doi:10.1006/jmbi.2001.4554 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) 307, 1419–1426

JMB

Van der Waals Locks: Loop-n-lock Structure of Globular Proteins

Igor N. Berezovsky* **and Edward N. Trifonov**

Department of Structural Biology, The Weizmann Institute of Science, PO Box 26 Rehovot 76100, Israel

**Corresponding author*

In a globular protein the polypeptide chain returns to itself many times, making numerous chain-to-chain contacts. The stability of these contacts is maintained primarily by van der Waals interactions. In this work we isolated and analysed van der Waals contacts that stabilise spatial structures of nine major folds. We suggest a specific way to identify the tightest contacts of prime importance for the stability of a given crystallized protein and introduce the notion of the van der Waals lock. The loops closed by the van der Waals interactions provide a basically novel view of protein globule organization: the loop-n-lock structure. This opens a new perspective in understanding protein folding as well: the consecutive looping of the polypeptide chain and the locking of the loop ends by tight van der Waals interactions.

Keywords: protein structure; closed loops; van der Waals locks; major folds; protein folding

Introduction

The formation of hydrophobic nuclei by ''van der Waals forces drawing together the hydrophobic side-chains of the amino acids'' was theoretically hypothesized already in 1944,[1,2] before the sequencing of the proteins. In this work Bresler & Talmud predicted the existence of globular proteins with a primarily hydrophobic interior and hydrophilic exterior provided by the balance between the amounts of the respective residues.[2] The balance is responsible for the size of the protein globule, as estimated by its surface energy. This conclusion was later confirmed by analysing the first 15 protein crystal structures.[3] Hydrophobic nucleation has been shown to have a central role in many of the subsequent protein structure and folding studies.[4-10]. In examining the folding process, it remains uncertain which of the nuclei in the final spatial structure actually participated in the folding. Moreover it is not clear which of the kinetic nuclei of the folding intermediates survived in the final fold. At the same time, the hydrophobic nuclei of the crystallized structures undoubtedly do contribute substantially if not crucially to the stability of the final fold. There are no studies that we know of that will pinpoint which of the multiple contacts in a given protein globule are the main contributors to its stability. Here we concentrated on exploring the purely structural aspects of van der Waals contacts, namely, where in the folds are the tightest van der Waals contacts located, and is there any order in the organization of such contacts.

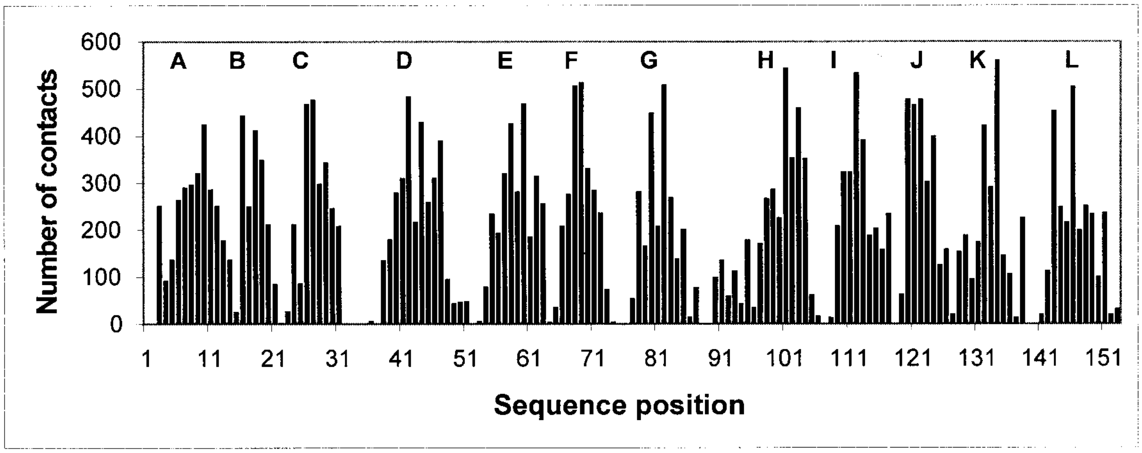
Results

There is no straightforward way to ascribe energy values to the close contacts. The standard approach for the calculation of van der Waals interactions indirectly includes all types of interactions.[11,12] Clearly, van der Waals forces would be a major component. However, polar and ion interactions would also result in the formation of close contacts. Moreover, such forces provide closer distances than van der Waals interactions. Notably, the calculation of van der Waals energy for sites with such contacts would give infinitely repulsive terms, whereas the site as a whole may well be a true tight van der Waals contact. This is why here we have chosen a simpler approach: operating with number of atom-to-atom contacts, rather than with energy integrals of van der Waals and other interactions. In this way we adequately describe tight van der Waals contacts, which indirectly includes the contributions from other interactions as well. We started by calculating the total number of contacts a given residue has with the other atoms of the rest of the protein globule. Detailed analyses of the distribution of the tight atom-to-atom contacts in 3D and along the sequence

E-mail address of the corresponding author:
Igor.Berezovsky@weizmann.ac.il

indicate that the distribution of the contacts is highly non-uniform. Figure 1(a) presents an example of such a distribution along the sequence (Trefoil fold, 1i1b). Here, for every residue along the polypeptide chain, the total number of atom-to-atom contacts this residue makes is calculated.

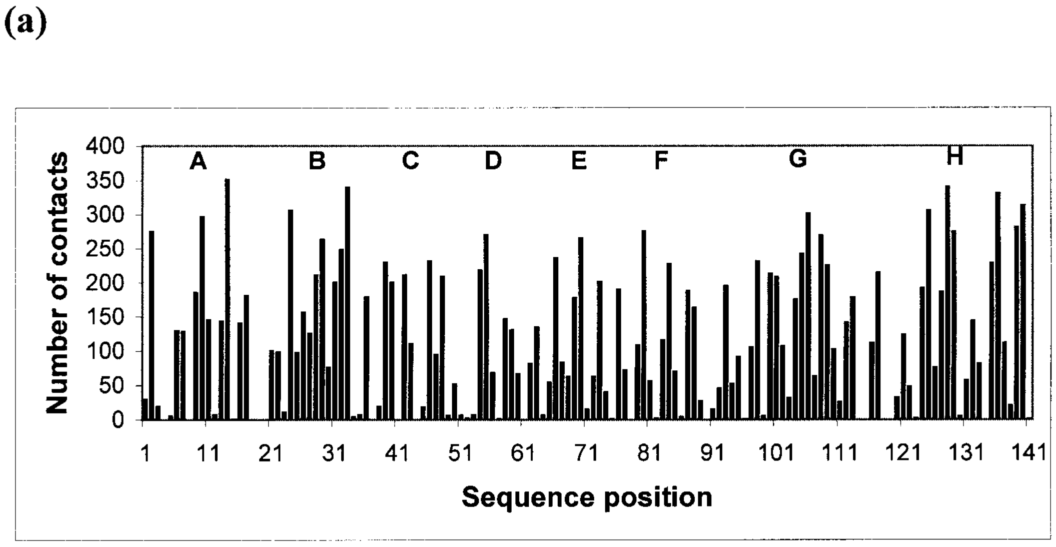
(a)



(b)

	A	B	C	D	E	F	G	H	I	J	K	L
A		139	36	977	78	1	0	152	123	92	0	1266
B			911	319	32	51	12	0	0	114	97	49
C				176	34	92	298	0	0	44	735	1
D					1203	56	1	355	32	0	0	84
E						297	51	913	106	26	0	2
F							1184	533	160	30	47	0
G								158	90	228	326	0
H									1015	92	18	3
I										418	109	499
J											892	522
K												122
L												

Figure 1. (a) An example of the plot of the total number of atom-to-atom contacts every residue of the Trefoil fold makes with other atoms of the rest of the protein globule. The contacts between the nearest five residues along the chain are not included, whereas contacts between residues i and beyond $i+5$ are counted. (b) The matrix of interactions between the multicontact clusters (A to L) for the Trefoil fold (1i1b).



(b)

	A	B	C	D	E	F	G	H
A		114	0	0	592	0	257	620
B			469	475	164	0	1192	0
C				423	0	0	328	0
D					0	0	1	0
E						243	86	432
F							77	914
G								894
H								

Figure 2. The contact density map (a) for the Globin fold (1thb) and the matrix of cluster-to-cluster interactions (b).

Thus, the non-uniformity of the distribution is clearly seen. The site F (residues 65-77) makes many contacts with others; however, at the same time the region 32-37 hardly makes any distant contacts. Typically, the contact density diagrams consist of several distinct clusters of contacts as in Figure 1. The clusters of high contact density in Trefoil fold (1i1b) are A (3-14), B (16-21), C (23-31), D (38-51), E (54-63), F (65-73), G (77-87), H (90-105), I (109-117), J (119-126), K (128-138), and L (141-151). Such diagrams, however, do not indicate which residues interact with a given cluster and contribute to the contact density. This information can be obtained by a separate calculation of the contact matrices. Figure 1(b) shows an example of such a matrix for the above Trefoil fold. The elements of the matrix indicate how many contacts are shared by any two clusters. The element A*D (977 atom-to-atom contacts), for example, corresponds to the interaction of the sites 3-14 and 38-

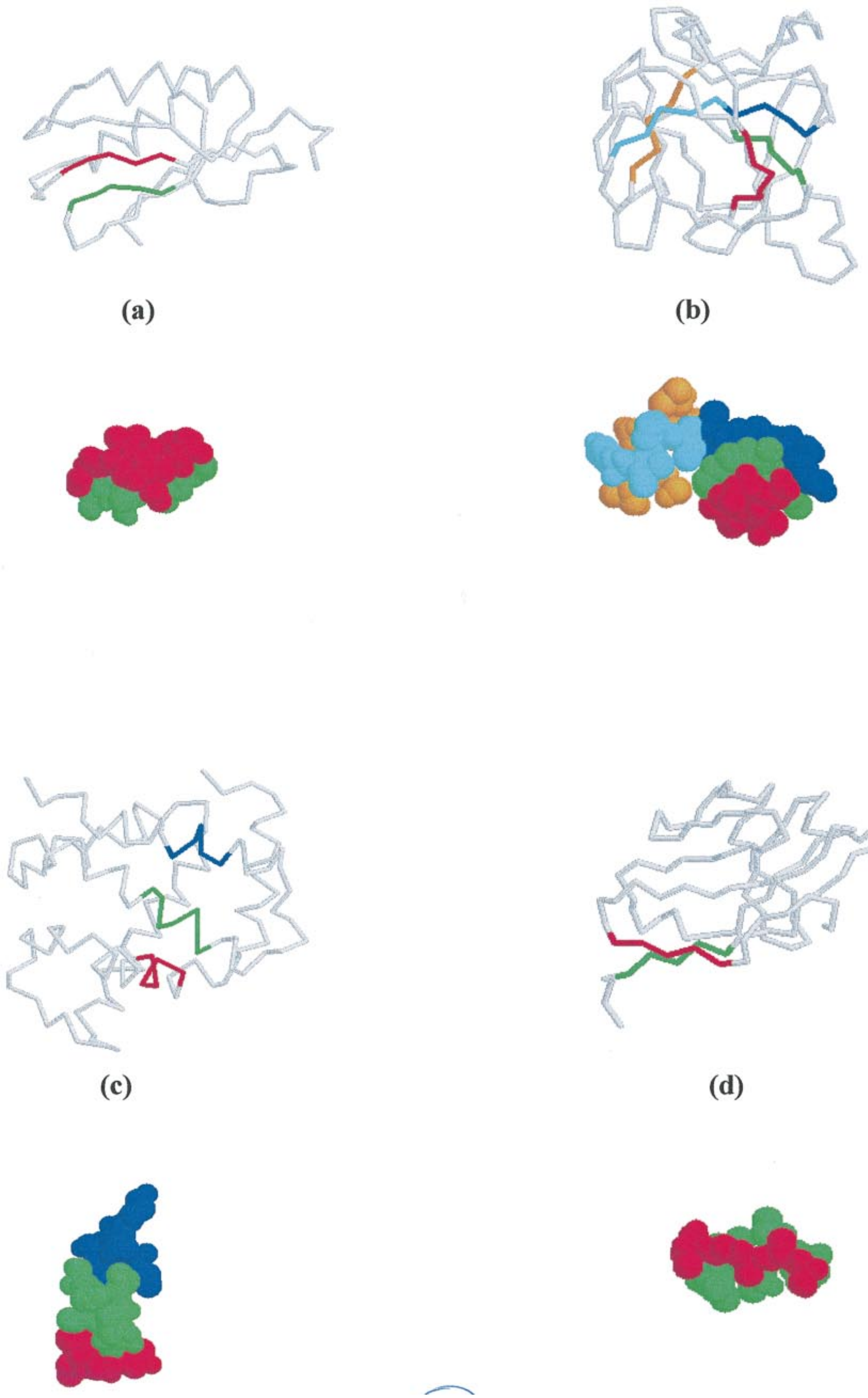


Figure 3 (*legend shown on page 1424*)

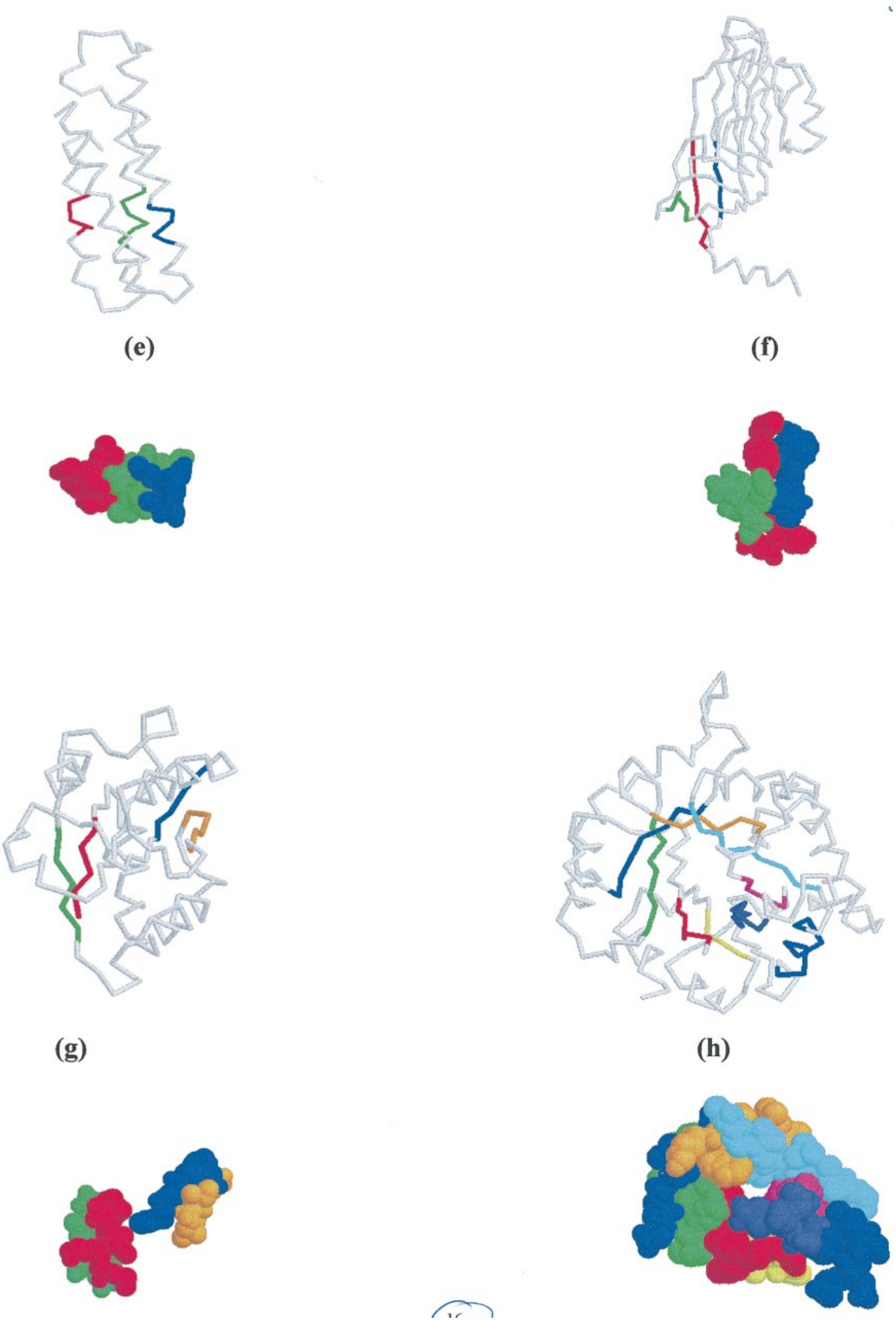


Figure 3 (*legend shown on page 1424*)

51. Several such strong pairs are seen in the matrix. Thus, the matrix comprises major contacts stabilizing the protein molecule. Further analysis conducted with more detailed matrix of contacts (data not shown) specifically indicated those residues that contribute most to every given cluster-to-cluster contact. For every pair of clusters, key sections of three to five amino acid residues can be determined. Such a pair of tightly contacting sections makes what we call a van der Waals lock. The van der Waals lock is thus defined as a pair of short interacting sections of the polypeptide chain locally making a high concentration of atom-to-atom contacts. The locks frequently form associations, bringing together more than two sections of the chain, which we call composite locks. As the matrix in Figure 1(b) suggests, the Trefoil fold contains on the order of 10-15 locks with an appreciable number of atom-to-atom contacts. Another eight superfolds[13] have a similar concentration of tight locks. A second example, with less obvious clusters of contacts, the Globin fold 1thb, is shown in Figure 2. Here, the contact diagram (Figure 2(a)) is subdivided into clusters A (6-17), B (21-36), C (38-50), D (54-60), E (62-77), F (79-89), G (91-113), and H (120-140). In the matrix (Figure 2(b)) the cluster-to-cluster interactions are displayed. Note that the details of subdivision in the clusters are not essential for the final result; it only simplifies the calculation. The locks can be identified directly in the full residue-to-residue matrix of contacts, as for example, 5*5 boxes with the highest scores.

Rather than displaying all these locks, Figure 3 presents several diverse examples of simple and composite locks. Note that in the space-filling presentation, almost every fragment of a protein molecule appears as a tight pack similar to the locks shown in Figure 3. That would be, however, an illusion, since only a few of the contacts in the globule are indeed as tight as those shown in Figure 3. Figure 3(a), αβ Sandwich (1aps), demonstrates van der Waals locks from sites 11-15 (red) and 45-49 (green). The protruding residues Glu12-Val13 "embrace" the site 45-49, whereas residues Tyr11 and Gly15 cover the ends of the segment 45-49. Both parts of the lock belong to β-strands.

The next two illustrations (Figure 3(b),(c)) reflect examples of composite locks where five pieces of β-strands, in the first case, and three pieces of α-helices, in the second one, are involved. In the Trefoil fold (1i1b) in Figure 3(b) a "central bar" is seen, consisting of residues 55-58 (cyan), continued by residues 59-63 (blue). This site provides contacts between sequentially and spatially distant parts of the structure by locking the residues 55-58 (cyan) with 100-104 (orange), and 59-63 (blue) with 40-44 (green), respectively. Residues 40-44, in turn, make contact with the beginning of the chain, residues 6-10 (red). Three α-helical pieces in the Globin fold (1thb, Figure 3(c)) form spatially a "waist" in the middle of the globule, consisting of the parallel α-helical pieces 24-28 (red) to 104-108 (green), and 102-106 (green) to 125-129 (blue), respectively (see also Figure 2). The example of the Immunoglobulin fold (2rhe) shows the van der Waals contacts locking the beginning of the chain (residues 9-13, unstructured chain, red) and its end (106-110, a piece of β-strand, green). Similar to Figure 3(c), Figure 3(e) shows the waist, consisting of three pieces of α-helices for the Up-down fold (256b), where residues 10-14 (red) interact with residues 29-33 (green), and residues 30-34, in turn, make a tight contact with site 72-76 (blue). In both waist cases the side-chains of the contacting residues interpenetrate, thus stabilizing the van der Waals lock. For example, residue Thr108 (green) penetrates the red region near residues Tyr24, Glu27, and Ala28, and residue Leu129 (blue) penetrates the green region, making contact with Ser102, Leu105, and Leu106 in the Globin fold. Similarly, Met33 in the Up-down fold enters the part 10-14 (red) near residue Leu10, as well as Ile72 from the site 72-76 spatially makes contact with residues Leu30, Met33, and Arg34 (green). In the Jelly roll fold (2stv, Figure 3(f)) sites 25-29 and 27-31 (both marked by red), belong to the β-strand and form a "bar", similar to that in the Trefoil fold (see Figure 1(b)). Here, the bar makes tight contact with the adjacent unstructured part of the chain, 57-61 (green) and the end of the β-strand 188-192 (blue), respectively. Interestingly, sites 57-61 and 25-31 have an almost perpendicular orientation, whereas the sites 188-192 and 27-31, which belong to the β-strands 27-38 and 182-193, respectively, are (anti)parallel. Doubly Wound fold (4fxn), presented in Figure 3(g), shows van der Waals locks connecting distant parts of the globule: residues 1-5 (red) and 30-34 (green) in the first lock, and

Figure 3. Diverse examples of simple and composite locks. Backbone presentation of the whole structure and space-filled images of van der Waals locks. Note that the backbone presentation does not show the amino acid residues, whereas the space-filling presentation does. (a) αβ Sandwich (1aps). Sites 11-15 (red) and 45-49 (green). (b) Trefoil (1i1b). Composite lock, comprising five pieces of β-strands: a "central bar" is seen, consisting of residues 55-58 (cyan) continued by residues 59-63 (blue). Other sites involved: 100-104 (orange), 40-44 (green), and 6-10 (red). (c) Globin (1thb). Composite lock of sites: 24-28 (red), 102-108 (green), and 125-129 (blue). (d) Immunoglobulin fold (2rhe). Lock of sites 9-13 (red) and 106-110 (green). (e) Up-down fold (256b). Sites 10-14 (red), 29-34 (green), and 72-76 (blue). (f) Jelly roll fold (2stv). Sites 25-31 (red), 57-61 (green), and 188-192 (blue). (g) Doubly Wound fold (4fxn). Sites 1-5 (red), 30-34 (green), 85-89 (blue), and 115-119 (orange). (h) TIM barrell fold (7tim). Large composite lock of sites: 3-7 (red), 36-42 (green), 58-64 (blue), 88-94 (orange), 121-129 (cyan), 162-166 (magenta), 182-189 (blue), 202-209 (purple), and 226-230 (yellow).

residues 85-89 (blue) and 115-119 (orange) in the second one. Both pairs of sites are parts of the corresponding β-strands. The last example of the TIM barrell fold (Figure 3(h)) is a large composite lock spanning in a circle through the globule from the beginning to the end of the chain. This lock involves nine discontinuous sites: 3-7 (red), 36-42 (green), 58-64 (blue), 88-94 (orange), 121-129 (cyan), 162-166 (magenta), 182-189 (blue), 202-209 (purple), and 226-230 (yellow). These include five elements of β-strands, two α-helices, and two segments of the chain with an undefined secondary structure. All sites excluding the first and the last ones contain extended segments interacting simultaneously with two others: 36-40 and 38-42 (both green) interact with 3-7 (red) and 58-62 (blue); 88-92 and 90-94 (both orange) interact with 60-64 (blue) and 121-125 (cyan); 162-166 (magenta) with 125-129 (cyan) and 182-186 (blue); 202-206 and 205-209 (both pruple) with 185-189 (blue) and 226-230 (yellow). The space-filling view of this composite lock clearly shows the ring-like core, apparently the major stabilizing structure of the fold.

Discussion

The traditional description of a protein structure is dominated by its elements of secondary structure. These elements definitely contribute to the protein's stability, not least by their involvement in the locks. In addition, many of the tight contacts stabilizing the protein globule involve structurally undefined sections as well, which should not be ignored. Thus, the protein globule can be viewed as a multitude of returning loops of the polypeptide chain closed by the chain-to-chain contacts, van der Waals locks: a loop-n-lock structure. Note that the ''closed loops'' mean the loops closed by the end-to-end contacts, rather than the term ''loops'', in its traditional use in protein science, as connectors between various elements of secondary structure.[14-17]

The locks can be formed by a large variety of combinations of amino acid residues. A comparison of related folds with high and low sequence identity (data not shown) indicates both conservation and high variability of the lock sequences, whereas the loop end positions are very much conserved.

The loop fold nature of a protein also suggests a simple straightforward scenario for protein folding: primary loops of standard size 25-30 amino acid residues[18] are formed consecutively during protein's synthesis, cotranslational folding, and stabilized by the end-to-end locking. The secondary (more distant) contacts are formed as well, either simultaneously or subsequently. The detection of the locks in general and the detection of the primary locks in particular are perhaps the most important aspects of studying protein structure and folding.

Thus, more extensive studying of the locks and their interactions is required. In particular, we expect that extended composite lock structures would be frequently serving as construction force elements of the overall structure. Examples of such force elements are presented in the Trefoil fold (see Figure 2(b), blue-cyan bar), in the Jelly roll fold (Figure 2(f), red bar), and in the TIM barrell fold (Figure 2(h), ring of nine elements).

Site-directed mutagenesis[19] of the sequences involved in identified locks would be highly promising in studies on the protein's stability, folding, and protein design.

Materials and Methods

Nine protein folds (Globin (1thb), Trefoil (1i1b), Up-down (256b), Immunoglobulin (2rhe), αβ Sandwich (1aps), Jelly roll (2stv), Doubly Wound (4fxn), UB αβ roll (1ubq), and TIM barrell (7tim)), selected[13] as major representatives, were analysed. The total number of contacts for each amino acid residue (see, for example Trefoil fold (1i1b), Figure 1(a)) included all contacts of the atoms of this residue. Only the contact distances between 2.5 and 5.0 Å were considered[11]. The numbers of contacts were calculated for atoms belonging to residues separated by at least five amino acid residues along the polypeptide chain. That is, residues which are nearest along the chain were not considered.[12] From the contact plots, the clusters with the highest number of contacts were selected and their connectivity established. Typical matrices of the interactions between the multicontact clusters (marked by letters) are presented for the Trefoil fold in Figure 1(b) and for the Globin fold in Figure 2(b). The next step was a detailed analysis of the clusters with the aim of determining key consecutive residues in each site, giving the maximal number of contacts (typically three to five amino acid residues). Each pair of these interacting sites was considered as a van der Waals lock (for the full definition of the van der Walls locks, see Results).

Acknowledgements

I.N.B. is a post-doctoral Fellow of the Feinberg Graduate School at the Weizmann Institute of Science.

References

1. Bresler, S. E. & Talmud, D. L. (1944). The nature of globular proteins. *Comp. Rend. Acad. Sci. URSS*, **43**, 310-314.
2. Bresler, S. E. & Talmud, D. L. (1944). A few consequences of the new hypothesis. *Comp. Rend. Acad. Sci. URSS*, **43**, 349-350.
3. Chothia, C. (1975). Structural invariants in protein folding. *Nature*, **254**, 304-308.
4. Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, **261**, 552-558.
5. Sneddon, S. F. & Tobias, D. J. (1992). The role of packing interactions in stabilizing folded proteins. *Biochemistry*, **31**, 2842-2846.
6. Pande, V. S., Grosberg, A. Yu., Tanaka, T. & Rokhsar, D. S. (1998). Pathways for protein folding:

is a new view needed? *Curr. Opin. Struct. Biol.* **8**, 68-79.

7. Poupon, A. & Mornon, J. P. (1998). Populations of hydrophobic amino acids within protein globular domains: identification of conserved "topohydrophobic" positions. *Proteins: Struct. Funct. Genet.* **33**, 329-342.
8. Poupon, A. & Mornon, J. P. (1999). Predicting the protein folding nucleus from sequences. *FEBS Letters,* **452**, 283-289.
9. Fersht, A. R. (2000). Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl Acad. Sci. USA,* **97**, 1525-1529.
10. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (2000). Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol.* **296**, 1183-1188.
11. Berezovsky, I. N., Namiot, V. A., Tumanyan, V. G. & Esipova, N. G. (1999). Hierarchy of the interaction energy distribution in the spatial structure of globular proteins and the problem of domain definition. *J. Biomol. Struct. Dynam.* **17**, 133-155.
12. Berezovskii, I. N. & Tumanyan, V. G. (1995). Objective method for isolating the domains of globular proteins. *Biophysics (Biofizika, Moscow),* **40**, 1181-1187.
13. Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature,* **372**, 631-634.
14. Leszczynski, J. F. & Rose, G. D. (1986). Loops in globular proteins: a novel category of secondary structure. *Science,* **234**, 849-855.
15. Martin, A. C. R., Toda, K., Stirk, H. J. & Thornton, J. M. (1995). Long loops in proteins. *Protein Eng.* **8**, 1093-1101.
16. Kwasigroch, J. M., Chomilier, J. & Mornon, J. P. (1996). A global taxonomy of loops in globular proteins. *J. Mol. Biol.* **259**, 855-872.
17. Oliva, B., Bates, P. A., Querol, E., Aviles, F. X. & Sternberg, M. J. E. (1997). An automated classification of the structure of protein loops. *J. Mol. Biol.* **259**, 814-830.
18. Berezovsky, I. N., Grosberg, A. Y. & Trifonov, E. N. (2000). Closed loops of nearly standard size: common basic element of protein structure. *FEBS Letters,* **466**, 283-286.
19. Matouschek, A. & Fersht, A. R. (1991). Protein engineering in analysis of protein folding pathways and stability. *Methods Enzymol.* **202**, 82-112.

Edited by J. Thornton

(Received 12 December 2000; received in revised form 16 February 2001; accepted 16 February 2001)