Protein Science (1996), 5:212-220. Cambridge University Press. Printed in the USA.

PROMOTIF—A program to identify and analyze structural motifs in proteins

E. GAIL HUTCHINSON AND JANET M. THORNTON

Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, United Kingdom

(RECEIVED July 24, 1995; ACCEPTED November 16, 1995)

Abstract

We describe a suite of programs, PROMOTIF, that analyzes a protein coordinate file and provides details about the structural motifs in the protein. The program currently analyzes the following structural features: secondary structure; β- and γ-turns; helical geometry and interactions; β-strands and β-sheet topology; β-bulges; β-hairpins; β-α-β units and ψ-loops; disulphide bridges; and main-chain hydrogen bonding patterns. PROMOTIF creates postscript files showing the examples of each type of motif in the protein, and a summary page. The program can also be used to compare motifs in a group of related structures, such as an ensemble of NMR structures.

Keywords: algorithm; analysis; protein structure; structural motif

Protein structures are arranged in hierarchical fashion, starting with the fundamental amino acid building blocks (primary structure) that form helices and strands. These secondary structural units are assembled in various ways to form the final tertiary structure of the protein. Although the structures of a large number of proteins are now known, they display a much smaller repertoire of patterns at the secondary, supersecondary, and tertiary levels. At the tertiary level, it has been observed that there are relatively few distinct domain folds (e.g., Sander & Schneider, 1991; Chothia, 1992; Orengo et al., 1993) in the Brookhaven Protein Data Bank (Bernstein et al., 1977). At lower levels in the structural hierarchy, a relatively small number of motifs has been observed to occur frequently, even in unrelated proteins. These include the basic units of secondary structure—α- and 3_{10} helices, β-strands, and β- and γ-turns, as well as supersecondary structures such as β-hairpins, β-α-β units, and β-sheet topologies.

A more detailed understanding of the location and structure of these commonly occurring motifs in proteins is important for a number of reasons. It may give insight into the relationships between proteins and their possible evolutionary origins. It also deepens our understanding of the relationship between the amino acid sequence and the tertiary structure. This in turn can be used to aid modeling by homology, ab-initio prediction of structure from sequence, and design of novel proteins. Supersecondary structures are also good candidates for nucleation sites in protein folding. Finally, many motifs, such as β-turns and β-bulges, are functionally important, in that they have been found to be involved in active sites and ligand binding surfaces.

For these reasons, detailed analyses of many of these motifs have been carried out (e.g., β-hairpins [Sibanda & Thornton 1985, 1989, Efimov, 1987]; β-turns [Wilmot & Thornton 1988, 1990; Hutchinson & Thornton, 1994]; β-bulges [Richardson, 1981; Chan et al., 1993]), and classification schemes have been devised to describe the conformations in which they occur. Clearly, the large number of protein structures now available means that it is important for the methods to identify and analyze such motifs to be automated. In this paper, we describe a new suite of programs, called PROMOTIF, that analyzes a protein coordinate file and provides details of the structural motifs in that protein.

Results and discussion

This first version of PROMOTIF provides information about the following structural features in proteins: secondary structure; β- and γ-turns; helical geometry and interactions; β-strands and β-sheet topology; β-bulges; β-hairpins; β-α-β units; ψ-loops; disulphide bridges; and main-chain hydrogen bonding patterns. Postscript files are created for each type of motif, and the program also produces a summary page, which gives a briefer description of each motif found in the protein. Most of the motifs are analyzed and classified according to rules defined in published papers. A more detailed description of the analysis and output for each motif is given below, using the protein bovine DNASE I (Brookhaven code 3DNI [Oefner & Suck, 1986]) as an example. For this purpose, this protein has the advantages

Reprint requests to: Janet M. Thornton, Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, United Kingdom; e-mail: thornton@uk.ac.ucl.bioc.bsm.

of being reasonably short and having examples of most of the motifs currently analyzed by PROMOTIF.

Secondary structure

The secondary structure of the protein is calculated using a local implementation (D.K. Smith, unpubl. data) of the DSSP algorithm of Kabsch and Sander (1983). In the standard DSSP algorithm, a residue is included in a secondary structure only if its NH and CO groups form the appropriate hydrogen bonds or, alternatively, for β-sheets only, if the CO_{i-1} and NH_{i+1} groups are involved in the appropriate hydrogen bonds. This gives assignments that broadly agree with IUPAC rule 6.2 (1970), which states that, to be involved in a particular secondary structure, a residue should have ϕ and ψ values close to the ideal values for that secondary structure.

The slightly modified algorithm used in this suite of programs conforms to IUPAC convention rule 6.3, according to which a residue is considered part of a β-sheet or α-helix if either its NH or CO groups are involved in the appropriate hydrogen bonds. In practice, this means that one extra residue is added to the ends of each strand and helix where possible and designated as lower case "h" and "e." This rule is the one most commonly used among crystallographers. The secondary structure thus calculated provides the raw data used for the remainder of the analyses.

Turns

β-Turns

A β-turn is defined as four consecutive residues (denoted by i, $i+1$, $i+2$, $i+3$) where the distance between the C_α atoms of residues i and $i+3$ is less than 7 Å and where the central two residues are not helical (Lewis et al., 1973). β-Turns are classified according to the ϕ,ψ angles of residues $i+1$ and $i+2$ (Venkatachalam, 1968; Richardson, 1981). The ideal angles for each of the turn types are shown in Table 1. The ϕ,ψ angles are allowed to vary by $\pm 30°$ from these ideal values, with the added flexibility of one angle being allowed to deviate by as much as 40°. Turns that do not fit any of the above criteria are classified as type IV.

Figure 1 shows an example of the color postscript schematic diagram produced by PROMOTIF for the β-turns in DNASE I. The program produces a Ramachandran plot and schematic diagram for each β-turn identified in the protein. The protein has a total of 20 β-turns, of which 8 are illustrated here.

The program also produces a black and white table (Table 2A) giving more detailed information about each β-turn. For the β-turns and all other motifs in the protein, machine-readable flat files (not shown) giving most of the information are also created.

γ-Turns

A γ-turn is defined as three consecutive residues (i, $i+1$, $i+2$) with a hydrogen bond between residues i and $i+2$ and where the ϕ,ψ angles of residue $i+1$ fall within 40° of the standard angles of either of the two classes: classic and inverse (Table 1) (Rose et al., 1985; Milner-White et al., 1988).

As for β-turns, the PROMOTIF output for γ-turns consists of a color diagram showing a Ramachandran plot and schematic diagram for each γ-turn identified in the protein DNASE I (Fig. 2). More detailed information is given in a table (Table 2B) and flat file.

Table 1. *Ideal angles for β- and γ-turn types*

Turn type	ϕ,ψ $(i+1)$	ϕ,ψ $(i+2)$
β-Turns		
I	−60°, −30°	−90°, 0°
II	−60°, 120°	80°, 0°
VIII	−60°, −30°	−120°, 120°
I′	60°, 30°	90°, 0°
II′	60°, −120°	−80°, 0°
VIa1[a]	−60°, 120°	−90°, 0°
VIa2[a]	−120°, 120°	−60°, 0°
VIba	−135°, 135°	−75°, 160°
IV	Turns excluded from above categories	
γ-Turns		
Classic	75°, −64°	–
Inverse	−79°, 69°	–

[a] Also requires *cis*-proline at position $i+2$. The ϕ,ψ angles for type VI β-turns were defined originally by Richardson (1981). Types VIa1 and VIa2 β-turns were used by Hutchinson and Thornton (1994) to distinguish between the two subclasses of type VIa β-turns with ϕ,ψ of residue $i+1$ in the β_E and β_P regions of the Ramachandran plot. γ-Turns have ϕ,ψ angles defined only for position $i+1$ and require a hydrogen bond between residues i and $i+2$.

Disulphide bridges

Disulphide bridges are identified for two cysteine residues whose sulphur atoms are less than 3 Å apart. Richardson (1981) iden-

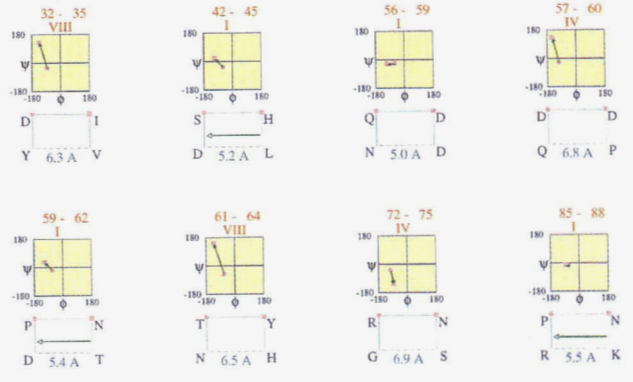


Fig. 1. Color postscript output generated by PROMOTIF for the first eight β-turns in DNASE I. Ramachandran plots show the ϕ and ψ angles of residues $i+1$ (purple circle) and $i+2$ (purple square) in the turn. These are connected by an arrow from $i+1$ to $i+2$. Beneath each Ramachandran plot is a schematic diagram showing the one-letter amino acid codes at each position in the turn, with residues $i+1$ and $i+2$ indicated by circle and square as in the Ramachandran diagram. The distance in Å between the C_α atoms of residues i and $i+3$ is indicated by the blue dotted line and figure at the bottom of the diagram. Where a hydrogen bond exists between the NH of residue $i+3$ and the CO of residue i, this is indicated by a green arrow. The sequence numbers of the residues involved in each turn and the turn type are indicated above each Ramachandran plot.

Table 2. *β- and γ-Turns in DNASE I*[a]

A. PROMOTIF listing of the first eight β-turns in DNASE I

Residue numbers	Sequence	Turn type	(i + 1) φ	(i + 1) ψ	(i + 2) φ	(i + 2) ψ	φ,ψ region	Chi1 (i + 1)	Chi1 (i + 2)	i to i + 3 Distance
32–35	Y D I V	VIII	−83.7	−34.2	−132.7	128.4	AB	−57.0	−56.3	6.3
42–45	D S H L	I	−57.6	−35.8	−110.3	17.2	AA	−76.8	−42.4	5.2
56–59	N Q D D	I	−59.3	−24.3	−106.5	−26.1	AA	92.9	−62.1	5.0
57–60	Q D D P	IV	−106.5	−26.1	−145.2	127.2	AB	−62.1	−170.8	6.8
59–62	D P N T	I	−66.3	−15.8	−114.7	32.0	Aa	32.9	−62.6	5.4
61–64	N T Y H	VIII	−66.2	−53.0	−130.4	138.7	AB	−45.4	−76.3	6.5
72–75	G R N S	IV	−103.0	−38.8	−85.6	−124.2	A	−161.2	−54.8	6.9
85–88	R P N K	I	−67.1	−16.8	−83.6	−14.6	AA	19.8	53.1	5.5

B. Details of all γ-turns in DNASE I from PROMOTIF

Residue numbers	Sequence	Turn type	(i + 1) φ	(i + 1) ψ	i to i + 3 Distance
170–172	N A D	Inverse	−75.5	38.9	5.9
209–211	C A Y	Inverse	−94.1	70.7	5.6
233–235	F D F	Inverse	−77.2	96.1	5.8

[a] **A:** From left to right are listed sequence numbers of the first (*i*) and last (*i* + 3) residues in each turn, the one-letter amino acid codes for each of the four residues in the turn, the turn type, the φ and ψ angles of residues *i* + 1 and *i* + 2, the regions of the Ramachandran plot occupied by residues *i* + 1 and *i* + 2, as classified by Efimov (1991) (A, core α-helical region; a, the region immediately surrounding this, occupied by less well-defined α-helical regions; B, β-sheet region; P, polyproline region; L, left-handed α-helical region (α_L); G, γ_L part of the left-handed α-helical region; E, the ε region in the bottom right-hand corner of the Ramachandran plot, which is occupied mainly by glycine residues). A blank space in the columns indicates that the φ,ψ values of the residue fall outside the major allowed regions of the Ramachandran plot. The final columns show the χ_1 angles of residues *i* + 1 and *i* + 2 and the distance between the C_α atoms of residues *i* and *i* + 3. **B:** From left to right are listed the sequence numbers of the first (*i*) and last (*i* + 2) residues in the turn, the one-letter amino acid codes for each of the three residues in the turn, the turn type (classic or inverse), the φ and ψ angles of residue *i* + 1, and the distance between the C_α atoms of residues *i* and *i* + 2.

tified several categories of disulphides based on their internal χ angles, in particular χ_2, χ_3, and χ_2' angles. She found that the majority of disulphides could be classed as left-handed spirals or right-handed hooks. We have loosely classified disulphides based only on the signs of these χ angles, into four categories (Table 3A) where appropriate.

PROMOTIF produces a table (Table 3B) that gives details of each disulphide bridge found in the protein. DNASE I has two disulphide bridges, one of which is classified as a short right-handed hook, having central χ angles with (−,+,−) signs, which is quite unusual. The second disulphide is even more unusual in that its central χ angles have (−,−,+) signs and it does not therefore fall into any of the above categories.

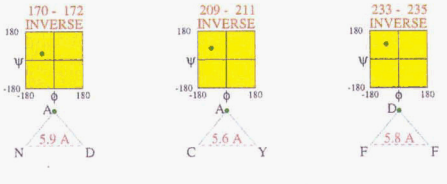


Fig. 2. Color postscript diagram generated by PROMOTIF for the γ-turns in DNASE I. Ramachandran plots show the φ and ψ angles of residue *i* + 1 (green circle) in the turn. Schematic diagrams underneath show the one-letter amino acid codes at each position in the turn. The distance in Å between the C_α atoms of residues *i* and *i* + 2 is indicated at the bottom of the diagram. The sequence numbers of the residues involved in the turn and the turn type are indicated above the Ramachandran plot.

Helices

PROMOTIF generates a table giving basic information about each helix identified by the secondary structure assignment program (Table 4). Helical wheels and nets are drawn for each helix, assuming that there are 3.6 residues per turn (Fig. 3 shows these wheels and nets for helices 3–6 in the protein). Finally, PROMOTIF provides information about the interacting pairs of helices in the protein (Table 5). Two helices are defined as "interacting" if they contain one or more atoms within 4.5 Å of one or more atoms of the other helix.

β-Strands and β-sheets

More basic data are provided for each β-strand (Table 6), with a second table (Table 7) giving information about each β-sheet. The topology of the sheet is given using the nomenclature of Richardson (1981). This assigns a number to the connection between each pair of sequential strands in the sheet. The number represents the number of strands the connection traverses in the sheet, and in which direction, with an "X" added for crossover connections. Thus, a β-hairpin would have a "+1" connection, and a β-α-β unit a "+1X" connection.

Table 3. *Disulphide bridges in DNASE I*[a]

A. Disulphide bridges identified by Richardson (1981)

Disulphide type	χ_2	χ_3	χ_2'
Left-handed spiral	−(−90°)	−(−90°)	−(−90°)
Right-handed hook	+(+120°)	+(90°)	−(−50°)
Right-handed spiral	+(+110°)	+(90°)	+(40°)
Short right-handed hook	−(−120°)	+(90°)	−(−110°)

B. Disulphide bridges identified by PROMOTIF

Cysteine 1	Cysteine 2	Chi1	Chi2	Chi3	Chi2′	Chi1′	C_α distance	Type
101	104	−100.5	−97.4	133.1	−40.4	−98.1	4.7	Short right-
173	209	−72.1	166.0	−100.3	−70.9	−57.1	5.9	hand hook

[a] **A:** The χ_2 and χ_2' values can be interchanged because they merely reflect the "starting" cysteine. Only the signs of the χ values were used in the classification of the disulphide bridges. Values given in the table represent typical values for each angle read from the graph given in Richardson (1981). **B:** From left to right are listed the two cysteine residues involved in the disulphide bridge, the internal χ angles (χ_1, χ_2, χ_3, χ_2', and χ_1'), the distance between the C_α atoms of the two cysteines, and the classification according to Table 3A.

β-Bulges

β-Bulges are regions of irregularity in a β-sheet, where the normal pattern of hydrogen bonding is disrupted, e.g., by the insertion of an extra residue. PROMOTIF uses the algorithm recently described by Chan et al. (1993) to identify and classify bulges in proteins. Figure 4 and Table 8 show the bulges found in DNASE I. Bulges are classified as parallel or antiparallel, depending on the relative orientation of the two β-strands involved. Within each of these categories, bulges are further subdivided into classic, wide, bent, G1, and special types, depending on the number of residues involved and the hydrogen bonding pattern. Thus, a bulge could, for example, be described as antiparallel classic. Classic and wide bulges both involve an extra residue on one β-strand relative to its neighboring strand. In antiparallel β-sheet, the classic bulges occur where the extra residue is between two narrowly spaced pairs of hydrogen bonds, whereas in the case of the wide bulges, the extra residue is between the widely spaced pairs of hydrogen bonds. Corresponding hydrogen bonding patterns for parallel classic and wide bulges can be found in Chan et al. (1993). Bent bulges occur much less frequently, and have one extra residue on both strand partners. G1 bulges occur only in antiparallel sheets; in these cases, residue 1 is in the α_L conformation and is therefore usually glycine. This usually occurs at the end of a β-strand. Special bulges involve a larger insertion of up to three residues in one strand and, like bent bulges, are very rare.

Figure 4 shows a Ramachandran plot and schematic hydrogen bonding diagram for each bulge in DNASE I. The sequence information and detailed ϕ, ψ values for residues X, 1, and 2 are shown in Table 8. Further information about residues 3 and 4, where present, is given in a flat file.

Table 4. *Helices in DNASE I*[a]

Helix number	Start residue	End residue	Helix type	No. of residues	Length (Å)	Unit rise (Å)	Residues per turn	Pitch (Å)	Deviation from ideal (°)	Sequence
1	13	16	α	4	6.30	1.44	3.89	5.60	39.9	ETKM
2	19	29	α	11	16.67	1.48	3.67	5.41	9.9	ATLASYIVRIV
3	46	55	α	10	15.23	1.49	3.55	5.29	7.4	VAVGKLLDYL
4	137	139	3_{10}	3	–	–	–	–	–	PSD
5	140	158	α	19	28.09	1.48	3.79	5.59	26.2	AVAEINSLYDVYLDVQQKW
6	178	183	3_{10}	6	11.07	1.92	3.16	6.05	12.2	SSQWSS
7	185	188	α	4	5.79	1.23	4.05	5.00	42.3	RLRT
8	219	224	α	6	9.34	1.47	3.47	5.10	5.4	SLLQSS
9	235	239	α	5	7.66	1.47	3.72	5.46	21.9	FQAAY
10	243	249	α	7	11.33	1.56	3.56	5.56	9.3	NEMALAI

[a] Helices are numbered consecutively from the N terminus (left-hand column) to allow helix interactions to be described. The start and end residue of each helix, helix type (α or 3_{10}), number of residues, and amino acid sequence are given for each helix. Geometrical parameters calculated are the length of the helix and the unit rise (both in Å), the number of residues per turn (ideally 3.6 for α-helices), the helix pitch in Å, and a measure of the deviation of the helix geometry from an ideal helix (in degrees). This latter value should be 0 for a perfect helix. These parameters are not calculated for helices with less than four residues.

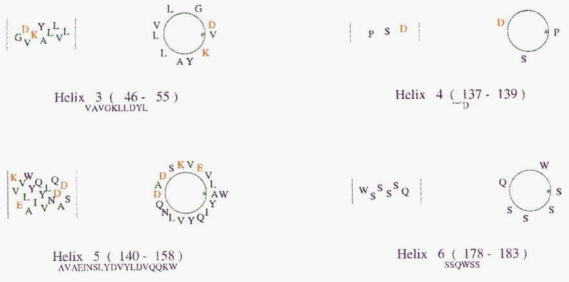


Fig. 3. Helical nets and helical wheels drawn for helices 3–6 in DNASE I. In both cases, the residues are indicated by their one- letter amino acid codes and color-coded for hydrophobic (green), polar (blue), and charged (red) amino acid types. The N-terminal residue of each helix is at the bottom left-hand corner of the helical net. In the helical wheel, the first residue is indicated by an asterisk and subsequent residues are plotted at 100° intervals, going around the circle in a clockwise fashion. The complete sequence and the residue numbers of the start and end of the helix are shown underneath the corresponding pair of diagrams.

Table 6. *β-Strands in DNASE I*[a]

Strand number	Start residue	End residue	Sheet label	No. of residues	Sequence
1	2	11	A	10	KIAAFNIRTF
2	34	40	A	7	IVLIQEV
3	64	67	A	4	HYVV
4	79	84	A	6	RYLFLF
5	89	96	B	8	VSVLDTYQ
6	114	120	B	7	AVVKFSS
7	127	132	B	6	EFAIVA
8	163	168	B	6	VMLMGD
9	193	194	B	2	QW
10	212	217	B	6	DRIVVA
11	231	232	A	2	AP
12	255	258	A	4	VEVT

[a] Strands are numbered consecutively from the N-terminus of the protein. Start and end residue numbers for each strand, the β-sheet to which it belongs (these are labeled A, B, C . . . from the first β-sheet found in the protein), the number of residues in the strand, and the amino acid sequence are given.

β-Hairpins

β-Hairpins consist of two β-strands that are antiparallel and hydrogen bonded together. The hairpins are classified as in Sibanda et al. (1989) using two numbers X:Y, which denote the number of residues in the loop between the two strands defined using two different IUPAC conventions (1970). If the end strand residues form two hydrogen bonds, then X = Y. If the distal hydrogen bond is not formed, the number of residues in the loop depends on which IUPAC definition of strands is used. In practice, if the end hydrogen bond is not formed, then Y = X + 2.

For the smaller loops, the hairpins are dominated by the formation of β-turns (usually I′ and II′) (Sibanda & Thornton, 1985). The 3:5 hairpins are dominated by one well-defined conformation, which can be described as a type I turn followed by a G1 bulge. The most common class among the 4:4 hairpins contains a type I β-turn. Where these particular conformations occur, they are indicated by appropriate letters after the main classification (e.g., 2:2I′; 3:5IG; 4:4I).

PROMOTIF plots a schematic diagram for each hairpin found in the protein, displaying the residue numbers and hairpin class (Fig. 5). The sequence and hydrogen bonding patterns involved in the strands and loop are shown on the right-hand side. This information is summarized in tabular form (Table 9).

β-α-β Motifs and ψ-loops

Preliminary information is given for β-α-β units and ψ-loops located in the protein. β-α-β Units consist of two parallel hydrogen bonded β-strands connected by an α-helix. A simple table gives information about the locations of these (Table 10).

ψ-Loops consist of two antiparallel strands connected by a "+2" connection, i.e., with one strand in between, hydrogen bonded to both of them (Tang et al., 1978). In contrast to β-α-β units and β-hairpins, these occur very rarely in proteins (Hutchinson & Thornton, 1990). Again, PROMOTIF gives simple information about the location of any ψ-loops that occur. No ψ-loops are found in DNASE I, but, where they are found, the

Table 5. *Pairwise interactions between helices*[a]

Helix 1	Helix 2	Distance (Å)	Omega (°)	No. of interacting residues		
				Total	Helix 1	Helix 2
1	2	5.5	−94.8	5	2	3
1	3	6.9	31.8	6	2	5
1	10	10.4	−130.1	2	1	2
2	3	10.4	−64.1	7	5	3
2	9	12.1	101.7	4	3	2
2	10	9.6	−75.7	8	5	3
5	6	9.8	−26.2	3	2	2
5	7	7.0	53.3	7	6	2
6	7	4.3	−71.4	4	3	3
9	10	9.7	125.7	4	2	3

[a] Helix numbers (as in Table 6) for the two helices involved, their distance of closest approach (in Å), and interaction angle are provided, as well as (from left to right) the number of interacting pairs of residues and the number of residues in each of the two helices involved in the interaction.

Table 7. *β-Sheets present in the protein*[a]

Sheet letter	No. of strands	Sheet type	Topology
A	6	Mixed	−1X −2X 1 4 −1
B	6	Mixed	1 1 1X 2X −1

[a] Sheet letters correspond to those used for the strands in Table 6. Information is provided about the number of β-strands and the sheet type (parallel, antiparallel, or mixed). If the sheet forms a closed β-barrel, this is also mentioned under Sheet type. Sheet topology described using the nomenclature of Richardson (1981) is also listed.

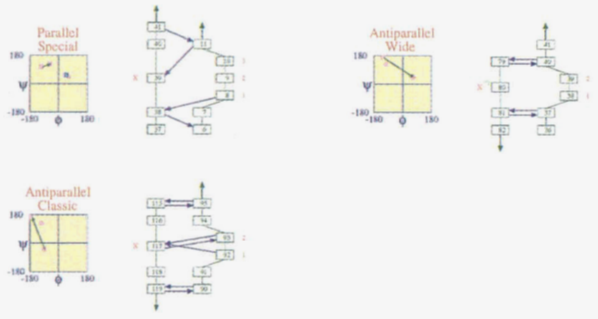


Fig. 4. Color postscript diagrams showing the three β-bulges in DNASE I. Ramachandran plot for each bulge indicates the ϕ, ψ values for residues X (purple circle), 1, and 2 (purple squares linked by a green arrow from residue 1 to 2). If there are further residues involved in the bulge, these are indicated by blue squares. The bulge type is indicated above the Ramachandran plot. To the right of each Ramachandran there is a schematic diagram showing the main-chain hydrogen bonding pattern in and around the bulge.

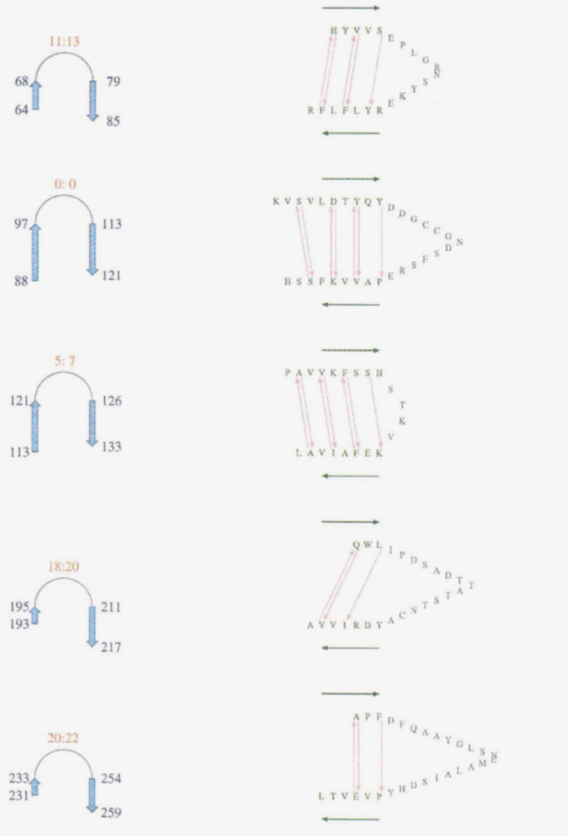


Fig. 5. Color postscript diagram illustrating the β-hairpins in DNASE I. Residue numbers corresponding to the ends of the β-strands involved and the hairpin class are shown in the left-hand diagram in each case. (The second hairpin has not been classified – it's classification is given as 0:0 because there is a chain break in the loop between the strands.) The right-hand diagram in each case shows the one-letter codes of the residues involved in the strands (green) and loop (purple) of each hairpin. Main-chain hydrogen bonds in the strands and at the ends of the loop are shown by pink arrows.

table generated by PROMOTIF is similar in format to that produced for β-α-β motifs (Table 10).

Main-chain hydrogen bonding patterns

PROMOTIF will plot a color schematic diagram drawn by the program HERA (Hutchinson & Thornton, 1990) to illustrate the main-chain hydrogen bonding patterns in the β-sheets and helices of the protein. This should help to locate the motifs described above in the context of the whole protein (Fig. 6).

Summary page

The above detailed information for each motif can be quite lengthy and is not always required. PROMOTIF also produces a summary page for use when only a brief summary of the motifs in a protein is required (Fig. 7). The top part of the page summarizes the secondary structural information, and the remainder of the summary page consists of mini-tables giving the location and classification of the examples of each motif type found in the structure.

Multiple protein version

PROMOTIF may also be run on a list of protein structures. This multiple protein version can be used in two ways. The first mode of operation generates composite flat files containing lists of all the examples of each motif in the list of proteins. This could, for example, be used to produce a list of β-turns in a nonhomologous protein data set. The list could then be further processed by the user to yield, e.g., sequence preferences for the different β-turn types (cf. Hutchinson & Thornton, 1994).

The second mode of operation generates a set of figures and tables as described in the preceding paragraphs for each protein

Table 8. *β-Bulges in DNASE I*[a]

Residue numbers			Sequence			Bulge	X		1		2	
X	1	2	X	1	2	type	ϕ	ψ	ϕ	ψ	ϕ	ψ
39	8	9	E	I	R	P S	66.9	43.0	−104.7	105.2	−51.1	125.0
80	38	39	Y	Q	E	A W	−101.2	123.7	−119.9	166.5	66.9	43.0
117	92	93	K	L	D	A C	−103.0	124.7	−88.1	−37.7	−166.9	164.2

[a] Residue numbers and one-letter amino acid codes for residues X (on the normal strand) and residues 1 and 2 (on the bulged strand) are provided. The bulge type is described using two letters: the first letter is P or A depending on whether the bulge involves parallel or antiparallel β-strands; the second letter can be C(lassic), W(ide), G1, B(ent), or S(pecial). ϕ and ψ angles for residues X, 1, and 2 are listed.

Table 9. *β-Hairpins in DNASE I*[a]

Strand 1		Strand 2		Number of residues		Hairpin class
Start	End	Start	End	Strand 1	Strand 2	
64	68	79	85	5	7	11:13
88	97	113	121	10	9	0:0
113	121	126	133	9	8	5:7
193	195	211	217	3	7	18:20
231	233	254	259	3	6	20:22

[a] Start and end residues of the two β-strands involved in the hairpin, the number of residues in each of these strands, and the hairpin class are listed.

Table 10. *β-α-β Motifs found in the protein*[a]

Strand 1		Strand 2		Strand 1 length	Strand 2 length	Loop length	Helix length
Start	End	Start	End				
2	11	34	40	10	7	22	15
127	132	163	168	6	6	30	19

[a] Information about each motif includes start and end residue numbers of each of the strands involved, and the number of residues involved in the two strands, the intervening loop, and the helical part of the loop.

in the list. In addition, for cases where a set of related proteins is being used, there is also the option to produce a color picture in which the motifs in the proteins are compared. This could be used for comparing a set of highly homologous proteins such as several hemoglobins. Here the positions of the motifs in each protein are very similar and, indeed, even the details such as the turn types and the ϕ, ψ angles are strongly conserved. Alternatively, the program, run in this mode, could be used to compare a set of models from an ensemble of solved NMR structures of the same protein. These models can differ considerably even in the location of the motifs, and details such as β-turn types at a given position are rarely conserved across the whole ensemble.

An example showing the use of the program to compare an ensemble of NMR structures is shown in Figure 8. The figure compares the motifs found in the first five members of the ensemble of NMR structures of the heat shock transcription factor from *Drosophila melanogaster* (Brookhaven code 1HKT, Vuister et al., 1994). For brevity, only the first 76 residues are shown. At the moment, the program assumes that the structures in the data set are aligned with identical sequence numbers at

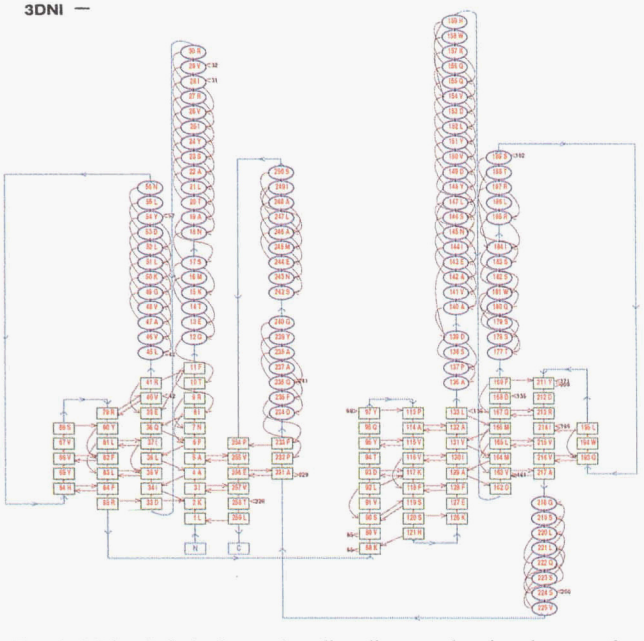


Fig. 6. Main-chain hydrogen bonding diagram showing the secondary structures in DNASE I (Brookhaven code 3DNI). Residues in the β-sheet are indicated by green rectangles and those in helices by blue circles. Within each box, the sequence number and one-letter amino acid code are shown for each residue. Main-chain hydrogen bonds defined by the secondary structure algorithm are indicated by brown arrows.

ENDONUCLEASE DEOXYRIBONUCLEASE I DNASE I E.C.3.1.21.1 BOVINE BOS TAURUS

Sequence:

LKIAAFNIRTFGETKMSNATLASYIVRIVRRYDIVLIQEVRDSHLVAVGKLLDYLNQDDPNTYHYVVSEPLGRNSYKERY
LFLFRPNKVSVLDTYQYDDGCCGNDSFSREPAVVKFSSHSTKVKEFAIVALHSAPSDAVAEINSLYDVYLDVQQKWHLND
VMLMGDFNADCSYVTSSQWSSIRLRTSSTFQWLIPDSADTTATSTNCAYDRIVVAGSLLQSSVVPGSAAPFDFQAAYGLS
NEMALAISDHYPVEVTLT

Number of residues observed = 258	Number of disordered residues = 2	Number of chains is: 1
Number of strands = 12	Number of alpha helices = 8	Number of 3,10 helices = 2
% strand residues = 26.4	% alpha helical residues = 25.6	% 3,10 helical residues = 3.1

Secondary Structure Summary:

EHHTETHTEETEEEGHETGHTETEHTEHHE

SHEETS

Number strands	Name	Type	Barrel	Topology
6	A	MIX	N	-1X -2X 1 4 -1
6	B	MIX	N	1 1 1X 2X -1

STRANDS

Location	Sheet
2 - 11	A
34 - 40	A
64 - 67	A
79 - 84	A
89 - 96	B
114 - 120	B
127 - 132	B
163 - 168	B
193 - 194	B
212 - 217	B
231 - 232	A
255 - 258	A

HELICES

Location	Type
13 - 16	H
19 - 29	H
46 - 55	H
137 - 139	3
140 - 158	H
178 - 183	3
185 - 188	H
219 - 224	H
235 - 239	H
243 - 249	H

DISULPHIDES

Cys 1	Cys 2
101	104
173	209

B-TURNS

Location	Seq	Type
32 - 35	YDIV	VIII
42 - 45	DSHL	I
56 - 59	NQDD	I
57 - 60	QDDP	IV
59 - 62	DPNT	I
61 - 64	NTYH	VIII
72 - 75	GRNS	IV
85 - 88	RPNK	I
86 - 89	PNKV	I
109 - 112	FSRE	VIII
120 - 123	SHST	I
123 - 126	TKVK	IV
125 - 128	VKEF	IV
172 - 175	DCSY	II'
173 - 176	CSYV	I
174 - 177	SYVT	VIII
189 - 192	SSTF	IV
194 - 197	WLIP	VIII
197 - 200	PDSA	I
226 - 229	VPGS	II

G-TURNS

Location	Seq	Type
170 - 172	NAD	INV
209 - 211	CAY	INV
233 - 235	FDF	INV

BULGES

Location		Type	Seq
39	8	9	P S EIR
80	38	39	A W YQE
117	92	93	A C KLD

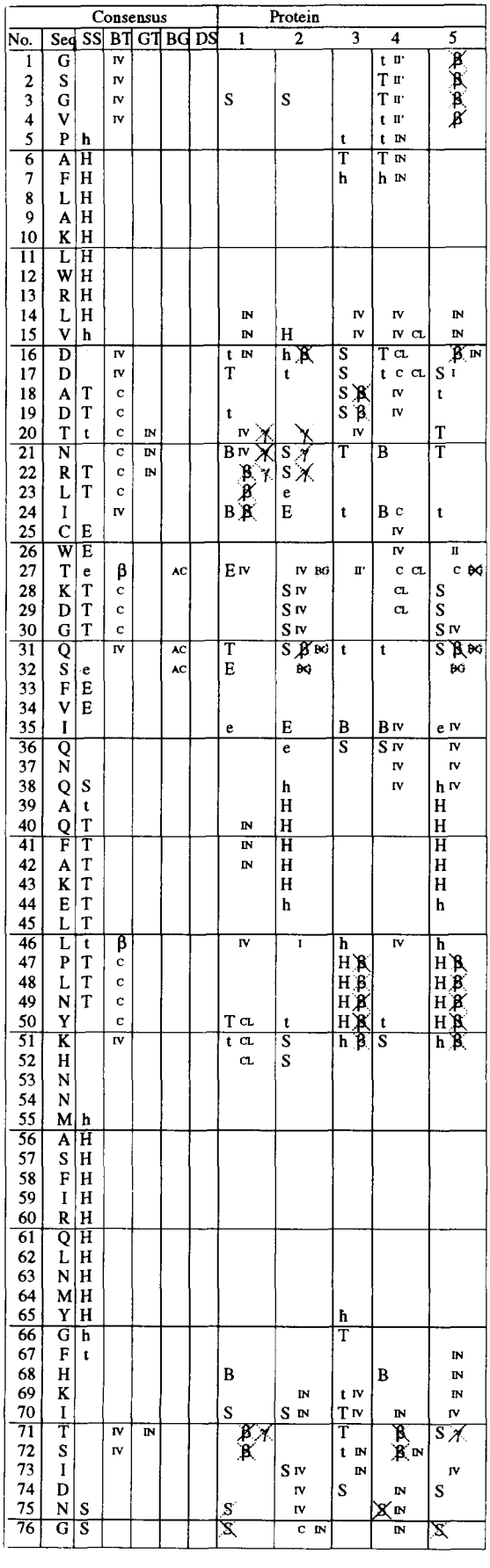
HAIRPINS

Location	Class
64 - 85	11:13
88 - 121	0: 0
113 - 133	5: 7
193 - 217	18:20
231 - 259	20:22

BETA ALPHA BETA UNITS

Start	End
2	40
127	168

Fig. 7. Summary page generated by PROMOTIF for DNASE I. The top part of the diagram shows the name of the protein, the amino acid sequence, and the number of residues and chains. The number of residues and sequence refer to those residues actually observed in the electron density. Any difference between this and the actual sequence is indicated by the number of disordered residues. The number and percentage of residues in each secondary structural type are shown below this. A string of characters indicates the sequence of secondary structural units as assigned by the Kabsch and Sander algorithm (E, strand; H, α-helix; G, 3_{10} helix; T, β-turn). For turns in particular, this may not correspond exactly to the assignments by PROMOTIF, because the Kabsch and Sander assignments are based solely on hydrogen bonding criteria. In the second part of the figure, mini-tables summarize the location of each type of motif in the protein as found by PROMOTIF.



equivalent positions in all structures. A consensus structure is displayed alongside the sequence and residue numbers on the left-hand side of the figure. By default, the consensus is calculated using all motifs that occur in more than half the set of structures. The fraction of structures used to calculate the consensus can be varied by the user. For clarity, only differences from this consensus are displayed for each protein in the data set. The secondary structure is mostly conserved across the five structures shown. However, helices are found from residues 38 to 44 in just two structures (2 and 5) and from residues 46 to 51 in structures 3 and 5. Turns are less conserved; no β- or γ-turn is conserved in location and class in all five structures. For example, the type IV turn defined in the consensus from residues 1–4 is a type II′ turn in structure 4 and is absent from structure 5. The consensus inverse γ-turn from residues 20 to 22 is not found in structures 1 and 2. There is a classic β-bulge defined in the consensus for residues 27, 31, and 32, but this does not occur in structures 2 and 5. Further information about each of these motifs can be found by consulting the more detailed plots produced for each member of the ensemble.

This provides a quick way to compare the motifs present in a set of very similar protein structures. We hope in the future to generalize the multiple protein version of the program by allowing less similar proteins to be used, given their sequence alignment.

Conclusion

The data presented show the range of information that is generated by the PROMOTIF program. The package will automatically identify most of the types of small structural motifs

Fig. 8. Comparison of the structures of the first 76 residues for five members of the ensemble of NMR structures of the heat shock transcription factor from *Drosophila melanogaster* (Brookhaven code 1HKT; Vuister et al., 1994). Left-hand columns show the residue number (No.), amino acid sequence (Seq), and consensus secondary structure assignments (SS) derived using the modified Kabsch and Sander algorithm (h/H, α-helix; t/T, turn; e/E, β-strand; and S, bend). The remaining four columns of the consensus structure indicate the locations of β-turns (BT), γ-turns (GT), β-bulges (BG), and disulphide bridges (DS) in the consensus structure. Where present, these motifs are indicated by the class of the motif in the appropriate column for a given residue. For β- and γ-turns, in addition to the various turn types (I, I′, II, II′, IV, VIII, VIa1, VIa2, and VIb for β-turns and IN(VERSE) and CL(ASSIC) for γ-turns), a residue can also be classified as part of a composite turn (C) if it is involved in more than one turn or simply β or γ if the consensus structure has a turn but there is no dominant turn type. Bulges are indicated by two letters, A or P, depending on whether the strands are antiparallel or parallel, and C(lassic), W(ide), S(pecial), or B(ent), depending on the pattern of hydrogen bonds. For each structure in the data set, indicated by the numbers 1–5 at the top of the columns, differences from the consensus structure are indicated as follows. The left-hand column of the data for each protein highlights differences in secondary structure—extra secondary structure is indicated by the appropriate letter and secondary structure missing with respect to the consensus is indicated by the consensus structure with an X through it. The remaining space indicates differences in the turns, bulges, and disulphides. If one of these is present in a particular structure and absent in the consensus, or if the motif type is different from the consensus, the residue is marked with the motif type. If a motif present in the consensus is absent from an individual structure, this is indicated by a cross through the motif (β, β-turn; γ, γ-turn; BG, bulge). Data for the remaining residues in the structure are not shown.

commonly found in proteins and display them in a way that is easy to read and understand. Where possible, the motifs have been defined and classified according to published methods and rules. The package is flexible and can be expanded easily to include further motifs as standardized methods for their identification become available. Obviously, it would be useful to include algorithms for domain classification and larger topological motifs (e.g., TIM barrels). This work is in progress. PROMOTIF should be a useful tool for anyone interested in analyzing and comparing protein structures and should help X-ray crystallographers and NMR spectroscopists to describe new structures as they are solved.

Methods

The programs are written in Fortran 77 and will run on either Unix or VMS platforms. Individual programs are linked via a command file, making the package easy to use. PROMOTIF can be run in one of three possible modes. The first takes as input a single Brookhaven format file and produces a series of output files for each motif. These include flat files, black and white postscript tables, and color schematic diagrams. The remaining two modes take as input a file containing a list of protein structures. The package can be used to produce flat files containing lists of all the examples of each motif in the set of proteins. Alternatively, the program can be used to produce output files corresponding to those in the single file input mode and, optionally, to compare the motifs present in a set of related, aligned protein structures. The user can control the mode of operation, the outputs produced, and the colors to be used by editing the standard parameters file. There is also the option to produce only black and white output.

Availability

The program is freely available for academic users from our anonymous ftp server (address 128.40.46.11), following signature of a licence agreement. The files are in the /pub/promotif directory. Alternatively, contact the authors via e-mail at one of the following addresses: gail@bsm.bioc.ucl.ac.uk or thornton @bsm.bioc.ucl.ac.uk. Further examples of the data produced by the program can be found on the World Wide Web (address http://www.bioc.ucl.ac.uk).

Acknowledgments

Some of the programs in the package are modified versions of programs written by D.K. Smith, A.W.E. Chan, T.P. Flores, and F.M.G. Richardson-Pearl. These programs were developed originally to provide data for the IDITIS protein structure database, which is an on-going collaborative project involving the BSM unit at UCL and Oxford Molecular. We are especially grateful to R. Laskowski, who provided much help with the postscript and whose excellent program, PROCHECK, provided an ideal to which we held PROMOTIF. We are very grateful to users of the program who have made helpful suggestions for improvements to the package. The work was supported financially by grants from the SERC and MRC.

References

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer based archival file for macromolecular structure. *J Mol Biol 112*:535–542.

Chan AWE, Hutchinson EG, Harris D, Thornton JM. 1993. Identification, classification and analysis of β-bulges in proteins. *Protein Sci 2*:1574–1590.

Chothia C. 1992. One thousand folds for the molecular biologist. *Nature 257*:543–544.

Efimov AV. 1987. Pseudo-homology of protein standard structures formed by 2 consecutive β-strands. *FEBS Lett 224*:372–376.

Efimov AV. 1991. Structure of α-α hairpins with short connections. *Protein Eng 4*:245–250.

Hutchinson EG, Thornton JM. 1990. HERA—A program to draw schematic diagrams of protein secondary structure. *Proteins Struct Funct Genet 8*:203–212.

Hutchinson EG, Thornton JM. 1994. A revised set of potentials for β-turn formation in proteins. *Protein Sci 3*:2207–2216.

IUPAC-IUB Commission on Biochemical Nomenclature. 1970. Abbreviations and symbols for the description of the conformation of polypeptide chains. *J Mol Biol 52*:1–17.

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

Lewis PN, Momany FA, Scheraga HA. 1973. Chain reversals in proteins. *Biochem Biophys Acta 303*:211–229.

Milner-White EJ, Ross BM, Ismail R, Belhadj-Mastefa K, Poet R. 1988. One type of γ-turn, rather than the other, gives rise to chain reversal in proteins. *J Mol Biol 204*:777–782.

Oefner C, Suck D. 1986. Crystallographic refinement and structure of DNASE I at 2 Å resolution. *J Mol Biol 192*:605–632.

Orengo CA, Flores TP, Taylor WR, Thornton JM. 1993. Identification and classification of protein fold families. *Protein Eng 6*:485–500.

Richardson JS. 1981. The anatomy and taxonomy of protein structure. *Adv Prot Chem 34*:167–339.

Rose GD, Gierasch LM, Smith JA. 1985. Turns in peptides and proteins. *Adv Prot Chem 37*:1–109.

Sander C, Schneider R. 1991. *Proteins Struct Funct Genet 9*:56–58.

Sibanda BL, Thornton JM. 1985. β-Hairpin families in globular proteins. *Nature 316*:170–175.

Sibanda BL, Blundell TL, Thornton JM. 1989. Conformation of β-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J Mol Biol 206*:759–777.

Tang J, James MNG, Hsu IN, Jenkins JA, Blundell TL. 1978. Structural evidence for gene duplication in the evolution of the acid proteases. *Nature 271*:618–621.

Venkatachalam CM. 1968. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of 3 linked peptide units. *Biopolymers 6*:1425–1436.

Vuister GW, Kim SJ, Wu C, Bax A. 1994. NMR evidence for similarities between the DNA-binding regions of *Drosophila melanogaster* heat shock factor and the helix-turn-helix and HNF-3/forkhead families of transcription factors. *Biochemistry (USA) 33*:10–16.

Wilmot CM, Thornton JM. 1988. Analysis and prediction of the different types of β-turn in proteins. *J Mol Biol 203*:221–232.

Wilmot CM, Thornton JM. 1990. β-turns and their distortions. A proposed new nomenclature. *Protein Eng 3*:479–493.