FEBS Letters 430 (1998) 150–153 FEBS 20438

Hypothesis

Origins of globular structure in proteins

Nobuhide Doi[a,b], Hiroshi Yanagawa[a,*]

[a] *Mitsubishi Kasei Institute of Life Sciences, 11 Minamiooya, Machida, Tokyo 194, Japan*
[b] *Graduate School of Environmental Earth Sciences, Hokkaido University, Sapporo, Hokkaido 060, Japan*

Received 5 May 1998

Abstract Since natural proteins are the products of a long evolutionary process, the structural properties of present-day proteins should depend not only on physico-chemical constraints, but also on evolutionary constraints. Here we propose a model for protein evolution, in which membranes play a key role as a scaffold for supporting the gradual evolution from flexible polypeptides to well-folded proteins. We suggest that the folding process of present-day globular proteins is a relic of this putative evolutionary process. To test the hypothesis that membranes once acted as a cradle for the folding of globular proteins, extensive research on membrane proteins and the interactions of globular proteins with membranes will be required.

Key words: Exon shuffling; Membrane-bound enzyme; Molecular chaperone; Molten globule state; Protein evolution; Protein folding

1. Introduction

How proteins fold into their well-ordered structures is one of the fundamental problems of molecular biology. The physico-chemical aspects of protein folding have been extensively studied, and recent findings indicate the importance of molten globule states [1] and molecular chaperones [2]. However, we may also consider the problem of protein folding from an evolutionary point of view, i.e. what changes in amino acid residues might have occurred in primitive random polypeptide structures in the course of molecular evolution in order to allow the emergence of present-day folded globular structures?

It has been proposed that the origin of protein structure is closely related to the origin of introns [3]. Eukaryotic genes are often interrupted by introns, that would facilitate shuffling or duplication of exons [4]. So far, there has been much debate about whether exon shuffling occurred early or late in gene evolution [3–9]. At the present time, the idea of late exon shuffling as an important mechanism to produce a large variety of multi-domain proteins in eukaryotic cells is widely accepted [5,6]. On the other hand, it remains unclear whether exon shuffling also contributed to the origin and early evolution of each protein domain in primitive cells. In the so-called early view of exon shuffling, it is thought that the present genes encoding globular proteins were constructed by assembly of mini-genes encoding small building blocks [4,7]. There are two main objections to this idea. (i) Early genes would have been small, because the effective sizes of the genetic molecules are determined roughly by the inverse of the mutation rate [10]. However, such small mini-genes or exons would not produce protein components capable of folding on their own; a short peptide sequence usually needs auxiliary stabilization to fold into a stable conformation [6]. (ii) All mini-genes or exons can be shuffled, but only one-third of such shuffled units are expected to be in-phase [5]. Thus, even if long open reading frames (ORFs) emerged, almost all the products would not fold and would have no function.

To solve these problems, we propose the importance of a scaffold for stabilizing the structure of unevolved proteins in early cells, and we present here a possible scenario for the origin and early evolution of protein domains. In this model, soluble globular proteins can evolve as a whole chain on a scaffold, and thus discrete small folding units with soluble structures are not needed. Evidence that globular proteins did indeed arise from primitive flexible polypeptides on a scaffold survives in the structural features of present-day globular proteins. The consequences of such an evolutionary process of globular proteins for the folding process are discussed.

2. A model for protein evolution

2.1. From RNA world to RNP world

In 'the exon theory of genes', Gilbert hypothesized an RNA world within membranes, in which the first protein synthesis would be started [4]. In the RNA world, primitive replication and translation reactions are catalyzed by ribozymes [4]. We will also start from the same situation. We are not concerned herein with the origins of life, i.e. the problem of whether self-replicating ribozymes or membranous vesicles [11] came first. We assume that the properties of the first membrane in primitive cells were similar to those of the present-day cell-membrane because of the principle of continuity.

The first question that we have to ask here is what types of peptides were encoded by mini-genes. Gilbert proposed three roles of such early peptides [4]: (i) to enhance the likelihood of RNA being wrapped in membranes; (ii) to serve as pores through membranes; and (iii) to support the three-dimensional structure of ribozymes. Here we will focus on the first two roles of peptides that interacted with membranes (transmembrane (TM) peptides), i.e. to serve as membrane materials and pores. It is known that short peptides readily form helical structures in hydrophobic microenvironments such as sodium dodecyl sulfate micelles [12], and that many peptides assemble in membranes to form pores and channels [13]. It seems reasonable to suppose that the primitive cell, which depended on the environmental prebiotic soup as a source for its require-

*Corresponding author. Fax: (81) (427) 24-6317.

Abbreviations: AP, alkaline phosphatase; ORF, open reading frame; TM, transmembrane

PII: S0014-5793(98)00674-7

ments, first improved on the cell-membrane to acquire necessary components more specifically and/or efficiently.

The primitive TM peptide sequences would presumably have contained intrinsic patterns of hydrophobic amino acids, just like the TM region of present-day membrane proteins, i.e. hydrophobic α-helices, β-strands, and amphipathic α-helices [14]. There are two merits in the idea that such TM peptides were first encoded by mini-genes when the primitive translation system and the genetic code were established. First, the fact that degenerate NUN triplets code for hydrophobic amino acids (where the N implies that all four bases are possible) could contribute to conservation of the hydrophobic residue patterns owing to the decrease in apparent mutation rate. Second, the use of only L-amino acids for construction of peptides could be rationalized as the result of selective pressure for helical structures. Several authors have previously proposed that the first set of coding sequences consisted of repeats of nucleotide oligomers encoding periodical polypeptides such as α-helical- and β-sheet-forming segments [15,16].

2.2. From TM peptides to membrane proteins

The second point to be considered is how mini-genes assembled and long ORFs emerged. Senapathy proposed a role of RNA splicing in eliminating stop codons in order to produce a long ORF from short reading frames [8]. Here we adopt Senapathy's 'stop-codon walk' mechanism except in the following two respects: (i) Senapathy was not concerned with the function of short reading frames, but we are, as mentioned above. In the early stage of evolution, the splicing of a primary RNA encoding several functional peptides may have led to the synthesis of long polypeptides without definite biological function. Thus, (ii) although Senapathy [8] and Cavalier-Smith [9] pointed out the necessity of a nuclear boundary in early cells to prevent the translation of unspliced primary RNAs, we rather favor the idea of slow splicing to prevent the loss of functional peptide genes encoded by unspliced RNAs in early cells. In other words, the original information in the RNA genes could be conserved, because only a part of the many copies of each RNA gene would be spliced, via the slow process.

If long ORFs were produced by splicing of primary RNAs encoding TM peptide genes, one-third of the new long ORFs would be expected to be in the original phase, i.e. to encode hydrophobic TM regions. As a result, the translated long polypeptides would be inserted into membranes, yielding new membrane proteins. The question then arises: can membrane proteins easily be constructed only with TM segments joined together? In a study of de novo design of multi-spanning integral membrane proteins, Whitley et al. [17] demonstrated that highly simplified membrane proteins can be efficiently inserted into the inner membrane of *E. coli*.

Popot et al. [18] have already suggested that the present integral membrane proteins were constructed by duplication and shuffling of the TM peptides. This model is attractive for several reasons. (i) Analyses of the gene structure of present-day membrane proteins reveal that introns tend to be located in the regions coding for extramembrane loops [19]. (ii) A frame shift of the cluster of NUN triplets coding for hydrophobic residues to that of UNN triplets can account for the Cys, Trp, Ser, Phe, or Tyr-rich regions which are often observed in extramembrane regions of present-day membrane proteins. (iii) Recent analyses for the prediction of TM helices

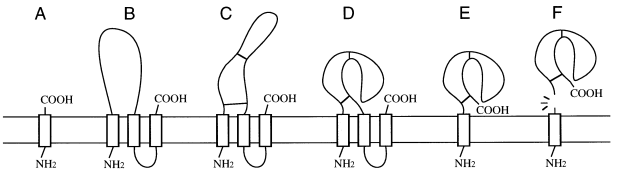

Fig. 1. A simple scheme for protein evolution on a membrane. The white boxes indicate TM regions. A: TM peptides are first encoded by mini-genes [4]. B: Long ORFs are produced from the mini-genes [8], yielding new membrane protein genes [18]. C,D: Unfolded polypeptides with membrane-bound forms have evolutionary benefits for acquiring new functions [36] and gradually fold in the course of optimization of the functions [26]. E,F: The TM regions are eliminated by gene editing or post-translational processing [33], so that the new folded domains are detached from membranes. This process naturally yields the N- and C-terminal proximity of evolved proteins, as is observed for many natural proteins [37–42].

in complete genomes suggested that there is a roughly monotonous reduction of the number of membrane proteins from one-helix proteins to highly polytopic proteins (see [20] and references cited therein).

Since a large variety of membrane proteins exists in eukaryotic cells and these proteins control cell-to-cell interactions, one might consider that membrane proteins are new proteins that have recently emerged. However, it seems that at least some of them are of ancient origin [21–23]. For example, ion channels can be divided into distinct families: there is more structural similarity among members of a given family from different species than among ion channels belonging to several families in a given species [21]. This fact indicates that the origin of a set of ion channels was earlier than differentiation of these species. Furthermore, "recent studies have established that most eukaryotic integral TM solute-transport proteins possess homologous prokaryotic counterparts" [22].

2.3. From unfolded loops to folded domains

In the case of a membrane protein of ancient origin, where one-third of the ORF encodes TM regions, how did the remaining two-thirds encoding extramembrane loops evolve? Nakashima and Nishikawa [24] examined the amino acid compositions of large (>50 residues) extramembrane segments of membrane proteins. They found that the amino acid compositions of cytoplasmic and extracellular peptides of membrane proteins corresponded well to those of intracellular and extracellular types of soluble proteins, respectively [24]. Although other interpretations are possible, one attractive hypothesis is that soluble proteins originated from the extramembrane polypeptides of membrane proteins, as shown in Fig. 1.

It is not difficult to imagine that long unfolded loops of membrane proteins might acquire simple functions (for example, intracellular loops bind useful molecules to store them in cells, or extracellular loops bind harmful molecules to prevent them entering the cells). In fact, biological functions of unfolded proteins have been noted (see [25] and references cited therein). Such unfolded sequences can gradually fold for optimization of their functions, as simulated with a simple spin-glass-like model [26]. In our opinion, membranes would have played an important role as a scaffold in the gradual evolution from flexible polypeptides to well-folded proteins (see Section 3). A further merit of this scheme is that even if unevolved

polypeptides have no function they could have survived as a part of the functional membrane proteins, and hence might have been given a chance to acquire new function and structure by trial and error. For instance, long flexible loops of active channel proteins may have acquired enzymatic activities. ''Aside from evolutionary considerations, enzymes and ion channels can no longer be treated as separate and non-overlapping groups of proteins'' [23].

Can the membrane proteins carry such a long unfolded polypeptide in the extramembrane region? Charbit et al. [27] demonstrated that an outer membrane protein, LamB, retains its biological activity with insertions of up to 60 residues of heterologous peptides into an extracellular loop. Recently we found that roughly 10% of random sequences of 120–130 residues can be inserted into the surface loop region of a water-soluble protein, *E. coli* RNase HI [28]. It is highly probable that long unfolded sequences can also be inserted into the loop region of membrane proteins. On the other hand, there are many examples of folded sequences being successfully fused to membrane proteins [29]. A 'sandwich' fusion was constructed in which a water-soluble protein, AP, was inserted into the loop region of a membrane protein, MalF [30]. The high activity of the sandwich fusion protein was somewhat surprising, since AP acts as a dimer [30]. However, this result is natural, if AP has passed through such a membrane-bound form in the course of evolution.

Recently a TM receptor was reengineered and converted to a soluble receptor without loss of stability and activity after excision of the TM regions [31]. In nature, several methods would be possible to detach membrane-bound proteins from membranes. (i) A part of the RNA genes encoding TM regions can be eliminated by alternative splicing. For example, the μ heavy chain of B lymphocyte tumor cell is converted from the membrane-bound form to the soluble form by exchanging the 3′ end of the mRNA [32]. (ii) The TM peptides can be eliminated by post-translational processing just as in the case of signal peptides of secretory proteins. ''Signal peptides are simply a slightly more 'highly evolved' variety of a basic TM peptide design that most likely is very ancient'' [33]. (iii) The conversion between the membrane-bound form and the soluble form may be realized by conformational change, such as in the case of colicin A [34]. Increasing numbers of water-soluble proteins have been found to interact with membranes under various conditions [14].

3. Does protein folding recapitulate protein evolution?

The model proposed in the previous section provides two theoretical benefits for evolution of protein structure. (i) Unfolded polypeptides are conformationally constrained because of the proximity of their N- and C-termini anchored on the scaffold, and thus are stabilized by the reduction of conformational entropy (reviewed in [35]). (ii) In an early study, Adam and Delbrück [36] indicated that the diffusional encounter between enzyme and substrate can be enhanced by reducing the dimensionality in which diffusion takes place from three-dimensional space to two-dimensional surface diffusion, suggesting that membrane-bound enzymes are evolutionarily fit. Is it possible to find evidence for these ideas in present-day proteins?

First, a preference for N- and C-terminal proximity in protein structure has long been observed for many natural proteins [37,38]. The protein structures have no knots [39] and the active centers of proteins are far from the terminal regions [37]. Ptitsyn [40] proposed a model for protein folding, in which the protein first bends roughly in half near its middle point, resulting in terminal proximity. Indeed, early interactions between the termini during folding were observed for certain proteins [41,42]. Globular structures with these properties may not only depend on thermodynamical stability, but also reflect a stage of evolution, as shown in Fig. 1.

Second, present-day globular proteins may pass through membrane-bound forms in the folding process as a relic of the evolutionary process. Bychkova et al. [43] proposed that the folding intermediate states (or molten globule states) may be suitable candidates for protein translocation across membranes, and molten globule-like states can be achieved under conditions which may mimic those near membrane surfaces [44]. Unlike such secretory globular proteins, however, cytoplasmic proteins do not translocate across membranes and are not always located near membranes during folding. If a membrane-like structural complex is present in a cell, this 'pseudo-membrane' may bind to the folding intermediates of globular proteins, and may promote the folding of proteins. This idea is supported by studies of molecular chaperones which recognize a diverse range of unrelated proteins [2]. The analogy between the membrane and a chaperone, GroEL, was recently supported by Hoshino et al. [45].

4. Concluding remarks

We propose an important role of membranes as a scaffold in the origin and early evolution of protein structure. This scaffold hypothesis is based on a combination of proposals or suggestions by many authors (see the legend of Fig. 1), and overcomes some of the weak points of these hypotheses. The scaffold model also avoids some of the conflict between early and late models of exon-shuffling (Section 2), and provides new insights into the relationship between protein folding and evolution (Section 3). This hypothesis is also consistent with the available experimental results, though more data are needed to test it more rigorously.

Recent analyses of complete genome sequences suggested that many major families of membrane proteins still remain to be characterized [20]. Thus, extensive research on membrane proteins is required for phylogenetic analysis. It may be difficult, however, to find primary-sequence homologies between globular proteins and membrane proteins, since the emergence of proteins would have occurred at a very early stage in the evolution of life. Structural studies on membrane proteins are thus important to confirm that all globular proteins possess homologous counterparts in membrane protein structures (at least one example was recently reported [46]). As other approaches, membrane protein engineering [17,47] and cell-surface engineering [27,48] are useful for clarifying the plasticity of membrane protein structures, and for simulating the artificial evolutionary process of new domains from random polypeptides displayed on the surface of scaffold proteins [35,49].

Although we have focused on the membrane as a scaffold in this paper, nucleic acids may also be available as a scaffold instead of membranes. For example, RNA-binding proteins may have been constructed by assembly of RNA-binding peptides, whose unfolded loops may have gradually folded on

RNA during optimization of functions. In this case, the nucleic acid may not only be a scaffold, but also a substrate for proteins, so that the nucleic acid-binding region would not have been eliminated from the proteins. If this is so, some details of the scaffold model proposed in this paper may have to be changed in the future. Nevertheless, we think that this 'membrane-scaffold' model, even if oversimplified, sheds light on the origin of the globular proteins (especially secretory proteins) which require adaptation to the appropriate location through interaction with membranes.

Acknowledgements: We thank Drs. T. Yomo, M. Itaya, Y. Kikuchi and S. Tokura for discussions.

References

[1] Kuwajima, K., Semisotnov, G.V., Finkelstein, A.V., Sugai, S. and Ptitsyn, O.B. (1993) FEBS Lett. 334, 265–268.
[2] Ellis, R.J. (1997) Biochem. Biophys. Res. Commun. 238, 687–692.
[3] Dibb, N.J. (1993) FEBS Lett. 325, 135–139.
[4] Gilbert, W. (1987) Cold Spring Harbor Symp. Quant. Biol. 52, 901–904.
[5] Patthy, L. (1991) Curr. Opin. Struct. Biol. 1, 351–361.
[6] Doolittle, R.F. and Bork, P. (1993) Sci. Am. 269, 50–56.
[7] Go, M. (1981) Nature 291, 90–93.
[8] Senapathy, P. (1988) Proc. Natl. Acad. Sci. USA 85, 1129–1133.
[9] Cavalier-Smith, T. (1991) Trends Genet. 7, 145–148.
[10] Eigen, M., Gardiner Jr., W.C. and Schuster, P. (1980) J. Theor. Biol. 85, 407–411.
[11] Yanagawa, H., Ogawa, Y., Kojima, K. and Ito, M. (1988) Orig. Life Evol. Biosph. 18, 179–207.
[12] Wu, C.-S.C. and Yang, J.T. (1988) Biopolymers 27, 423–430.
[13] Marsh, D. (1996) Biochem. J. 315, 345–361.
[14] Gennis, R.B. (1989) Biomembranes. Molecular Structure and Function, Springer-Verlag, New York, NY.
[15] Brack, A. and Orgel, L.E. (1975) Nature 256, 383–387.
[16] Ohno, S. (1987) J. Mol. Evol. 25, 325–329.
[17] Whitley, P., Nilsson, I. and von Heijne, G. (1994) Nat. Struct. Biol. 1, 858–862.
[18] Popot, J.-L., de Vitry, C. and Atteia, A. (1994) in: Membrane Protein Structure. Experimental Approaches (White, S.H., Ed.) pp. 41–96, Oxford University Press, New York, NY.
[19] Jennings, M.L. (1989) Annu. Rev. Biochem. 58, 999–1027.
[20] Jones, D.T. (1998) FEBS Lett. 423, 281–285.
[21] Ranganathan, R. (1994) Proc. Natl. Acad. Sci. USA 91, 3484–3486.
[22] Saier Jr., M.H. (1994) BioEssays 16, 23–29.
[23] Jan, L.Y. and Jan, Y.N. (1992) Cell 69, 715–718.
[24] Nakashima, H. and Nishikawa, K. (1992) FEBS Lett. 303, 141–146.
[25] Plaxco, K.W. and Groß, M. (1997) Nature 386, 657–658.
[26] Saito, S., Sasai, M. and Yomo, T. (1997) Proc. Natl. Acad. Sci. USA 94, 11324–11328.
[27] Charbit, A., Molla, A., Saurin, W. and Hofnung, M. (1988) Gene 70, 181–189.
[28] Doi, N., Itaya, M., Yomo, T., Tokura, S. and Yanagawa, H. (1997) FEBS Lett. 402, 177–180.
[29] Boyd, D. (1994) in: Membrane Protein Structure. Experimental Approaches (White, S.H., Ed.) pp. 144–163, Oxford University Press, New York, NY.
[30] Ehrmann, M., Boyd, D. and Beckwith, J. (1990) Proc. Natl. Acad. Sci. USA 87, 7574–7578.
[31] Ottemann, K.M. and Koshland Jr., D.E. (1997) Proc. Natl. Acad. Sci. USA 94, 11201–11204.
[32] Alt, F.W., Bothwell, A.L.M., Knapp, M., Siden, E., Koshland, M. and Baltimore, D. (1980) Cell 20, 293–301.
[33] von Heijne, G. (1990) J. Membr. Biol. 115, 195–201.
[34] Lakey, J.H., Baty, D. and Pattus, F. (1991) J. Mol. Biol. 218, 639–653.
[35] Doi, N. and Yanagawa, H. (1998) Cell. Mol. Life Sci. 54, in press.
[36] Adam, G. and Delbrück, M. (1968) in: Structural Chemistry and Molecular Biology (Davidson, N. and Rich, A., Eds.) pp. 198–215, Freeman, San Francisco, CA.
[37] Thornton, J.M. and Sibanda, B.L. (1983) J. Mol. Biol. 167, 443–460.
[38] Christopher, J.A. and Baldwin, T.O. (1996) J. Mol. Biol. 257, 175–187.
[39] Mansfield, M.L. (1994) Nat. Struct. Biol. 1, 213–214.
[40] Ptitsyn, O.B. (1981) FEBS Lett. 131, 197–202.
[41] Roder, H., Elöve, G.A. and Englander, S.W. (1988) Nature 335, 700–704.
[42] Jennings, P.A. and Wright, P.E. (1993) Science 262, 892–896.
[43] Bychkova, V.E., Pain, R.H. and Ptitsyn, O.B. (1988) FEBS Lett. 238, 231–234.
[44] Bychkova, V.E., Dujsekina, A.E., Klenin, S.I., Tiktopulo, E.I., Uversky, V.N. and Ptitsyn, O.B. (1996) Biochemistry 35, 6058–6063.
[45] Hoshino, M., Kawata, Y. and Goto, Y. (1996) J. Mol. Biol. 262, 575–587.
[46] Neuwald, A.F. (1997) Protein Sci. 6, 1764–1767.
[47] Popot, J.-L. and Saraste, M. (1995) Curr. Opin. Biotechnol. 6, 394–402.
[48] Georgiou, G., Stathopoulos, C., Daugherty, P.S., Nayak, A.R., Iverson, B.L. and Curtiss III, R. (1997) Nat. Biotechnol. 15, 29–34.
[49] Doi, N., Yomo, T., Itaya, M. and Yanagawa, H. (1998) FEBS Lett. 427, 51–54.