

REVIEWS

UDC 575.83

Evolution of Seed Storage Globulins and Cupin Superfamily¹

A. D. Shutov and I. A. Kakhovskaya

Laboratory of Plant Biochemistry, State University of Moldova, MD-2009, Chişinău, Moldova;

e-mail shutovandrei@yahoo.com

Received October 25, 2010; in final form, November 8, 2010

Abstract—An extensive superfamily of cupins (clan cl09118) currently combines thousands of functionally and structurally diverse prokaryote and eukaryote proteins, which contain a β -barrel of antiparallel β -strands (cupin module). Possible ways of the formation of the cupin superfamily were suggested based on the comparison of primary and tertiary structures of proteins from several conserved families of cupins including seed storage globulins and plant oxalate oxydases (germins), and bacterial oxalate decarboxylases, gentisate dioxygenases and epimerases. The origin of the two-domain structure of seed storage globulins from cyanobacterial two-domain oxalate decarboxylases has been deduced. The evolutionary pathway of single-domain germins previously suggested to be immediate progenitors of storage globulins was traced back. Common evolutionary roots of germins and oxalate decarboxylases descend from recent bacterial and archaeobacterial proteins whose primitive structure is restricted to the cupin module. These root proteins reflect the hypothetical structure of a pro-cupin that probably gave rise to at least a part of the total diversity of members of the cupin superfamily (for instance, to the cupin module of gentisate dioxygenases). The major dilemma for the description of the cupin superfamily is distinguishing evolutionary divergence from convergence. The structural convergence can be exemplified by formation of a β -barrel inside of the extremely conserved structures of the otherwise unrelated epimerases from *Archaea* and bacteria.

DOI: 10.1134/S0026893311030162

Keywords: seed storage globulins, germins, oxalate decarboxylases, cupin superfamily, molecular evolution

The terms legumin and vicilin were introduced at the end of the XIXth century by T.B. Osborn to designate two types of storage globulins from seeds of four legume species. Now it has become evident that both storage globulins are characteristic of seeds from all spermatophytes (see [1]). Comparisons of primary [2, 3] and then tertiary [4] structures indicated that legumin and vicilin share a common evolutionary precursor. A fern spore-specific vicilin-like protein caa91187 exhibits both legumin- and vicilin-like sequence features and thus can be regarded as a model for the progenitor of both types of seed storage globulins [5–7].

Each storage globulin subunit contains two structurally equivalent domains. Both domains consist of a β -barrel of eight basic antiparallel β -strands BCDE-FGHI followed by a group of two or three α -helices [2, 4]. Amino acid sequences of these domains are homologous. Evidently, they were evolved from a common single-domain ancestor, which has been duplicated during evolution [2]. A single-domain germin (plant oxalate oxidase) homologous to storage globulins was suggested to represent such an ancestor [8] (hypothesis I). This hypothesis stimulated a further search for evolutionary progenitors of storage globulins, which

resulted in the finding of an extensive superfamily of so called cupins [9, 10]. Two nodular sequence motives from germin structure [11] that correspond to two truncated halves of the β -barrel ($Gx_5HxHx_{3,4}Ex_6G$, β -strands CDE and $Gx_5PxGx_2Hx_3N$, β -strands FGH) were used as queries.

Among cupins, bacterial oxalate decarboxylases (OD) are most interesting from the viewpoint of storage globulin evolution. The OD structure consists of two domains homologous to the storage globulin domains. Therefore, it was suggested that the bacterial OD might be the direct and sole ancestor of seed storage globulins [12] (hypothesis II). It cannot be excluded that a gene coding for a cyanobacterial OD was inherited by plants symbiotically. Further, it might have been transferred into the nucleus via endosymbiosis [13] and evolved into the present-day storage globulins.

Modern databases contain many thousands of amino acid sequences of proteins relegated to cupin superfamily (clan cl09118). The presence of a β -barrel structure, which is considered to be characteristic for all cupins, was shown for tertiary structures of more than hundred bacterial, archaeobacterial and eukaryote proteins. Common ancestry of at least 18 subclasses of cupins was suggested [10]. However, homology of short amino acid sequences among these subclasses is

¹ The article was translated by the authors.

Abbreviations: OD, oxalate decarboxylase; GD, gentisate-1,2-dioxygenase; EP, dTDP-4-dehydrorhamnose 3,5-epimerase.

not always evident. In some cases, the similarity between primary structures of cupins might be a result of convergence because of the formation of similar tertiary structures.

In the context of above mentioned alternative hypotheses of the origin of seed storage globulins [8, 12], we followed in this review the evolutionary pathway of legumin and vicilin and compared primary and higher order structures of legumin, germin and bacterial OD. An attempt to estimate the probability of divergent and/or convergent evolution of some other proteins from the cupin superfamily was made.

METHODS AND STRATEGY OF ANALYSIS

The following programs were used: BLAST (<http://www.ncbi.nih.gov/>) for the search for homologous protein amino acid sequences, ClustalW2 for their alignment and TREECON [14] for their evolutionary analysis; DeepView/Swiss-Pdb Viewer for analysis of protein tertiary structures, construction of structural alignments and contouring of ribbon diagrams.

To provide a statistically significant topology of evolutionary trees of storage globulins and some other cupins we used the following strategy.

(1) The sequence region selected for the analysis should be as wide as possible. Alignment of an extensive preliminary sequence set of cupins (including those of known tertiary structure) indicated that such a region covers at least the entire β -barrel sequence (thus all eight basic β -strands can be used).

(2) In the course of the selection of a suitable sequence set (a most critical point) only those sequences that occupy a position close to the general evolutionary stem should be used. The selection has to be based on the statistical certainty of sequence clustering and branching of clusters. On the other hand, the taxonomical diversity of species within each of the basic clusters should be represented.

(3) The sequence suitable as a root of an evolutionary tree has to belong to the analyzed protein family but should exhibit the least similarity to all other sequences.

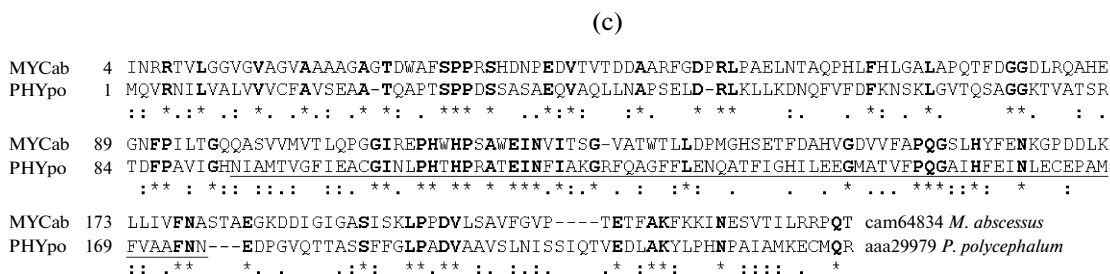
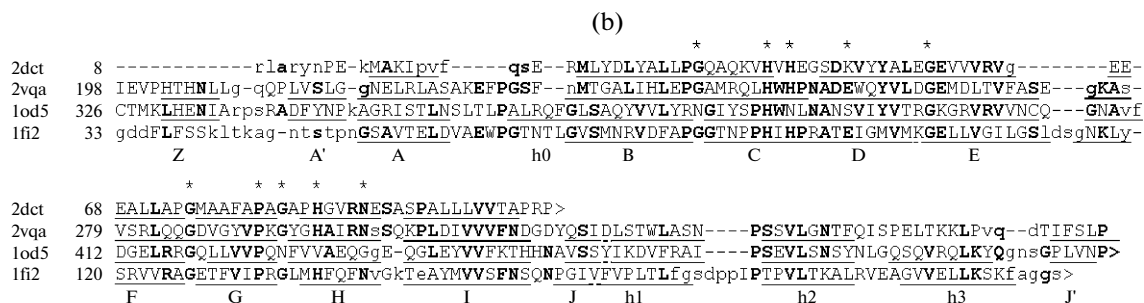
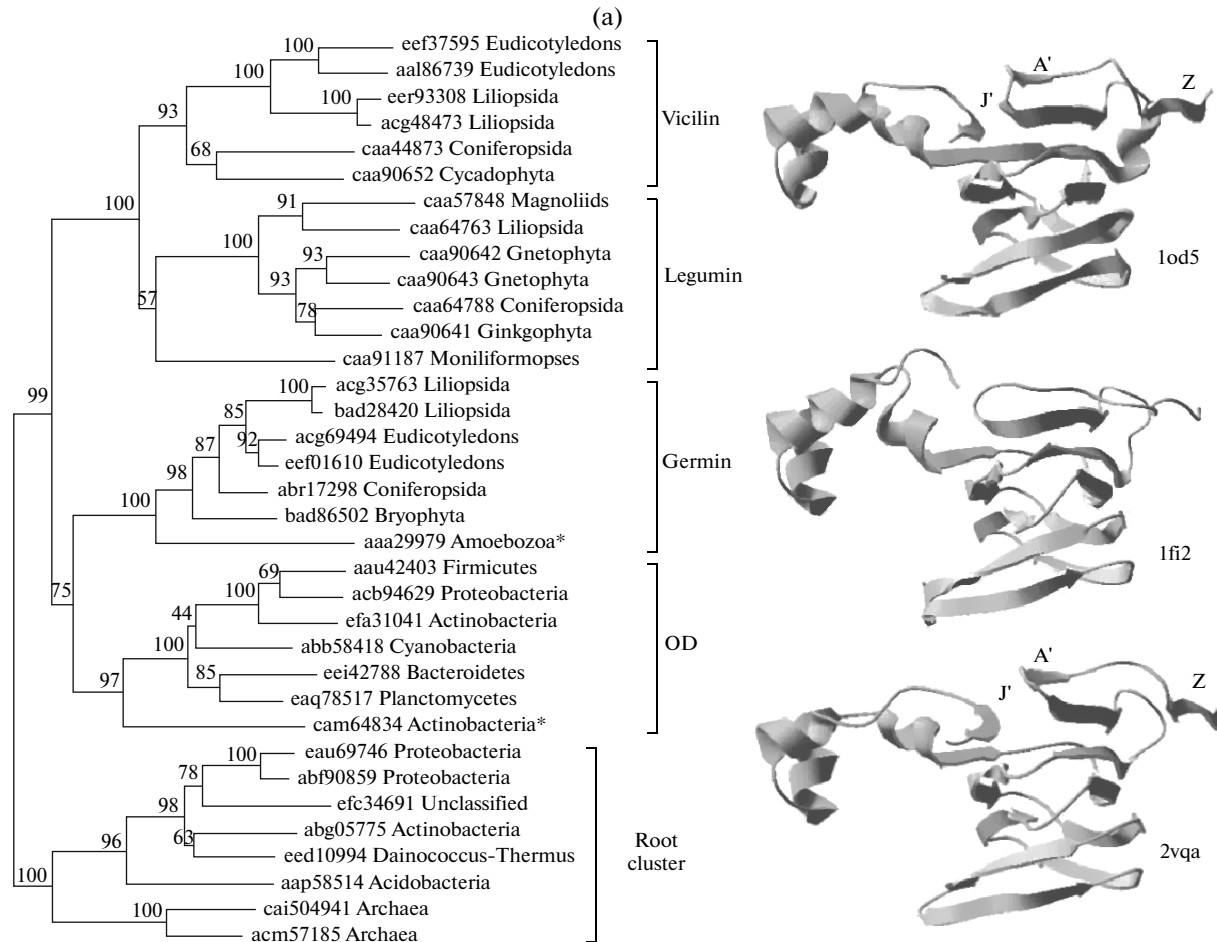
(4) In many cases, the comparative analysis of cupin β -barrel sequences is complicated because of the presence of a specifically extended variable loop between β -strands E and F, which is devoid of secondary structures [7, 10]. To identify the borders of this loop, we used the structural alignment constructed on the basis of a superimposition of C^α atoms from the backbone structures of cupins in their pair wise combinations. The level of coincidence of the structural alignment and the sequence alignment used for the tree construction was taken as a measure of trustworthiness of the latter.

(5) Single-domain proteins of known tertiary structures from the selected collection of cupins form homodimers that are structurally equivalent to homologous domains of two-domain proteins (bicupins). This circumstance allows the usage of pooled sequences for alignment and subsequent evolutionary analysis, namely two sequence parts are pooled together: the N- and C-terminal β -barrel sequences of bicupins and reduplicated β -barrel sequences of single-domain cupins. The justifiability of such an expedient is supported by the identity of the topology of evolutionary trees constructed on the basis of independent analyses of sequences from either N-terminal or C-terminal β -barrels of bicupins and β -barrels of single-domain cupins.

EVOLUTIONARY PATHWAY OF SEED STORAGE GLOBULINS

Following the strategy described above we detected five clusters that form statistically reliable evolutionary branches illustrated by proteins of known tertiary structures (Fig. 1a). A common root of the tree is probably represented by a group of single domain proteins, whose subunit structure is almost exclusively composed of a β -barrel (Figs. 1a, 1b). All the products of the further evolution acquired a C-terminal extension that forms a group of α -helices in the proteins of known tertiary structure. The structure of the two-domain OD was probably formed due to duplication of their single-domain precursor exemplified by a single-domain OD from *Mycobacterium abscessus* cam64834 (Fig. 1).

Fig. 1. Evolution of seed storage globulins. (a) The evolutionary tree. The sequence region subjected to the analysis corresponds to the sum of two cupin modules present in two-domain proteins and in the homo-dimers of single-domain proteins (156 alignment positions). Numbers along the branches refer to bootstrap values (% from 1000 replicates). Asterisks emphasize sequences of the single-domain OD and spherulin. Ribbon diagrams of the subunit tertiary structures of legumin lod5 from *Glycine max* [16] and OD 2vqa from *Synechocystis* sp. [17] (C-terminal domains), and germin 1fi2 from *Hordeum vulgare* [11] are shown at the right part of the figure. (b) Structural alignment of sequences representing legumin, germin and OD clusters and the root cluster (2dct from *Thermus thermophilus*). The alignment is based on a superimposition of C^α atoms from protein backbone structures. Low case letters indicate residues non-matching in any pair-wisely superimposed structures. Symbol > indicates C-terminal amino acids of proteins. A high similarity of the structural and sequence alignments in the region of the cupin module (strands BCDE-FGHI) should be noted. Asterisks indicate positions containing conserved amino acid residues characteristic of cupins [10]. (c) Homology of almost full-length sequences of the single-domain OD (MYCab) and spherulin (PHYpo). The region of the cupin module is underlined.



We detected homology of almost full-length sequences of the OD from *M. abscessus* and a single-domain germin-like spherulin from *Physarum polycephalum* aaa29979 (Fig. 1c). Up to now, this is a single direct evidence of previously suggested [15] origin of germins from single-domain bacterial proteins via classic divergent evolution. It should be noted that the presence of genes of spherulin (a certain precursor of germins [10]) in *P. polycephalum* (Amoebozoa) genome rejects the involvement of symbiogenesis in the evolutionary pathway of germin [15].

The two-domain structure of storage globulins was either formed via independent duplication of a single-domain plant germin-like precursor (hypothesis I [8, 15]) or it was directly inherited from two-domain OD (hypothesis II). According to the topology of the evolutionary tree, germins and OD form joint cluster (Fig. 1a) and thus might equally be estimated as storage globulin precursors. To choose between these two hypotheses the comparison of full primary and higher order structures of OD, germins and storage globulins might be helpful. Following facts argue in favor of the hypothesis II. In contrast to germins, the domain structure of both OD and storage globulins is stabilized by additional β -strands A' and J'; furthermore, additional β -strand Z stabilizes inter-domain interactions (Fig. 1a). The unusually high level of similarity of primary, tertiary and quaternary structures of the cyanobacterial OD and plant storage globulins might be due to the involvement of symbiogenesis in the evolutionary process.

Thus, two-domain OD can be estimated as the most probable evolutionary precursors of two-domain seed storage globulins. Single domain OD probably gave rise to both two-domain OD and germin-like proteins (*Amoebozoa* spherulins), and thus to germins.

CUPIN SUPERFAMILY

A common feature that allowed the unification in the cupin superfamily of proteins, though diverse in functions, amino acid sequences and domain organization, was the presence of a β -barrel containing the nodular sequence motifs (more or less eroded) characteristic of germins [10] in their structure (either known or expected). It is hardly possible to ascertain the reliable phylogenetic relationships between distinct members of the cupin superfamily when based only on the comparison of multiple aligned short sequence segments restricted to the nodular motives (52 amino acid residues, β -strands CDE-FGH, see Fig. 1b).

The cupin superfamily consists of numerous families of functionally related homologous proteins. Therefore, the intrafamily evolutionary analysis usually does not meet with difficulties. To ascertain the evolutionary relationships between different cupin families is much more complicated especially when

the interfamily similarity is restricted only to the region of the cupin module. The phenomenon of the presence of the cupin β -barrel inside the structures of functionally distinct proteins might be rationalized by four reasons described below.

(1) The evolution of two or several cupin families developed in accordance with canons of classic divergent evolution. Their primary ancestor could be a pro-cupin with a structure wholly consisting of a β -barrel. Further evolution of a pro-cupin led to the formation of a common secondary ancestor of elaborate structure. In this way the amino acid sequences of different families retained homology features far from limits of the cupin module. The evolution of seed storage globulins appears to be an example (up to now the only proved example) of such a divergent evolution of cupin families (Fig. 2a).

(2) Similarly, a pro-cupin could be a common primary ancestor of two or several cupin families. However, its further evolution can follow two or several different scenarios creating different secondary ancestors of two or several families. Therefore, different families retain homology features exclusively in the region of the cupin β -barrel. The evolutionary relationships between the root cluster of the branch of storage globulins and the family of gentisate dioxygenases (GD) probably exemplifies this very case (Figs. 2b, 2c).

(3) The β -barrel of many cupin families might serve as an example of evolutionary mobile module. Such a module can be so successful in its structural and functional adaptation that it appears in numerous proteins in very different structural settings [18]. In a new setting, the mobile module inserted into a foreign protein undergoes modification that impedes the ascertainment of its origin. Meanwhile, the general task should be to describe the origin of the fontal cupin module and to trace back the history of the nascence of the module in different families. Modification of the cupin module targeted into a foreign protein depends on time. When the event of the insertion is recent, it can be detected easily (Fig. 2d). It should be noted that we use the term "mobile module" in a wide meaning irrespectively of genetic mechanisms responsible for its mobility.

(4) Up to now, there are no reliable evidences of common ancestry of the β -barrel module from all the members of the cupin superfamily, as has been implied [10]. It seems very probable that the similarity features, detectable only within the cupin modules of two or several cupin families appeared as a result of convergent evolution due to formation of similar β -barrel structures. Possible independent evolution of two cupin families can be exemplified by a comparison of cupins from the root cluster of the storage globulin branch and epimerases (Figs. 2e, 2f).

The four presumed possibilities of the development of the β -barrel structure within the functionally dis-

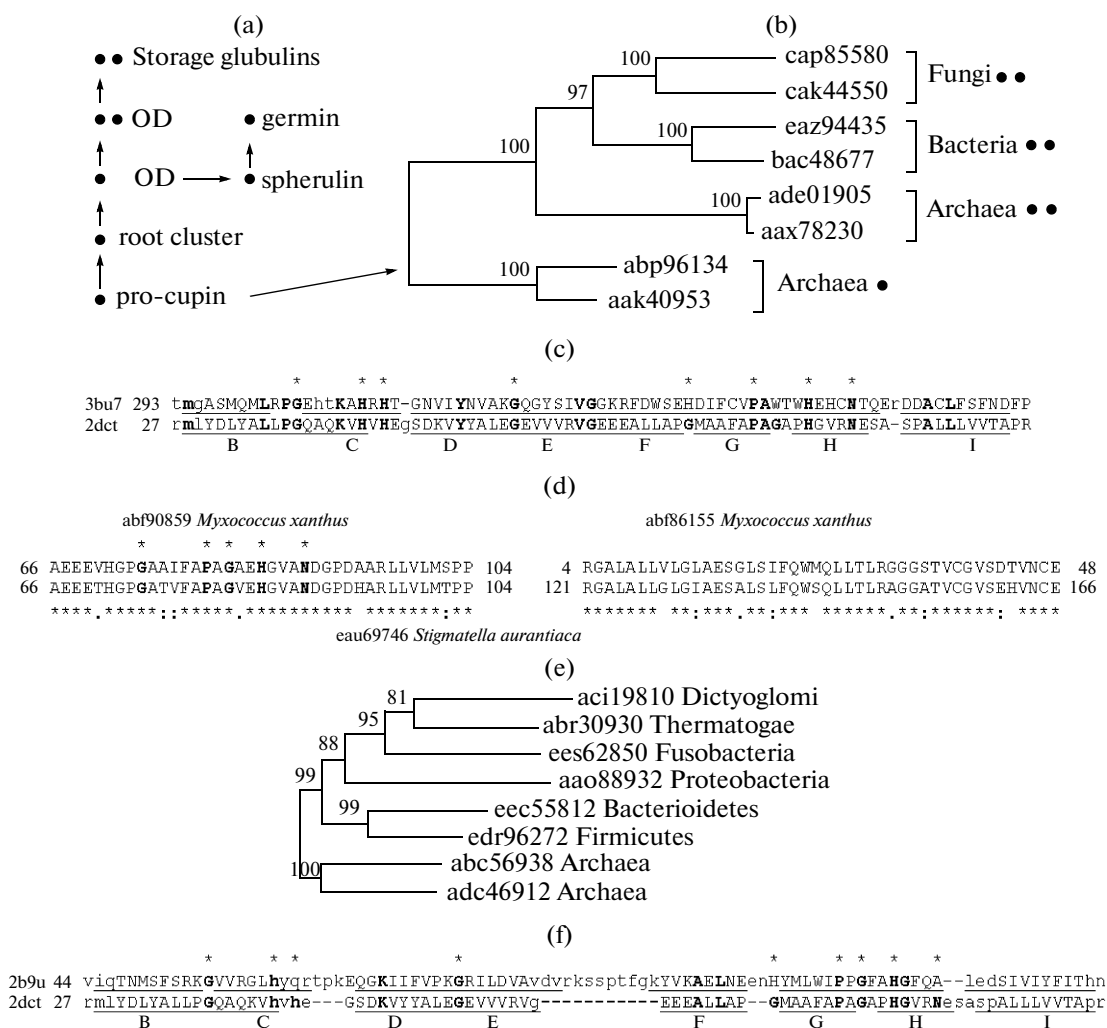


Fig. 2. Possible evolutionary pathways of the β -barrel structures from functionally distinct cupin families. (a) Evolutionary branches of storage globulins and germins. Symbols ● and ●● indicate single-domain and two-domain proteins, respectively. (b) Two-domain GD were probably formed due to a duplication of a single-domain GD (the root cluster). Sequence region subjected to the analysis corresponds to the sum of full-length domain sequences from two-domain proteins and from homo-dimers of single-domain proteins (249 alignment positions). (c) Structural alignment of GD 3bu7 from proteobacteria *Ruegeria pomeroyi* and cupin 2dct from *T. thermophilus*. (d) sequence of vitamin K epoxide reductase from *Stigmatella aurantiaca*, which contains a cupin module (its C-terminal part shown), exhibits homology to the sum of self-dependent sequences of cupin and vitamin K epoxide reductase from *M. xanthus*. (e) Phylogenetic relationships of bacterial and archaeal EP. Full-length sequences were analyzed (178 alignment positions). (f) Structural alignment of EP 2b9u from the archaeon *Sulfolobus tokodaii* and cupin 2dct from *T. thermophilus*. Asterisks indicate positions, containing conserved amino acid residues characteristic of cupins [10]. Low case letters indicate residues non-matching in superimposed structures.

tinct cupin families is illustrated by examples shown below.

The evolutionary branch of seed storage globulins. Cupins form a relatively conserved root cluster which reflects a primitive β -barrel structure of a hypothetical pro-cupin as a common ancestor of two or more cupin families. Single-domain OD derived from the root cluster acquired an elaborate structure further developed in accordance with two independent scenarios generating (a) single-domain spherulins and then ger-

mins, and (b) two-domain OD and then storage globulins (Fig. 2a).

The family of gentisate dioxygenases (GD) contains archaeobacterial, bacterial and fungal proteins whose structure is composed of two structurally equivalent domains. Far from limits of the cupin β -barrel, the full-length domain sequences of two-domain GD exhibit homology to the full-length sequences of single-domain GD from *Archaea* (Fig. 2b). Evidently, the region of homology between GD and OD is restricted to the β -barrel sequence. Hence, the two-domain GD

and OD structures have been generated due to two independent duplications of different single-domain progenitors (rather than due to a single duplication event as it was suggested earlier [10]). Structural and sequence alignments of GD 3bu7 from proteobacteria *Ruegeria pomeroyi* and cupin 2dct from *Thermus thermophilus* (see Fig. 1) almost coincide in the region of the β -barrel (Fig. 2c). Similarity of both higher order and primary structures of the cupin module of GD and proteins from the root cluster of the storage globulin branch allows suggesting that they have been derived from a common evolutionary progenitor (a pro-cupin that existed at early stages of life).

Cupin β -barrel as an evolutionary mobile module.

As an example, let us use the sequence eau69746 from proteobacteria *Stigmatella aurantiaca* (note, the cupin module from this sequence was demonstrated to belong to the root cluster of the OD branch, see Fig. 1a). The *S. aurantiaca* sequence consists of two parts, a cupin β -barrel and a vitamin K epoxide reductase. Both parts exist as independent proteins in many bacteria. In particular, the sum of sequences of relevant proteins abf90859 and abf86155 from proteobacteria *Myxococcus xanthus* exhibits a high degree of homology (only 7% of non-conserved substitutions) to the full-length sequence from *S. aurantiaca* (Fig. 2d). Therefore, it seems very probable that the genes coding for the respective independent proteins exist in the genome of *S. aurantiaca* as well. Formation of the multidomain protein eau69746 from *S. aurantiaca* is apparently a result of a duplication of a gene coding for a short cupin sequence and its insertion into the 5'-region of the vitamin K epoxide reductase gene. The multiplicity of the events of such a copying (duplication) of nucleotide sequences during evolution of eukaryote genomes [20] allows us to suggest that the duplication of the cupin module represents a general mechanism of the β -barrel formation inside manifold structures of proteins from the cupin superfamily.

The family of epimerases (EP). An extended cluster of single-domain EP contains archaeobacterial, bacterial (Fig. 2e), and animal proteins. The cupin β -barrel region covers approximately one third of total EP sequences, which are extremely conserved. In the β -barrel region, structural and sequence alignments of the EP 2b9u from the archaeon *Sulfolobus tokodaii* and the cupin 2dct from *T. thermophilus* (see Fig. 1) are different (Fig. 2f). Therefore, there are no reasons to suspect common ancestry of EP and OD families of cupins. For instance, matching of Gly and Pro residues in sequence and structural alignments might be due to convergence because of the presence of similar β -barrel structures, were these residues are responsible for the formation of turns between β -strands B/C, D/E and G/H.

In this review, we intended to show how difficult it is to describe the phylogenetic relationships between

protein families currently united into the cupin superfamily. An expedient way to solve this task probably consists of the identification of the evolutionary precursors of each of the families and in the comparison of primary and higher order structures of these precursors. Hopefully, this way will be helpful to distinguish between divergent and convergent evolution as a source of similarity between cupin β -barrels of different protein families and will give the possibility to follow the path of the relocation of the cupin module as an evolutionary mobile structural element.

REFERENCES

1. Shewry P.R., Casey R. 1999. *Seed Proteins*. Dordrecht: Cluwer.
2. Lawrence M.C., Izard T., Beuchat M., Blagrove R.J., Colman P.M. 1994. Structure of phaseolin in 2.2 Å resolution. *J. Mol. Biol.* **238**, 748–776.
3. Shutov A.D., Kakhovskaya I.A., Braun H., Bäumlein H., Muntz K. 1995. Legumin and vicilin-like seed storage proteins: Evidence for a common single-domain ancestral gene. *J. Mol. Evol.* **41**, 1057–1069.
4. Adachi M., Takenaka Y., Gidamis A.B., Mikami B., Utsumi S. 2001. Crystal structure of soybean proglycinin A1aB1b homotrimer. *J. Mol. Biol.* **305**, 291–305.
5. Shutov A.D., Braun H., Chesnokov Yu.V., Bäumlein H. 1998. A gene encoding a vicilin-like protein is specifically expressed in fern spores. Evolutionary pathway of seed storage globulins. *Eur. J. Biochem.* **252**, 79–89.
6. Shutov A.D., Bäumlein H. 1999. Origin and evolution of seed storage globulins. In: *Seed Proteins*. Eds. Casey R., Shewry P. Dordrecht: Kluwer, pp. 543–561.
7. Shutov A.D., Blattner F.R., Bäumlein H., Muntz K. 2003. Storage and mobilization as antagonistic functional constraints of seed storage globulin evolution. *J. Exp. Bot.* **54**, 1645–1654.
8. Bäumlein H., Braun H., Kakhovskaya I.A., Shutov A.D. 1995. Seed storage proteins of spermatophytes share a common ancestor with desiccation proteins of fungi. *J. Mol. Evol.* **41**, 1070–1075.
9. Dunwell J.M., Khuri S., Gane P.J. 2000. Microbial relatives of the seed storage proteins of higher plants: Conservation of structure and diversification of function during evolution of the cupin superfamily. *Microbiol. Mol. Biol. Rev.* **64**, 153–179.
10. Dunwell J.M., Culham A., Carter C.E., Sosa-Aguirre C.R., Goodenough P.W. 2001. Evolution of functional diversity in the cupin superfamily. *Trends Biochem. Sci.* **26**, 740–746.
11. Woo E.J., Dunwell J.M., Goodenough P.W., Marvier A.C., Pikersgill R.W. 2000. Germin is a manganese containing homohexamer with oxalate oxidase and superoxide dismutase activities. *Nature Struct. Biol.* **7**, 1036–1040.
12. Dunwell J.M., Gane P.J. 1998. Microbial relatives of seed storage proteins: Conservation of motifs in a functionally diverse superfamily of enzymes. *J. Mol. Evol.* **46**, 147–154.
13. Martin W., Herrmann R.G. 1998. Gene transfer from organelles to the nucleus: How much, what happens, and why. *Plant Physiol.* **118**, 9–17.

14. Van de Peer Y., De Wachter R. 1994. TREECON for Windows: A software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput. Appl. Biosci.* **10**, 569–570.
15. Shutov A.D., Blattner F.R., Bäumlein H. 1999. Evolution of structurally conserved protein module from Archaea to plants. *Trends Genet.* **15**, 348–349.
16. Adachi M., Kanamori J., Masuda T., Yagasaki K., Kitamura K., Mikami B., Utsumi S. 2003. Crystal structure of soybean 11S globulin: Glycinin A3B4 homoheptamer. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 7395–7400.
17. Tottey S., Waldron K.J., Firbank S.J., Reale B., Bessant C., Sato K., Cheek T.R., Gray J., Banfield M.J., Dennison C., Robinson N.J. 2008. Protein-folding location can regulate manganese-binding versus copper- or zinc-binding. *Nature*. **455**, 1138–1142.
18. Doolittle R.F. 1995. The multiplicity of domains in proteins. *Ann. Rev. Biochem.* **64**, 287–314.
19. Chen J., Li W., Wang M., Zhu G., Liu D., Sun F., Hao N., Li X., Rao Z., Zhang X.C. 2008. Crystal structure and mutagenic analysis of GDOsp, a gentisate 1,2-dioxygenase from *Silicibacter pomeroyi*. *Protein Sci.* **17**, 1362–1373.
20. Thomas E.E., Srebro N., Sebat J., Navin N., Healy J., Mishra B., Wigler M. 2004. Distribution of short paired duplications in mammalian genomes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 10349–10354.