Three-Dimensional Structural Feature Search of Proteins

Hiroaki Kato and Yoshimasa Takahashi*

Laboratory for Molecular Information Systems, Department of Knowledge-Based Information Engineering,
Toyohashi University of Technology, Tempaku-cho, Toyohashi 441

(Received November 8, 1996)

This paper describes a three-dimensional (3D) structural feature search, or motif search, of proteins using SS3D-P. Our present approach is based upon the use of a substructure searching program for small organic molecules, SS3D, reported in our previous work. In SS3D-P, the 3D structure of a protein is approximated with the coordinates of α-carbon atoms of the main chain, and treated as a molecular graph that is represented using a distance matrix. The program allows us to specify the tolerance value of the inter-residual distances to be compared. Several characterization schemes are also available for the different amino acid residues that can be distinguished. The program we have developed was validated by search trials with 3D query patterns of known protein motif structures: the P-loop, the EF-hand, and the Zinc finger motif using Protein Data Bank (PDB) structure files.

It is well known that the three-dimensional (3D) structure of a protein is closely related to the function of the protein itself. This is especially true for particular structural features called motifs which have specific geometric arrangements within protein molecules. They are considered to be well-reserved sites in the genomic sequences. So to examine such motifs or 3D structural features in a more general sense is one of most important problems, not only in structure-function studies of proteins, but also in genome informatic studies. Additionally, with the increasing number of known 3D structures of proteins, the protein structure database is more and more important as a key element in attempts being made to derive molecular structural information that is of molecular biological interest. It is almost impossible to manually search the 3D structural motifs within proteins because of their large and complex structures. Computational methods are required for a systematic search for the 3D features of proteins in such a database.

The early work of 3D substructure searching or geometric searching of proteins has been done by Lesk.[1] The approach was based on a preprocessing to reduce the potential matches between the query atoms and database atoms that would be considered and tested for isomorphism by trying to rotate a set of candidate atoms from the database atoms to get a possible match. There are two types of approaches for the 3D structural feature searching of proteins: atom-level (or residue-level) geometric searching based on Cα atoms that form the backbone of a protein, and secondary structure motif searching. It is clear that Lesk's approach is one of the former case. Another atom-level 3D substructure searching of proteins has been reported by Brint et al.[2] Their approach is based on the Ullman's subgraph isomorphism algorithm.[3] Recently, for the latter approaches, many computational methods have been proposed using structure comparison studies to analyse the 3D similarity of proteins, in which protein structures to be compared are described by lines or vectors in 3D space that involve the α-helix and β-strand secondary structure elements.[4—10] However, most of them are mainly for aligning a pair of globally similar structures and identifying similar fragments at secondary structure level, and are not always suitable for the full database search with a user-defined 3D query for the protein substructure searching. To derive useful information from such a database, some powerful tool which can be practically applied to fully analyse the structural features of proteins is necessary, because the studies on the atom-level or the residue-level substructure searching are still few.

In this paper, we describe a computer program SS3D-P for 3D structural feature searching at the residue-level, which allows a search of the Protein Data Bank to be carried out to identify all occurrences of a user-defined 3D query pattern or a 3D motif consisting not only of chain-based peptide segments but also of a set of disconnected amino acid residues.

Methods

The basic idea for the algorithm presented here was based on that of SS3D (three-dimensional substructure search system for small organic molecules) which has been developed in our earlier work.[11] In SS3D, the 3D structure of a molecule is treated as a set of points that correspond to constituent atoms in 3D space. The set of points is described by a matrix representation of which each element involves the inter-atomic distance within the molecule. Thus the set of points can be regarded as an edge-weighted complete graph. In other words, we can represent the structural information of a molecule including the 3D geometry with a weighted (labeled) graph of which the nodes correspond to the constituent atoms. Therefore, the 3D substructure searching can be treated as one of subgraph matching problems. The process of the subgraph matching in SS3D consists of two major parts. They are (1) the generation of the docking graph of two molecular graphs which come from a 3D query substructure (a set of structural fragments with a particular spatial arrangement) and a molecule, and (2) the identification of the maximal cliques of the docking graph produced. The docking graph used here is an attempt at a pairwise map of atomic nodes of the two original molecular graphs. The mapping works if the distances between any two atoms are within set tolerances. A clique in the docking graph corresponds to a grouping of atoms in the original graphs, where all the intragrouping distances are the same in both original graphs. Thus, the 3D-query substructure searching is equivalent to examining the presence or absence of the clique(s) with the same number of atoms within the query substructure. For more detail about this algorithm, please refer to the reference.[11)]

In principle, the SS3D can be applicable to the 3D structural feature search of proteins. But in the case of protein structures, the number of constituent atoms is large and thus the docking graph to be considered also becomes quite large. This causes an high demanding of computational time for the clique finding process. In the present work, in order to apply the algorithm to the structural feature analysis of proteins, a reduced representation has been adopted to describe the 3D structural information of a protein. In this way, each amino acid residue of the protein is regarded as a pseudo-atom (super-atom) and it's 3D coordinates are approximated by those of α-carbon (Cα) atoms of the residues. This approximation can considerably decrease the size of the graphs that have to be treated. The different amino acid residues are distinguished by the node-weights of the pseudo-atoms. Physical and chemical properties of the amino acid residues can also be taken into account by a similar incorporation into the weighting scheme of the pseudo-atoms. According to the algorithms, we have developed a three-dimensional structural feature search system for proteins, called SS3D-P. It was implemented on a Silicon Graphics Unix workstation using ANSI-C language.

Results and Discussion

In order to test the validity and performance of our program, we prepared two test databases of protein structures for the trial use. This data was taken from PDB files[12)] (the release version of January, 1995). The first database contains 543 proteins (605 chains) whose the structures were determined by X-ray crystallographic analysis. They were selected under the conditions that the resolution is 2.80 Å or of greater accuracy and the number of residues is less than 500. All of these proteins are a subset of the data set recommended by Hobohm et al.,[13)] which involves structurally diverse proteins. The second one contains 90 proteins whose structures were determined by NMR spectroscopy. They also have less than 500 amino acid residues. The following section describe the results of the search trials with three different types of 3D queries: a P-loop motif, an EF-hand motif, and a Zinc finger motif.

3D Search for the P-Loop Motif Segment: The first search trial was carried out with the 3D query for the P-loop motif[14,15)] segment shown in Fig. 1. The P-loop motif consists of eight residues which are consecutive in the sequence. It is well known that the P-loop motif is an ATP/GTP-binding site. The 3D coordinate data of the query for the P-loop motif in this trial is taken from the corresponding site (A-chain, G10-S17) of cH-p21 Ras protein (6Q21) in the PDB file. At first, it was verified that the SS3D-P correctly found the reference site of 6Q21 which is used as the query substructure for itself. Alternatively, the 3D search for the P-loop motif segment was carried out for the 633 proteins contained in both the two databases that had been prepared in advance. The following search conditions were used for this trial: the tolerance value of the distance is 1.6 Å, no different residues are considered, and the eight residues must be consecutive in the sequence. Twelve sites from 12 proteins were detected in the trial. The results are summarized in Table 1. It is shown that five sites from five proteins, labeled with '*' in the last column of the table have the same pattern as the query, in terms of the sequence pattern, nevertheless, the type of amino acid was not considered in the search. SS3D-P correctly found the corresponding site of an alternative cH-p21 Ras protein, which has the same motif site. In PROSITE,[16)] the sequence pattern is described with [AG]–x(4)–G–K–[ST], where square parentheses '[ ]' show the acceptable amino acids for a given position and 'x' stands for a position where any amino acid is accepted (x(4) corresponds to x–x–x–x). The program also correctly found the corresponding site of 1EFM, of which the site is known in PROSITE. For those of the other three proteins (1GKY, 1TNDA, and 2REB), there were not the PDB cross reference codes of them in the PROSITE database. However, the PROSITE document file lists guanylate kinase, transducin, and rec A protein as protein families in which P-loop motif is found. It was verified by the survey of other database, SCOP,[17)] that the detected sites of guanylate kinase 1GKY (G8-S15), transducin 1TNDA (G36-S43) and rec A protein 2REB (G66-T73) are also the corresponding sites. In addition to this, our program identified the similar structural features within other proteins. Table 1 shows that three adenylate kinases (1AK3A, 1AKEA, and 3ADK) have similar peptide segments to the query, as described in PROSITE, in which there is a single mutation from the P-loop sequence pattern: in the last position G is found instead of S or T. The pub-

3-D QUERY MOTIF (P-LOOP)

(ATP/GTP-binding site motif A).

Protein	: cH-p21 Ras protein (6Q21A)								
Res. No	:	10	11	12	13	14	15	16	17
A.A.	:	G	A	G	G	V	G	K	S

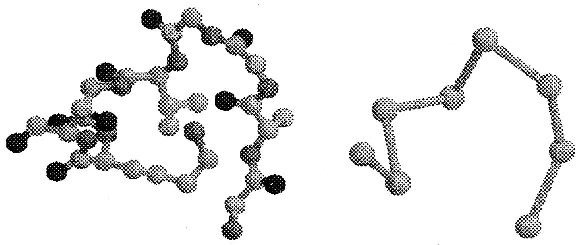


Fig. 1. The P-loop motif from cH-p21 Ras protein that was used as the first query pattern in the 3D structural feature search by SS3D-P: (a) an atom-level representation of the 3D motif structure, (b) a reduced representation based on Cα atoms.

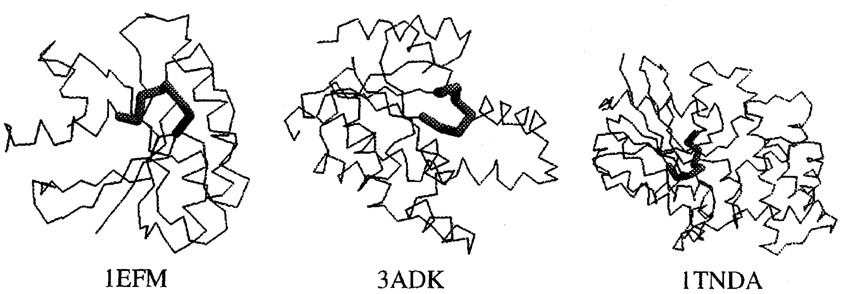
Table 1. Results of the 3D Search for the P-Loop Motif Segments by SS3D-P

ID No.	PDB	Protein	Total	Hit site								
	code		residues	(amino acide identifier)								
[13-01]	1AK3A	Adenylate kinase	(214)	G 12	A 13	P 14	G 15	S 16	G 17	K 18	G 19	†
[14-01]	1AKEA	Adenylate kinase	(214)	G 7	A 8	P 9	G 10	A 11	G 12	K 13	G 14	†
[95-01]	1DHR	Dihydropteridin reductase	(236)	A161	G162	K163	N164	S 165	G166	M167	P168	
[106-01]	1EFM	Elongation factor Tu	(158)	G 18	H 19	V 20	D 21	H 22	G 23	K 24	T 25	*
[141-01]	1GKY	Guanylate kinase	(186)	G 8	P 9	S 10	G 11	T 12	G 13	K 14	S 15	*
[217-01]	1LPFA	(Dihydro) lipoamide dehydrogenase	(472)	L286	A287	A288	D289	S290	G291	V292	T293	
[256-01]	1NSCA	Influenza neuraminidase	(390)	R373	E372	T371	K370	S369	M368	T367	R366	
[371-01]	1TNDA	Transducin	(323)	G 36	A 37	G 38	E 39	S 40	G 41	K 42	S 43	*
[413-01]	2BAT	Influenza neuraminidase	(388)	R371	L370	D369	K368	S367	I 366	T365	R364	
[489-01]	2REB	RecA protein	(303)	G 66	P 67	E 68	S 69	S 70	G 71	K 72	T 73	*
[511-01]	3ADK	Adenylate kinase	(194)	G 15	G 16	P 17	G 18	S 19	G 20	K 21	G 22	†
[576-01]	5P21	cH-p21 Ras protein	(166)	G 10	A 11	G 12	G 13	V 14	G 15	K 16	S 17	*

The hit sites labelled with '*' exactly have the same sequence pattern as that of the query, [AG]–x(4)–G–K–[ST] that was described in the PROSITE database. Those labelled with '†' exist in a slightly different form: in the last position glycine (G) is found instead of serine (S) or threonine (T).

lications have suggested that the detected sites of 1AK3A (G12-G19), 1AKEA (G7-G14), and 3ADK (G15-G22) are the sites related to binding to phosphate group.[18] It should be noted that the present search was carried out using only geometrical information of the set of points (pseudo-atoms) in 3D space, and no labeling of the name and the type of amino acid residue has ever been used. In Fig. 2, some graphical views of the hit proteins are displayed with the sites corresponding to the 3D query segment. The total computational time required for the present 3D search is about 18 min; this is not the real central processing unit (CPU) time, but the elapsed time for the total process.

3D Search for the EF-Hand Motif Segment: The second search trial was carried out with the 3D query for the EF-hand motif segment. One of the EF-hand motifs is well known as the active site of calcium binding proteins, e.g. Troponin C.[19] The motif region consists of thirteen residues, as shown in Fig. 3. The 3D coordinate data for the query of the EF-hand motif is taken from the corresponding site (D142-F154) of Troponin C (1TOP) stored in the PDB file. Obviously, SS3D-P correctly found the reference site of 1TOP used as the query pattern in the trial. First, the search was carried out for 543 proteins (605 chains) which were determined by X-ray analysis. The search conditions used for the present search trial are as follows: The tolerance value of the distance is 3.0 Å, no different residues are considered, and the 13 residues must be consecutive in the sequence. Here, it should be noted that no information on the calcium atom is included in the query pattern. The result is shown in Table 2. Some graphical views of the selected proteins from Table 2 are shown with their hit sites in Fig. 4. 1PAL and 1OSA have multiple sites within each molecule (two for 1PAL and four for 1OSA, respectively). In PROSITE, the sequence pattern of the EF-hand motif site is given by the description of D–x–[DNS]–{ILVFYW}–[DENSTG]–[DNQGHRK]–{GP}–[LIVMC]–[DENQSTAGC]–x(2)–[DE]–[LIVMFYW], where a pair of curly brackets '{ }' show the amino acids that are not accepted at a given position (see the previous section for others).[16] Table 2 and Fig. 4 show that the program correctly found all of the hit sites of them. The reader should notice again that only the spatial arrangement of the pseudo-atoms of the query motif segment is considered like a set of points in the 3D space, and no other characterization of the residues was employed for the present search. Nevertheless,



1EFM 3ADK 1TNDA

Query: P-loop (ATP/GTP-binding site motif A; G10-S17 in 6Q21A);

Search condition: all residues are connected, all kinds of residues are equivalent, and the tolerance value for the distance is 1.6Å.

Fig. 2. Graphical views of the hit proteins and the detected sites obtained by P-loop motif search using SS3D-P.

3-D QUERY MOTIF (EF-HAND)

Protein : Troponin C (1TOP)

Res. No.	: 142	143	144	145	146	147	148	149	150	151	152	153	154
A.A.	: D	K	N	N	D	G	R	I	D	F	D	E	F

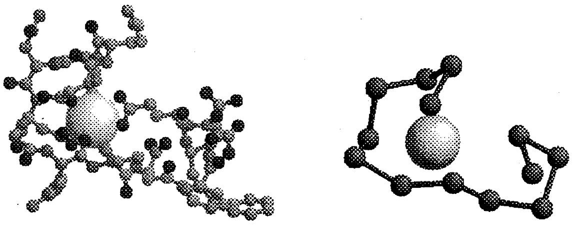
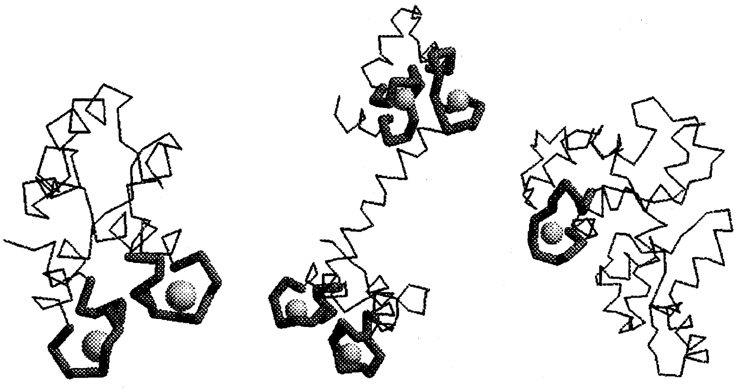


Fig. 3. The EF-hand motif from Troponin C that was used as the second query pattern in the 3D structural feature search by SS3D-P: (a) an atom-level representation of the 3D motif structure, (b) a reduced representation based on Cα atoms.



Query: EF-hand (D142-F154 in 1TOP);

Search condition: all residues are connected, all kinds of residues are equivalent, and the tolerance value for the distance is 3.0Å.

Fig. 4. Graphical views of the hit proteins and the detected sites obtained by EF-hand motif search using SS3D-P.

as shown in Table 2, all of the sites found by the SS3D-P program are also consistent, in terms of the sequence pattern, with the EF-hand motifs described in PROSITE. This is a quite interesting fact and that is also true for the result for the 90 proteins obtained from NMR method. The search result for the 90 proteins using the same query and the same conditions is show in Table 3. These results suggest that the real motif that is closely related to a certain biological function may have a highly restricted spatial arrangement or conformation. The total computational time required for the former search is about 87 min, and 2 min or less for the latter search.

3D Search for the Zinc Finger Motif: Another search trial was carried out with the 3D query pattern for the Zinc finger motif. The Zinc finger motif was first described in the analysis of the amino acid sequence of a transcription factor, TFIIIA.[20] It is considered that a Zinc finger is closely related to forming the DNA-binding region of the related proteins. Several different families of the Zinc finger motifs have been reported.[21] In the present search, a "classic" Zinc finger that has two cysteine residues and two histidine residues bound to zinc was used as the 3D query pattern. The 3D coordinate data to describe the query pattern are taken from the finger site of C106, C109, H122, H126 of a transcription factor (1ARD) in the PDB file, of which the structure was based on NMR analysis. The search was carried out for the 533 proteins used in the above. The following conditions were used for the present search trial: the tolerance value for the distance was 1.6 Å, different types of the amino acid residues were considered in the 3D query pattern matching. It should be also noted that no information on the zinc atom is included in the query pattern. SS3D-P found five sites from three different proteins: bacteriophage T7 lysozyme (1LBA), ZIF286 (1ZAAC), and influenza neuraminidase (2BAT). The computational time required for the present search was 158 seconds. The result is summarized in Table 4. Excepting 1ZAAC, no proteins used in the present search were detected in PROSITE. 1LBA is a zinc amidase.[22] The PDB structure file shows that 1LBA contains a zinc atom within the molecule and the detected site is located close to the zinc atom. But, the detected site (C18, C80, H122, H68) is different from the classic Zinc finger motif, in terms of the definition of the motif with the related loop region. Cheng et al.[22] suggested that the ligands to a zinc atom are H17, Y46, H122, C130. The site of 2BAT is also considered as noise in the search. On the other hand, for 1ZAAC, three Zinc

Table 2. Results of the 3D Search for the EF-Hand Motif Segments for the 543 Proteins (by X-Ray Analysis) Using SS3D-P

ID No.	PDB	Protein	Total	Hit site												
	code		residues	(amino acide identifier)												
[66-01]	1CDP	Parvalbumin	(108)	D 51	Q 52	D 53	K 54	S 55	G 56	F 57	I 58	E 59	E 60	D 61	E 62	L 63
[66-02]	1CDP	Parvalbumin	(108)	D 90	S 91	D 92	G 93	D 94	G 95	K 96	I 97	G 98	V 99	D100	E101	F102
[72-01]	1CLL	Calmodulin	(144)	D 20	K 21	D 22	G 23	D 24	G 25	T 26	I 27	T 28	T 29	K 30	E 31	L 32
[72-02]	1CLL	Calmodulin	(144)	D 56	A 57	D 58	G 59	N 60	G 61	T 62	I 63	D 64	F 65	P 66	E 67	F 68
[72-03]	1CLL	Calmodulin	(144)	D 93	K 94	D 95	G 96	N 97	G 98	Y 99	I 100	S101	A102	A103	E104	L105
[72-04]	1CLL	Calmodulin	(144)	D129	I 130	D131	G132	D133	G134	Q135	V136	N137	Y138	E139	E140	F141
[242-01]	1MYSB	Myosin	(138)	D 35	Q 36	N 37	R 38	D 39	G 40	I 41	I 42	D 43	K 44	D 45	D 46	L 47
[263-01]	1OSA	Calmodulin	(148)	D 20	K 21	D 22	G 23	D 24	G 25	T 26	I 27	T 28	T 29	K 30	E 31	L 32
[263-02]	1OSA	Calmodulin	(148)	D 56	A 57	D 58	G 59	N 60	G 61	T 62	I 63	D 64	F 65	P 66	E 67	F 68
[263-03]	1OSA	Calmodulin	(148)	D 93	R 94	D 95	G 96	N 97	G 98	L 99	I 100	S101	A102	A103	E104	L105
[263-04]	1OSA	Calmodulin	(148)	D129	I 130	D131	G132	D133	G134	H135	I 136	N137	Y138	E139	E140	F141
[267-01]	1PAL	Parvalbumin	(107)	D 51	Q 52	D 53	K 54	S 55	G 56	F 57	I 58	E 59	E 60	D 61	E 62	L 63
[267-02]	1PAL	Parvalbumin	(107)	D 90	K 91	D 92	G 93	D 94	G 95	M 96	I 97	G 98	V 99	D100	E101	F102
[319-01]	1REC	Recoverin	(185)	D110	V111	D112	G113	N114	G115	T116	I 117	S118	K119	N120	E121	V122
[329-01]	1RRO	Oncomodulin	(108)	D 51	N 52	D 53	Q 54	S 55	G 56	Y 57	L 58	D 59	G 60	D 61	E 62	L 63
[329-02]	1RRO	Oncomodulin	(108)	D 90	N 91	D 92	G 93	D 94	G 95	K 96	I 97	G 98	A 99	D100	E101	F102
[331-01]	1RTP1	Parvalbumin	(109)	D 51	K 52	D 53	K 54	S 55	G 56	F 57	I 58	E 59	E 60	D 61	E 62	L 63
[331-02]	1RTP1	Parvalbumin	(109)	D 90	K 91	D 92	G 93	D 94	G 95	K 96	I 97	G 98	V 99	E100	E101	F102
[337-01]	1SCMB	Myosin	(138)	D 28	V 29	D 30	R 31	D 32	G 33	F 34	V 35	S 36	K 37	E 38	D 39	I 40
[495-01]	2SAS	Sarcoplasmic calcium-binding protein	(185)	D 70	I 71	N 72	K 73	D 74	D 75	V 76	V 77	S 78	W 79	E 80	E 81	Y 82
[495-02]	2SAS	Sarcoplasmic calcium-binding protein	(185)	D115	V116	S117	G118	D119	G120	I 121	V122	D123	L124	E125	E126	F127
[496-01]	2SCPA	Sarcoplasmic calcium-binding protein	(174)	D104	T105	N106	E107	D108	N109	N110	I 111	S112	R113	D114	E115	Y116
[496-02]	2SCPA	Sarcoplasmic calcium-binding protein	(174)	D138	T139	N140	N141	D142	G143	L144	L145	S146	L147	E148	E149	F150
[578-01]	5PAL	Parvalbumin	(109)	D 51	K 52	D 53	Q 54	S 55	G 56	F 57	I 58	E 59	E 60	E 61	E 62	L 63
[578-02]	5PAL	Parvalbumin	(109)	D 90	S 91	D 92	H 93	D 94	G 95	K 96	I 97	G 98	A 99	D100	E101	F102
[581-01]	5TNC	Troponin C	(161)	D106	K107	N108	A109	D110	G111	F112	I 113	D114	I 115	E116	E117	L118
[581-02]	5TNC	Troponin C	(161)	D142	K143	N144	N145	D146	G147	R148	I 149	D150	F151	D152	E153	F154

All of the detected sites have the same sequence pattern as that of the query, D–x–[DNS]–{ILVFYW}–[DENSTG]–[DNQGHRK]–{GP}–[LIVMC]–[DENQSTAGC]–x(2)–[DE]–[LIVMFYW] that was described in the PROSITE database. For the detail of the description of the sequence pattern, see the text.

Table 3. Results of the 3D Search for the EF-Hand Motif Segments for the 90 Proteins (by NMR) Using SS3D-P

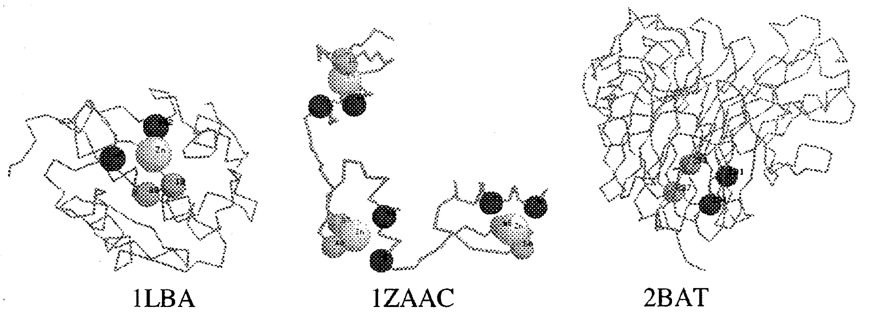
ID No.	PDB code	Protein	Total residues	Hit site (Amino acide identifier)
[24-01]	1CTDA	Troponin C	(34)	D 14 K 15 N 16 A 17 D 18 G 19 Y 20 I 21 D 22 I 23 E 24 E 25 L 26
[85-01]	3PAT	Parvalbumin	(109)	D 51 A 52 D 53 A 54 S 55 G 56 F 57 I 58 E 59 E 60 E 61 E 62 L 63
[85-02]	3PAT	Parvalbumin	(109)	D 90 K 91 D 92 G 93 D 94 G 95 K 96 I 97 G 98 I 99 D 100 E 101 F 102

All of the detected sites have the same sequence pattern as that of the query, see the foot note in Table 2.

Table 4. Results of the 3D Search for the Zinc-Finger Motif Segments for the 543 Proteins (by X-Ray Analysis) Using SS3D-P

ID No.	PDB code	Protein	Total residues	Hit site (Amino acide identifier)
[202-01]	1LBA	Bacteriophage T7 lysozyme	(146)	C 18 C 80 H122 H 68
[400-01]	1ZAAC	ZIF268	(85)	C 7 C 12 H 25 H 29 *
[400-02]	1ZAAC	ZIF268	(85)	C 37 C 40 H 53 H 57 *
[400-03]	1ZAAC	ZIF268	(85)	C 65 C 68 H 81 H 85 *
[413-01]	2BAT	Influenza neuraminidase	(388)	C230 C237 H184 H191

*Detected in PROSITE.



Query: Zinc-finger (C106, C109, H122 and H126 in 1ARD);

Search condition: diffemt kinds of residues are distinguished, and the tolerance value for the distance is 1.6Å.

Fig. 5. Graphical views of the hit proteins and the detected sites obtained by Zinc finger motif search using SS3D-P.

fingers were detected in PROSITE. Our program correctly found all of these sites. The graphical views of the detected sites of these proteins are shown in Fig. 5.

Obviously, the current approach doesn't require a connected substructure or a consecutive peptide segment as the query of the 3D structural feature search, in other words, the SS3D-P allows us to specify an arbitrary set of amino acid residues and/or peptide segments, as the query pattern to be searched, in which the amino acid residues have a particular spatial arrangement in 3D space. These results show that the present approach is successfully applicable for the 3D motif search of proteins.

This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas 'Genome Informatics', from The Ministry of Education, Science, Sports and Culture. We used computer resources at the Human Genome Center, Institute of Medical Science, The University of Tokyo.

References

1) A. M. Lesk, *Commun. ACM*, **22**, 219 (1979).

2) A. T. Brint, H. M. Davies, E. M. Mitchell, and P. Willett, *J. Mol. Graphics*, **7**, 48 (1989).

3) J. R. Ullman, *J. Assoc. Comput. Mach.*, **23**, 31 (1976).

4) S. J. Remington and B. W. Matthews, *J. Mol. Biol.*, **140**, 77 (1980).

5) W. R. Taylor and C. A. Orengo, *J. Mol. Biol.*, **208**, 1 (1989).

6) E. M. Mitchell, P. J. Artymiuk, D. W. Rice, and P. Willett, *J. Mol. Biol.*, **212**, 151 (1989).

7) R. A. Abgyan and V. N. Maiorov, *J. Biol. Struct. Dyn.*, **6**, 1045 (1989).

8) H. M. Grindley, P. J. Artymiuk, D. W. Rice, and P. Willett, *J. Mol. Biol.*, **229**, 707 (1993).

9) Y. Matsuo and M. Kanehisa, *Comput. Appl. Biosci.*, **9**, 153 (1993).

10) K. Mizuguchi and N. Go, *Protein Eng.*, **8**, 353 (1995).

11) H. Kato and Y. Takahashi, *Bull. Chem. Soc. Jpn.*, **70**, 123 (1997).

12) F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Mayer

Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, **112**, 535 (1977).

13) U. Hobohm, M. Scharf, R. Schneider, and C. Sander, *Protein Sci.*, **1**, 409 (1992).

14) J. E. Walker, M. Saraste, M. J. Runswick, and N. J. Gay, *EMBO J.*, **1**, 945 (1982).

15) M. Saraste, P. R. Sibbald, and A. Wittinghofer, *Trends Biochem. Sci.*, **15**, 430 (1990).

16) A. Bairoch, *Nucleic Acids Res.*, **19**, 2241 (1991).

17) A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Cothia, *J. Mol. Biol.*, **247**, 536 (1995).

18) A. Heil, G. Muller, L. Noda, T. Pinder, H. Shirmer, I. Shirmer, and I. Zabern, *Eur. J. Biochem.*, **43**, 131 (1974).

19) N. D. Moncrief, R. H. Krestinger, and M. Goodman, *J. Mol. Evol.*, **30**, 522 (1990).

20) A. Klug and D. Rhods, *Trends Biochem. Sci.*, **12**, 464 (1987).

21) C. Branden and J. Tooze, "Introduction to Protein Structure," Garland Publishing, New York (1991).

22) X. Cheng, X. Zhang, J. W. Pfligrath, and F. W. Studier, *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 4034 (1994).