BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 11 2007, pages 1331–1338
doi:10.1093/bioinformatics/btm121

Structural bioinformatics

Searching for three-dimensional secondary structural patterns in proteins with ProSMoS

Shuoyong Shi[2], Yi Zhong[2], Indraneel Majumdar[2], S. Sri Krishna[2,†] and Nick V. Grishin[1,2,*]

[1]Howard Hughes Medical Institute and [2]Department of Biochemistry, University of Texas Southwestern Medical Center, 5323, Harry Hines Blvd, Dallas, TX 75390-9050, USA

Received and revised on March 2, 2007; accepted on March 18, 2007

Advance Access publication March 24, 2007

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Many evolutionarily distant, but functionally meaningful links between proteins come to light through comparison of spatial structures. Most programs that assess structural similarity compare two proteins to each other and find regions in common between them. Structural classification experts look for a particular structural motif instead. Programs base similarity scores on superposition or closeness of either Cartesian coordinates or inter-residue contacts. Experts pay more attention to the general orientation of the main chain and mutual spatial arrangement of secondary structural elements. There is a need for a computational tool to find proteins with the same secondary structures, topological connections and spatial architecture, regardless of subtle differences in 3D coordinates.

Results: We developed ProSMoS—a Protein Structure Motif Search program that emulates an expert. Starting from a spatial structure, the program uses previously delineated secondary structural elements. A meta-matrix of interactions between the elements (parallel or antiparallel) minding handedness of connections (left or right) and other features (e.g. element lengths and hydrogen bonds) is constructed prior to or during the searches. All structures are reduced to such meta-matrices that contain just enough information to define a protein fold, but this definition remains very general and deviations in 3D coordinates are tolerated. User supplies a meta-matrix for a structural motif of interest, and ProSMoS finds all proteins in the protein data bank (PDB) that match the meta-matrix. ProSMoS performance is compared to other programs and is illustrated on a β-Grasp motif. A brief analysis of all β-Grasp-containing proteins is presented.

Program availability: ProSMoS is freely available for non-commercial use from ftp://iole.swmed.edu/pub/ProSMoS.

Contact: grishin@chop.swmed.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

*To whom correspondence should be addressed.

†Present address: Joint Center for Structural Genomics, Burnham Institute for Medical Research, La Jolla, CA 92037, USA.

1 INTRODUCTION

Evolutionary classification of proteins provides important insights into their biological function (Andreeva *et al*., 2004; Cheek *et al*., 2005). Many rather distant, but functionally relevant connections between proteins were discovered through comparison of spatial structures (Gibrat *et al*., 1996; Koehl, 2001). Structure similarity search programs typically rely on superposition and closeness of either spatial coordinates or inter-residue contacts to define substructures in common between proteins (Eidhammer *et al*., 2000). However, experts who work on structure classification rarely compare two structures to each other the way computer programs do, but instead look for specific 3D motifs in them (Andreeva *et al*., 2004; Lesk, 1995; Pearl *et al*., 2003). Such motifs are defined by the main-chain topology, and general orientation and packing of secondary structural elements. Many recurring 3D motifs have been named, for instance, OB-fold, triple-stranded β-helix, α/β-plait and double-ψ β-barrel. Experts frequently resort to pattern recognition in detection of 3D motifs with the 3D patterns being recognized manually. Despite the emphasis experts put on motifs, currently there is no readily available program that allows a user to search for pre-defined patterns in spatial packing of secondary structural elements. In contrast, such pattern-based approaches have been widely available for sequences (Altschul *et al*., 1997; Sigrist *et al*., 2002). Recently, several automatic structural pattern-matching algorithms have been developed, but they each address a specialized subtask of the general problem (Boutonnet *et al*., 1998; Zotenko *et al*., 2006), or still rely on the structural alignment generated by superposition (Shapiro and Brutlag, 2004). The most widely used pattern search method TOPS also falls short, since, by definition, it finds topological matches, but ignores other important spatial properties (Michalopoulos *et al*., 2004; Torrance *et al*., 2005).

We developed a program, ProSMoS (Protein Structure Motif Search) that searches a library of protein structures for user-defined 3D patterns of secondary structural elements (Fig. 1). *First*, the set of atomic coordinates for a protein is reduced to a set of secondary structural elements (SSEs). Such a simplified representation is frequently used (Camoglu *et al*., 2003; Krissinel and Henrick, 2004; Madej *et al*., 1995), since it avoids highly specific details of individual structures to help in detecting very distant structural relationships. In addition, the

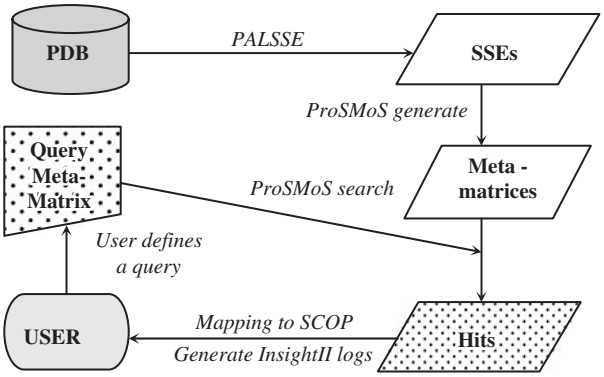

Fig. 1. Workflow of ProSMoS analysis.

most general structural similarity is defined in terms of folds. According to SCOP definition (Andreeva *et al*., 2004), proteins are considered to share the same fold, if they have the same major secondary structures in similar spatial arrangement and with the same topological connections. Since ProSMoS is geared towards detection of fold-level similarities, it is logical to operate on the 3D packing of SSEs. The main difficulty with using SSEs is the reliability of SSE definition. Since regions not included in the secondary structures are not used in a search, we opted for a program that defines SSEs more liberally, and used PALSSE (Majumdar *et al*., 2005), which was developed by our group for this purpose. PALSSE is robust to coordinate errors up to 1.5 Å. SSEs identified by it cover an average of ~85% of residues in structures and mostly agree with expert definition. Classic programs, such as DSSP (Kabsch and Sander, 1983) and Stride (Frishman and Argos, 1995) that were designed primarily for the precise identification of the structural state of each residue, do not necessarily cater to the requirements of SSE-based pattern search. *Second*, for each structure in the protein data bank (PDB) (Berman *et al*., 2000), the SSEs are stored in a meta-matrix (Richards and Kundrot, 1988), which contains information about the type (e.g. α-helix, β-strand) and length of SSEs, the type of contact between these elements (e.g. parallel, anti-parallel, H-bonds, no interaction, any interaction) and the chirality of connections (left- or right-handed). Thus, ProSMoS is not sensitive to finer structural details and can be used to find fold-level structural similarities, or to search for the presence of structural motifs. *Third*, for a query structural pattern provided by the user and defined as a meta-matrix, ProSMoS will search the database of meta-matrices from PDB to find proteins that contain the query structural pattern. The procedure is binary, rather than probabilistic, thus all the returned hits contain the exact query meta-matrix as a submatrix, and all other proteins do not match the pattern. A number of options are provided in ProSMoS to regulate the stringency of the search, such as limits on SSE lengths, contact distance, whether to include circular permutations, etc.

Early versions of ProSMoS have previously been used by our group to classify CASP5 targets (Kinch *et al*., 2003) and other protein structures (Qi and Grishin, 2005). Here, we formally describe this software and demonstrate its features by searching for the β-Grasp motif, which we find present in a number of SCOP folds. Furthermore, we perform comprehensive comparison among TOPS (Michalopoulos *et al*., 2004), SSM (Krissinel and Henrick, 2004) and ProSMoS on a series of widespread protein structural patterns.

2 METHODS

Workflow of ProSMoS analysis is shown on Figure 1. PDB files are processed with PALSSE to generate secondary structure elements (SSEs) and with ProSMoS to generate a database of meta-matrices. User defines a query meta-matrix and searches the database. Resulting hits can be post-processed with several scripts and may be mapped to SCOP or can be visualized in InsightII.

2.1 Database of meta-matrices

Each PDB file is pre-processed to generate a meta-matrix. Meta-matrix contains the following information: SSE types, coordinates of SSE starts and ends, types of interactions between SSEs and β-sheet definitions. Handedness is calculated from the meta-matrix on the fly during searches. The methodological details are briefly outlined below and are explained in the Supplementary Material.

- As defined by PALSSE (Majumdar *et al*., 2005), three types of SSEs are used, namely H, E and L. H is a helical conformation, which includes all types of helices, such as α, 3_{10} and π. E is a β-strand and L is a linker between two consecutive parallel β-strands. The linker is usually a stretch of polypeptide chain in extended conformation (Eswar *et al*., 2003), which does not have a neighbor to form hydrogen bonds with, and thus cannot be called a β-strand. We introduced the linker element, because it is important for the topology. Two parallel β-strands, consecutive in sequence, require a connection between them, and this connection is sometimes supplied by a loop. This loop is recorded in the meta-matrix as a linker L.
- Each SSE is stored as a vector, specified by coordinates of its start and end points. For α-helices, rotational fit method is used to define the helical axis (Christopher *et al*., 1996). C_α atoms of an α-helix (residues i to $i+n-1$) are superimposed with C_α atoms of the same α-helix, but with one-residue shift (residues $i+1$ to $i+n$). Rotation axis of this superposition is taken as the helical axes, and C_α atoms are projected onto it with the first and the last residue projections marking the start and the end of the vector. A different approach is used to define the axes of β-strands. A β-strand is divided into three semi-equal parts, and coordinates of the start and the end of the middle part are used to define the axis. More specifically, for a strand of length N, the axes are defined by two midpoints: between C_α atoms k and $k+1$, and between C_α atoms $N-k$, $N-k+1$, where k is the integer part of $N/3$. The vector is obtained by projecting the coordinates of the first and the last C_α atom of the β-strand onto this axis.
- We define six types of interactions. They are **c**, **t**, **u**, **v**, **N** and **-**. Types **c** and **t** refer to hydrogen-bonded parallel and antiparallel β-strands, respectively. For all other SSE pairs, the presence or absence of interaction is defined by the distance (default <11 Å) overlap (default >2.5 Å) and β-sheet information. The distance is the shortest distance between the C_α coordinates on the two elements. The overlap is the intersection of projections of the two vectors representing SSEs on the line passing through the midpoint of vectors' starts and the midpoints of vectors' ends. No interaction (**-**) is recorded if either distance or overlap criterion fails, or both SSEs are non-hydrogen-bonded β-strands in the same β-sheet. In case both criteria (distance and overlap) are satisfied, the angle φ between the vectors is calculated.

The interactions are **u**, **v** and **N** for $0 \leq \phi < 85^\circ$, $85^\circ \leq \phi < 95^\circ$ and $95^\circ \leq \phi < 180^\circ$, respectively.

- For each β-sheet, β-sheet definition lists all β-strands that are part of the sheet.

2.2 Query meta-matrix

Any meta-matrix from the database can be taken as a query meta-matrix. However, we found that it is useful to define additional criteria for the query. Detailed description of the format is given in the Supplementary Materials. Briefly, 10 types of interactions can be used. In addition to six types described above, we use: **X**, which matches all six symbols of the database meta-matrix, and **x**—matches five symbols with no interaction (-) being a mismatch. **T** and **C** match {v,t} and {u,c}, respectively. The query meta-matrix usually contains information about handedness of connections between SSEs (e.g. SSEs #1,#3,#4—right-handed), desired length ranges for some or all SSEs, and β-sheet definitions. For best results, query meta-matrix should be carefully constructed by user. However, to simplify this process, we provide a script for computing the first approximation to a meta-matrix for a given PDB. This approximation should be further edited to remove some SSEs, to modify some interactions or to introduce other desired changes.

2.3 Search algorithm

Graph path search algorithm (Weiss, 1997) is used to find all possible submatrices in each database meta-matrix that match the query meta-matrix exactly. Thus only the hits that match all the parameters defined in the query file, such as interaction matrix, handedness specifications, length ranges for SSEs and sheet definitions are reported. For each database meta-matrix, the handedness between the elements, for which handedness is defined in the query meta-matrix, is calculated at run time. By default, ProSMoS searches for the motif of interest in every individual chain of the PDB file. Options are available to search for motifs that span through several chains, and for matches with circular permutations. Thus, permuted versions of the query matrix may be generated by ProSMoS and used as queries. The output of ProSMoS is a series of files, one file per PDB that matched the query structure pattern. In each file, residue ranges for the matching motif are listed. Scripts were developed (ftp://iole.swmed.edu/pub/ProSMoS/) to reduce this information to SCOP superfamily representatives and to generate input files to display matches in InsightII or PyMOL for visual study.

3 RESULTS AND DISCUSSION

Although traditional 3D geometric similarity search followed by superposition is indispensable for finding evolutionary relatives, 3D secondary structure pattern matching may be more useful in discovering weak similarities, divergent or convergent in origin. ProSMoS enables a user to find proteins that share a pre-defined structural motif. Such a search may be the first step in structure classification studies, or may be attempted if traditional structure search methods fail to find matches.

3.1 ProSMoS search for the β-Grasp pattern

We illustrate performance of ProSMoS using β-Grasp (ubiquitin-like) fold as an example. The β-Grasp has a two-layer $\alpha+\beta$ architecture with a $\beta\beta\alpha\beta\beta$ core, in which four β-strands form a mixed β-sheet with a strand order 2143. Strands 1 and 4 are parallel, and an α-helix connects strands 2 and 3 to form a right-handed $\beta\alpha\beta$ unit. The query meta-matrix of the β-Grasp 3D pattern (Fig. 2a) encodes these structural features. Elements are numbered consecutively from 1 to 5, regardless of their type. To avoid matches to very short helices, we require the helix to be no less than eight residues.

The database of meta-matrices was derived from 38 156 files in PDB (December 2006). A total of 712 hits matching the β-Grasp query were found. Out of them, 402 were classified in first seven classes of SCOP1.69 (Andreeva *et al.*, 2004), and we focus only on these hits. A total of 28 SCOP1.69 superfamilies from classes 1 to 7 grouped in 16 folds contain at least one representative with a β-Grasp motif. We manually assign each superfamily into one of the three categories defined by the relationship between the β-Grasp motif and the domain structural core (Table 1).

(1) Domains, in which β-Grasp motif is the structural core, i.e. SSEs of the β-Grasp form the center of the structure, and SSEs that are not part of β-Grasp, if present, are smaller, appear secondary and are peripheral.

(2) Domains, in which β-Grasp motif is only a part of the structural core, i.e. gregarious folds (Harrison *et al.*, 2002), which contain β-Grasp motif, but SSEs outside this motif appear equally important evolutionarily and structurally. Thus the core of these domains covers all SSEs of β-Grasp, but, in addition, includes other SSEs as well.

(3) Domains, in which some SSEs of the β-Grasp motif are not part of the structural core, i.e. domains with β-Grasp motif formed by structural drift (Krishna and Grishin, 2005) and only partially overlapping with the core. If some β-Grasp SSEs are absent in most homologs of the domain, the domain is placed in this group.

For the purpose of evolutionary structure classification, domains from the first category could be classified in a β-Grasp fold. Domains from the other two categories contain β-Grasp motif, but their fold is not β-Grasp, because the cores of these domains include SSEs that are not part of the β-Grasp motif. ProSMoS does not make fold assignments, because such assignments require 'fold core' definition. ProSMoS finds all domains containing the user-defined pattern (β-Grasp motif in this example). Then it is up to the user to decide whether the motif is the core, is part of the core, or partially overlaps with the core. Since the core definition is quite subjective and dependent on the library of protein structures available for analysis, assignment of a domain to a particular category may be debated, but the presence of β-Grasp motif (as defined by the query meta-matrix) in all of these domains is a fact.

3.1.1 Category 1: Domains with the β-Grasp core In SCOP, **β-Grasp fold** is typified by ubiquitin (Fig. 2b) (Walters *et al.*, 2004). Ubiquitin-like proteins play key roles in modulating various cellular processes by acting as modifiers or signaling messengers that control many cellular functions, such as cell proliferation, apoptosis, the cell cycle and DNA repair (Hoeller *et al.*, 2006). Along with ubiquitin, some other well-known proteins, such as immunoglobulin-binding domains of proteins G and L, 2Fe-2S ferredoxins and staphylokinase are also placed

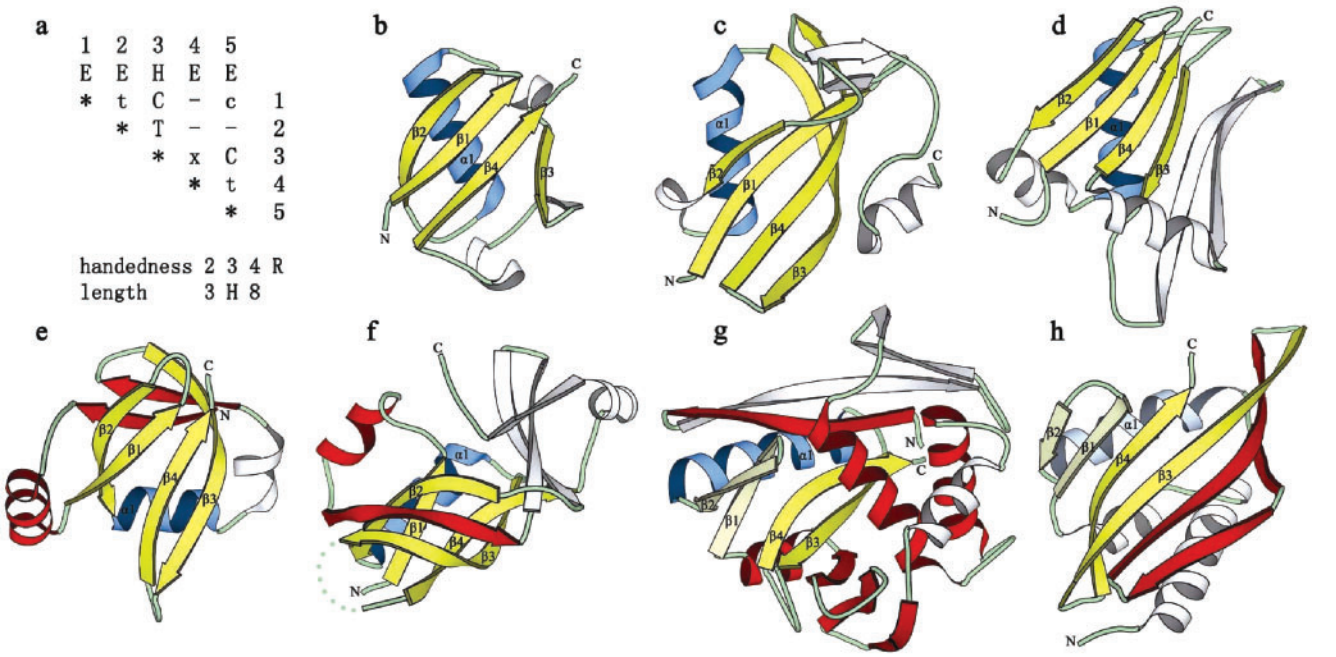

Fig. 2. β-Grasp motif in protein structures. (**a**) The query meta-matrix defining the β-Grasp motif. Secondary structures are consecutively numbered. E (β-strand) and H (helix) indicate the type of secondary structural element (SSE). Lower case letters c and t symbolize parallel and antiparallel hydrogen-bonding interactions between β-strands, respectively. Upper case letters C and T indicate parallel and antiparallel interactions between SSEs, respectively, not considering hydrogen bonds. Lower case letter x stands for the interaction being present regardless of orientation (parallel or antiparallel). Symbol '–' means no interaction between SSE. In 'handedness' line, upper case letter R defines right-handed chirality in a triplet of secondary structures specified by their numbers. The last line in the meta-matrix means that the third element (α-helix) should be longer than seven residues. See Materials and Methods section and Supplementary Materials for further details. Ribbon diagrams of (**b**) d1ubq__, (**c**) d1mut__, (**d**) d1aorb2, (**e**) d1d6ka_, and (**f**) d1vkba_, (**g**) d1su4a3 and (**h**) d1e3va_ were produced using the program MOLSCRIPT (Kraulis, 1991). β-Strands shown in yellow and α-helices shown in blue comprise the core of the β-Grasp motif. These core elements are labeled. Non-β-Grasp structural core SSEs are colored red. Non-core SSEs are white. β-Grasp structural core SSEs that are insertions into the domain core (i.e. β-Grasp formed by structural drift) are shown in lighter colors (hairpin $\beta1\beta2$ in g and h, and helix in h).

in the β-Grasp fold. Proteins in SCOP β-Grasp fold are grouped into 12 superfamilies. All of these superfamilies were found by ProSMoS (Table 1). Thus, we see that ProSMoS does not miss fold-level similarities and finds structurally diverse proteins with the same core motif. Diversity of proteins in the β-Grasp fold manifests in the length variation of the core SSEs and insertion of SSEs or (sub)domains between the core SSEs. Most of these insertions are seen between the strands 3 and 4 of the β-Grasp motif.

In addition to proteins from the SCOP β-Grasp fold, ProSMoS finds four other SCOP folds with the β-Grasp motif forming the core. For three of these folds, SCOP description explicitly mentions the presence of the β-Grasp motif. The mixed sheet of the **nudix fold** contains β-Grasp as its central piece (Fig. 2c). Emphasizing possible importance of non-β-Grasp SSEs in nudix proteins, in particular the C-terminal α-helix covering the other side of the β-sheet thus forming the third layer in nudix architecture, SCOP reserves a separate fold for these proteins. **Anthrax protective antigen** is a multidomain fold, and the domain III (Supplementary Fig.3a) is annotated in SCOP as β-Grasp-like, but the fold is defined to include a composite assembly of all four domains. **Oxidoreductase molybdopterin-binding domain fold** is described in SCOP as 'unusual' and containing β-Grasp like motif. Catalytic domain of sulfite oxidase indeed contains many insertions in the β-Grasp motif (Supplementary Fig. 3b), but we think that evolutionary origin of this domain may be rooted in β-Grasp proteins. The largest insertions to the β-Grasp core are placed at both termini and between the strands 3 and 4, location typical for insertions in ubiquitin-like domains. The fourth fold, in which ProSMoS detected the β-Grasp motif, is the **N-terminal domain of aldehyde ferredoxin oxidoreductase**. This SCOP fold is formed by duplication. Two duplicates assemble through face-to-face interaction between their β-sheets, forming a β-sandwich-like architecture. Each duplicate contains β-Grasp motif, in which a β-hairpin followed by a short α-helix is inserted between strands 3 and 4 (Fig. 2d).

3.1.2 Category 2: β-Grasp motif forms part of the core Folds displaying significant, but partial, structural similarity to other folds were termed gregarious (Harrison *et al.*, 2002). Gregarious folds often contain commonly reoccurring super-secondary structural motifs, which are matching such motifs in other folds. Although we did not find β-Grasp motif to be particularly common, ProSMoS detected it in five SCOP folds, where the β-Grasp SSEs are only a subset of the core SSEs of the fold (Table 1). These gregarious folds

Table 1. SCOP1.69 superfamilies with the β-Grasp motif

Category	SCOP Folds	SCOP Superfamilies	SCOP ID	Residue range in PDB
(1) β-Grasp core	β-Grasp fold (ubiquitin-like)	Ubiquitin-like (d.15.1);	d1ubq__	1–76
		CAD & PB1 domains (d.15.2);	d1c9fa_	8–77
		MoaD/ThiS (d.15.3);	d1fm0d_	1–76
		2Fe-2S ferredoxin-like (d.15.4);	d1frd__	1–92
		Staphylokinase/ streptokinase (d.15.5);	d2sak__	22–135
		Superantigen toxins, C-domain (d.15.6);	d1an8_2	102–208
		Immunoglobulin-binding domains (d.15.7);	d1pgx__	13–69
		Translation factor IF3, N-domain (d.15.8);	d1tif__	15–61
		Glutamine synthetase, N-domain (d.15.9);	d1f1hl1	15–94
		TGS-like (d.15.10);	d1qf6a2	2–61
		Doublecortin (d.15.11);	d1mfwa_	56–133
		TmoB-like (d.15.12)	d1t0qc_	3–85
	Nudix	MutT-like (d.113.1)	d1mut__	2–92
	Anthrax protective antigen	Anthrax protective antigen (f.11.1)	d1acc__	488–594
	Oxidoreductase domain	Sulfite oxidase domain (d.176.1)	d1soxa3	152–166, 218–243, 302–310
	AF oxidoreductase domain	Aldehyde ferredoxin oxidoreductase (d.152.1)	d1aorb2	6–115 and 119–205 (duplication)
(2) Gregarious folds	Ribosomal protein L25-like	Ribosomal protein L25-like (b.53.1)	d1d6ka_	25–94
	BtrG-like	BtrG-like (d.269.1)	d1vkba_	1–7, 59–118
	RNA-polymerase	DNA-dependent RNA-polymerase (e.29.1)	d1i3qb_	577–628
	HP Yml108w	Hypothetical protein Yml108w (d.263.1)	d1n6za_	6–86
	QueA-like	QueA-like (e.53.1)	d1vkyb_	26–51, 285–335
(3) Structural drift	Ferredoxin-like	4Fe-4S ferredoxins (d.58.1)	d1h0hb_	2–8, 160–203
	TIM barrel	Enolase C-domain-like (c.1.11);	d1e9ia1	149–212
		(Trans)glycosidases (c.1.8)	d1fhla_	22–40, 316–325
	α/α-Toroid	Six-hairpin glycosyltransferases (a.102.1)	d1ut9a1	353–377, 434–493
	Metal ATPase domain	Metal cation-transporting ATPase (d.220.1)	d1su4a3	526–601
	Cystatin-like	NTF2-like (d.17.4)	d1e3va_	36–65, 94–120
	Knottins	Growth factor receptor domain (g.3.9)	d1igra3	238–275

ProSMoS results are grouped by SCOP superfamily. Only one representative with β-Grasp motif is shown per superfamily. Approximate residue ranges for the motif are given. β-Grasp motif may cover the entire structural and evolutionary core of the domain (Category 1), be part of the core (Category 2), or partially overlap with the core (Category 3). Since in Category 3 proteins some SSEs of the β-Grasp motif are not part of the domain core, not all superfamily members may have β-Grasp motif.

include two β-barrels, namely **ribosomal protein L25** (Fig. 2e) and **BtrG-like** (Fig. 2f) domains. In addition to β-Grasp motif, SSEs that complete the barrels and α-helices that connect the β-strands in the barrels are essential core elements in these folds (shown in red on Fig. 2). Interestingly, in RPL25, the β-sheet surface of the β-Grasp is convex and is exposed, but in BtrG-like proteins this surface is concave and is buried inside the barrel. BtrG-like structure shows some topological similarity to the ferredoxin-like fold (Andreeva *et al*., 2004), but it is wrapped in a barrel and the β-strand is inserted in the β-sheet between the first and the last β-strands of the typical ferredoxin-like structure. This last β-strand is parallel to the first and is also the last strand of the β-Grasp motif (Fig. 2f). Domain in a β subunit of **DNA-dependent RNA polymerase** (1i3q, residue B545–B633) and **hypothetical protein Yml108w** are related by circular permutation and are likely to be homologous. These proteins adopt β-Grasp topology with an additional N- or C-terminal α-helix, which interacts with the β-sheet and completes the hydrophobic core of the structure (Supplementary Fig. 3c

and d). The larger domain of the **queuosine biosynthesis protein QueA** is structurally complex, but it contains β-Grasp motif on one side of its 9-stranded β-sheet (Supplementary Fig. 3e).

3.1.3 Category 3: β-Grasp motif partially overlaps with the core Part of a domain structure forming in evolution with contribution from insertions may display similarity to evolutionarily unrelated structures. We called this phenomenon structural drift (Krishna and Grishin, 2005), and it is a special case of gregariousness. A protein experiencing structural drift can be described as a hybrid of two overlapping substructures: one is similar to the ancestors of the protein and thus represents the evolutionary core, while the other one covers only part of this core plus some other elements added by insertions. This second substructure may show convergent similarity to other proteins. ProSMoS searches revealed such 'secondary' β-Grasp motifs formed by structural drift in seven SCOP superfamilies. In these examples, some SSEs of β-Grasp motif are part of the main protein core, while other β-Grasp SSEs are insertions and are not present in many related proteins. Category 3 examples are not difficult to detect, since the majority of proteins in the same superfamily and fold will not have β-Grasp motif, which is restricted to a few proteins having the insertions. The drift is not very common, thus most examples occur among the proteins with widespread and populated folds, such as TIM-barrel and ferredoxin-like folds.

As described previously (Krishna and Grishin, 2005), structural drifts resulted in the formation of a β-Grasp motif in one **4Fe-4S ferredoxin** domain from formate dehydrogenase (Supplementary Fig. 3f). Other ferredoxins, or even other superfamilies in the ferredoxin-like fold, do not share this β-Grasp motif. Some members of two TIM β/α-barrel superfamilies: **(Trans)glycosidases** and **enolase-like** barrels, independently developed secondary β-Grasp motifs (Supplementary Fig. 3g and h). Most β-strands in these motifs are apparent insertion to the β/α-barrel core, while the α-helix is the barrel core element. β-Grasp motif is similarly assembled from insertions in **α/α-toroid** fold of cellulose 1,4-β-cellobiosidase CbhA (Supplementary Fig. 3i). ATP-binding domain of **calcium ATPase** has a complex fold (Fig. 2g) somewhat resembling QueA in architecture. The first β-hairpin of the β-Grasp motif is an insertion in the ATPase evolutionary core (compare with potassium-transporting ATPase, PDB:2a29). Similar scenario describes **the nuclear transport factor-2** (NTF2) (Fig. 2h), which belongs to the cystatin-like fold, and only the second β-hairpin of the β-Grasp motif is considered to be part of the α-β_4 core of the fold, as specified in SCOP description. An unusual example is seen in **growth factor receptor domain (**Supplementary Fig. 3j), which is a tandem repeat of many knottin-like disulfide-connected β hairpins. Typically, a loop or a strand-like structure connects these hairpins, but in one case an 8-residue α-helix is present instead. This α-helix, together with two neighboring hairpins, matches the β-Grasp 3D pattern.

3.2 Comparing ProSMoS with TOPS on β-Grasp

We compared the results of ProSMoS with TOPS, an established topology search server (Torrance *et al*., 2005). Since the TOPS database (Michalopoulos *et al*., 2004) is pre-computed to include only domains from the SCOP version 1.61, we restricted our comparison to the first seven classes of this version. A search for the ubiquitin roll-like structural pattern using TOPS found 95 hits to SCOP domains. These domains are from 18 superfamilies, 9 of which belong to the β-Grasp fold. On the same database, ProSMoS found 240 hits that belong to 21 SCOP superfamilies (Supplementary Table 1s). Compared to ProSMoS, TOPS fails to find 6 SCOP superfamilies. Two of these missed superfamilies (Glutamine synthetase N-domain and TGS-like) are from the SCOP β-Grasp fold, for sulfite oxidase domain SCOP mentions the presence of β-Grasp motif, and remaining 3 superfamilies (4Fe-4S ferredoxins, enolase and NTF2-like) we categorized as structural drifts. All these examples are discussed in the previous section. Moreover, TOPS finds ubiquitin-like structure pattern in 3 superfamilies (Ribosome inactivating proteins RIP—d1dm0a_, POZ domain—d1dsxa_ and Porins—1fep) not found by ProSMoS. In RIP superfamily (d1dm0a_), the α-helix does not pack against the β-sheet and the five SSEs found by TOPS do not form a compact hydrophobic core typical for the β-Grasp motif (Supplementary Fig. 3k). In POZ domain (d1dsxa_), the α-helix is distant from the β-sheet and is not packed along it (Supplementary Fig. 3l). Thus RIP and POZ do not have 3D packing of β-Grasp motif, just similar topology. An interesting case is a 'plug' subdomain of transmembrane ferric enterobactin receptor FepA (1fep, Supplementary Fig. 3m). ProSMoS does not detect FepA domain since the α-helix of the β-Grasp motif in it is 7 residues long, as defined by PALSSE. It is just one residue shorter than the length cutoff given in the query matrix. However, if we change the helix lower length limit to 7, FepA will be found. FepA example illustrates that ProSMoS allows more freedom in designing query patterns, and additional options can be used to restrict hits and thus to regulate the stringency of the patterns.

These results do not undermine the TOPS server; they simply demonstrate that 3D structure patterns depend not only on topology, but also on other geometric characteristics of the polypeptide chain. Search for topology alone is insufficient, and ProSMoS is better suited for 3D pattern searches than TOPS.

3.3 Comparing ProSMoS with TOPS and SSM on widespread structure patterns

To illustrate ProSMoS performance further, we carried out a search for eight structure patterns using three programs: ProSMoS, TOPS and SSM (Krissinel and Henrick, 2004) (Table 2). The eight structure patterns are those defined on the TOPS website (http://www.tops.leeds.ac.uk). TOPS search results for these patterns on SCOP 1.61 database have been precomputed by the TOPS authors. However, more importantly, these predefined patterns include some of the most widespread structural motifs known in proteins (Holm and Sander, 1996), and thus offer a good representative sample for our analysis. To construct a meta-matrix for ProSMoS, we took the coordinates of the first hit found by TOPS and selected

Table 2. Comparison among ProSMoS, TOPS and SSM on widespread structure patterns

Structure pattern	P	T	S	P+T	P+S	T+S	P+T+S	P only	T only	S only	R	P/R	T/R	S/R
Greek key	40	21	9	20	3	3	2	19	0	5	24	22/24	13/24	3/24
Immunoglobulin	7	4	5	2	2	2	2	5	2	3	10	6/10	4/10	4/10
Jellyroll	2	2	2	2	1	1	1	0	0	1	2	2/2	1/2	1/2
Key barrel	5	5	5	4	3	3	3	1	1	2	0	0/0	0/0	0/0
Rossmann-like	43	28	10	28	10	8	8	13	0	0	12	12/12	11/12	2/12
Plait (ferredoxin)	62	30	18	29	14	9	9	25	1	1	39	38/39	27/39	16/39
TIM-barrel	21	15	3	17	3	3	3	7	1	0	22	21/22	15/22	3/22
β-Grasp	21	18	12	15	5	5	5	6	3	6	12	12/12	10/12	5/12

Eight structure patterns defined on TOPS website (http://www.tops.leeds.ac.uk, see Supplementary Material for structure pattern diagrams and meta-matrices) were used to compare the performance of the three programs: ProSMoS (P), TOPS (T) and SSM (S). Default parameters were used for TOPS and SSM, except that 100% structure match option was set in SSM to make the results comparable with the other two programs. Hundred percent match requires that all secondary structural elements of the query have matches in a hit. Numbers of SCOP 1.61 superfamilies found by each of the programs (P, T and S) for each pattern are given. Symbol '+' means that both programs found the pattern. 'Only' columns show the number of superfamilies found by a single program only. 'R' gives the number of superfamilies, out of the total number found by all programs, for which SCOP explicitly mentions the target pattern in the structure description. The ratio columns show the fraction of 'R' superfamilies found by a specified program to the total number of R superfamilies.

only those SSEs matched by TOPS. Most of the eight patterns follow the standard definition, except that the immunoglobulin pattern has been defined by TOPS authors to include eight strands instead of the classic 7. Although SSM performs structure similarity search, rather than a pattern search, we also included SSM in comparison, because it is based on secondary structure matching, and an option is provided to match all the SSEs in a query. We restrict the database to first seven true classes in SCOP 1.61, because TOPS results have been precomputed on this database.

The results are summarized in Table 2 and detailed in Tables 3–10 in Supplementary Materials. It is clear that ProSMoS finds more matches than other two programs, and SSM finds the least number of matches. However, SSM produces a large fraction of unique hits that other programs do not find, consistently with SSM being a similarity search rather than a pattern-search approach. Detection of a large number of hits by ProSMoS is important in a structure classification project, which should not miss potential similarities, but the final decision of interpreting pattern-match results should rest with the user. Example of how this should be done is given above for the β-Grasp motif.

It is apparent that finding more hits is good only if the hits are correct. Due to shortage of space, we cannot present analysis as detailed as for the β-Grasp motif, so we resorted to counting SCOP superfamily and fold descriptions that specifically mention motifs in questions. It should be noted that the absence of a motif mention in SCOP description does not mean that the motif is absent. Table 2 shows that the majority of motifs mentioned in SCOP are found by ProSMoS, but not by other programs that miss many of them. As an example, we take TIM-barrel pattern. All TIM-barrels are unified in one SCOP fold and ProSMoS found all 21 TIM-barrel superfamilies, but four, with no hits outside the fold. Out of these four, the only superfamily found by TOPS and not found by ProSMoS (PLP-binding barrel, c.1.6) is a circular permutation according to SCOP comment, and is an α/β-barrel instead of a β/α-barrel, and thus does not match the query meta-matrix, which was used without allowing for circular permutations. Other three TIM-barrel superfamilies are not found by any of the three programs, because the pattern in them is incomplete: e.g. 7-stranded instead of 8-stranded, or one of the α-helices is replaced by a loop, or the barrel is open. TOPS and SSM miss more TIM-barrels (Table 2), and TOPS, as expected, found c.1.6, which does not match the 3D pattern, but matches topogy only.

Our analysis of results also reveals that superfamilies found by other two programs, but not found by ProSMoS, frequently do not contain the 3D patterns in question. For instance, TOPS found two additional Greek key superfamilies (b.15.1, b.28.1). However, those two superfamilies are reported to contain other types of Greek key motifs, not the 'complete Greek key' 3D pattern used as a query (Zhang and Kim, 2000). Table 2 shows that for Rossmann-like, β/α-Plait and Greek-key patterns ProSMoS finds many more hits in addition to those mentioned in SCOP. Manual examination of these structures reveals the presence of motifs sought in all superfamilies found by ProSMoS. In agreement with our findings, many of these hits for Rossmann pattern have been placed among Rossmann-like proteins previously (Aravind *et al*., 2002; Aravind *et al*., 1998; Burroughs *et al*., 2006; Leipe *et al*., 2003).

4 CONCLUSION

We introduced ProSMoS, a program to search for 3D patterns in protein structures. Such searches are particularly useful for detection of distant structure matches. Since only CA atoms are used in the search and substantial deviations in 3D coordinates are tolerated, theoretical models and low-resolution structures are no obstacle for ProSMoS. Our program will be helpful in analysis of protein domains that adopt similar folds with the purpose of functional prediction and in structure classification projects. ProSMoS can easily uncover convergent similarities resulting from structural drift. The origin of such similarities is not evolutionary, but structural, so their analysis may shed light on the rules of folding and assist in structure modeling.

NOTE ADDED IN PROOF

While this manuscript has been in the works, another study dealing with pattern recognition in protein structures has been released [Kamat AP and Lesk AM, ''Contact patterns between helices and strands of sheet define protein folding patterns.'' Proteins.2007; Epub ahead of print]. Conceptually, ProSMoS is similar to the ''tableau'' approach of Kamat & Lesk, however, there is a number of methodological differences. For instance, ProSMoS allows the usage of handedness, can work with Cα coordinates only and uses less restrictive definitions of SSEs and interactions between them. On the other hand, Kamat & Lesk use a clever algorithm of defining relative orientation of SSEs, in contrast to a more simplistic way implemented in ProSMoS.

ACKNOWLEDGEMENT

This work was supported in part by NIH grant GM67165 to N.V.G.

Conflicts of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Andreeva,A. *et al.* (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–229.

Aravind,L. *et al.* (1998) Toprim–a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. *Nucleic Acids Res.*, **26**, 4205–4213.

Aravind,L. *et al.* (2002) Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA. *Proteins*, **48**, 1–14.

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Boutonnet,N.S. *et al.* (1998) Structural classification of alphabetabeta and betabetaalpha supersecondary structure units in proteins. *Proteins*, **30**, 193–212.

Burroughs,A.M. *et al.* (2006) Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. *J. Mol. Biol.*, **361**, 1003–1034.

Camoglu,O. *et al.* (2003) PSI: indexing protein structures for fast similarity search. *Bioinformatics*, **19** (Suppl. 1), i81–i83.

Cheek,S. *et al.* (2005) A comprehensive update of the sequence and structure classification of kinases. *BMC Struct. Biol.*, **5**, 6.

Christopher,J.A. *et al.* (1996) Algorithms for finding the axis of a helix: fast rotational and parametric least-squares methods. *Comput. Chem.*, **20**, 339–345.

Eidhammer,I. *et al.* (2000) Structure comparison and structure patterns. *J. Comput. Biol.*, **7**, 685–716.

Eswar,N. *et al.* (2003) Stranded in isolation: structural role of isolated extended strands in proteins. *Protein Eng.*, **16**, 331–339.

Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment. *Proteins*, **23**, 566–579.

Gibrat,J.F. *et al.* (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.

Harrison,A. *et al.* (2002) Quantifying the similarities within fold space. *J. Mol. Biol.*, **323**, 909–926.

Hoeller,D. *et al.* (2006) Ubiquitin and ubiquitin-like proteins in cancer pathogenesis. *Nat. Rev. Cancer*, **6**, 776–788.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.

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

Kinch,L.N. *et al.* (2003) CASP5 target classification. *Proteins*, **53** (Suppl. 6), 340–351.

Koehl,P. (2001) Protein structure similarities. *Curr. Opin. in Struct. Biol.*, **11**, 348–353.

Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallogr.*, 946–950.

Krishna,S.S. and Grishin,N.V. (2005) Structural drift: a possible path to protein fold change. *Bioinformatics*, **21**, 1308–1310.

Krissinel,E. and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr*, **60**, 2256–2268.

Leipe,D.D. *et al.* (2003) Evolution and classification of P-loop kinases and related proteins. *J. Mol. Biol.*, **333**, 781–815.

Lesk,A.M. *et al.* (1995) Systematic representation of protein folding patterns. *Journal of molecular graphics*, **13**, 159–164.

Madej,T. *et al.* (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.

Majumdar,I. *et al.* (2005) PALSSE: a program to delineate linear secondary structural elements from protein structures. *BMC Bioinformatics*, **6**, 202.

Michalopoulos,I. *et al.* (2004) TOPS: an enhanced database of protein structural topology. *Nucleic Acids Res.*, **32**, D251–254.

Pearl,F.M. *et al.* (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.

Qi,Y. and Grishin,N.V. (2005) Structural classification of thioredoxin-like fold proteins. *Proteins*, **58**, 376–388.

Richards,F.M. and Kundrot,C.E. (1988) Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins*, **3**, 71–84.

Shapiro,J. and Brutlag,D. (2004) FoldMiner: structural motif discovery using an improved superposition algorithm. *Protein Sci.*, **13**, 278–294.

Sigrist,C.J. *et al.* (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinformatics*, **3**, 265–274.

Torrance,G.M. *et al.* (2005) Protein structure topological comparison, discovery and matching service. *Bioinformatics*, **21**, 2537–2538.

Walters,K.J. *et al.* (2004) Ubiquitin family proteins and their relationship to the proteasome: a structural perspective. *Biochim. Biophys. Acta.*, **1695**, 73–87.

Weiss,M.A. (1997) Graph Algorithms. In: Shanklin,J.C. and Hyde,T. (eds.) *Data Structures and Algorithm Analysis in C*. Addison Wesley, pp. 283–345.

Zhang,C. and Kim,S.H. (2000) A comprehensive analysis of the Greek key motifs in protein beta-barrels and beta-sandwiches. *Proteins*, **40**, 409–419.

Zotenko,E., O'Leary,D.P. and Przytycka,T.M. (2006) Secondary structure spatial conformation footprint: a novel method for fast protein structure comparison and classification. *BMC Struct. Biol.*, **6**, 12.