Eur Biophys J (2000) 29: 7–16

ARTICLE

Zhen-Yu Xuan · Lun-Jiang Ling · Run-Sheng Chen

A new method for protein domain recognition

Received: 9 November 1998 / Revised version: 20 December 1999 / Accepted: 20 December 1999

Abstract A fuzzy cluster method is presented to recognize protein domains. This algorithm can identify domains globally. A protein structure set was used to test the algorithm. Among 219 proteins, 66.7% yielded results that agreed with the reference definitions, 30.6% showed minor differences, and only 2.7% (six proteins) showed major differences with the reference. The new method is more than 20 times fast than previous algorithms.

Key words Fuzzy cluster analysis · Protein domain · Domain database · Protein structure

Introduction

The three-dimensional (3D) structure of a protein is crucial for understanding its precise function. It is also important for drug design and other biotechnological applications. To date, there are more than 8000 3D structures in the Protein Data Bank (PDB), and this number is increasing with an exponential speed. It is important to analyze these structures and collect useful information for predicting 3D structures of proteins.

Although there still is no strict and widely accepted definition of a domain, the concept of the domain has been widely used to simplify and classify protein structures. Many algorithms have been presented based on different definitions of a domain, such as distance-mapping (Liljas and Rossmann 1974; Nichols and Rose 1995), clustering (Crippen 1978), minimization of interface area (Wodak and Janin 1981; Janin and Wodak 1983), minimization of specific volume (Lesk and Rose 1981), maximization of compactness (Zehfus and Rose 1986; Zehfus 1987, 1993, 1994), and use of a cutting plane (Rose 1979). However, some of these methods can only deal with a single segment (continuous) domain, and some are too computationally expensive. Siddiqui and Barton (1995) applied a more effective algorithm to identify domains, but like most others this method is also a residue-by-residue cutting search procedure. Holm and Sander (1994) presented a global measure to identify domains in protein structures. By using matrix translation and calculating the oscillation time τ, they could recognize domains very quickly. However, many of the domains' definitions disagree with those found in the literature.

Liljas and Rossmann (1974) suggested that a domain has many short residue-residue distances within itself, but few with the rest of the protein. Based on this, here we present a new method using the "similarity of contact environment" (explained below) to build a relationship among residues, then identify domains by means of fuzzy mathematics.

Methods

For identifying a domain automatically we defined a vector $\boldsymbol{E}_i = (\boldsymbol{e}_{i1}, \boldsymbol{e}_{i2}, \boldsymbol{e}_{i3}, \ldots)$ for the ith residue, to show the distance relationship between the ith residue and each of all the residues. Here

$$\boldsymbol{e}_{ij} = \begin{cases} 0 & \text{if distance between the } i\text{th residue} \\ & \text{and the } j\text{th residue} > \text{BORD} \\ 1 & \text{if distance between the } i\text{th residue} \\ & \text{and the } j\text{th residue} \leq \text{BORD} \end{cases}$$

If $\boldsymbol{e}_{ij} = 0$, it means that the distance between the ith and jth residue is too large and no contact exists; if $\boldsymbol{e}_{ij} = 1$, it means that there exists one contact between the ith and the jth residue (see Fig. 1a). We simply use the distance between two C^{α} atoms to represent the distance of the

Z.-Y. Xuan · L.-J. Ling · R.-S. Chen (✉)
Protein Engineering Laboratory, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, P.R. China
e-mail: chenrs@sun5.ibp.ac.cn

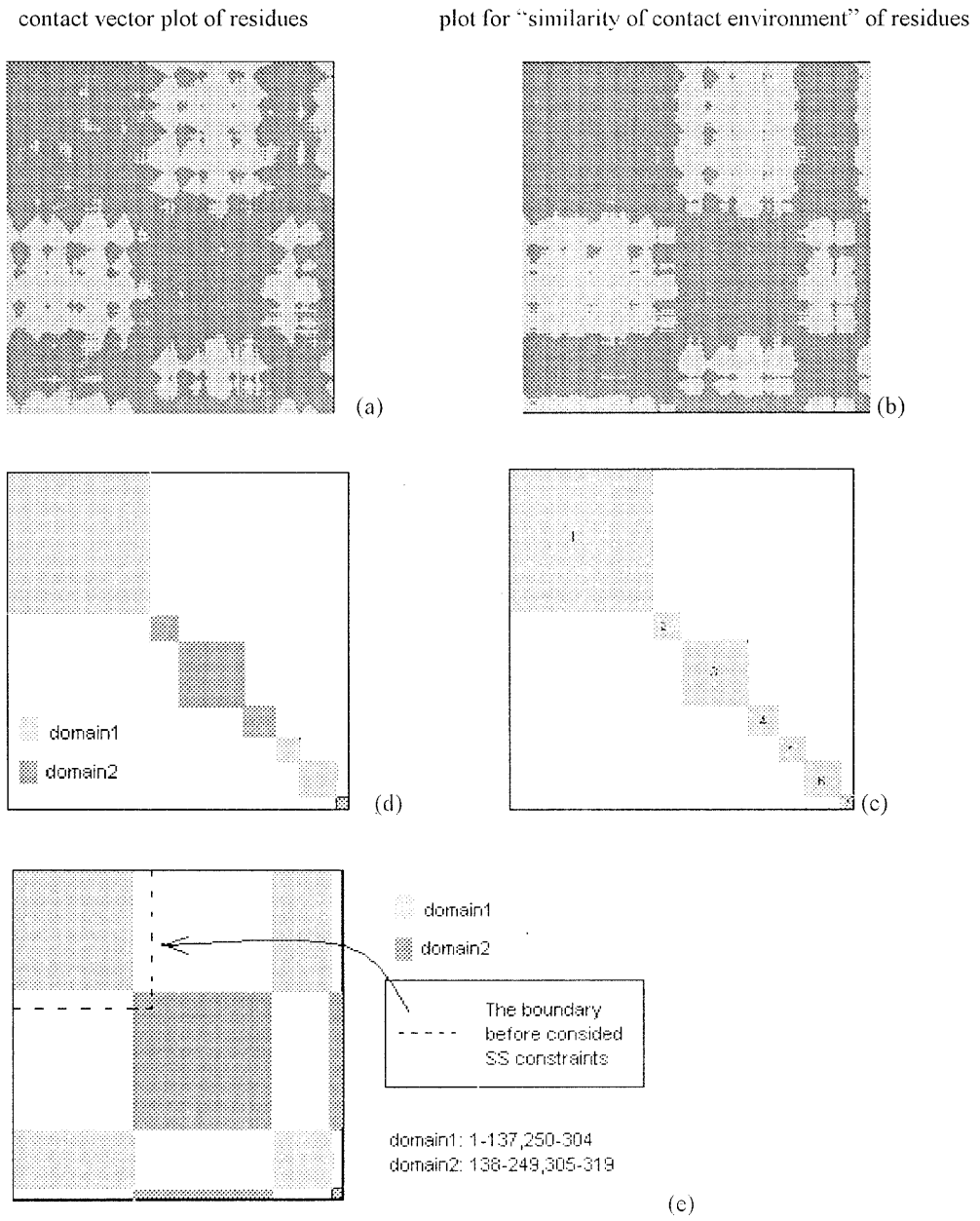

Fig. 1a–e Step-by-step procedure for identifying domain (1pfkα-chain). **a** In the contact vector plot of residues, each line (or column) is a vector, composed of 0 and 1, to show the contact relationship among one residue and others (*light*: no contact, 0; *dark*: existing contact, 1). **b** The plot for "similarity of contact environment" of residues; each line (or column) is a vector, which is composed of 0 and 1, to show the similarity of contact environment between one residue and others (*light*: similarity < CUTOFF, 0; *dark*: similarity ≥ CUTOFF, 1). **b** → **c** Clustering and forming fragments. **c** → **d** Gathering and marking all fragments **d** → **e** Assembling each fragment to obey all constraints and identify the domains

two corresponding residues in order to calculate quickly. BORD is a variable parameter used to define whether there exists contact between two residues (its value is be discussed below).

In order to describe the residue's environment, the concept of "similarity of contact environment" of residues is introduced. This concept is expressed as a vector $\boldsymbol{S}_i = (\boldsymbol{s}_{i1}, \boldsymbol{s}_{i2}, \boldsymbol{s}_{i3}, \ldots, \boldsymbol{s}_{ij}, \ldots)$ where $i = 1, 2, 3 \ldots$ and $j = 1, 2, 3 \ldots$:

$$\boldsymbol{s}_{ij} = \sum_{k=1}^{\mathrm{AANUM}} \delta(\boldsymbol{e}_{ik}, \boldsymbol{e}_{jk})/\mathrm{AANUM} \quad 0 \le \boldsymbol{s}_{ij} \le 1.0 \qquad (1)$$

where

$$\delta(\boldsymbol{e}_{ik}, \boldsymbol{e}_{jk}) = \begin{cases} 0 & \text{if } \boldsymbol{e}_{ik} \ne \boldsymbol{e}_{jk} \\ 1 & \text{if } \boldsymbol{e}_{ik} = \boldsymbol{e}_{jk} \end{cases}$$

AANUM is the total number of residues in one peptide chain.

Each vector $\boldsymbol{S}_i$ shows the similarity of contact environment between the residue and each of the other residues. So, these vectors include almost all the structural information of one peptide chain. In this way a domain is an assembly of residues which have a similar contact environment. The next step is how to recognize the domains from these vectors. We do this according to following steps: (1) collecting and forming fragments; (2) marking the fragments; (3) dynamic assembly.

Collecting and forming fragments

For simplifying the algorithm, we first converted the float vector $\boldsymbol{S}_i$ to an integer vector of "similarity of

contact environment" $\boldsymbol{N}_i = (\boldsymbol{n}_{i1}, \boldsymbol{n}_{i2}, \boldsymbol{n}_{i3} \ldots)$ which is composed of 0 and 1:

$$\boldsymbol{n}_{ij} = \begin{cases} 0 & \text{if } \boldsymbol{S}_{ij} \leq \text{CUTOFF} \\ 1 & \text{if } \boldsymbol{S}_{ij} > \text{CUTOFF} \end{cases}$$

Figure 1b shows all the $\boldsymbol{n}_{ij}$ of one protein chain. Then some residues, which are adjacent in sequence and have same similarity of contact environment, are collected together to form one fragment. This means that the residues from n_1 to n_2, whose $\boldsymbol{n}_{ij} (n_1 \leq i \leq n_2, n_1 \leq j \leq n_2)$ all equal 1, can be collected on to the same fragment. After this, the residues in the peptide chain are collected to produce many fragments (see Fig. 1c).

The "similarity of contact environment" of each fragment is represented as a vector $\boldsymbol{W}_l = (\boldsymbol{w}_{l1}, \boldsymbol{w}_{l2}, \boldsymbol{w}_{l3}, \ldots, \boldsymbol{w}_{li}, \ldots)$, $i = 1, 2, 3, \ldots, \text{AANUM}$; $l = 1, 2 \ldots, m$ where m is the number of fragments. In all k residues (from n_1 to $n_1 + k - 1$) of the lth fragments

$$\boldsymbol{w}_{li} = \begin{cases} 1 & \text{if } \sum_{j=n_1}^{n_1+k-1} \boldsymbol{n}_{ij} > \frac{k}{2} \\ 0 & \text{if } \sum_{j=n_1}^{n_1+k-1} \boldsymbol{n}_{ij} \leq \frac{k}{2} \end{cases}$$

Marking the fragments

For finding the fragments which are included in the same domain, fuzzy cluster analysis is used (Zadeh 1965; Dubois and Prade 1980) (see Appendix below). After these fragment are clustered, all of the fragments in the same domain will be marked with the same symbol.

In the fuzzy cluster analysis, each fragment is regarded as an object to be clustered. Then we have defined

$$\boldsymbol{R}_{l_1 l_2} = \sum_{k=0}^{\text{AANUM}} \frac{\delta(\boldsymbol{w}_{l_1 k}, \boldsymbol{w}_{l_2 k})}{\text{AANUM}} \quad l_1, l_2 = 1, 2, 3, \ldots, \boldsymbol{m}$$

$$0 \leq \boldsymbol{R}_{l_1 l_2} \leq 1.0 \quad (2)$$

as the matrix elements to build the fuzzy similarity matrix $\{\boldsymbol{R}_{l_1 l_2}\}_{m \times m}$. Here

$$\delta(\boldsymbol{w}_{l_1 k}, \boldsymbol{w}_{l_2 k}) = \begin{cases} 0 & \text{if } \boldsymbol{w}_{l_1 k} \neq \boldsymbol{w}_{l_2 k} \\ 1 & \text{if } \boldsymbol{w}_{l_1 k} = \boldsymbol{w}_{l_2 k} \end{cases}$$

Based on the theory of fuzzy mathematics, a fuzzy equivalence matrix was also obtained from a fuzzy similarity matrix. At each clustering level λ (see Appendix below), we can divide all objects (fragments) into different clusters. The fragments that belong to the same cluster (which also means the same domain) are marked with same symbol (see Fig. 1d).

Dynamic assembling

In order to determine the domain rationally, some constraints are needed and are added to the procedure. The minimum domain size should not be smaller than the MDS (minimal size of domain); the minimum part of one domain, if it is at the end, should be larger than MSSe (minimal size of fragment at the end of chain), and if it is in the middle, should be larger than MSSm (minimal size of fragment in the middle of chain) (Siddiqui and Barton 1995). The secondary structural limitation should also be considered. The split site between the adjacent two domains should not be within a secondary structure and should be at one end of the secondary structure. In our algorithm, the secondary structure definition of a protein in DSSP (Kabsch and Sander 1983) was adopted.

Considering these constraints, we use a link list to simulate the peptide chain. Each node in the link list stands for a fragment. If one fragment did not obey the constraints, it was assembled to the adjacent fragment. If the node was at the end, it was assembled to the adjacent node directly. If the node was in the middle of link list, it was compared to the adjacent two nodes first, then the pair of nodes which has the larger $\boldsymbol{R}_{l_1 l_2}$ will be merged together.

This process is continued until all the fragments obey the constraints; then the domains are identified from the peptide chain. One domain is composed of those nodes that have the same symbol (see Fig. 1e).

In order to evaluate the reliability of the results, we calculated an index V_{split} for each protein chain. If one chain includes more than one domain, using the same definition of V_{split} in Siddiqui and Barton (1995) we calculated each pair of domains' V_{split} and chose the minimum one as the chain index V. If one chain only has one domain, we regarded the amino acid which is at the end of the chain as one "domain" and regarded the remainder as another domain. Then V_{split} was calculated.

Results

In order to compare our results with other groups, we chose the same set of protein structures as Siddiqui and Barton (1995) used. There were a total of 230 protein structures, but 11 of them were not found in the PDB (release #77, July 1996); hence 219 structures were involved in our study. Among these 219 proteins, there were 66.7% of them whose derived domains agreed with the reference definitions, 30.6% showed minor differences, and only 2.7% (six proteins) showed very different definitions (see Tables 1 and 2). The six proteins which show very different definitions compared with the references are listed in Table 3.

The distribution of protein size

Figure 2 shows the size distribution of all the proteins that we used. Most are smaller than 400 residues.

Table 1 Domain list[a]

A	B	C	D	E
1aak			1–151	All
1aap	A		All	All
1aaq	B		1–99	All
1aar	A		All	All
1aba			1–87	All
1ace			1–534	All
1ake	A		1–214	All
1alc		1	1–122	38–104
		2		1–37, 105–122
1ald		1	1–108, 303–338	All
		2	109–234	
		3	235–302, 339–363	
1ama		1	3–50, 313–410	15–47, 326–410
		2	51–312	48–325
1apk			143–260	All
1aps			1–98	All
1atn	A	1	1–137, 353–372	1–144, 338–375
		2	138–352	145–337
1atn	D		1–260	All
1avr		1	3–87, 248–319	14–86
		2	88–247	87–160
		3		161–246
		4		247–318
1bbh	A		1–131	All
1bbp	A		2–178	All
1bbt	1	1	1–29, 197–208	
		2	30–196	31–189
1bbt	2		9–218	All
1bbt	3	1	1–40	
		2	41–220	42–214
1bia		1	1–75	1–60
		2	76–317	61–271
		3		272–317
1bic			3–261	All
1bov	A		1–69	All
1bpk			261–379	All
1brd			8–226	All
1caa			All	All
1cbx			1–307	All
1cc5			5–87	All
1cd8			1–114	All
1cho	E	1	1–245	1–16, 124–233
		2		28–123, 234–245
1cho	I		All	All
1cmb	A		1–104	All
1cob	A		1–151	All
1col	A		5–201	All
1cpc	A		1–174	All
1csc		1	1–271, 422–433	1–274, 381–437
		2	272–421	275–380
1cse	I		All	All
1ctf			53–120	All
1drf			1–186	All
1eca			1–136	All
1epi			All	All
1etu			5–200	All
1ezm		1	1–135	1–135
		2	136–298	136–298
1fha			5–184	All
1fia	A		10–98	All
1fkb			1–107	All
1flx			All	All
1fxd			All	All
1fxi	A		1–96	All
1gal		1	3–63, 226–326, 513–383	3–108, 225–327, 514–583
		2	64–144	109–225
		3	145–225, 327–361, 455–512	328–513
		4	362–454	
1gdl	O	1	1–177	0–148, 318–333
		2	178–333	145–317
1gfl			All	All
1gky		1	1–186	1–30, 81–186
		2		31–80
1gmf	A		5–123	All
1gmp	A		1–96	All
1gox		1	1–63	All
		2	64–227	
		3	228–359	
1gp1	A	1	10–194	All
1gpb		1	19–841	19–484, 814–831
		2		485–813
1gpr			4–161	All
1grc	A		1–209	All
1gst	A	1	1–127	1–81
		2		90–217
1hbg			1–147	All
1hcl		1	5–170	1–177
		2	171–217	178–400
		3	218–396	
		4	397–653	401–653
1hcc			All	All
1hip			1–85	All
1hiv	A		1–99	All
1hoe			All	All
1hom			All	1–68
1hrh	A		427–556	All
1hsa	A	1	1–182	1–175
		2	183–276	182–276
1hsa	B		1–99	All
1ilb			3–153	All
1ifc			1–131	1–131
1ipd		1	1–134, 252–345	All
		2	135–251	
1l53		1	1–164	1–69
		2		70–164
1lap		1	1–173	1–160
		2	174–484	162–353
		3		354–484
1lfi		1	1–100, 248–310 435–597, 665–691	1–90, 252–320
		2	101–247	91–251
		3	311–434, 598–664	345–433, 596–663
		4		434–595
1lld	A	1	7–319	7–146
		2		147–319
1lmb	3		6–92	All
1lpe			23–166	All
1lts	A		4–188	All
1lts	D		1–103	All
1mba			1–146	All
1mlp	A		All	All
1myg	A		1–153	All
1nsb	A	1	76–465	108–173
		2		174–214
		3		215–267
		4		268–314
		5		315–394
		6		395–459
1ofv			1–169	All
1p11	E	1	15–244	15–117, 232–244
		2		118–230
1paz			1–120	All
1pba			1–81	1–81
1pfk	A	1	1–137, 250–304	0–139, 256–304
		2	138–249, 305–319	140–255, 305–319

Table 1 (Contd.)

A	B	C	D	E
1pgx			All	All
1pha		1	10–77, 293–333	10–101, 296–355
		2	78–120	102–295, 355–414
		3	121–168, 252–292, 334–414	
		4	169–251	
1phh		1	1–41, 101–181, 269–327	All
		2	42–100	
		3	182–268, 328–394	
1pi2			All	All
1pii		1	1–254	1–255
		2	255–452	256–452
1pk4			1–80	All
1plc			1–99	All
1ppn		1	1–212	1–9, 112–206
		2		10–111
1prc	C	1	1–145, 261–333	33–143, 315–332
		2	146–260	144–314
1prc	H	1	1–87	All
		2	88–258	
1prc	M	1	1–51, 198–259	1–51
		2	52–197, 260–323	52–190
		3		191–323
1r69			All	All
1rat			1–124	All
1rbp			1–175	All
1rcb			1–129	All
1rhd		1	1–153	1–158
		2	154–293	159–293
1rn4			1–104	All
1rnb	A		2–110	All
1rop	A		All	All
1rve	A		2–245	All
1sgt		1	16–245	16–22, 129–229
		2		23–128, 230–245
1snc			7–141	All
1snw	A	1	114–264	114–177
		2		178–264
1stp			13–133	All
1tgs	I		All	All
1thm			1–279	1–127
				128–208
1tie			1–170	All
1tlk			33–135	All
1tme	l		1–256	31–250
1tnf	A		6–157	All
1trb		1	1–114, 244–316	1–116, 245–316
		2	115–243	117–244
1ula		1	1–116, 263–289	All
		2	117–262	
1utg			All	All
1vsg	A	1	1–21, 116–240	1–32, 87–254
		2	22–115, 241–362	33–86, 255–362
1xis		1	2–15, 218–323	2–317
		2	16–217	
		3	324–338	
1ycc			1–103	All
256b	A		1–106	All
2aai	A	1	1–219	1–117
		2		118–210
		3	220–267	211–267
2aai	B	1	1–138	1–135
		2	139–262	136–262
2aza	A		1–129	All
2c2c		1	1–112	1–62
		2		63–95
		3		96–112

Table 1 (Contd.)

2ccy	A		2–128	All
2cdv			1–107	All
2cpk	E	1	15–126, 320–350	33–126, 318–350
		2	127–319	15–31, 126–317
2ctx			All	All
2cwg	A	1	1–89	2–41
		2		42–88
		3	90–171	89–128
		4		128–171
2cyp		1	2–143, 256–294	All
		2	144–255	
2end			2–138	All
2fbj	H	1	1–124	1–118
		2	125–220	119–208
2fbj	L	1	1–102	1–104
		2	103–213	105–213
2fx2			2–148	All
2fxb			1–81	All
2gn5			1–87	All
2had		1	1–153, 229–310	1–155, 230–310
		2	154–228	156–229
2hip	A		All	All
2hmq	A		1–113	All
2hqr			2–88	All
2lig	A		25–181	All
2liv		1	1–123, 250–328	1–120, 250–328
		2	124–249, 329–344	121–249, 329–344
2ltn	A		1–181	All
2mcm			1–112	All
2mev	4		All	All
2mnr		1	3–129, 270–359	3–121, 344–359
		2	130–269	122–338
2msb	B		109–221	All
2npx		1	1–112, 224–323	1–114
		2	113–243	115–243
		3	324–447	244–324
		4		325–446
2pab	A		10–123	All
2pgd		1	1–176	33–344, 438–466
		2	177–473	345–437
2plv	1	1	6–76	83–202, 234–265
		2	77–302	
2plv	4	1	1–41	All
		2	42–235	
2por			1–301	All
2reb		1	3–264	23–268
		2	265–328	269–328
2rn2			1–155	All
2rsp	A		1–124	All
2scp	A		1–174	All
2sn3			All	All
2stv			12–195	26–195
2tgi			1–112	All
2tmd	A	1	1–383	1–383
		2	384–729	384–494, 649–733
		3		495–648
2tmv	P		1–154	All
2trx	A		1–108	All
2ts1		1	1–70, 214–319	248–319
		2	71–213	1–220
2tsc	A	1	1–38, 229–264	All
		2	39–228	
2wrp	R		5–108	All
2yhx		1	2–20, 247–371	2–19, 284–363
		2	21–246, 431–458	50–189, 433–451
		3	372–430	20–49, 190–282 364–432
351c			1–82	All
3b5c			3–88	All
3bc1		1	3–51, 223–358	All

Table 1 (Contd.)

A	B	C	D	E
		2	52–222	
3cd4		1	1–103	1–97
		2	104–178	98–178
3chy			2–129	All
3cla			6–219	All
3cox		1	5–44, 126–155 227–319, 445–506	5–44, 226–316 462–506
		2	45–125, 156–226 320–444	45–225, 317–461
3dfr			1–162	All
3dpv	A	1	37–71, 113–210 246–317, 363–405, 459–584	37–281, 336–410 446–584
		2	72–112, 211–245, 318–362, 406–458	282–335
		3		411–445
3ebx			All	All
3enl		1	1–34, 72–147, 300–436	1–142
		2	35–71, 148–299	143–420
3gap	A	1	1–112	1–125
		2	113–208	126–208
3gly		1	1–20, 392–430	1–20, 227–432
		2	21–226, 431–471	21–226, 433–471
		3	227–391	
3grs		1	18–62, 122–161 291–367	18–57, 108–158 293–363
		2	63–121, 162–290, 368–478	50–107, 159–291
		3		365–478
3il8			All	All
3pgk		1	1–198, 404–415	1–98, 388–478
		2	199–403	199–387
3pgm		1	1–230	1–88, 148–230
		2		89–148
3pmg	A	1	1–190	1–188
		2	191–561	189–297, 379–408
		3		298–378
		4		420–562
3psg		1	1–44	1–170
		2	45–156	
		3	157–326	180–327
3rub	S		1–123	All
3sdp	A		5–190	All
3sic	I		7–113	All
4blm	A	1	31–291	31–70, 217–291
		2		71–216
4bp2			1–123	All
4fgf			20–143	All
4gcr		1	1–42	1–80
		2	43–85, 132–174	83–174
		3	86–131	
4icb			All	
4mdh	A	1	1–81	1–151
		2	82–154	
		3	155–333	152–333
4sbv	A		62–260	All
4sgb	I		All	All
4tnc		1	3–106	3–88
		2	107–162	101–162
5fbp	A	1	6–201, 247–335	1–201
		2	202–246	202–335
5fd1			1–106	All
5p21			1–166	All
5rub	A	1	2–136, 297–363	1–139
		2	137–296, 363–457	140–457
6abp		1	2–109	2–109, 254–286
		2	110–306	110–253, 295–306

Table 1 (Contd.)

6edx	A		All	All
7cat	A	1	3–147, 207–431	3–68
		2	148–206, 432–500	68–433
		3		434–500
7tim	A		2–248	All
8acn		1	2–67	2–201
		2	68–154	
		3	155–200	
		4	201–517	202–511
		5	518–754	532–754
8adh		1	1–186, 317–374	1–175, 319–374
		2	187–316	176–318
8atc	A	1	1–130, 292–310	1–143, 291–310
		2	131–291	144–290
8atc	B	1	8–98	8–100
		2	99–153	101–153
8rxn			All	All

[a] Abbreviations used for the headings of each column: A, PDB code; B, chain index; C, domain number; D, derived definition from this work; E, reference definition. ''All'' indicates that the whole protein chain is a single domain

Table 2 Result statistics

	Agreed (I)	Minor different (II)	Major different (III)
Number	146	67	6
Percent	66.7%	30.6%	2.7%
Total		219 protein structures	

Table 3 Proteins which had major different domain definitions

PDB code	Chain	Protein name
1gal		Oxidoreductase (flavoprotein)
1vsg	A	Glycoprotein
2mnr		Racemase
1prc	C	Photosynthetic reaction center
2pgd		Oxidoreductase [CHOH(D) – NADP + (A)]
7cat	A	Oxidoreductase (H_2O_2 acceptor)

The distribution of domain size

Figure 3 shows the size distribution of all 310 domains. Most are smaller than 300 residues. The distribution is more concentrated than that of the proteins.

The distribution of the domain number of one protein chain

Figure 4 shows that most chains only include one domain. Some protein chains include 2–3 domains, but few have more than 4 domains.

The program was run on a Silicon Graphics Indigo R4400 workstation. It took only 38 min to complete the calculations for all 219 protein structures, which is faster than other algorithms (see Table 4).

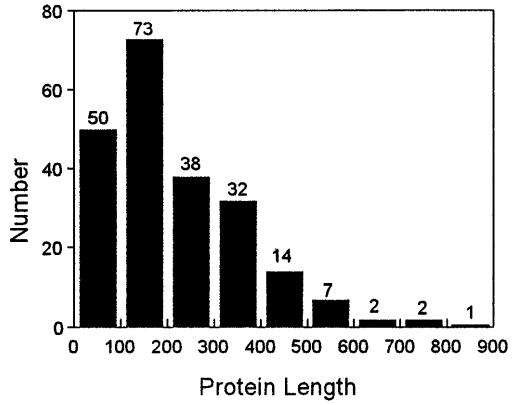

Fig. 2 The size distribution of the proteins used. Most are smaller than 400 residues

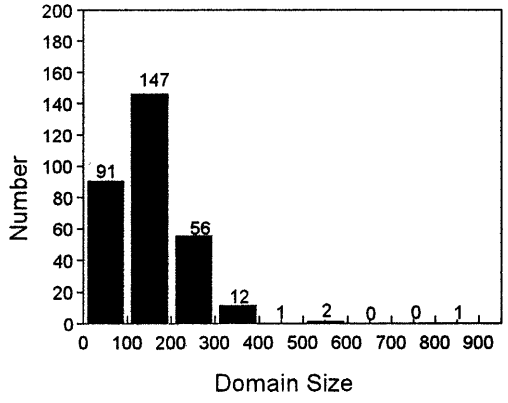

Fig. 3 The size distribution of all 310 domains. Most are smaller than 300 residues

Discussion

Domain identification

Although domain identification is basically a mathematical problem, yet domain-classified information from the research of protein domain recognition is very important for protein design. Comparing these results with previous methods, our algorithm possesses the same precision, but the consumed CPU time is very economized.

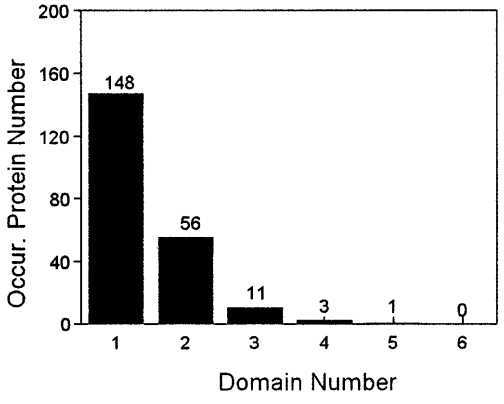

Fig. 4 The domain number distribution of proteins. Many proteins only include one domain. Some had two or three domains. Few proteins have more than four domains

"Gray area" in domain identification

Siddiqui and Barton (1995) had found that there existed a "gray area" (the uncertainty in the split site) while the domain is identified. In our work, we also met this problem (see Fig. 5). The structure of the A-chain of wheat germ agglutinin isolectin (denoted by its PDB code: 2cwg) seemed to consist of two domains (I, II), but each domain can also be regarded as two subdomains (I_a, I_b and II_a, II_b). If the proper parameters were chosen, such as LAMD in our algorithm, it may be helpful for eliminating the "gray area".

Secondary structure constraint

In our method, secondary structure information was a very important constraint to correct the split site. We thought that the secondary structure is a relatively substantive part of a domain, so a split site should not be within a secondary structure (including helix and sheet), but could be at the end of a helix or a sheet. For example, the structure of the contractile system protein (4tnc) had two domains linked by a long helix (from residue 75 to residue 106) (see Fig. 6). In our opinion, the split site should be in the upper stream of residue 75 or down stream of residue 106. We set it at residue 106 after considering the secondary structure constraint; others have set it at residue 88 and some at residue 90 (Siddiqui and Barton 1995).

Table 4 Comparison of some methods

	Total structures	Set I	Set II	Set III	Max. accept	Time cost	Ave. time
This work	220	146 (67%)	67 (30%)	6 (3%)	213 (97%)	38 min	10.38 s
Siddiqui and Barton (1995)	230	161 (70%)	41 (18%)	28 (12%)	202 (88%)	16.5 h	258 s
Holm and Sander (1994)	330	Many of their published domain definition disagree with those found in the literature				40 min	7.27 s

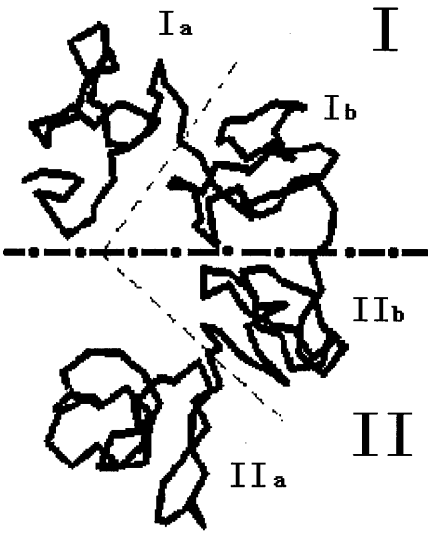

Fig. 5 Plot of gray area of domains. From the 3D structure of protein 2cwg (wheat germ agglutinin isolectin), A-chain, two domains (domain I, II, split by dash-dotted line) can be identified, but it can also be thought that this chain includes four domains ($I_a, I_b,$ and II_a, II_b, split by dashed line and dash-dotted line). So there is a "gray area" in searching the domain. The figure was produced using RasMol 2.5 (Sayle 1995)

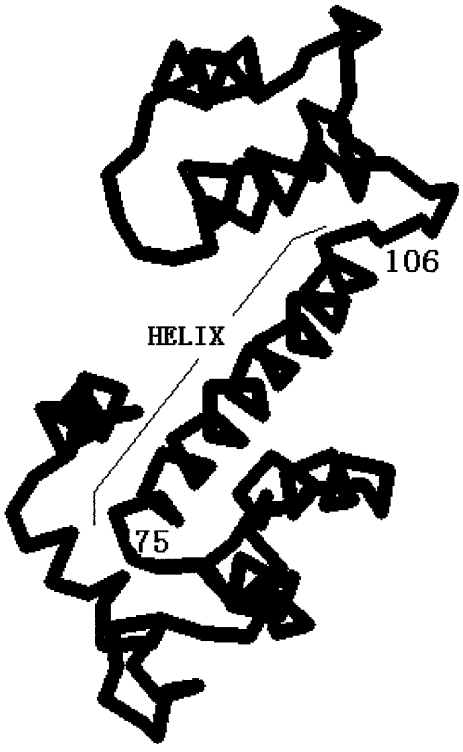

Fig. 6 3D structure of 4tnc. This contractile system protein seems to consist of two domains, and there is a very long α-helix between the two domains. The split site should not be in the middle of the helix. In our method, residue 106 is considered as the split site of two domains from two end sites of this long helix, 75 and 106. The figure was produced using RasMol 2.5 (Sayle 1995)

The V_{split} index of protein domain recognition

The statistics of V_{split} index are shown in Table 5. The domain definition is correct if the V_{split} index is larger than 100; we would doubt the result if the V_{split} index was less than 50. In our algorithm, if the V_{split} index is less than 10, the two domains from which the V_{split} is calculated would be merged. Eight cases which have V_{split} less than 10 were encountered in our calculation, four in set I (1ace, 1bic, 1rhd, 2tgi), three in set II (2npx, 2plv_1, 3dpv_A), and one in set III (7cat_A). We found that this V_{split} index is very useful for estimating the results.

Choosing reasonable parameters

In our method, several parameters were used.

BORD distinguishes whether contact exists between two residues. Too small a value would create too many small domains, while too large a value could not correctly identify domains. We set BORD = 28 Å in this work. CUTOFF is used to calculated the "similarity of contact environment" between two residues. It equals 0.5 in our algorithm. LAMD is a parameter needed in cluster analysis. We found that if it is too large, the fragments would be too small, and if it is too small, many structural details would be concealed. After some experiments, we set LAMD equal to 0.8. The other parameters, MDS, MSSe, and MSSm, are used with the same values as in Siddiqui and Barton (1995).

In short, cluster analysis and fuzzy mathematics are very effective tools in domain recognition. Using these methods, we could identify the domains of proteins globally. The methods are also very fast and accurate.

Table 5 Index statistics[a]

	Set I (%)	Set II (%)	Set III (%)
$V < 50$	1.78	27.87	20.00
$V < 100$	20.54	50.82	100.00
$100 \le V \le 1000$	75.89	45.90	0.00
$V > 1000$	3.57	3.28	0.00

[a] The results for which one chain has a single domain are not included in these statistics

Acknowledgements We would like to thank the National Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics & Scientific and Engineering Computing, Chinese Academy of Sciences, who allowed free use their computers and for their kind help. This work is supported by grants 39392900 and 39830070 of the Chinese National Scientific Foundation.

Appendix: fuzzy cluster analysis

Fuzzy set theory, introduced by Zadeh (1965), is a generalization of abstract set theory. It has a wider scope of applicability than abstract set theory in solving

problems that involve, to some degree, subjective evaluation.

Fuzzy set

Intuitively, a fuzzy set is a class that admits the possibility of partial membership in it. Let $X = \{x\}$ denote a space of objects. Then a fuzzy set A in X is a set of ordered pairs:

$$A = \{(x, \chi_A(x))\}, \quad x \in X$$

where $\chi_A(x)$ is termed "the grade of membership of x in A". We assume for simplicity that $\chi_A(x)$ is a number in the interval [0, 1], with the grades 1 and 0 representing, respectively, full membership and nonmembership in a fuzzy set.

Examples of some fuzzy sets are:

Fuzzy set A, labeled "integers approximately equal to 5", may be defined as

$$A = \frac{0.1}{2} + \frac{0.4}{3} + \frac{0.9}{4} + \frac{1.0}{5} + \frac{0.9}{6} + \frac{0.4}{7} + \frac{0.1}{8}$$

Fuzzy set A, labeled "real numbers clustered around 5", may be defined by the grade of membership function

$$\chi_A(x) = \left\{1 + \left[\frac{1}{4}(x-5)\right]^2\right\}^{-1}$$

or as

$$A = \int_R \frac{\left\{1 + \left[\frac{1}{4}(x-5)\right]^2\right\}^{-1}}{x} \qquad R = \{\text{real number}\}$$

Fuzzy relationship

If X is the Cartesian product of n universes of discourse $X_1, \ldots, X_n$, then an n-ary fuzzy relation, R, in X is a fuzzy subset of X. R may be expressed as the union of its consistent fuzzy singletons $\chi_R = (x_1, \ldots, x_n)/(x_1, \ldots, x_n)$, that is

$$R = \int_{\chi_1 \times \cdots \times \chi_n} \frac{\chi_R(x_1, \ldots, x_n)}{(x_1, \ldots, x_n)}$$

where χ_R is the membership function of R.

Fuzzy cluster analysis

Fuzzy cluster analysis is based on a fuzzy equivalence relationship, but this relationship is hard to obtain directly so we obtain it from a fuzzy similarity relationship. If $U = \{u_1, u_2, \ldots, u_n\}$ is a set of the objects discussed, and $u_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ represents the relationship between each object u_i and m kinds of elements, the fuzzy similarity relationship, which can be represented as a fuzzy similarity matrix $R = (S_{ij})_{n \times n}$, has many methods to consider. Our method is shown as Eq. (2) above.

A fuzzy equivalence matrix can be obtained by multiplying R with itself 2^m times. The product operator here is a special one which obeys the rule of fuzzy mathematics:

$$R \to R^2 \to R^4 \to \ldots \to R^{2^m}$$

$$m = \begin{cases} \log_2 n & \log_2 n \in \{\text{integer}\} \\ 1 + \log_2 n & \log_2 n \notin \{\text{integer}\} \end{cases}$$

From the fuzzy equivalence matrix R^{2^m} at each clustering level λ we first create a cut-off matrix $R_\lambda^{2^m}$

$$(R_\lambda^{2^m})_{nk} = \begin{cases} 1 & (R^{2^m})_{nk} > \lambda \\ 0 & (R^{2^m})_{nk} < \lambda \end{cases}$$

then we divide all objects u_i into clusters from this matrix.

Here is an example:

$$X = \{x_1, x_2, x_3, x_4, x_5\},\ x_1 = (5, 5, 3, 2),\ x_2 = (2, 3, 4, 5),$$
$$x_3 = (5, 5, 2, 3),\ x_4 = (1, 5, 3, 1),\ x_5 = (2, 4, 5, 1)$$

then a fuzzy similarity matrix is obtained through the fuzzy relationship

$$S_{ij} = 1 - 0.1 \times \sum_{k=1}^{4} |x_{ik} - x_{jk}| \quad (i, j = 1\text{–}5)$$

$$m = 1 + [\log_2 5] = 3$$

so a fuzzy equivalence matrix can be created through $R \to R^2 \to R^4 \to R^8$:

$$R = \begin{bmatrix} 1 & 0.1 & 0.8 & 0.5 & 0.3 \\ & 1 & 0.1 & 0.2 & 0.4 \\ & & 1 & 0.3 & 0.1 \\ & & & 1 & 0.6 \\ & & & & 1 \end{bmatrix}$$

$$R^* = R^8 = \begin{bmatrix} 1 & 0.4 & 0.8 & 0.5 & 0.5 \\ & 1 & 0.4 & 0.4 & 0.4 \\ & & 1 & 0.5 & 0.5 \\ & & & 1 & 0.6 \\ & & & & 1 \end{bmatrix}$$

At last, with each λ given, all clusters can be obtained from each cut-off matrix:

$\lambda = 1, X$ is divided into 5 clusters: $\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}$

$\lambda = 0.8, X$ is divided into 4 clusters: $\{x_1, x_3\}, \{x_2\}, \{x_4\}, \{x_5\}$

$\lambda = 0.6, X$ is divided into 3 clusters: $\{x_1, x_3\}, \{x_2\}, \{x_4, x_5\}$

$\lambda = 0.5, X$ is divided into 2 clusters: $\{x_1, x_3, x_4, x_5\}, \{x_2\}$

$\lambda = 0.4$, all x_i are in one cluster: $\{x_1, x_2, x_3, x_4, x_5\}$

Some cut-off matrix at each λ:

$$\begin{bmatrix} 1 & & & 0 & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ 0 & & & & 1 \end{bmatrix}_{\lambda=1}, \quad \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}_{\lambda=0.8},$$

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}_{\lambda=0.6}, \ldots$$

References

Crippen GM (1978) The tree structural organization of domains in globular proteins. J Mol Biol 126: 315–332

Dubois D, Prade H (1980) Fuzzy sets and systems – theory and application. Academic Press, New York

Holm L, Sander C (1994) Parser for protein folding units. Protein 19: 256–268

Janin J, Wodak SJ (1983) Structural domains in proteins and their role in the dynamics of protein function. Prog Biophys Mol Biol 42: 21–78

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Bioploymers 22: 2577–2637

Lesk AM, Rose GD (1981) Folding units in globular proteins. Proc Natl Acad Sci USA 78: 4304–4308

Liljas A, Rossman MG (1974) X-ray studies of protein interactions. Annu Rev Biochem 43: 475–507

Nichols WL, Rose GD (1995) Rigid domains in proteins: an algorithmic approach to their identification. Protein 23: 38–48

Rose GD (1979) Hierarchic organization of domains in globular proteins. J Mol Biol 134: 447–470

Sayle RA, Milner-White EJ (1995) RASMOL: biomolecular graphics for all. Trends Biochem Sci 20: 374

Siddiqui AS, Barton GJ (1995) Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definition. Protein Sci 4: 872–884

Wodak SJ, Janin J (1981) Location of structural domains in proteins. Biochemistry 20: 6544–6552

Zadeh LA (1965) Fuzzy sets. Inf Control 8: 338–353

Zehfus MH (1987) Continuous compact protein domains. Protein 2: 90–110

Zehfus MH (1993) Improved calculation of compactness and a reevaluation of continuous compact units. Protein 16: 293–300

Zehfus MH (1994) Binary discontinuous compact protein domains. Protein Eng 7: 335–340

Zehfus MH, Rose GD (1986) Compact units in proteins. Biochemistry 25: 5759–5765