Representation of amino acid sequences in terms of interaction energy in protein globules

Igor N. Berezovsky*, Vladimir G. Tumanyan, Natalia G. Esipova

Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 117984, Russia

Received 25 August 1997; revised version received 6 October 1997

Abstract We suggest a new simple approach for comparing the primary structure of proteins and their spatial structure. It relies on the one-to-one correspondence between each residue of the polypeptide chain and the energy of van der Waals interactions between the regions of the native globule flanking this residue. The method obviates the sophisticated geometrical criteria for estimating similarity between spatial structures. Besides, it permits one to analyze structural units of different scale.

Key words: Hierarchy; Amino acid sequence; Protein spatial structure; van der Waals interaction; Ribonuclease

1. Introduction

The comparison of protein structures at different levels of the hierarchic organization of the globule is a significant problem of structural biology [1–4]. This problem requires understanding of the structural properties of proteins in relation to their primary structures [5–9]. In particular, one has to formulate the physical criteria to delineate energy-independent parts of the spatial structure [10–13]. This would make it possible to establish the hierarchy of structural units of different scale in the protein globule [14,15]: regions with high energy content [12], modules, subdomains, domains [16,17]. This system of terms is applied in subsequent analysis of spatial structure at different levels of protein globule hierarchy, as well as in the case of multiglobular structures.

The aim of this work is to describe any protein amino acid sequence by a system of energy parameters suitable for graphical comparison of the protein spatial structures, delineation of the domain (module) structure, and detection of similar regions in various proteins. This may provide a basis for comparisons of spatial and primary structures, locations of insertions and deletions, and analysis of standard motifs which are common to different proteins. We have performed a computational analysis of the distribution of van der Waals interaction energy in the spatial structure of proteins. A novel approach (see Section 2) to estimating the interaction energy allows us to sort out interactions taking place at different hierarchy levels. Most of the current approaches to the analysis of the protein globule in the structural and energetic aspects require preliminary subjective information such as direct visual analysis or sophisticated parameters chosen a priori [16,17]. We try to overcome these shortcomings by developing an obvious simple method. This method requires only one parameter (barrier value, for details see Section 2) which may be varied throughout the entire range of its possible values. Moreover, by varying this parameter one can observe a different scale of structural and energetic hierarchy. Actually, we subdivide the sequence into regions corresponding to the elements of the hierarchy and perform an analysis of these regions independently of one another. The method is applied to nine known 3D structures of various ribonucleases. Two of them, bovine pancreatic RNase A (3rn3; here and below the code in the Brookhaven Protein Data Bank is given) and bovine seminal fluid RNase BS (1bsr), are close structural homologues. The second pair of close structural homologues are bacterial RNases of *Bacillus amyloliquefaciens* (barnase) and *Bacillus intermedius* 7P (binase). The third pair is composed of RNases H from *Escherichia coli* (2rn2) and *Thermus thermophilus* (1ril). The last group is a triplet of homologues: RNase MS from *Aspergillus saitoi* (1rms), RNase F1 from *Fusarium moniliforme* (1fus), and RNase T1 from *Aspergillus oryzae* (1rls). Analysis of the sets of close structural homologues gives us a possibility to observe regions of spatial similarity corresponding to specific motifs of primary structure. On the other hand, we can see regions of primary structure which are responsible for differences between the spatial structures under comparison.

2. Materials and methods

The programs used in the calculations were written in Borland C and run on an IBM/PC Pentium-133.

The essence of the approach is the idea of one-to-one correspondence between each residue of the polypeptide chain and the energy of van der Waals interactions between the regions of the native globule flanking this particular amino acid. The energy of interactions is calculated with the Lennard-Jones 6–12 potential [18], using the Sheraga parametrization [19]. The energy of the paired interactions is calculated for atoms belonging to residues separated by at least two amino acid residues. This is sufficient [11,12] for estimating the interactions between the adjacent regions of the globule.

The following procedure is aimed at defining the hierarchy of sequence fragments corresponding to the particular structural regions of the protein globule. First, we plot the curve of interaction energy between the regions of the native globule flanking each residue. Second, the minimum of interactions between parts of the globule is assumed to correspond to the local maximum on the curve. The null value of the interaction energy implies complete energetic independence of the adjacent regions from each other. We assume that a particular maximum coincides with the point of minimal interactions at the corresponding hierarchical level. The point of minimal interaction is determined using the value which we call 'potential barrier'. If the differences between some maximum on the energy interaction curve and the previous and next minima exceed a fixed threshold, then this maximum corresponds to the point of minimal interactions (at the particular level of hierarchy). 'Barrier' values are normalized by the value of the global minimum (E_0) on the first curve. Since the total energy of the compared proteins may differ, we introduce the following coefficient K for the value of the barrier: B2 = KB1, K = Min2/Min1. Let us compare protein 1 against protein 2. In this

*Corresponding author. Fax: (7) (095) 1351405.
E-mail: ber@imb.imb.ac.ru

PII S0014-5793(97)01346-X

case, Min1 and Min2 are the values of global minima on the first curve corresponding to protein 1 and protein 2, respectively. B1 and B2 are the respective values of the potential barriers. Several iterations for the set of barrier values of the minimal energy are performed in the following manner. We assign a set of barrier values: $0.3E_0$, $0.25E_0$, $0.2E_0$, $0.15E_0$, $0.1E_0$, and $0.05E_0$ of global minimum on the first curve (E_0) for the protein with the lowest global minimum (protein 1). Then, for protein 2 the barrier values are: $K0.3E_0$, $K0.25E_0$, $K0.2E_0$, $K0.15E_0$, $K0.1E_0$, and $K0.05E_0$, respectively. All data in Table 1 and Figs. 1 and 2 are presented only for the barriers which permit dividing the sequence into energy-independent parts at the given hierarchical level. Then, the amino acid residues coinciding with a certain minimal value of the interaction energy are defined as boundaries of the amino acid sequence regions. In addition the values of the interaction energy for isolated regions of the sequence and the interaction energy between them are computed.

3. Results

Figs. 1 and 2 are the graphical representation of two sets of structural homologues: RNases H and RNases MS, F1, and T1. In Table 1 the results are summarized for the two other sets of homologues: RNases A and BS, and binase and barnase. Calculations are performed for different values of barriers, corresponding to various levels of hierarchy.

The similarity of regions (1–36) versus (6–40) in RNases H from *E. coli* and *T. thermophilus* is evident (Fig. 1), as well as for regions (62–121) versus (67–125) (some difference in the region (121–end) may be explained by the absence of the 19 C-terminal residues from the X-ray structure). RNases MS, F1, and T1 exhibit similarities mainly in regions (1–30(31), (32(33)–58(57)), and (90–104(105,106)) (Fig. 2). Regions of similarity for RNase A and RNase BS, as one can see from Table 1, include residues (77–107) versus (83–107), and (109–124) versus (109–124). Subsequences (1–21), (23–32), and (55–110) versus (1–20), (22–32), and (54–110) are similar for barnase versus binase. Fragment 109–124 is common to both RNases A and BS for all values of barriers (Table 1). Similarly, fragment 55–110 (54–110) is common to binase and barnase. Only at the level with the lowest barrier value ($0.05E_0$; here and below barrier value for the protein with the lowest global minimum is presented, see Section 2) it splits into 55–83 and 85–110 in barnase and 54–94, 96–102, 104–110 in binase. Similarities in RNases MS, F1, and T1 are found

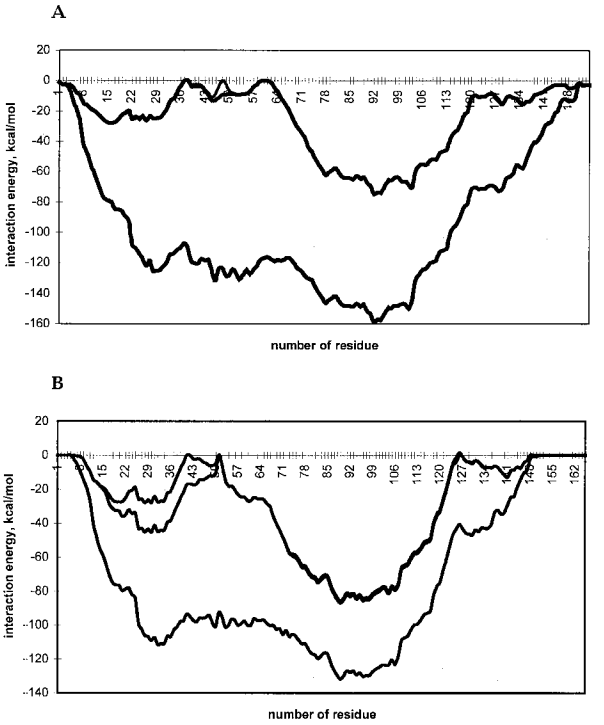

Fig. 1. Contribution of individual regions to the interaction energy. A: RNase H from *E. coli* (2rn2); (B) RNase H from *T. thermophilus*. The lower curve presents the interaction energy between regions of the native globule flanking each residue. A particular maximum coincides with the point of minimal interactions. Other curves correspond to energy hierarchical levels and were produced by varying of the 'potential barrier' value ($0.1E_0$, $0.05E_0$ for RNase H from *E. coli* and $K0.15E_0$, $K0.05E_0$ for RNase H from *T. thermophilus*, K = 0.83). Clear similarity of regions (1–36) and (62–121) versus (6–40) and (67–125), respectively, can be observed.

for the following subsequences: (i) (1–30), (32–58),(60–88), and (90–106(104)) in the F1 and T1 (barrier value $0.1E_0$); (ii) (33(32)–57(58)) and (90–105(106,104)) in MS, F1, and T1 (barrier value $0.05E_0$); (iii) (1(3)–31(30)) in MS and T1, as well as (60–72) and (74–88) in F1 and T1 (barrier value $0.05E_0$). Thus, the data presented reveal fragments of primary structure which correspond to differences between compared

Table 1
Regions of primary structure defined by the minima of the interaction energy

B1	Ribonuclease A	Ribonuclease BS K=0.85	Barnase	Binase K=0.9
25	–	–	–	**1–52, 54–110**
20	–	1–12, 14–107, **109–124**	**1–53, 55–110**	**1–52, 54–110**
15	1–107, **109–124**	1–12, 14–69, 71–107, **109–124**	**1–53, 55–110**	**1–20, 22–52, 54–110**
10	1–47, 49–107, **109–124**	1–12, 14–69, 71–73, 75–107, **109–124**	**1–21, 23–53, 55–110**	**1–20, 22–52, 54–110**
5	1–40, 42–47, 49–69, 71–75, **77–107, 109–124**	1–12, 14–32, 34–40, 42–59, 61–69, 71–73, 75–81, **83–107, 109–124**	**1–21, 23–32, 34–53**, 55–83, 85–110	**1–20, 22–32, 34–52**, 54–94, 96–102, 104–110

RNase A versus RNase BS and barnase versus binase are compared. B1 are the values for determining the hierarchical levels in the protein with the deepest global minimum (E_0, for details of the procedure see Section 2). To compare hierarchical levels in the RNase A versus RNase BS and in the barnase versus binase, B2 was calculated (see Section 2). The K values for the comparison are given in the first row. Similar regions of primary structure in the compared proteins are shown in bold.

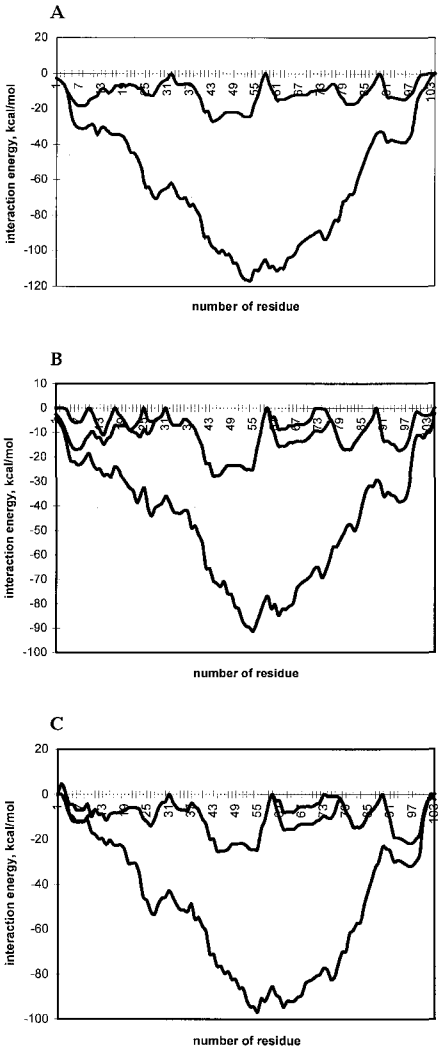

Fig. 2. Contribution of individual regions to the interaction energy. A: RNase MS from *A. saitoi* (1rms); B: RNase F1 from *F. moniliforme* (1fus); C: RNase T1 from *A. oryzae* (1rls). Barrier which permit dividing the sequence into energy-independent parts are $0.05E_0$ for RNase MS and $K0.1E_0$, $K0.05E_0$ for F1 ($K=0.78$) and T1 ($K=0.83$). Subsequences (32(33)–58(57)) and (90–105(106,104)) of RNases MS, F1, T1 correspond to the fragments of energy curves with similar behavior at all hierarchical levels. Subsequence (1–30) of RNase F1 differs from the same regions in the RNase MS and T1 at the hierarchical level with lowest barrier value ($0.005E_0$), and subsequence (59–88) of RNase MS differs from regions (60–72) and (74–88) in RNases F1 and T1.

spatial structures. Obviously, the influence of barrier values may be traced in this case.

4. Discussion

In this work we suggest plotting the interaction energy against the amino acid sequence. Thus, we recognize regions of the primary structure corresponding to the parts of the spatial structure with high energy capacity (minima on the curves) and boundaries between these parts (maxima on the curves). This approach reflects some features of the protein spatial structures and their thermodynamic properties. One can see that the shape of the curves (Figs. 1 and 2) and the location of the boundaries between regions of primary structure (Table 1) reflect the structural homology of proteins.

The existence of substitutions in homologues leads to varying subdivisions of primary structure, as follows, for example, from the analysis of RNases A and BS (Table 1). These data agree with the fact that more than half of the substitutions are in regions 16–20 and 28–40. As to the data for binase and barnase presented in Table 1, we may conclude that the delineated segments show a striking correspondence to cooperative units defined in microcalorimetric experiments [20]. Besides, our results are in general accordance with the modular structure (1–24, 25–52, 53–73, 74–88, 89–98, and 99–110) revealed on the basis of the so-called centripetal profile approach [21] and studied in biochemical experiments [22,23]. As one can see, Table 1 presents fairly similar patterns of the primary structure division.

In addition, the interaction energy curves can be used as an approach to alignment by matching similar parts of curves and points of null energy between them. Thus from the comparison of Figs. 1A and 1B one can suggest a shift of the curve in Fig. 1B to the left leading to greater similarity of the energy graphic representations. As a result, 21 of the first 36 amino acids are identical (five of the remaining 15 positions contain physically similar residues). It is easy to mark regions 1–31(32) and 32(33)–58(57), (90–105(106,104)) for RNases MS, F, and T1, as well as regions (60–72) and (74–88) for RNases F1 and T1 (see Fig. 2). These regions can be used for alignment based on similarity of the corresponding parts of the energy curve. The described regions of primary structure constitute a system of segments of secondary structure which interact to form the skeleton of the RNase T1 protein fold: the α-helix revealed in the X-ray experimental structure [24] corresponds to subsequence 1–32; β-strands 39–42, 56–59 and 76–80, 89–91 in the X-ray structure belong to delineated regions of primary structure 32–58, 74–88 and 89–91, respectively (see Fig. 2C).

Besides, it is interesting to analyze the results for RNase H from *E. coli* and RNase H from *T. thermophilus* (Fig. 1). The main trough on the curve (residues 48–121) is deeper in the protein from the thermophilic organism as compared to the mesophilic analogue. This should be a result of exchange of smaller residues in the mesophilic protein for bulky residues in the thermophilic protein. Moreover, one can see the correspondence between regions of primary structure detected by our approach and the general protein fold. The major domain RNase H from *E. coli* observed in [25] comprises the following regions of secondary structure: α-helices 43–58, 71–79, 100–112, 127–141 and β-strands 5–14, 19–28, and 31–39. Region 1–36 (Fig. 1A) corresponds to tightly packed β-strands 5–14, 19–28 and 31–39; region 38–60 to the large α-helix 43–58 in the X-ray structure; region 60–155 consists of α-helices 71–79, 100–112, and 127–141.

Finally, this approach is, in fact, a kind of projection of the 3D structure of a globular protein on the 1D structure of its sequence. This method allows us to determine the regions of primary structure corresponding to the parts of the spatial structure with high energy content. Thus, one can compare these regions of primary structure at different hierarchical levels. This numerical sequence of interaction energy values corresponding to the amino acid sequence can be used as an input for further formal procedures.

Acknowledgements: The authors are grateful to G. Frank, A. Galkin, M. Gelfand, V. Ivanov, and L. Kisselev for critical reading of the

manuscript and valuable comments. We are pleased to thank D. Hartley (National Institute of Diabetes and Gastric Disorders, Bethesda, MD, USA) for making available for us the coordinates of barnase and binase.

References

[1] Chothia, C. (1992) Nature 357, 543–544.
[2] Orengo, C.A., Jones, D.T. and Thornton, J.M. (1994) Nature 372, 631–634.
[3] Russell, R.B. and Barton, G.J. (1994) J. Mol. Biol. 244, 332–350.
[4] Pawlowski, K., Bierzynski, A. and Godzik, A. (1996) J. Mol. Biol. 258, 349–366.
[5] Lesk, A.M. and Chothia, C. (1980) J. Mol. Biol. 136, 225–270.
[6] Chothia, C. and Lesk, A.M. (1986) EMBO J. 5, 823–826.
[7] Chelvanayagam, G., Roy, G. and Argos, P. (1994) Protein Eng. 7, 173–184.
[8] Hinds, D.A. and Levitt, M. (1996) J. Mol. Biol. 258, 201–209.
[9] Russell, R.B., Saqi, M.A.S., Sayle, R.A., Bates, P.A. and Sternberg, M.J.E. (1997) J. Mol. Biol. 269, 423–439.
[10] Maiorov, V.N. and Crippen, G.M. (1992) J. Mol. Biol. 227, 876–888.
[11] Berezovsky, I.N. and Tumanyan, V.G. (1995) Biofizika (Moscow) 40, 1181–1187.
[12] Berezovsky, I.N., Esipova, N.G. and Tumanyan, V.G. (1997) Biofizika (Moscow) 42, 567–576.
[13] Shoemaker, B.A., Wang, J. and Wolynes, P.G. (1997) Proc. Nat. Acad. Sci. USA 94, 777–782.
[14] Crippen, G.M. (1978) J. Mol. Biol. 126, 315–332.
[15] Rose, G.D. (1979) J. Mol. Biol. 134, 447–470.
[16] Siddiqui, A.S. and Barton, G.J. (1995) Protein Sci. 4, 872–884.
[17] Islam, S.A., Luo, J. and Sternberg, M.J.E. (1995) Protein Eng. 8, 513–525.
[18] Dunfield, L.G., Burgess, A.W. and Sheraga, H.A. (1978) J. Phys. Chem. 24, 2609–2616.
[19] Nemethy, G., Pottle, M.S. and Sheraga, H.A. (1983) J. Phys. Chem. 87, 1883–1887.
[20] Protasevich, I.I., Platonov, A.L., Pavlovsky, A.G. and Esipova, N.G. (1987) J. Biomol. Struct. Dyn. 4, 885–893.
[21] Go, M. (1983) Proc. Nat. Acad. Sci. USA 80, 1964–1968.
[22] Yanagawa, H., Yoshida, K., Torigoe, C., Park, J.-S., Sato, K., Shirai, T. and Go, M. (1993) J. Biol. Chem. 268, 5861–5865.
[23] Yoshida, K., Shibata, T., Masai, J., Sato, K., Noguti, T., Go, M. and Yanagawa, H. (1993) Biochemistry 32, 2162–2166.
[24] Hakoshima, T., Itoh, T., Tomita, K., Goda, K., Nishikawa, S., Morioka, H., Uesugi, S., Ohtsuka, E. and Ikehara, M. (1992) J. Mol. Biol. 223, 1013–1028.
[25] Katayanagi, K., Miyagawa, M., Matsushima, M., Ishikawa, M., Kanaya, S., Nakamura, H., Ikehara, M., Matsuzaki, T. and Morikawa, K. (1992) J. Mol. Biol. 223, 1029–1052.