. 2002) and correspond to conserved sequence units (Elofsson and Sonnhammer 1999). Therefore, large-scale and modular rearrangements of protein structures can be found by sequence analysis of the rearrangements of domains defined by the conservation of the protein sequence. Most domains tend to be very tightly associated with other specific domains and always occur in the same combination (Apic, Gough, and Teichmann 2001; Vogel et al. 2004). However, some domains associate with a diverse set of other domains, and yet others occur mostly as singletons (for a recent review, see Bornberg-Bauer et al. 2005). Either way, the basic mechanisms of modular evolution are gene duplication, gene fusion, and domain loss. The last may happen either by introduction of new start or stop codons or by slow erosion at the sequence level (Bornberg-Bauer et al. 2005). Basic types of rearrangements can happen sequentially, causing more complex rearrangements such as circular permutations (CPs) (Bateman et al. 2002).

CPs are rearrangements where the sequential order between two proteins is inverted (see fig. 1; Jeltsch 1999; Bujnicki 2002; Weiner, Thomas, and Bornberg-Bauer 2005). It is important to know whether proteins undergoing a CP event can retain their functionality if their domains remain intact. Such knowledge will improve our understanding on the structure-function relationship and will give specific clues on the functionality of domains.

Artificially created CPs were shown to maintain catalytic activity of proteins in many cases, or at least the catalytic center was not destroyed (Hennecke, Sebbel, and Glockshuber 1999; Cheltsov, Barber, and Ferreira 2001). Consequently, CPs show that a rearrangement, detected at the sequence level, may have little or no impact on protein structure and function. It can be argued that such a situation is possible in two cases. First, the rearrangement does not influence the structure of the folded protein or, alternatively, if the domains are structurally independent of each other, they can fold and fulfill their biological function in spite of a change in the spatial arrangement.

From an evolutionary point of view, a thorough understanding of the mechanisms leading to CPs also enables a more quantitative estimate of how often and in which epochs the elementary operations leading to CPs occur. This in turn will show how important such rearrangements are in increasing the genetic variability upon which natural selection can act.

The existence of CPs also has important implications for database searches. All widely used algorithms are built on the assumption that the major operations during protein evolution are substitutions, insertions, and deletions of amino acids. In essence, this leaves the order of the homologous fragments constant. If the order of fragments is inverted, as is the case in CPs, these algorithms will only identify a short homologous subfragment, and, consequently, proteins with a close evolutionary relationship may thus be missed.

Proposed Mechanisms of CPs

Three main mechanisms are thought to be responsible for circularly permutated proteins (Jeltsch 1999; Bujnicki 2002; Weiner, Thomas, and Bornberg-Bauer 2005). The mechanism we call duplication/deletion relies on independent events of gene duplication and partial deletion of domains at the termini (fig. 1*a*). Given a protein with the domain sequence ABC, it can first undergo duplication, so that the resulting gene has the domain structure ABCABC. Next, domains at one of the termini, for example the N-terminal domain A, are deleted, and the intermediary form BCABC arises. Finally, the deletion of C-terminal domains BC results in the CP of the initial protein with the structure BCA. If a CP is caused by this mechanism, we expect the existence of intermediate forms of CPs, designated as intermediate CPs (iCPs; Weiner, Thomas, and Bornberg-Bauer 2005).

In an "independent fusion," two multidomain proteins are created independently in different organisms or different genomic locations from single-domain fragments or proteins with fewer domains (fig. 1*b*). For example, two independent genes consisting, respectively, of domains AB

Key words: circular permutation, domains, modular evolution, rearrangement.

E-mail: january@uni-muenster.de.

Mol. Biol. Evol. 23(4):734–743. 2006
doi:10.1093/molbev/msj091
Advance Access publication January 23, 2006

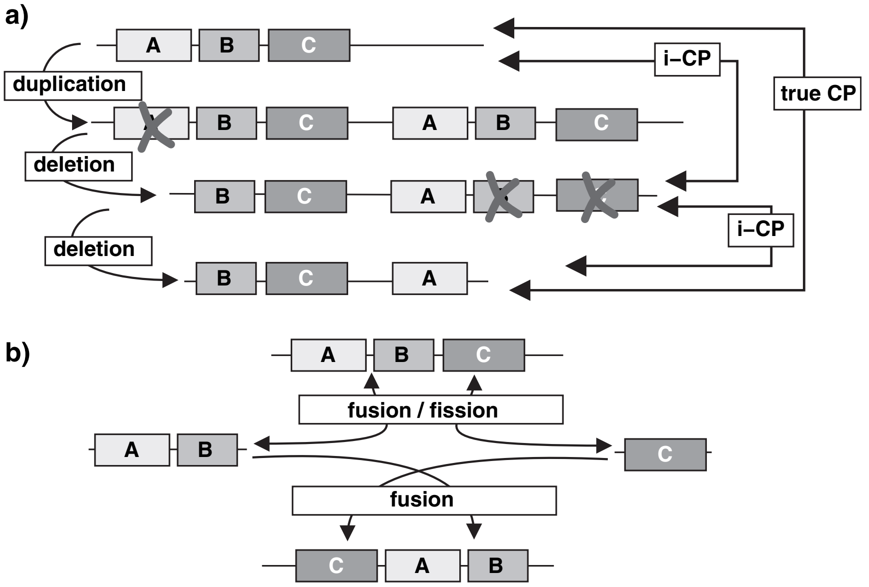

FIG. 1.—Mechanisms of CPs: (*a*) duplication/deletion and (*b*) independent fusion. A, B, and C denote protein domains. X: domain deletion; i-CP: intermediate CP, a result of the duplication/deletion mechanism.

and C can form, through fusion, two new genes: ABC or CAB. These genes will be, in consequence, related by a CP. Alternatively, in a second variant of this mechanism, one protein can undergo fission, resulting in two distinct genes corresponding to the N- and C-terminal fragments of the parent protein. Subsequently, these proteins can be reassembled in a different order. For example, given the initial protein with the domain structure ABC, it can undergo a fission to form genes AB and C, which are then fused again to form the circularly permutated variant of the initial gene with the domain structure CAB.

This is in concordance with the view that a circularly permutated protein will have unchanged biological function if the rearranged domains can fold and function independent of each other. This condition is obviously fulfilled if the terminal fragments can act as single proteins.

We argue that the absence of "partial" proteins corresponding to both the N- and the C-terminal parts is evidence against the fusion/fission model, as is the existence of iCPs.

On the other hand, if the partial proteins correspond to a promiscuous domain, then a fusion event seems likely, and the model is further corroborated.

The two mechanisms have different predictions concerning the phylogeny of the domain sequences (fig. 2). In the case of the two variants of the independent fusion mechanism, the phylogeny will be straightforward and relatively easy to reconstruct because the phylogeny is not obscured by the existence of iCPs. If, on the other hand, the rearrangements are a result of alternating duplication and deletion events, the phylogeny reconstruction based on the amino acid sequences of the whole proteins will not necessarily reflect the actual phylogeny of the proteins (fig. 2).

A third mechanism has been termed as cut and paste by Bujnicki (2002). Methyltransferases (MTs) are found to be components of the restriction-modification (RM) system. The invasion of a cell by the RM system encoded on a plasmid may lead to a fragmentation of the MT gene by the host endonucleases. This is followed by a reassembly of the gene and changes the order of the MT domains. The MTs are the sole example in which an underlying genetic mechanism of a CP has been proposed in detail and demonstrated on specific examples (Bujnicki 1999, 2002). This mechanism can be viewed as one of the two mechanisms proposed before, in which all stages happen at one time. However, the underlying genetics requires a specific setup. An RM system must be present, and the rearranged genes have to be on a DNA strand susceptible to restriction, for example, encoded on a plasmid.

In this paper, we investigate in full detail the results of a domain-based search for CPs in respect to evolutionary issues. We seek to shed light on which evolutionary process generating CPs is more frequent. We ask how the mechanisms work in detail and, finally, when, in evolutionary terms, most of these events took place. In the following, we present examples for CPs which have arisen by alternative fusion, followed by those that arose by fusion and domain loss, and finally briefly recapitulate those having arisen by other mechanisms.

Methods

The whole ProDom database version 1/2004 (Corpet et al. 2000) was scanned, and the results were confirmed using the Pfam database (Bateman et al. 2002) version 18.0. These databases contain all proteins from the SwissProt/TrEMBL databases automatically annotated in terms of domains. The Pfam database is in part manually annotated and has generally a much higher quality. However, ProDom has a higher coverage of the annotated sequences in terms of amino acid percentage and shorter domains (fig. 3*E* and *F*). Both databases contain approximately the same number of sequences with two or more domains.

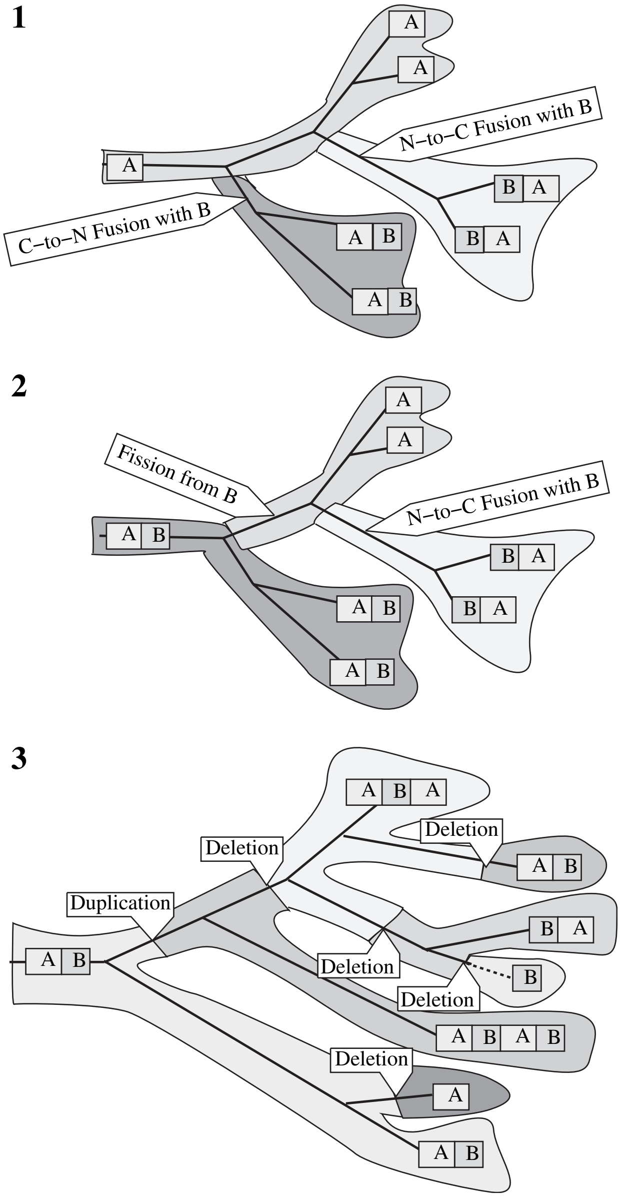

FIG. 2.—Phylogeny as consequence of the different models of CPs. The figure shows a scheme of the phylogeny obtained from only one domain, A. This domain rearranges with a second domain, denoted B. In the independent fusion model (1) and fusion/fission model (2), unless fusion events are very frequent, a simple phylogenetic tree is to be expected. The branches A, AB, and BA can be clearly separated from the initial partial genes. In the case of duplication/deletion model (3), a single duplication event combined with several deletions can lead to a more complicated phylogeny. Dotted line shows the deletion of domain A.

RASPODOM (Weiner, Thomas, and Bornberg-Bauer 2005) was used to retrieve CPs. This algorithm relies on domain annotations of sequences and can also be used to find iCPs (intermediate forms of the duplication/deletion mechanism).

The identified CPs were clustered to estimate the actual number of CP events. For example, if there are four proteins P1, P2, P3, and P4 is circularly permutated, then RASPODOM will report three cases of CPs (P1-P4, P2-P4, P3-P4). These CPs constitute one cluster of CPs.

The estimation of the actual number of circularly permutated proteins in a given cluster proceeds as follows: Proteins from a cluster are divided into two sets. Given a protein from one set, there are only CPs between this protein and proteins from the other set. In the mentioned example, P1, P2, and P3 would form one set and P4 would remain in the other set. The estimated minimal number of circularly permutated proteins in a cluster is then the number of proteins in the smaller of the two sets.

Finally, all clusters have been manually scanned to identify false positives. For each cluster, both a graphical alignment of the domain structures of the corresponding proteins and dot plots with marked domains were produced (fig. 3). The full list of results, the corresponding cluster assignments, multiple alignments, and domain dot plots can be found in the Supplementary Material online.

Phylogenetic analyses were carried out with protein parsimony method using the protpars program from the PHYLIP package (Felsenstein 1989) or using the maximum likelihood method with the PHYML program (Guindon and Gascuel 2003). Bootstrapping was done with 100 repeats.

Results

Around 160 clusters of proteins containing "true" CPs were found in the ProDom and Pfam databases (table 1). Another 54 had clearly distinguishable iCPs but no full CPs (for a taxonomic classification of results, see table 2). There were 63,417 different proteins in the initial result set, which could be grouped in 971 clusters.

An automated assignment of the possible CP mechanism yielded 130 clusters, which do not contain any intermediate forms and for which single-domain proteins could be found. Furthermore, we found 30 clusters, which contain iCPs, and are thus attributed to the duplication/deletion mechanism (table 1).

All results, domain dot plots, and RASPODOM lattices are available in the Supplementary Material online. Here, we present examples in which the assignment to one of the mechanisms of CPs is well supported and illustrate the main conclusions of the paper.

CPs Caused by Independent Fusion of Subproteins

Many of the CPs found are proteins with few domains, which have been probably circularly permutated by means of an independent fusion of a gene with either another multidomain gene or, in some cases, a single mobile domain.

Among such CPs, there were two highly abundant groups of proteins. Eighteen clusters correspond to chitinases and cellulases and 14 clusters contained proteins with protein kinase activity. These proteins contained domains which are known to be highly mobile and are found in a large spectrum of proteins, for example, the protein kinase domain, the carbohydrate-binding domain of chitinases, or the cellulase domain.

Transhydrogenases

Transhydrogenases represent an omnipresent family of proteins. The enzyme has three major functional units: I, II (transmembrane region), and III (Hatefi and Yamaguchi

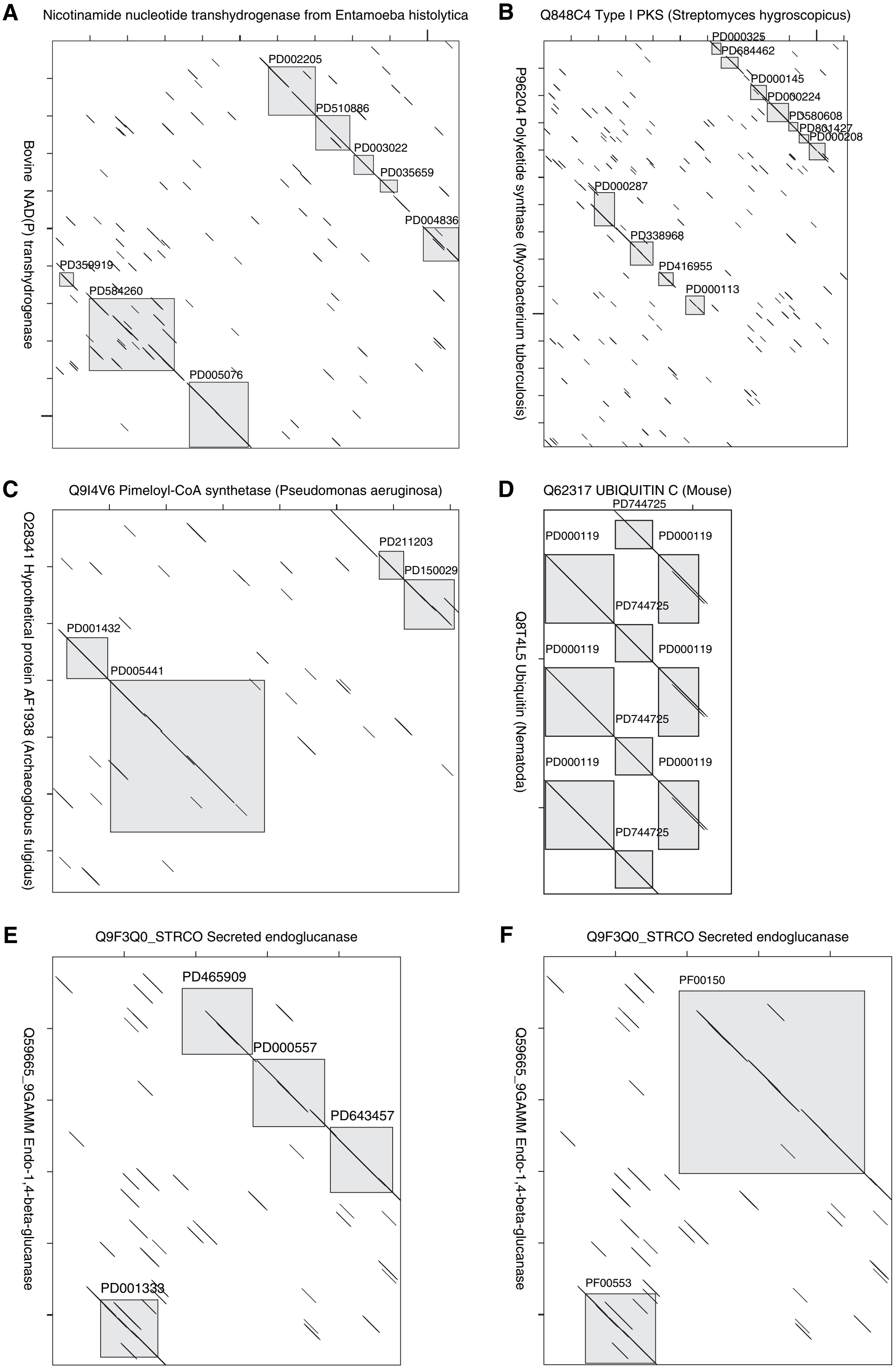

FIG. 3.—Domain and dot plot of (*A*) two transhydrogenases: vertebrate (bovine, NNTM_BOVIN) and protozoan (from *Entamoeba histolytica*, Q24813); (*B*) two bacterial type I polyketide synthases; (*C*) a bacterial protein containing the CoA-binding domain and a hypothetical archeal protein from *Archeoglobus fulgidus*; (*D*) two poly-Ub's from mouse and a nematode (*Acanthocheilonema viteae*); (*E*) two bacterial glucanases from *Cellvibrio japonicus* (Q59665) and *Streptomyces coelicolor* (Q9F3Q0); and (*F*) the same two proteins annotated by Pfam. Gray boxes indicate matching domains as annotated in ProDom or Pfam. Black diagonals show similarities in amino acid sequences. See text for further explanations.

Table 1
Results Statistics for the ProDom and Pfam Databases[a]

	CP[b]	iCP[c]	FP[d]	n.a.[e]	DD[f]	FF[g]
Both	98 (36,797)	17 (1,175)	76 (2,575)	93 (3,750)	24 (33,980)	74 (2,817)
Pfam only	47 (1,246)	13 (736)	454 (17,554)	467 (18,290)	5 (367)	42 (879)
ProDom only	15 (333)	24 (661)	227 (2,340)	251 (3,001)	1 (17)	14 (316)
Sum	160 (38,376)	54 (2,572)	757 (22,469)	811 (25,041)	30 (34,364)	130 (4,012)
			971 (63,417)			971 (63,417)

[a] Table cells contain the number of clusters (number of involved sequences).

[b] CP: clusters containing real CPs.

[c] iCP: clusters containing intermediary CP forms, but no complete CPs.

[d] n.a.: not attributed.

[e] FP: false positives, clusters not containing any CPs or iCPs.

[f] DD: clusters attributed to the deletion/duplication mechanism.

[g] FF: clusters attributed to the fusion/fission mechanism.

1996). In bacteria, these are split between two genes, α and β. The first contains units I and IIa (part of the transmembrane region) and the second units IIb (second part of the transmembrane region) and III. In higher eukaryotes, all units are encoded on one gene, in the order I-II-III, that is, α-β. However, in parasitic protozoans the order is reversed, and, although the units are contained within one single gene, the order is IIb-III-I-IIa or β-α. To test the hypothesis that the mechanism of the CP was an independent fusion of ancestral genes, we have analyzed the phylogenetic relationships independently for units α and β. If the protozoan genes arose through the duplication/deletion mechanism from ancestral α-β genes, then they should be more closely related to the genes from higher eukaryotes than to bacterial genes.

The phylogenetic analysis of the α and β subunits of the transhydrogenases from bacteria, protozoa, and higher eukaryotes showed that the three groups of domain arrangements in transhydrogenases are clearly separated (Supplementary Material online). This is consistent with the independent fusion/fission model: the α and β subunits of transhydrogenases were acquired and fused independently very early during the evolution of these groups of eukaryotes. This is further supported by the cellular location of the transhydrogenases and their function in both groups. Whereas in higher eukaryotes the transhydrogenases are located in mitochondria, in protozoans they are found in small organellae called mitosomes and present at much lower density (Weston, White, and Jackson 2001). Therefore, its role in connection with a proton gradient, as it is the case in other eukaryotes, is doubtful. This suggests that their function is different in protozoans.

Proteins with a Coenzyme A–Binding Domain

At least eight circularly permutated proteins from various bacterial and archeal organisms containing the coenzyme A (CoA)–binding domain (Pfam accession number: PF02629), a Rossmann fold, have been identified. All these eight proteins have been annotated as hypothetical. Interestingly, there are pairs of two circularly permutated proteins from the same genomes, as it is in the case of *Bradyrhizobium japonicum* or *Streptomyces avermilitis*.

Table 2
Summary of the Results

Protein Function	Organisms	Clusters[a]	CPs[b]	Total[c]
ABC transport system	Archaea (2.1%), Bacteria (36.3%), Eukaryota (60.3%), Viruses (1.3%)	9	7,750	28,290
Acetyltransferases	Archaea (21.2%), Bacteria (73.1%), Eukaryota (5.8%)	1	9	53
PTS transport system	Bacteria (100.0%)	3	277	598
Carboxylases	Archaea (1.8%), Bacteria (71.4%), Eukaryota (26.8%)	2	11	241
Chitinases/cellulases	Archaea (0.7%), Bacteria (49.1%), Eukaryota (47.8%), Viruses (2.4%)	11	193	940
Other hydrolases	Archaea (5.0%), Bacteria (75.5%), Eukaryota (19.5%)	12	47	405
Dehydrogenases/reductases/transhydrogenases	Archaea (10.2%), Bacteria (72.3%), Eukaryota (17.5%)	19	367	2,281
Kinases	Bacteria (18.4%), Eukaryota (81.6%)	11	30	138
MTs and RM specificity enzyme	Archaea (9.8%), Bacteria (79.3%), Eukaryota (5.8%), Viruses (4.7%), other sequences (0.4%)	5	51	282
Peptidases/proteases	Bacteria (58.2%), Eukaryota (41.8%)	6	47	275
Synthases	Archaea (3.1%), Bacteria (83.3%), Eukaryota (13.6%)	7	423	1,518
Transcription related/zinc finger	Bacteria (3.7%), Eukaryota (96.3%)	6	608	1,877
Other	Archaea (3.6%), Bacteria (35.1%), Eukaryota (61.3%)	46	162	1,334
Unknown	Archaea (2.4%), Bacteria (13.8%), Eukaryota (83.8%)	22	32	144
Total	Archaea (2.6%), Bacteria (41.2%), Eukaryota (55.1%), Viruses (1.1%)	160	10,010	38,376

[a] Clusters: number of clusters that were found to contain proteins of the given function.

[b] CPs: minimal estimate of the number of CP events.

[c] Total: total number of involved proteins.

These CoA-related proteins show a wide flexibility as to the domain arrangement; there were two distinct CPs identified at different positions in the modular arrangement of the proteins. All proteins have several common, yet unannotated, domains either at N- or C-terminus of the CoA domain. The proteins, in which the CoA-binding domain is at the N-terminus, can additionally have an acetyltransferase domain (AcTr) (PF00583) either at the N- or at the C-terminus. Therefore, there are four combinations of the domains: CoA-XYZ (where XYZ stands for unannotated domains), XYZ-CoA, AcTr-CoA-XYZ, and CoA-XYZ-AcTr (fig. 4).

There have been at least two independent CPs, as shown by the four different domain combinations. The duplicated instances of these proteins as well as iCPs—as predicted by the duplication/deletion mechanism—have not been found. On the other hand, the CoA-binding domain is very versatile and forms numerous combinations with other domains but can also occur as a singleton. This is in accordance with the independent fusion model, which is further supported by the phylogeny of the domains (fig. 4*A* and *B*).

Further Cases of the Independent Fusion/Fission Mechanism

Phosphocarrier Proteins (Bacteria). The phosphotransferase system (PTS) is a major carbohydrate transport system in bacteria (Postma, Lengeler, and Jacobson 1993). It consists of three main elements: the enzyme EI (phosphoenolopyruvate [PEP]-utilizing protein) transfers the phosphoryl group from PEP to the phosphocarrier protein HPR, which in turn transfers it to the permease unit. This unit consists of three distinct, sugar-specific subunits: EIIA, EIIB, and EIIC. We find many different combinations of the components of the PTS system (e.g., EIIA-HPR-EI or HPR-EI-EIIA; EIIA-EIIB-EIIC or EIIB-EIIC-EIIA). Because proteins which contain one domain only (e.g., the EIIA component or the EI enzyme) can also be found and no iCPs are present, we conclude that the mechanism for these CPs is independent fusion.

Adenosine triphosphate–Binding Casette. ABC transporters are a large family of proteins transporting molecules across membranes and consist of two copies of the ABC transporter domain and two copies of the ABC transmembrane domain (Schmitt and Tampe 2002); these may be on a single polypeptide chain. We find different arrangements between transporters from humans, *Drosophila melanogaster* and *Arabidopsis thaliana*. A CP is furthermore found between two ABC transporters from *A. thaliana*.

Glycosyl Hydrolases. This is a very widespread and diverse group of enzymes. Often these enzymes consist of two domains, N- and C-terminal glycosyl hydrolase domains. However, we have identified at least three bacterial glycosylases which have a swapped order of these domains. Some of these circularly permutated sequences have been described before (Lin et al. 1990; Ohmiya, Takano, and Shimizu 1990), but the unusual structure was not noted. Both domains are also found in bacteria as individual proteins (singletons). No three-dimensional structures of the permutated proteins are available.

Fatty Oxidation Complex. We find archeal proteins with a different arrangement of domains in this highly complex enzyme. The domains involved are HCDH (a CoA dehydrogenase) and ECH (enoyl-CoA hydratase). Again, proteins which have either one of the two domains are abundant.

More examples can be found in our Supplementary Material online.

CPs Caused by the Duplication/Deletion Mechanism

Protein families that are likely to have evolved CPs by this mechanism are usually large groups of long, multimeric proteins. CPs seem to be a side effect of internal gene duplication events which are common for some protein families. In several cases of the identified iCPs, we were not able to identify a corresponding true CP. These results have been excluded from table 2.

Nonribosomal Peptide Synthetases

A large family of multidomain proteins have been identified to take part in CP events in bacteria. Nonribosomal peptide synthetases (NRPS) often contain the following domains: adenylation, peptidyl carrier, and condensation domain. Several multimeric repetitions of these proteins have been found, as well as many iCPs. This suggests that the duplication/cleavage process has caused CP in this case. Many paralogs of these proteins are found on the same genome because they play a major role in the biosynthesis of bacterial toxins (Finking and Marahiel 2004). Because of the high number of repeats within one gene and the high number of paralogs within one genome, the phylogeny of these proteins turns out to be quite complex (fig. 5). As predicted for iCPs (see *Discussion*), different repeats of a domain from one and the same gene can have different evolutionary histories and, consequently, different phylogenies (fig. 5).

The duplication/deletion mechanism was identified in peptide and polyketide synthases as well as ubiquitins (Ub's). In these proteins, the modularity or multimerity fulfills a biological function and therefore is likely to be favored by natural selection. The bulk (around 75%) of identified CPs does not show traces of duplications, and no related iCPs could be found. On the other hand, in cases of the CPs attributed to the duplication/deletion mechanism, the majority of rearrangements reported by RASPODOM are iCPs, and true CPs constitute only a minor fraction, showing that duplications are a very common event in these groups of proteins.

Ubiquitin

Ub is essential as signal for protein degradation in eukaryotes (Pickart 2000). Poly-Ub chains attached to a lysine residue of substrate proteins mediate the proteolysis of the target substrate. The Ub domain is highly conserved in eukaryotes, but the Poly-Ub proteins differ in the number of Ub repeats.

We have not found any case of a true CP, despite finding many iCPs. The Ub domain corresponds, in most cases, to two ProDom domains: PD000119 and PD744725.

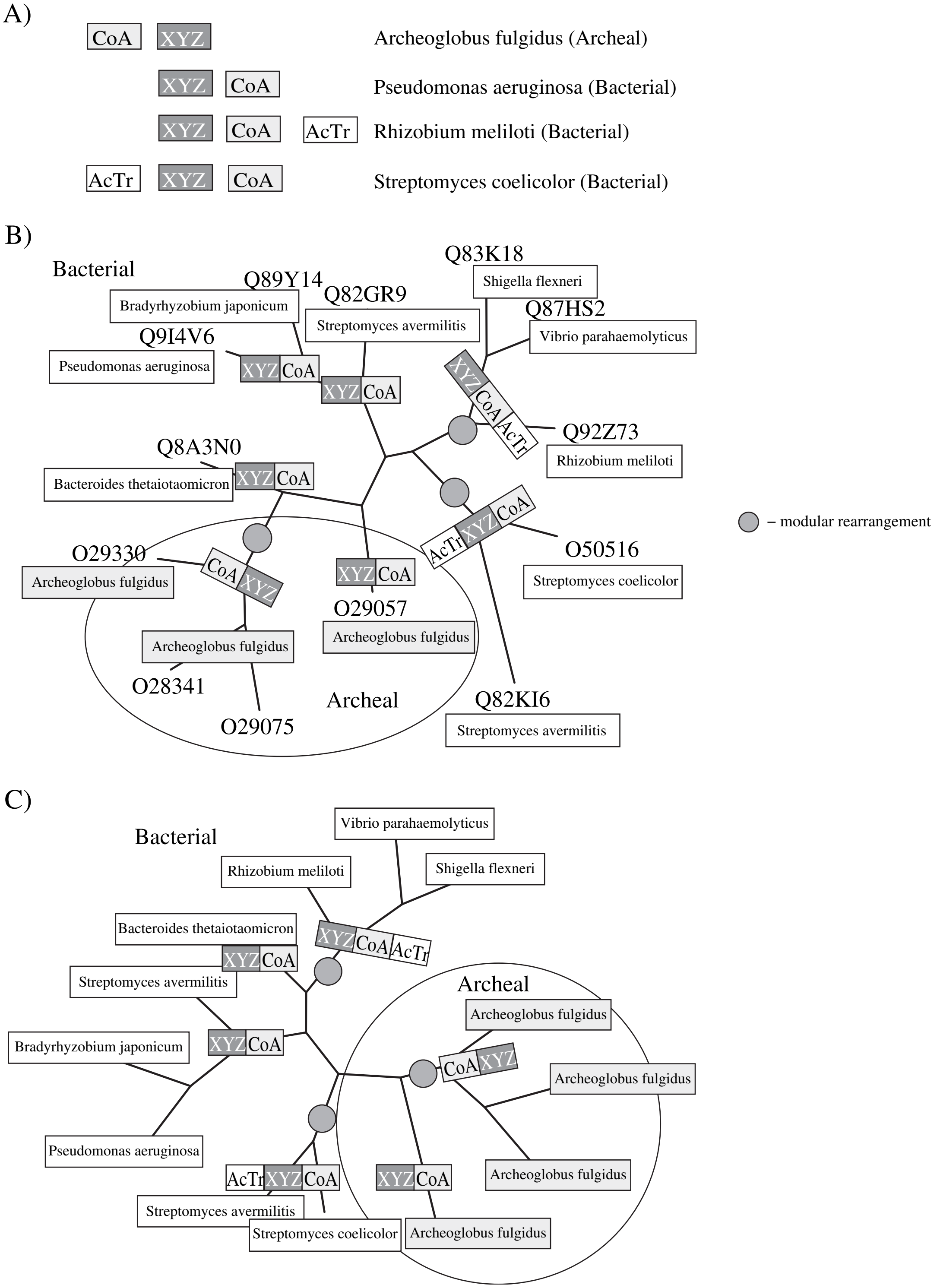

FIG. 4.—Circularly permutated proteins containing the CoA-binding domain. Domains are labeled as follows: CoA—CoA-binding domain, XYZ—-domain or domains of unknown function, and AcTr—acetyltransferase domain. (*A*) Schematic alignment of the four combinations found. Below, separate phylogenies of domains in a small subset of proteins are shown to demonstrate the differences in phylogenies and possible evolutionary history of the proteins. (*B*) Phylogenies of the CoA-binding domain in exemplary proteins. (*C*) Phylogeny of the unknown domains. Phylogeny of the acetyltransferase is not shown. Gray circles indicate modular rearrangement events.

All Poly-Ub proteins contain both of them, however, the domain with which the proteins start or end may vary (fig. 3), hence the existence of the observed iCPs. The evolution of these proteins is well described; Perelygin et al. (2002) have shown that unequal crossing-over may have played an important role. This underlying genetic mechanism fits the duplication/deletion model: it results in a partial duplication, where the terminal fragments are lost.

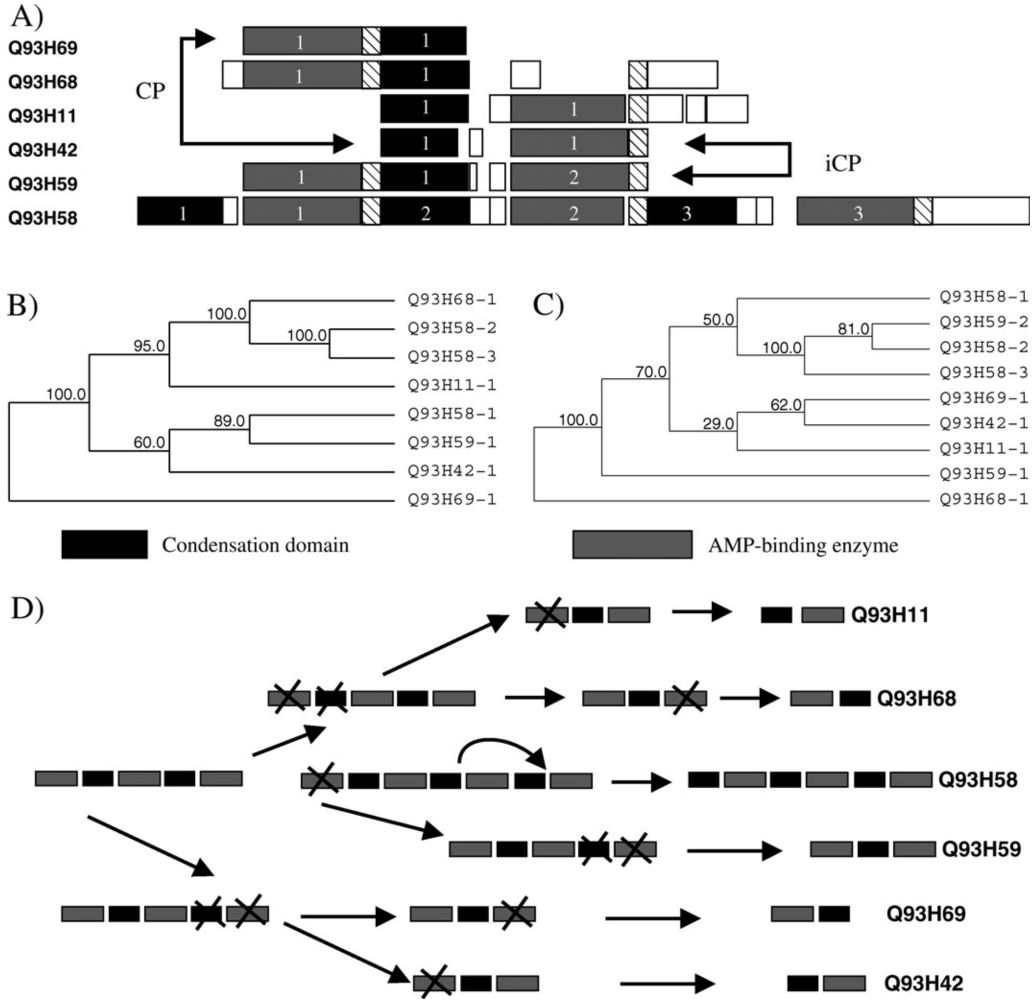

FIG. 5.—(*A*) Multiple alignment of domain arrangements of six related NRPS from *Streptomyces avermilitis*. Black rectangle: condensation domain; gray: adenosine-5′-monophosphate (AMP)–binding domain; and striped: phosphopantetheine-binding domain. A complete CP pair and an iCP pair are indicated. (*B*) and (*C*) show phylogenies of the condensation domains and AMP-binding domains, respectively. As some of the proteins contain multiple domains of the same type, the branches have been numbered accordingly. The phylogenies have been obtained with PHYML with 100 bootstrap replicates. Numbers next to the protein identifiers correspond to the consecutive number of a given domain in (*A*). (*D*) shows a speculative reconstruction of the domain rearrangements which could result in the presented phylogeny and domain arrangements.

Other Examples

Other protein families in which iCPs tend to occur are, for example, polyketide synthases (related to the NRPS), some permease proteins of ABC transporters, transferrin precursors, and some families of hypothetical proteins. However, in most of these cases a "clean" CP could not be identified.

CPs Caused by the RM System

This mechanism has been proposed by Bujnicki (2002) to explain the CPs in MTs. We were able to identify these CPs using RASPODOM (Supplementary Material online).

False Positives

Out of 971 created clusters, 160 contained true CPs, 54 contained iCPs only, and 757 contained only false positives. There were three main sources of this type of errors: (1) wrong annotation of domains, (2) high self-similarity of proteins, and (3) one wrong annotation in the sequence database. In several cases a CP was identified, but an inspection of the dot plot showed that, although one of the terminal domains was missing from the ProDom or Pfam annotation, it could still be observed on the sequence level (Supplementary Material online). This was especially the case for many of the multimeric proteins and iCPs. Multiple repeats of the same domain were also sometimes problematic for the RASPODOM algorithm, causing a fraction of false positives, especially in the case of iCPs, wrongly identified as such. In one case, a very clear-cut true CP turned out to be a misannotated sequence of some hepatitis B virus (HBV) strains, which were originally correctly described in Norder, Courouce, and Magnius (1994). In summary, the proportion of false positives was higher in the set of protein pairs annotated by RASPODOM as iCPs.

Discussion

The most important conclusion that can be drawn from the presented results is that such specific and profound rearrangements as CPs are widely spread in all groups of organisms, including archeal, bacterial, mitochondrial, and eukaryotic proteins. This stresses the importance of large-scale rearrangements. Most of the CPs were found in bacterial proteins, many also in eukaryotic or archeal proteins, and some in viral protein sequences (table 2).

Can the Different Mechanisms Be Distinguished?

It can be argued that our criteria for the distinction of the two considered models—duplication/deletion and fusion/fission—are not distinctive enough. First of all, both mechanisms can co-occur (see below). Furthermore, a duplicated intermediary form can disappear, and subsequent deletions can lead to the formation of the shorter variants, containing either N- or C-terminal part of the CP. Such a case would mimick the prediction of the independent fusion/fission mechanism. On the other hand, supposed duplications or iCPs could also be a result of independent fusion/fission mechanisms.

The presented examples show that our criteria are valid. We have not found cases where both mechanisms can be observed at the same time. While some shortened forms indeed exist in the case of peptide synthetases, they cannot form the observed circularly permutated proteins; furthermore, the predicted mechanisms agree well with the reconstructed phylogeny of the domains.

Can the Proposed Mechanisms Co-Occur?

So far, we have presented the two possible mechanisms as mutually exclusive. However, it is conceivable that a duplication event is followed by gene fusions and fissions or vice versa. To our judgment, such a "mixed" scenario is very rare and can often be ruled out.

For example, when we find the domain arrangements AB and CD in a species which is deeply rooted in the species tree and ABCD and CDAB in two species which have more recently split, it is most parsimonious to assume that both ABCD and CDAB have arisen by independent fusion events. If, however, we were to assume that ABCD has arisen from ABCD (or CDAB) via the intermediate fusion gene with the domain arrangement ABCDABCD (or CDABCDAB), followed by the losses of terminal domains, we still need to postulate that the initial fusion event, that is, ABCD (or CDAB) has arisen via fusion of AB and CD.

For all cases where the fusion/fission mechanism was proposed, we can explicitly dismiss such a mixed scenario because (1) the intermediate (ABCDABCD or remnants thereof, such as an iCP) was not observed and (2) this scenario would be less parsimonious.

Prevalence of the Fusion/Fission Mechanism

We have found both: CPs that arose through the duplication/deletion mechanism and others that are more likely to have arisen through independent fusion of smaller proteins. We found that in cases where a mechanism can be assigned to a CP, the independent fusion/fission is more common than duplication/deletion. This was unexpected. The independent fusion/fission mechanism requires two separate in-frame fusions or one fusion and one fission. At the first glance, these events seem to be less likely and require a more precise genetic mechanism than deletion of terminal fragments of a protein sequence. The duplication/deletion mechanism requires only one duplication and subsequent terminal deletions.

Possible explanations for this finding are as follows: The ratio of CPs caused by the fusion/fission mechanism remains stable and is, to a large extent, dependent on the "mobility" of domains, which can be found in all known genomes (see section on the "Fusion/fission mechanism"). As for the duplication/deletion mechanism, either (1) there exists a specific genetic mechanism causing a higher rate of duplications in the case of the aforementioned proteins or (2) duplications are highly disadvantageous under normal circumstances and their fixation depends on the selective pressure. This question remains yet to be answered.

Concluding Remarks

The analysis of CPs that are caused by different rearrangement events shows an important aspect of protein evolution. We have shown that such rearrangements can, to a large extent, produce misleading topologies in phylogenetic trees. It cannot be excluded that, in the course of evolution, the various rearrangements (duplications, fissions, fusions, unequal crossing-over, transposition, etc.) can happen independent of each other.

The complex path of modular evolution of proteins is well illustrated by the bacterial peptide synthetases. These proteins are responsible for nonribosomal synthesis of proteins, and their modularity is directly correlated with the amino acid sequence of the synthesized peptide. It is therefore likely that the evolution of these proteins is directed by the selective pressure on the resulting peptides, just as the evolution of a nucleic acid sequence is governed by the selective pressure acting upon proteins.

Literature Cited

Apic, G., J. Gough, and S. Teichmann. 2001. An insight into domain combinations. Bioinformatics **17**(Suppl. 1):S83–S89.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer. 2002. The Pfam protein families database. Nucleic Acids Res. **30**:276–280.

Bornberg-Bauer, E., F. Beaussart, S. Kummerfeld, S. Teichmann, and J. Weiner, III. 2005. The evolution of domain arrangements in proteins and interaction networks. Cell. Mol. Life Sci. **62**:435–445.

Bujnicki, J. 1999. Comparison of protein structures reveals monophyletic origin of the AdoMetdependent methyltransferase family and mechanistic convergence rather than recent differentiation of N4-cytosine and N6-adenine DNA methylation. In Silico Biol. **1**:175–182.

———. 2002. Sequence permutations in the molecular evolution of DNA methyltransferases. BMC Evol. Biol. **2**:3.

Cheltsov, A., M. Barber, and G. Ferreira. 2001. Circular permutation of 5-aminolevulinate synthase. Mapping the polypeptide chain to its function. J. Biol. Chem. **276**:19141–19149.

Corpet, F., F. Servant, J. Gouzy, and D. Kahn. 2000. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res. **28**:267–269.

Elofsson, A., and E. Sonnhammer. 1999. A comparison of sequence and structure protein domain families as a basis for structural genomics. Bioinformatics **15**:480–500.

Felsenstein, J. 1989. PHYLIP (phylogeny inference package). Version 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.

Finking, R., and M. A. Marahiel. 2004. Biosynthesis of nonribosomal peptides. Annu. Rev. Microbiol. **58**:453–488.

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52**:696–704.

Hatefi, Y., and M. Yamaguchi. 1996. Nicotinamide nucleotide transhydrogenase: a model for utilization of substrate binding energy for proton translocation. FASEB J. **10**:444–452.

Hennecke, J., P. Sebbel, and R. Glockshuber. 1999. Random circular permutation of DsbA reveals segments that are essential for protein folding and stability. J. Mol. Biol. **286**:1197–1215.

Jeltsch, A. 1999. Circular permutations in the molecular evolution of DNA methyltransferases. J. Mol. Evol. **49**:161–164.

Lin, L., E. Rumbak, H. Zappe, J. Thomson, and D. Woods. 1990. Cloning, sequencing and analysis of expression of a Butyrivibrio fibrisolvens gene encoding a beta-glucosidase. J. Gen. Microbiol. **136**:1567–1576.

Norder, H., A. Courouce, and L. Magnius. 1994. Complete genomes, phylogenetic relatedness, and structural proteins of six strains of the hepatitis B virus, four of which represent two new genotypes. Virology **198**:489–503.

Ohmiya, K., M. Takano, and S. Shimizu. 1990. DNA sequence of a beta-glucosidase from Ruminococcus albus. Nucleic Acids Res. **18**:671.

Perelygin, A., F. Kondrashov, I. Rogozin, and M. Brinton. 2002. Evolution of the mouse polyubiquitin-C gene. J. Mol. Evol. **55**:202–210.

Pickart, C. 2000. Ubiquitin in chains. Trends Biochem. Sci. **25**:544–548.

Postma, P., J. Lengeler, and G. Jacobson. 1993. Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria. Microbiol. Rev. **57**:543–594.

Schmitt, L., and R. Tampe. 2002. Structure and mechanism of ABC transporters. Curr. Opin. Struct. Biol. **12**:754–760.

Vogel, C., C. Berzuini, M. Bashton, J. Gough, and S. Teichmann. 2004. Supra-domains: evolutionary units larger than single protein domains. J. Mol. Biol. **336**:809–823.

Weiner, J., 3rd, G. Thomas, and E. Bornberg-Bauer. 2005. Rapid motif-based prediction of circular permutations in multidomain proteins. Bioinformatics **21**:932–937.

Weston, C., S. White, and J. Jackson. 2001. The unusual transhydrogenase of Entamoeba histolytica. FEBS Lett. **488**:51–54.

Claudia Schmidt-Dannert, Associate Editor

Accepted December 12, 2005