Pacific Symposium on Biocomputing 4:494-504 (1999)

PROTEIN EVOLUTION AND PROTEIN FOLDING: NON-FUNCTIONAL CONSERVED RESIDUES AND THEIR PROBABLE ROLE

O.B. PTITSYN
National Cancer Institute, NIH, Laboratory of Experimental & Computational Biology, Molecular Structure Section, Building 12B, Room B116, Bethesda, MD 20892-5677, USA; Institute of Protein Research, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russia

It is shown that there are two types of conserved residues in evolutionary and functionally related proteins whose sequences have been well diverged in evolution. The first group consists of residues forming the active center, while the second (first established in this work) has nothing to do with function and therefore should be related to protein structure and/or protein folding. The lattter group consists of 4 residues in c-type cytochromes and 6 residues in globins. All these residues belong to α-helices and occupy positions (i, i+4) or (i, i+3), stabilizing one helical turn in some helices. These residues form an interface between the N- and C-terminal helices in c-type cytochromes and helices A, G, H in globins. These helical complexes form early in protein folding and are relatively stable in both equilibrium and kinetic folding intermediates. The attractive hypothesis is that these helices form folding nuclei in protein in the frame of the nucleation-growth mechanism of protein folding.

1 Introduction

There is an old puzzle in the relation between protein sequences and their 3D-structures. Evolutionary and functionally related proteins with well diverged sequences still keep the same folding pattern, i.e. the crude mutual positions of main α-helices and β-strands.[1, 2] One of the best examples is the large family of c-type cytochromes[3]. It is one of the oldest protein families since cytochromes *c* serve for utilization of the external energy, which is the main feature of each form of life. The c-type cytochromes have about 1.5 billions years of evolution and their sequences have diverged so much that there is virtually no obvious homology between their different subfamilies[4]. Even chain lengths of these proteins vary from ~60 to more

than 200. Despite these very large difference in sequences the 3D-structures of all the c-type cytochromes are similar. The α-helices in c-type cytochromes from mitochondria, chloroplasts and different bacteria are the same and have the same mutual positions while all the differences in chain length are localized in the loops.

There may be two possible explanations of the similarity of 3D-structures for evolutionary and functionally related proteins with well diverged sequences. The first one is more or less traditional and suggests that the protein pattern depends not on whole details of its sequence but on some general features of sequences, which are conserved during evolution. However, logically it may be also another explanation according to which the folding pattern depends on several residues, which are conserved during evolution.

This explanation seemed to be senseless since there were no reasons to ascribe the key role in the coding of the folding pattern to some group of few residues. The situation changed however when experiments of Alan Fersht[5,6] and calculations of Eugene Shakhnovich[7] revived the old idea that proteins can fold according to the two-state principle and that their folding may follow the nucleation-growth mechanism[8]. According to this mechanism the unfolded chain fluctuates unless some definite set of residue occasionally come together. These residues form the nucleus of protein folding which is the rate-limiting step of the process, since being once formed the folding nucleus can grow very fast involving the whole protein molecule.

Strictly speaking the nucleation-growth mechanism refers only to those proteins which fold by the all-or-none process. However, there is little doubt that even in the more general case, when folding occurs through the kinetic molten globule-like intermediate, it still may follow the nucleation-growth mechanism. In this case nucleation may lie on the way from the unfolded to the intermediate state while the transition from the intermediate to the native state can be hindered by some non-principal events, like proline cis-trans isomerization or non-native liganding.

Thus, it may happen that the protein pattern is determined by its folding nucleus consisting of just few residues. We have tried to check this idea by considering the common crude 3D-structures of two large protein families, c-type cytochromes and globins.

2 The consensus principle and general approach

To this end we have to identify the common conserved residues in the protein of each of these families. A traditional approach is just to compare all the available sequences of the family and identify as conserved those identical or similar residues, which occupy the given position in the overwhelming majority of sequences in the whole family, for instance in all globins. However, the results of this approach face a very important shortcoming. Some subfamilies of the given family may be studied much more extensively than the others, being more available and/or more interesting. As an example 90% of all the available globin sequences belong to vertebrate hemoglobins and myoglobins leaving only 10% to all the other globin subfamilies. Since vertebrate globins are evolutionary much closer to each other than to insect, worm, plant and others globins, we can come to a wrong conclusion that all globins have a reasonably good homology while in fact this condition refers only to vertebrates globins.

Therefore, we propose another approach to the identification of conserved residues in protein families which can be called the «consensus principle». This principle demands that the really conserved residues are those which occupy a given position in *each* subfamily, for instance in myoglobins, insect globins, leghemoglobins and so on. Let us consider for instance two positions in the A helix of globins: A5 and A8. A5 is occupied by positively charged residues in 79% of the sequences and A8 is occupied by bulky aliphatic residues in 99% of the sequences. It seems that both these positions are conserved. However, application of the consensus principle shows that in four subfamilies position A5 has no Lys or Arg residues and therefore its apparent conservatism is due almost entirely to vertebrate hemoglobins and myoglobins. As a contrast, position A8 is occupied by bulky aliphatic residues in at least 89% of the sequences in each of these subfamilies and therefore is really conserved.

Therefore, our general approach to the finding of conserved residues is as follows. First, we have to divide the given protein family into biochemical or biological different subfamilies. Second, we have to align the sequences for each subfamily and to identify the conserved residues in the given subfamily. Third, we have to compare the alignments of different subfamilies and to identify their common conserved residues. The most interesting conserved residues are those which form clusters in 3D-structure, and therefore our next step is to identify the conserved residues, which are in contact, i.e., have the shortest distance between non-hydrogen atoms less than 5Å.

3 c-type cytochromes

The large protein family of c-type cytochromes can be divided into seven subfamilies each consisting of at least eight sequences. There are cytochromes *c* and c_1 from mitochondria, cytochromes c_6 and c_f from chloroplasts as well as bacterial cytochromes c_2, c_{551} and c_{550}. These 7 subfamilies have only seven positions, which are occupied by identical or similar residues (see Fig.1A). Three of these positions, Cys14, Cys17 and His18 bind the heme by covalent (Cys's) or coordination (His) links and therefore their conservatism is of functional origin. However there are four positions which do not bind the heme, but still are occupied by similar residues in all the seven subfamilies. These positions are position 4 (in notations of horse cyt *c*) occupied almost entirely by Gly in 6 subfamilies and by Ala in 1 subfamily, position 10 occupied by Phe in 6 subfamilies and by Tyr in 1 subfamily, position 94 occupied by Val or Leu in 6 subfamilies and by Phe in 1 subfamily, and position 97 entirely occupied by aromatic residues in all the 7 subfamilies. These four conserved residues contact with each other forming a cluster.

It is important to mention that all the four conserved residues which do not bind heme belong to the large N- and C-terminal α-helices and form turns (i, i+4) or (i, i+3) of these helices. Moreover, these four residues form an interface between helices, which are almost perpendicular to each others.

The most interesting aspect of these results is that the N- and C-terminal helices and especially their interface play a special role in cytochrome *c* folding. These helices became stable, i.e. partly protected from deuterium exchange, at the earliest stage of folding[9] and are stable in equilibrium acid (molten globule) state of cytochrome *c* (see [10]). Moreover, the mutations in positions 10, 94 and 97 in yeast iso-1-cytochrome *c* destabilize the molten globule state of cytochrome *c* exactly to the same extent that they destabilize its native state[11]. It is a good evidence that these residues are packed in the molten globule state as tightly as in the native state.

These experimental data suggest that the four non-functional residues which remain conserved in all the subfamilies of c-type cytochromes are related to protein folding. Therefore they can be called a *folding cluster*.

Fig.2A presents this folding cluster. There are two interesting points in its structure. The first is the strong contact between almost entirely aromatic position 10 in the N-terminal helix and the entirely aromatic position 97 in the C-terminal helix. This contact includes 13 interatomic contacts and is the strongest contact between the non-heme binding residues. It emphasizes the important role of interaction between aromatic residues in the folding and/or stability of c-type cytochromes. Another interesting contact is that between Gly (or Ala) 6 in the N-helix and hydrophobic group in position 94 in the C-helix. The point is that the backbone at position 94 comes very near to the « hole» in the N-helix formed by

glycine which leads to 6 backbone-backbone interatomic contacts in this part of the protein as compared with 2 contacts between side chain 94 and the backbone in Gly6.

4 Globins

Now let us turn to the second large protein family of globins[12]. Sixty positions are occupied by identical or similar residues in ≥90% of globin sequences[13]. However only 19 of these positions, i.e. about 12% are occupied by identical or similar residues in the each of the 12 globin subfamilies (see Fig.1B).

Six of these 19 residues (namely CD1, E11, F4, F8, FG5, G5) make very strong contacts with heme. Five other conserved residues (B10, B14, C2, CD4, E4) form a linear system of contacts linked at one point with the universal heme-binding CD1. The conserved residues B13 and H19 do not belong to this system but, still belong to the immediate surrounding of the heme.

However, there are six conserved residues forming a non-polar cluster, which has no contacts either with the heme or with its immediate surrounding. These residues belong to helices A, G, H (2 residues per helix) and occupy positions (i, I+4) in each of these helices.

It is well known that helices A, G, H in myoglobins, like the N- and C-terminal helices in cytochrome *c* play a special role in protein folding. They are also relatively stable, i.e. partially protected from deuterium exchange at the earliest detectable stage of folding[14]. They also remain relatively stable in acid (i.e. in the molten globule) state[15]. They also may be packed in the molten globule as tightly as in the native state[16]. Moreover, we have shown recently that these three helices really form the AGH complex in the molten globule state and that this complex has a native-like overall structure[17].

Thus, we can conclude that the conserved residues in A, G, H helices of globins, just like the conserved residues in the N- and C-terminal helices in c-type cytochromes, are important for protein folding and form a folding cluster (see Fig.2B). This cluster in globins includes entirely or almost entirely the aliphatic position A8, entirely aromatic position A12, the entirely aliphatic position G16 and the almost entirely aromatic position H8. Only two positions, G12 and H12, which are linked with the other conserved positions just by intrahelical contacts, are occupied by aliphatic and aromatic residues to a comparable extent. This suggests that this folding cluster is stabilized not only by hydrophobic interactions, but by

some specific van-der Waals contacts which are sensitive to the replacement of aromatic side chains by aliphatic ones and *vice versa*.

5 Discussion

Thus, our analysis of about a thousand sequences of c-type cytochromes and globins has shown that in both protein families there are non-polar clusters of non-functional residues. These clusters

- exist in the early kinetic folding intermediate
- exist in the equilibrium molten globule
- may be packed in the molten globule as tightly as in the native state.

This strongly suggests that these clusters are important for protein folding.

Up to this point I have simply described our results. Now some speculations. We have demonstrated the existence of a clear correlation between the conserved non-functional residues and the structure of folding intermediate. However, folding intermediates, as we now know, are usual but not obligatory features of protein folding. Therefore, we probably have to look for a more general role of these folding clusters.

This general role may be that these clusters represent folding nuclei common for all subfamilies of the given protein family. The point is that the folding nucleus should be the lowest part of the potential barrier for protein folding and therefore they should be conserved during their biological evolution. This idea has been tested in the paper by Shakhnovich *et al*. who considered the folding of simple models on a cubic lattice[18]. We have shown that all the residues, which are involved in the folding nucleus, are conserved, and that almost all the conserved residues are included into the folding nucleus. As a further step we have found the most conserved residues in chymotrypsin inhibitor 2 and have predicted its folding nucleus in a good coincidence with the experimental data partly obtained after our prediction[19].

It is interesting to discuss the possible role of α-helices in the formation of folding nucleus. We know that it is difficult to initiate an α-helix since the first hydrogen bond has to fix the conformations of three monomer units. As a contrast, it is very easy to grow it, since each subsequent hydrogen bond fixes the conformation of only one monomer unit. Having in mind this peculiarity, I can propose the following folding mechanism for helical proteins. Folding may start with an

occasional collision of a few definite residues, which form the folding nucleus and therefore are conserved. The energy of interactions of different possible folding clusters cannot differ very much and therefore the folding nucleus is probably not the most favorable cluster but one, which can grow without any problems. The ideal case would be conserved residues, which form turns in the α-helices. The collision of these residues is coupled with the nucleation of one turn in each of these helices, which leads to their simultaneous growth. This growth is free of charge and provides a large gain of interaction energy without any additional entropy cost. Therefore, the transition state of protein folding can include two or more helices with a set of contacts between the conserved residues forming the folding nucleus. Then a protein can either fall down from its transition state directly to its native state or can be trapped for a while in the molten globule-like kinetic intermediate before transforming to the native state.

Acknowledgments

The study of conserved residues in globins was performed together with Ting Kai-Li. The author expresses his thanks for her collaboration.

References

1. C. Chothia and A. Lesk, *EMBO J.* **5**, 823 (1986).
2. T.E. Creighton in *Proteins. Structure and Molecular Properties*, 2nd edition, p.244 (W.H. Freeman & Co, New York, 1993).
3. R.E. Dickerson *et al, J. Mol. Biol*. **100**, 479 (1976).
4. R.E. Dickerson, *Sci. Am*. **226**, 58 (1972).
5. S.E. Jackson *et al, Biochemistry* **32**, 11270 (1993).
6. D.E. Otzen *et al, Proc. Natl. Acad. Sci. USA* **91**, 10422 (1994).

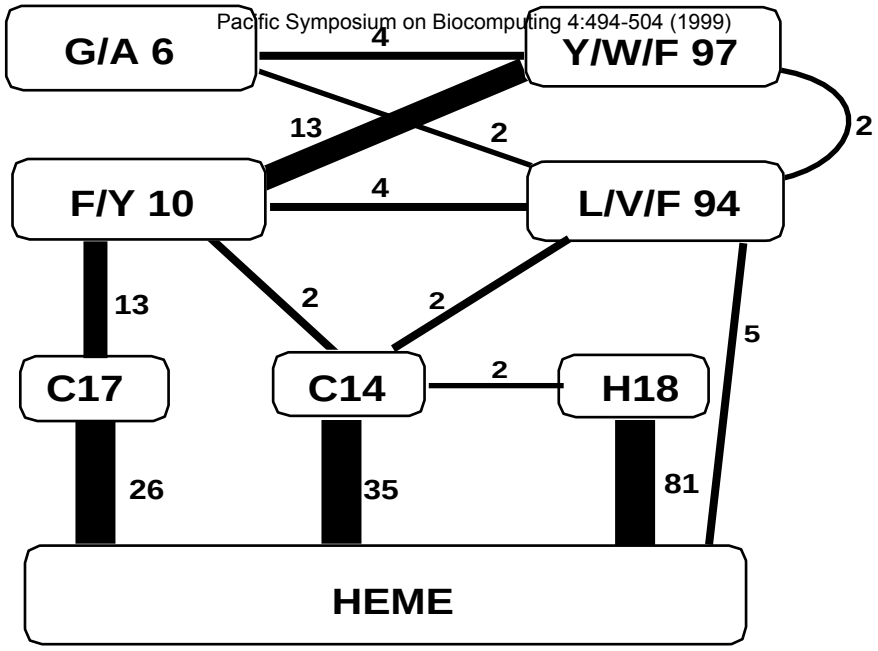
7. V.I. Abkevich *et al*, *Biochemistry* **33**, 10026 (1994).
8. T.V. Tsong *et al*, *J. Mol. Biol*. **63**, 457 (1972).
9. H. Roder *et al*, *Nature* **335**, 700 (1988).
10. M.-F. Jeng *et al*, *Biochemistry* **29**, 10433 (1990).
11. J.L. Marmorino & C.J. Pielak, *Biochemistry* **34**, 3140 (1995).
12. M.F. Perutz *et al*, *J. Mol. Biol*. **13**, 669 (1965).
13. D. Bashford *et al*, *J. Mol. Biol*. **196**, 199 (1987).
14. P.A. Jennings and P.E. Wright, *Science* **262**, 892 (1993).
15. F.M. Hughson *et al*, *Science* **249**, 1544 (1990).
16. M.S. Kay and R.L. Baldwin, *Nature Struct. Biol*. **3**, 439 (1996).
17. O.V. Tcherkasskaya and O.B. Ptitsyn, *J. Mol. Biol*, submitted.
18. E.I. Shakhnovich *et al*, *Nature* **379**, 96 (1996).
19. L. Itztaki *et al., J. Mol. Biol*. **254**, 260 (1995).

Figure 1: Residues conserved in all available subfamilies of c-type cytochromes (A) and globins (B). A corresponding number of atom-atomic contacts are shown near each line. Arcs show the contacts between neighbor turns of α-helix.

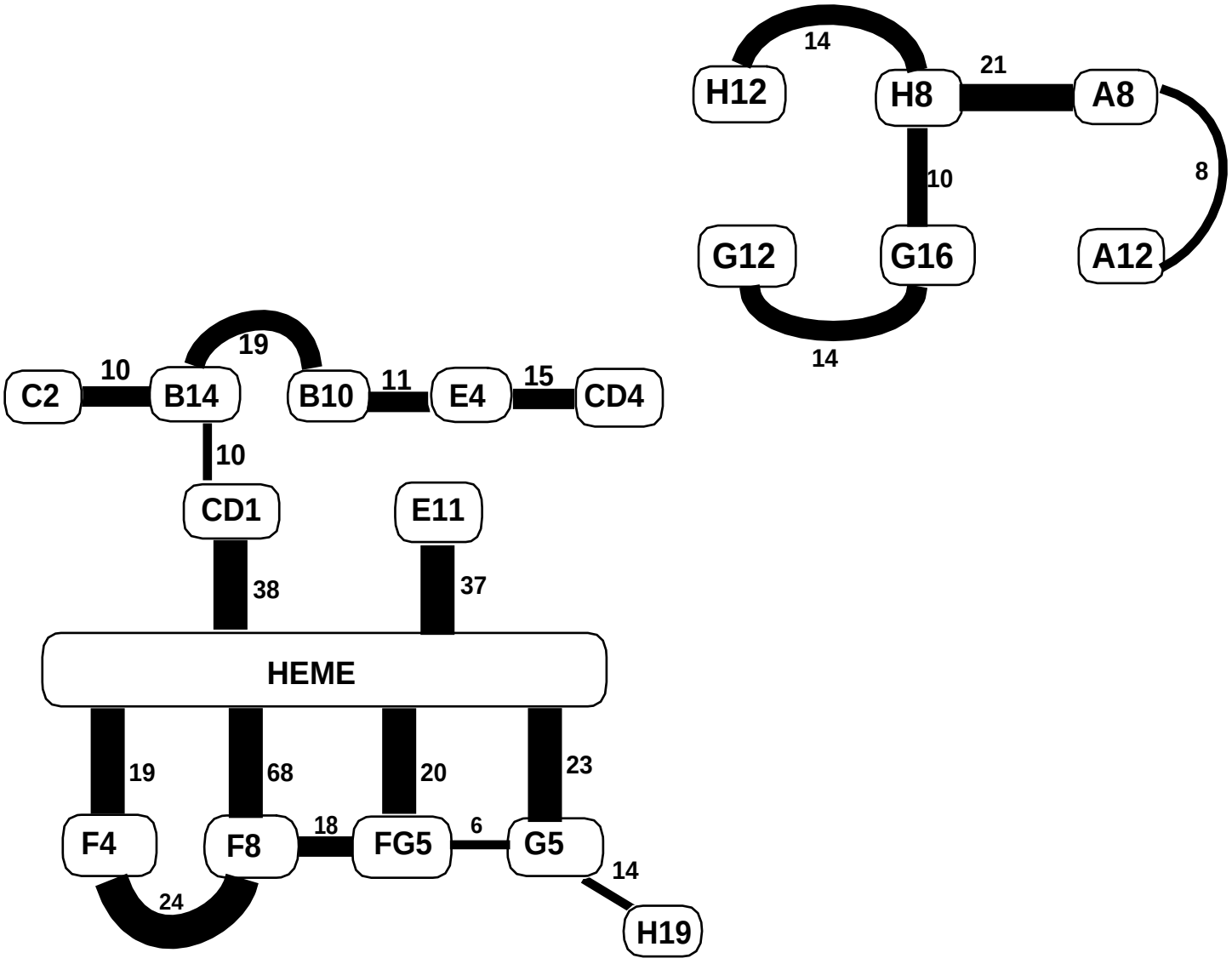
Figure 2: Clusters of conserved residues in c-type cytochromes (A) and globins (B). The rotations are the same as in Figure 1.

Pacific Symposium on Biocomputing 4:494-504 (1999)

A

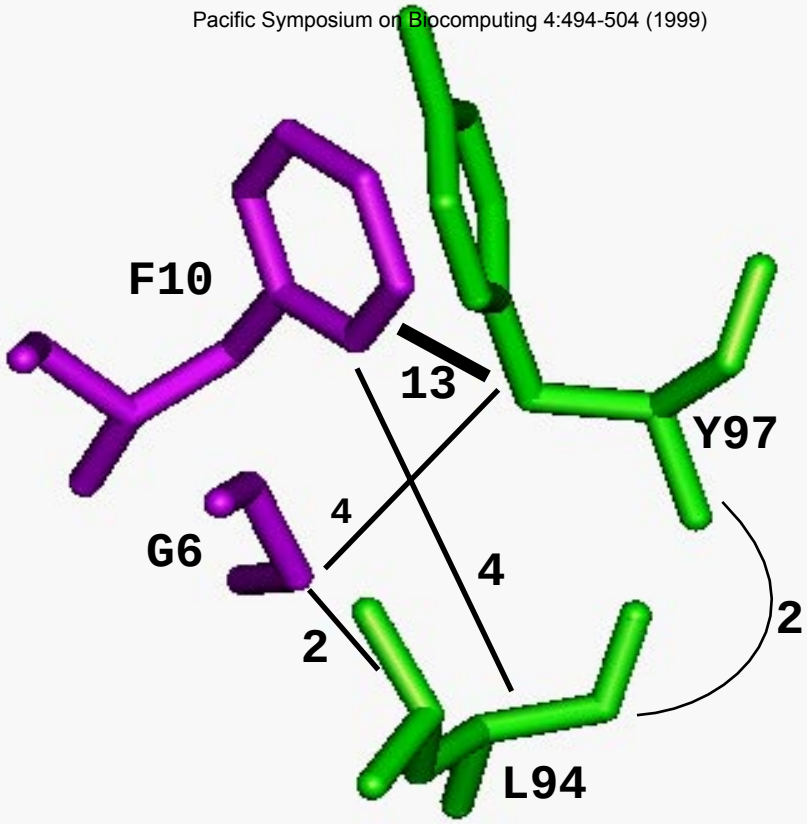


B



Pacific Symposium on Biocomputing 4:494-504 (1999)

A



B